2-2023

Empirical Investigations and Dataset Collection for American Sign Language-aware Personal Assistants

Abraham Trout Glasser
atg2036@rit.edu

Empirical Investigations and Dataset Collection for American Sign Language-aware Personal Assistants

by

Abraham Trout Glasser

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Computing and Information Sciences

B. Thomas Golisano College of Computing and
Information Sciences
Rochester Institute of Technology

February 2023

Empirical Investigations and Dataset Collection for American Sign Language-aware Personal Assistants

by

Abraham Trout Glasser

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Dr. Matt Huenerfauth Dissertation Advisor	Date
Dr. Kristen Shinohara Dissertation Committee Member	Date
Dr. Roshan Peiris Dissertation Committee Member	Date
Dr. Danielle Bragg (Microsoft Research) Dissertation Committee Member	Date
Dr. Dan Phillips Dissertation Defense Chairperson	Date

Certified by:

Dr. Pengcheng Shi Ph.D. Program Director, Computing and Information Sciences	Date

Empirical Investigations and Dataset Collection for American Sign Language-aware Personal Assistants

by

Abraham Trout Glasser

Submitted to the
B. Thomas Golisano College of Computing and Information Sciences Ph.D. Program in Computing and Information Sciences
in partial fulfillment of the requirements for the
Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

Proliferation of voice-controlled personal-assistant devices poses accessibility barriers for Deaf and Hard of Hearing (DHH) users. It is estimated that 128 million people use a voice assistant in the United States, and 4.2 billion digital voice assistants being used around the world, e.g. through over 60,000 different smart home devices that support Amazon's Alexa, which is one instance of virtual assistants [97, 134]. The ongoing (at the time of this dissertation) coronavirus pandemic has accelerated the uptake of home-based, voice-controlled devices. As artificial intelligence researchers and developers are working on American Sign Language (ASL) recognition technologies, Human-Computer Interaction (HCI) research is needed to understand what Deaf and Hard of Hearing (DHH) users may want from this technology and how to best design the interaction.

This dissertation's contributions are split into four parts:

1. Part I: DHH Interest

This part addresses the gap in knowledge about DHH users' experience with personal assistant devices, shedding light on DHH users' prior experience with this technology and their interest in devices that could understand sign-language commands. This insight supports ongoing technological advancement in sign-language recognition technologies.

2. Part II: Dataset Collection

With the context that there is a ASL data bottleneck for these technologies, this part investigates remote ASL data collection at scale, creating an online sign language data collection platform and testing its viability. Extending upon this, a continuous signing data collection platform was made and is tested. This part also employs a remote data collection protocol using a Wizard-of-Oz prototype personal assistant device, to allow DHH users to spontaneously interact with such a device in sign language. The data collected from this is described, and an in-person Wizard-of-Oz experiment is conducted to investigate aspects that were not possible through the remote protocol.

3. Part III: DHH Behavior

From the remote Wizard-of-Oz methodology in Part II, this part presents analysis of this data and addresses the gap of knowledge when it comes to ASL interaction with personal assistant devices. This part also describes analysis of the in-person Wizard-of-Oz experiment mentioned in Part II, showing what we have learned about the linguistic properties of in-person interaction.

4. Part IV: Privacy Concerns

This last part touches on a theme that occured during parts I-III; privacy concerns. This part utilizes state-of-the-art image processing technology and guides the development of ASL-optimized face technology to protect anonymity of DHH users' to their satisfaction. Lastly, a small interview study was appended to the in-person Wizard-of-Oz experiment to further confirm whether DHH users would be more comfortable in using a personal assistant device if face-disguise technology was embedded.

Acknowledgments

This endeavor would not have been possible alone. I am deeply indebted to my PhD advisor, Dr. Matt Huenerfauth, for his invaluable advice, patience, and feedback throughout my doctoral career. I am also grateful to my dissertation committee, who generously provided additional knowledge and expertise. Additionally, I'd like to recognize the funding bodies that supported my work, allowing me to recruit participants, purchase equipment, and travel to conferences to publish and present, disseminating my work internationally.

I am also thankful for my research lab - from professors and senior PhD students to undergraduate research assistants and research participants - where I could always find inspiration and motivation, and gained unique perspectives. I would be remiss if I did not mention my family, friends, and partner, who were always there to cheer me on and keep my spirits high in this process.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Broadly, with the proliferation of voice-controlled personal assistant devices comes accessibility barriers for Deaf and Hard of Hearing (DHH) users, and there has not been significant prior research on sign-language conversational interactions with technology. Furthermore, voice-control is becoming a ubiquitous interface to technology, and as this trend continues, the urgency of addressing accessibility challenges in this technology increases. For instance, since these personal-assistant conversational interface systems are often based in smart speakers that may be shared across multiple users in a household, these technologies are already appearing in the homes of people who are DHH, e.g., when hearing members of the household purchase these devices. It is estimated that 128 million people use a voice assistant in the United States, and 4.2 billion digital voice assistants being used around the world, e.g., through over 60,000 different smart home devices that support Amazon's Alexa, which is one instance of virtual assistants [97, 134]. In addition, the ongoing (at the time of this dissertation) coronavirus pandemic has accelerated the uptake of home-based, voice-controlled devices.

This dissertation addresses a lack of knowledge about DHH user interest and requirements for this technology, a lack of knowledge about how to collect training data of relevant examples of American Sign Language (ASL) commands for this context, and a lack of knowledge about the types of commands DHH ASL-signing users would like to perform. Further, this dissertation investigates the structure and

vocabulary of such commands, how users would wish to interact with the device, and how they physically approach or engage with personal assistant devices that can understand ASL.

Gathering such data and answering these questions provide important guidance to designers of the underlying AI technologies that would enable such interactions in the future, such as computer-vision technologies for automatic sign recognition of personal assistant commands. Additionally, this dissertation investigates the topic of privacy concerns when it comes to such data.

1.2 Structure of this Dissertation

Chapter 2 provides the reader with background information regarding personal assistant device technology. An overview of typical personal assistant interaction, device popularity, and image examples of popular devices are given. Also in this chapter is a critical literature review of prior work with several goals:

- First, since most personal assistant technologies are based on speech interaction, the state of the art in automatic speech recognition is discussed, especially in regard to the limits of such systems at recognizing the voices of people who are DHH – to establish that speech-based interaction with such systems is difficult for these users.

- Next, the limitations of current text-based-interaction workarounds for personal assistant systems are described, to further motivate the need for new ways of interacting with these systems.

- Since there has not been significant work investigating personal assistant systems among DHH users, I briefly survey related work in which user requirements and accessibility of such systems has been investigated among other groups of users, to determine which methods were used to understand users' needs and usage.

- Finally, the limitations in automatic sign-language recognition, especially in the context of personal assistant systems, is surveyed, to explain how the research described in this dissertation guides the creation of necessary datasets for that technology.

1.2.1 Part I: Investigating Current DHH Familiarity, Interest, and Imagined Usage of Personal Assistants that Can Understand Sign Language

As will be discussed in chapter 2, there is a lack of prior research on the usage of these devices by people who are DHH. Such research would shed light on DHH users' prior experience with this technology and their interest in devices that could understand sign-language commands. Understanding DHH user's priorities in how they would like to use such technology, as well as the types of commands they would like to use in ASL, is also necessary to support ongoing technological advancement in sign-language recognition technologies, for automatically identifying words and phrases performed by signers in video. Chapter 3 addresses the gap in knowledge about DHH users' experience with personal-assistant devices, and investigates the following research questions:

> **RQ1.1:** Have DHH ASL users used devices like this up to now, and what has been their experience?
>
> **RQ1.2:** If DHH ASL signers could interact with a personal-assistant device in ASL, would they be interested in its use?
>
> **RQ1.3:** What types of commands would DHH ASL signers imagine using with such a device?
>
> **RQ1.4:** How would DHH ASL users imagine waking up, interacting, and receiving responses from such a device?

Focusing on the device-activation phase of the personal assistant interaction, chapter 4 investigates:

> **RQ2:** What are the factors considered by DHH users to judge a wake-up interaction accessible and usable for a personal assistant device that understands ASL?

1.2.2 Part II: Dataset collection

Section 2.3.2 in chapter 2 explains how there is a data bottleneck for AI researchers working on sign language recognition. Prior work, e.g., traditional in-lab ASL data collection, has been most successful when users are provided with specific prompts of words or passages in English. Through an internship

with Microsoft Research, I investigated related issues in remote ASL data collection at scale, exploring the possibility of using a scalable online sign language data collection platform and testing its viability.

With the context that a corpus of labelled, isolated signs may help develop technologies involving individual sign recognition, e.g., digital personal assistants that respond to simple signed commands, chapter 5 investigates:

> **RQ3:** How can DHH and signing communities be enabled to curate sign language datasets that overcome limitations of traditional in-lab collection (e.g. limited demographics, controlled environments, limited size and quality, expensive post-processing and labeling)?

For real-world use cases of sign language processing, such as signing longer-than-one-word commands to personal assistant devices, there needs to be natural conversation with complete sentences, thus continuous sign language data. After identifying a viable platform for collecting individual signs in chapter 5, I explored whether this methodology could be extended to also generate a continuous signing dataset, while supporting bilingual content. In chapter 6, I investigate:

> **RQ4.1:** How can everyday signers efficiently contribute to continuous sign language datasets?
>
> **RQ4.2:** Ensuring that the DHH community is involved in the process, how would the platform be designed? What are the design criteria?
>
> **RQ4.3:** How would DHH users respond to crowd-generated content?
>
> **RQ4.4:** Can the platform incentivize contributors by being a sign language bilingual resource?

Through these data collection efforts, I gained insight about how this could be done for personal assistant commands, which may benefit AI researchers in the future. Part I addresses the lack of knowledge about DHH user interest and requirements for this technology, asking users to imagine interacting with a device that understands ASL. However, participants were never provided the opportunity to actually try doing so, throughout the interview and survey studies. Similarly, for the "wake-up" prototypes in chapter 4, participants were shown videos of potential interaction approaches to activate such a device, but participants did not get a chance to actually try them out. Putting the knowledge learned from Part

I into practice, I employed a remote data collection protocol, using a Wizard-of-Oz prototype personal assistant device that could understand ASL. Chapter 7 describes our experimental set-up allowing DHH users to spontaneously wake-up and interact with a personal assistant device in sign language, in a limited manner. In this chapter, I describe the methodological setup and logistics, and describe the dataset that was collected, starting with the general characteristics of the dataset itself. Then, I explain how we analyzed and annotated this data, including the specific properties that we recorded, and share the dataset publicly.

This data enabled us to learn many things, which will be discussed in Part III (specifically chapter 9). However, there were also limitations. For instance, participants were not able to connect the personal assistant device to other objects in their home, e.g., a smart TV or lights. Users were not able to change their location inside the room, and had to be situated in a Zoom videoconference call, in which the personal assistant device display was visible. Additionally, there were some linguistic features of natural ASL communication that were not able to be captured, such as referencing to other objects in the room, or utilizing three-dimensional signing space, an intrinsic property of ASL (ASL is a visuospatial language). At the end of this part, chapter 8 conducted an in-person, physical Wizard-of-Oz experimental set-up that allowed investigation of aspects not possible through the remote protocol described previously, such as the aforementioned linguistic properties of in-person interaction.

1.2.3 Part III: DHH Behavior with a Personal Assistant that appears to understand American Sign Language

Despite the limitations of the remote data collection protocol described in chapter 7, we were still able to gain valuable insight from this engagement with the DHH community. Starting in Part III, chapter 9 presents analysis of this data and addresses these research questions:

> **RQ5.1:** Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users instinctively "wake-up" the device or initiate a command?

> **RQ5.2:** Based on observation of the behavior of DHH ASL signers interacting with a personal

assistant device in sign language, what categories of commands/requests do users produce?

RQ5.3: What do these commands look like in ASL?

RQ5.4: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users recover or respond when there is an error or breakdown?

RQ5.5: After DHH ASL signers had the opportunity to interact with a personal assistant device in sign language, did their interest in such interaction increase, decrease, or stay the same?

While the research questions outlined above addressed the gap of knowledge when it comes to ASL interaction with personal assistant devices, these were not real-world conditions, where a person is typically physically in the same room as the device, in a residential-like setting. In this scenario, the user would be able to change their location inside the room and utilize their full signing space, i.e. users often "point" to items or people in their surrounding environment to refer to them in ASL, and not having their ASL commands confined to an integrated camera on a computer where they are in a Zoom videoconference call. It is unknown if DHH users would do this referencing while interacting with an inanimate device. During the remote study, while participants had the freedom to think of different commands to try out with the device, they were being visibly watched by a researcher from our team and had to improvise to continue thinking of more commands to try out. It is unknown how DHH users may spontaneously and naturally interact with a personal assistant while they are focusing on a typical event that occurs in the home (e.g. preparing a drink and snack). New types of uses are possible in an in-person, home-like setting where control of items in the room is possible, yet no prior work had explored how users linguistically construct such commands in ASL.

Chapter 8, as mentioned earlier, describes an in-person Wizard-of-Oz set-up that aims to emulate a residential living room and kitchen, and to collect data of a DHH user interacting with a personal assistant using ASL in a natural and instinctive manner while in this home-like environment. In this part, chapter 10 presents the analysis of the results and discusses the following research questions:

RQ6.1: What are some linguistic properties of in-person interaction with full signing space, such as referencing to other objects inside the room?

RQ6.2: Based on observation, how would a DHH user naturally position themselves in proximity to a personal assistant device in a residential-like living room and kitchen area?

RQ6.3: When given the opportunity to interact with a personal assistant device that is physically in the same room as them in a residential setting, what are DHH users' experience and opinion on this?

1.2.4 Part IV: Privacy Concerns

In chapter 3, participants mentioned that they had some privacy concerns. If a personal assistant uses a camera to recognize ASL, then that would mean DHH users need to have a personal assistant with an integrated camera and place it in their homes. Similar to existing privacy concerns by general personal assistant users who have concerns about integrated microphones picking up audio in their homes, DHH users showed concern about having such a device with a camera watching their home.

Investigating whether the state-of-the-art image processing technology could alleviate DHH users' concerns by "anonymizing" their ASL videos, I was involved in conducting a study on the criteria required to protect anonymity to DHH users' satisfaction. In chapter 11, an interview study was conducted to evaluate prototype face-disguise technology applied to videos of human ASL signers, and preserving facial expressions and head movements (which are essential characteristics and properties of ASL). Specifically, with the goal of guiding the development of ASL-optimized face technology and inform designers of future applications for these users (e.g. integrating this into personal assistant devices to ensure protection of the DHH user's privacy), chapter 11 addresses this research question:

RQ7.1: Is state-of-the-art face-disguise technology capable of perserving facial expressions and natural human appearance for sign language video?

RQ7.2: What are DHH users' interest in and impressions of this technology for protecting anonymity, including users' views of various dimensions of system performance?

This study presented quantitative and qualitative evaluation of understandability, naturalness, and anonymity-preservation, and compared prototypes varying in their appearance transformations, and found evidence of users' views on the acceptability of this technology and its potential uses. However, this study focused on the general usage of this to communicate online, e.g., through social media postings. This study had not considered the use of this technology during personal-assistant interactions.

Within the context of personal assistants that can understand ASL, there is also potential for face-disguise technology to be incorporated into such devices from a privacy perspective. For instance, ASL video recordings collected by a personal assistant could be "anonymized" *before* being processed by the device or transmitted across the Internet to a remote server, to protect the identity of the person appearing in the video. Chapter 12 appended a small interview study to the in-person Wizard-of-Oz experiment, and discussed:

> **RQ8:** Would using a state-of-the-art face-disguise technology to anonymize DHH users' ASL recordings (that are used for device processing) before they are processed by a personal assistant device alleviate privacy concerns?

Chapter 2

Background and Prior Work

This chapter provides the reader background on personal assistant devices, the underlying technology, current accessibility workarounds, and describes why these do not work for the DHH community. In particular, sections 2.2.3 and 2.3.2 discuss what prior methodologies have been used in studies related to DHH users interacting with personal assistants using ASL. Those sections also outline reasons why there is a need for additional collection of appropriate ASL data, giving motivation and describing rationale behind Parts I and II of this dissertation.

2.1 Personal Assistant Devices

Modern personal-assistant technologies are based on speech interaction, with Automatic Speech Recognition (ASR) transcribing verbal commands into text, which is then processed by the device. Typically, speaker-based personal-assistant devices are voice-activated and process a user's command if prefixed by an activation word, sometimes called a "wake-word" [77]. For example, a user needs to say, "Alexa," to get the attention of an Amazon Echo device, and then the user would speak the command. After processing the command, the device typically provides an audio response. However, some devices include a screen, so that they can provide both audio and visual responses [85].

In 2019, 72% of respondents of a global survey [128] reported using a digital assistant, and 45% reported owning one, with an additional 26% planning to purchase one soon. The report also covered

popular use of digital assistants for music (63%), lighting (57%), security cameras (38%), thermostats (37%) by homeowners. Specifically, people issued commands related to playing music, looking for directions, getting news and weather. Unlike mobile devices, personal-assistant devices are generally confined to one place in the house, which may influence how they are used. For instance, in a mixed-method study, Sciuto et al. found that device location affects use of such devices [145], and they identified where owners typically place devices: bedrooms, living rooms, kitchen, and home offices.

As evident through [60, 139, 179], personal assistants that are controlled using voice brings accessibility barriers for DHH users. This is becoming an increasingly urgent accessibility challenge, especially because many various "smart devices" (e.g. light bulbs, speakers, cameras, TVs, or coffee machines) are commercially available with voice-based personal assistant technologies embedded. Additionally, the coronavirus pandemic (current at the time of writing) has amplified the adoption of home-based technologies like these, and more of these technologies are appearing in the homes of people who are DHH, e.g. when hearing members of the household purchase these devices. Figure 2.1 shows some examples of current popular consumer personal assistant devices. Inside the figure, it can be seen that personal assistant devices come in various shapes and sizes, a very popular form being smart speakers (as small as a hockey puck or as large as a basketball ball), and more recently smart displays that have integrated cameras and screens for displaying results and output.

2.2 Current Personal Assistant Usability

2.2.1 Limits of speech recognition for DHH voices

Automatic Speech Recognition (ASR) is an underlying technology that supports users' speech-based interaction with personal assistant devices. ASR automatically transcribes verbal commands into text, which is then processed by the personal assistant device. The DHH population is very diverse, with the level of hearing and speaking skill varying widely among individuals [18, 65]. For DHH individuals who do not use their voice (or do not feel comfortable doing so in some social settings), voice-controlled devices are inaccessible. Even for those individuals who do use their voice, it may not be understand-

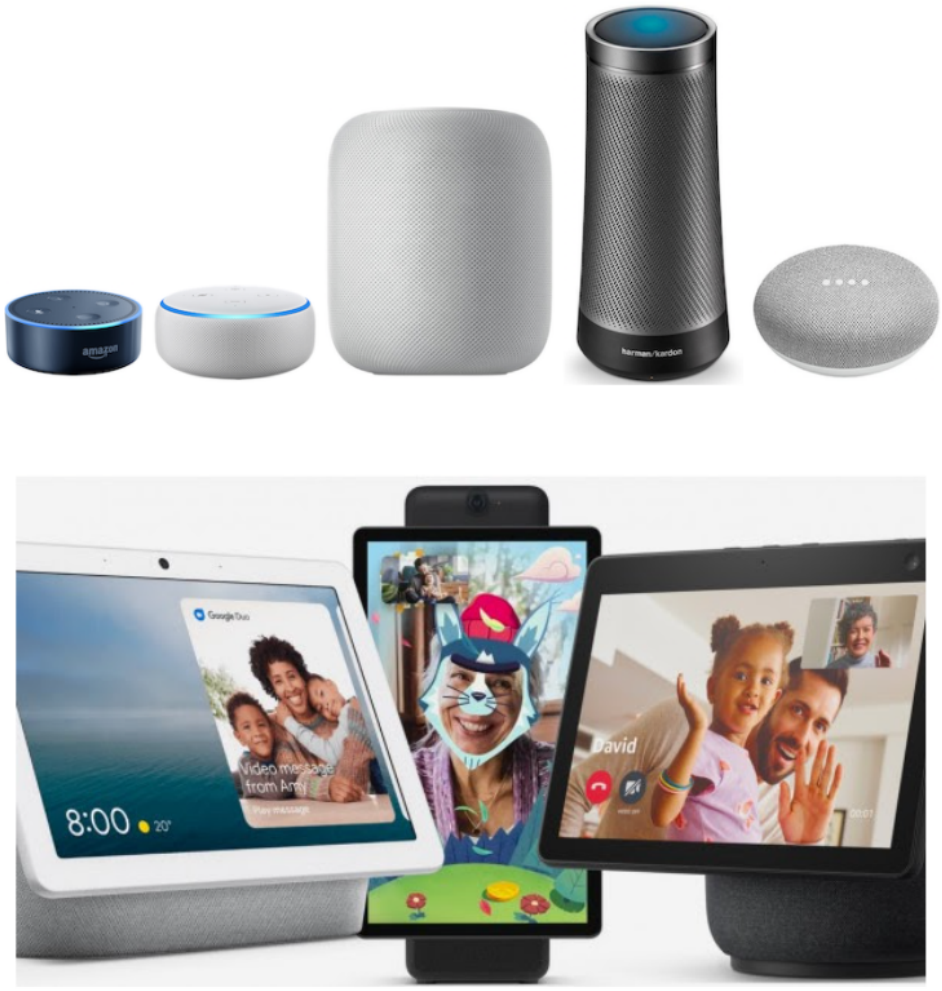


Figure 2.1: Examples of popular personal assistant devices
Sources: https://moniotrlab.ccis.neu.edu/smart-speakers-study-pets20/
https://www.the-ambient.com/buyers-guides/best-smart-speakers-displays-screens-2574

able to ASR technology. In a prior study (winning first place at the ACM CHI 2019 student research competition) [60], I found that even among the voices of DHH individuals whom professional speech pathologists and naive hearing listeners agreed were very understandable, modern ASR technology was unsuccessful at understanding the speech. This was a concerning finding, since it indicated that our human instincts about which voices among DHH individuals may be easy to understand may not be predictive of whether ASR technology will work successfully.

2.2.2 Text-input workarounds are insufficient

As a workaround for speech-based interaction, some modern voice-based personal assistant devices offer a text-based input option (in which the user is able to type English commands into the system using a touch screen on the device or wireless keyboard). Unfortunately, providing this alternative text-input option is not a complete (nor functionally equivalent) solution for personal assistant devices. There are many settings and scenarios in which text-input would be undesirable by the DHH user, such as when a user is across the room from a device or when the user's hands are messy during cooking in a kitchen setting. Also, there are many DHH individuals who prefer communication in ASL, and some of these users may have difficulty with an English text-based interface, e.g. due to literacy concerns.

2.2.3 Prior work on DHH users and personal assistants

In this section, background is given on circumstances leading to the decision of utilizing a semi-structured interview and survey-based methodology in Part I and a Wizard-of-Oz methodology in Part II of this dissertation.

Pradhan et al. [137] briefly discussed DHH users, as part of their survey of current personal-assistant-device users who have disabilities (not only DHH). They found a small percentage of device users had hearing loss, and these users still faced accessibility issues with personal assistants even though devices were marketed as accessible.

In a related prior work on DHH users' interactions with personal assistants devices [139], researchers conducted a small exploratory study that asked DHH participants to issue commands to a device (in a Wizard-of-Oz manner) using ASL, non-ASL gestures, and computer speech synthesized from text typed by the user. They developed a few tasks (e.g. "turn all lights on") and provided pre-defined ASL command and gesture participants should use for the task (that is, participants could not spontaneously ask different commands or experiment). Once they used these ASL commands or gestures, they had an ASL interpreter voice the commands to the Alexa device in a Wizard-of-Oz manner. Their participants found it awkward to learn and remember non-ASL gesture commands, and they preferred issuing commands in ASL. Participants expressed concerns about the system not recognizing alternate signs and

were consistently excited about the possibility of ASL recognition for personal assistant systems. They report that participants were satisfied with the TTS application, but they could not tell if their messages were poorly vocalized or misunderstood, as it is an inaccessible speech interface. The non-ASL gesture system devised by the researchers for this study was an unnatural, nonlingustic input methodology, requiring users to learn a new communication system only for Alexa. They report that participants were frustrated with the gesture system as it was unintuitive and they desired a system more similar to ASL or user-customizable gestures.

While this initial work has established some interest among the DHH community, the study had relatively few participants, and the authors discuss some limitations of the work in regard to the study design and whether their participants were sufficiently representative of the broader DHH community [139]. Their study had a large number of variables, and was meant "to serve as a first look at alternative inputs for spoken modality systems."

This same team of researchers conducted a follow-on study using similar methodology [179]. This work focused on comparing user experience with using ASL v.s. using tablet apps. They set up a primitive "living room" with a couch, TV, and lamps. Their experiment gave participant a list of tasks, and had to perform each task twice using different approaches: ASL and a tablet (which had the Fire TV and Philips Hue applications installed). For each task, they had three tries for the command, and could choose to repeat the same ASL sign or change them. Other than these tasks, participants were not given the opportunity to experiment and spontaneously interact with the Alexa device, and the other phases of interaction (e.g. device activation) were not discussed. Besides some technical and logistic limitations, they reported that their participants preferred using ASL than using the specific applications on the tablet. At the end of their report, they repeat that their updated study was still meant "to serve as a first look at alternative inputs for spoken modality systems," and they "strongly urge user interface developers to explore accessible interfaces for spoken modality systems, before the systems become truly ubiquitous in daily life."

These two studies [139, 179] did not deeply investigate DHH individuals' device preferences, their requirements and desired use cases, nor the full set of questions posed in chapter 1. While these studies

had provided initial insight into this research space, the design of those studies did not include a systematic qualitative analysis of observational data, as those studies were more exploratory in nature. Despite this, they revealed that a Wizard-of-Oz approach is possible for supporting users in issuing commands to a device, with an ASL interpreter translating the commands. These studies also revealed preliminary evidence of the excitement of DHH users in the potential use of ASL recognition technology for personal assistant devices, with users preferring to use their primary language without the need for additional devices such as tablets for text input.

Beyond these few survey-based and Wizard-of-Oz-based studies above, no prior empirical research has focused on DHH users in regard to personal assistants. There is a need for further engagement with the DHH community, and a need to learn what DHH users want and how they would use personal assistant devices.

Because of this limited DHH-specific prior work, it is useful to broaden our focus to consider prior work on other groups of users with disabilities, with a goal of understanding what HCI methodologies would be appropriate. One exploratory study [2] elicited viewpoints and practices of interacting with personal assistants among people who are blind. This semi-structured interview study, with many open-ended questions and follow-ups, examined several aspects of blind users' interaction with personal assistant devices and how the user experience could be improved. The previously mentioned Pradhan et al. [137] study used a survey-based methodology among users with diverse disabilities, and [139, 179] showed that Wizard-of-Oz observation studies are a possibility.

Prior work in surface gestures has thoroughly investigated how to best create gesture sets informed by user behavior. Wobbrock et al. presented a study of surface gestures, as well as presenting a taxonomy of surface gestures useful for surface computing [178]. In their work, they gained insight into the mental models of non-technical users and outlined implications for technology and design, yielding a user-defined gesture set based on participants' agreement of over 1080 gestures. Their methodology included showing users the effect of a gesture on a touch-screen surface computer, and then asked participants to portray the cause (eliciting the input from users). While this work was impactful for designers of touch-screen UIs, it is unknown if this translates to camera-based personal assistant devices

that utilize sign language recognition models for activation and interaction. Following this work, in a Wizard-of-Oz study with such devices, if an user is informed of what happens when a device is activated, then they can be asked to perform the gesture or sign they would use to do so.

The studies and methodologies described in this section have provided inspiration for the work presented in Part I and Part II of this dissertation.

2.3 Sign Languages

This section discusses prior work that has established the existence of a data bottleneck which is holding back the development of sign-language technologies, as well as calls for researchers to address this bottleneck. This prior work motivates efforts described in Part II of this dissertation.

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.

2.3.1 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 [57, 157]. Generation refers to generating signed content, for example through signing avatars [50, 80]. 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 [29, 44, 94, 186]. The state-of-the-art in sign language modeling has evolved significantly with the advent of deep learning (e.g. [94]). 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.

2.3.2 Sign language recognition needs datasets

There has been recent excitement among the DHH community and researchers in the area of sign-language technologies, as evidenced by research projects, hackathons, and workshops regarding this area [1, 23, 115, 175]. For personal assistants, sign language recognition would translate the video of the user performing sign language into written queries for the system's search engine [23]. Despite misleading media reports about personal assistant devices that can understand sign language, no current device can accurately understand sign-language commands. Demos of such prototypes are generally not robust, i.e. only working for a trivially small set of fixed commands or when the command is performed in an unnatural way [15, 35, 140]. While artificial intelligence researchers and developers are still making progress in the area of sign language recognition technologies, to enable more robust understanding of user's commands [23], it is important for HCI researchers to begin investigating the future interaction potential of this technology. There is a need to understand what users may want from this technology and how to best design the interaction. Rodolitz et al. called for HCI researchers to explore interaction methods for DHH with personal assistants before they become ubiquitous in daily lives [139].

A well-regarded recent review of sign-language technologies (best-paper award at the ACM ASSETS'19 conference) [23] found that data is the key bottleneck for artificial intelligence researchers working on sign language recognition. They hosted an interdisciplinary workshop with 39 domain experts with diverse backgrounds, where they reviewed the state-of-the-art, and listed calls to action for the research community. These calls included but were not limited to focusing on real-world applications and creating larger, more representative, public video datasets. They emphasized the current lack of data, cited as the biggest obstacle in sign language technology research. Data collection is difficult and costly, yet "without sufficient data, system performance will be limited and unlikely to meet the Deaf community's standards".

Despite these challenges, groups have worked on sign language data creation and curation. Datasets exist for many signed languages, including but not limited to German[1], American[2], Argentinean[3], Turkish[4] (more listed in [23, 37, 129, 130], etc.). The main parameters of sign language datasets include the number of subjects, samples, language level, type, and annotations/labels. As explained in [23], existing sign language datasets greatly limit the robustness of systems trained on them. Current datasets are not sufficiently large – typically containing fewer than 100,000 articulated signs.

Also, many existing datasets contain individual signs, which may not be as useful for real-world use cases of sign language processing. For real-world applications, there needs to be natural conversation with complete sentences, i.e. "continuous" sign language. Continuous sign language recognition and translation is challenging due to epenthesis effects (insertion of extra features into signs) and co-articulation (ending of a sign affecting the start of the next), among other difficulties. Solving these requires large amounts of continuous sign language data to learn from. Continuous signed sentences would also be useful for DHH individuals trying to understand content, especially new concepts, as it is natural and comfortable for them. There are some continuous signing datasets, such as [45], which help fill this void, however they are typically small and recorded in a studio environment rather than a natural setting, which makes generalization to diverse users and real-world environments difficult.

Currently available sign-language datasets are expensive to produce, due to the cost of annotating video of humans signing. While these datasets support linguistic research, given the complexity and diversity of the language within each, they are too diffuse. To support modern deep-learning methods for sign recognition, the datasets would need to be much larger (which is too expensive) or more narrowly focused on personal-assistant commands, to provide enough relevant training data for such learning. However, it is unknown what commands DHH users would want to perform, especially if the devices were able to understand ASL-based input. After learning about DHH users' interest in what they would like to say to these devices, then it would be possible to build a dataset of videos of such commands, which could be used by sign-language recognition researchers.

[1]https://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/
[2]https://github.com/YAYAYru/sign-lanuage-datasets
[3]http://sedici.unlp.edu.ar/handle/10915/56764
[4]https://www.cmpe.boun.edu.tr/pilab/BosphorusSign/home_en.html

Part I: Investigating Current DHH Familiarity, Interest, and Imagined Usage of Personal Assistants that Can Understand Sign Language

Prologue to Part I

There is a lack of knowledge about DHH user interest and requirements for personal assistant devices, and it is unknown how they currently use existing commercially-available consumer devices, such as the Google Nest Hub or Amazon Echo Show. These devices, along with other voice-controlled interfaces, are becoming ubiquitous in daily life, and it is becoming increasingly urgent to address the accessibility challenges they pose to DHH users. Motivated by chapter 1 and chapter 2, Part I of this dissertation directly engages with DHH individuals and investigates their familiarity, interest, and imagined usage of personal assistant technologies. The results from Part I subsequently serve as foundation and motivation for the remaining research efforts in this dissertation.

Since there has been no significant prior research on this topic, chapter 3 employs a mixed interview and survey study and chapter 4 uses formative interviews and video prototypes to begin addressing the gap in knowledge and to get initial reactions from the DHH community. Specifically, Part I of this dissertation investigates the following research questions:

RQ1.1: Have DHH ASL users used devices like this up to now, and what has been their experience?

RQ1.2: If DHH ASL signers could interact with a personal-assistant device in ASL, would they be interested in its use?

RQ1.3: What types of commands would DHH ASL signers imagine using with such a device?

RQ1.4: How would DHH ASL users imagine waking up, interacting, and receiving responses from such a device?

RQ2: What are the factors considered by DHH users to judge a wake-up interaction accessible and usable for a personal assistant device that understands ASL?

Chapter 3

DHH Users' Interest in Sign-Language Interaction with Personal-Assistant Devices[5]

3.1 Introduction

With their growing popularity, it is important to consider the accessibility of web-powered personal assistants, and the COVID-19 pandemic has magnified the importance of home-based technologies like this. As described in chapter 2, the recent proliferation of voice-based personal-assistant technology poses new accessibility barriers for many Deaf and Hard of Hearing (DHH) users [60].

This chapter will discuss how we conducted interviews and a survey to investigate whether DHH users use these systems currently, their interest in using devices capable of sign-language input, the types of commands they would like these systems to support, where they would use these devices in the home, and how they would prefer to interact with these devices. Interviews were conducted with 21 DHH ASL users, and interviews informed the design of an online survey with 86 DHH ASL users,

[5]The information in this chapter is based on a joint project with my advisor (Dr. Matt Huenerfauth), and a graduate student at RIT (Vaishnavi Mande) whom assisted me with qualitative data analysis. The results were published as a paper at the W4A'21 conference [62].

to gather more quantitative data. A small percentage of DHH ASL users reported experience with personal assistants, however they had great interest in sign-language interactions with these devices. Respondents were interested in traditional uses of personal assistants, e.g. setting alarms and timers, as well as applications more specific to the DHH experience, e.g. notifications of environmental sounds, or requesting ASL translations. Participants also shared where they would put devices in their home, and how they imagined interacting with these devices.

The main contribution the work described in this chapter is empirical: This study contributes knowledge of DHH signers' interest in ASL interaction with these devices, providing a motivation for future research on this technology. This study also reveals how the uses of interest to DHH users differ from those of the wider population, which is an essential foundation for any future efforts to create a representative dataset of the types of ASL commands DHH users may wish to give to such devices. Furthermore, this study reveals some concerns and preferences among DHH signers for how their interaction with personal-assistant devices should occur in ASL, which suggests key avenues for future HCI and accessibility research in this area.

3.2 Personal-Assistant Device Popularity

Before investigating DHH ASL signers' experience and interest in personal-assistant technologies, it is useful to examine prior research among the general population, for context. As mentioned in section 2.1, 72% of respondents of a global survey reported using a digital assistant, mainly using them for music (63%), lighting (57%), security cameras (38%), thermostats (37%), and typically placing the devices in bedrooms, living rooms, kitchen, and home offices.

Prior to considering how DHH users might interact with personal-assistant devices using ASL, it is useful to consider the typical stages of interaction for voice-based personal assistants. As described in section 2.1, personal-assistants are typically voice-activated by a "wake-word," followed by a command. Then, the device can provide both audio and visual responses. As DHH ASL signers envision how they may interact with personal-assistant devices capable of understanding ASL commands, we are specifically interested in users' preferences for these various stages of the interaction, which include:

waking up the device to get the devices' attention, followed by an ASL command, followed by the response. Thus, we are also interested in understanding what response modality DHH ASL users might prefer, e.g. text displayed on the device.

3.3 Research Questions

An analysis of prior work (chapter 2) has revealed a gap in knowledge about DHH users' experience with personal-assistant devices, including accessibility issues they may face with this technology. Furthermore, as technological advances may someday enable these devices to understand sign-language input commands, research is needed on whether users are interested in this interaction, what they would like to do with devices that support sign-language interaction, and their expectations or concerns. Digital assistants can be accessed in many different ways, e.g. via IoT devices (TVs, cars), smartphones, or home-based devices (smart speakers, displays). Given the shared use of smart devices among a household, we focus on one form-factor for clarity in our study: home-based personal assistant devices. Our research questions fall into two categories: RQ1.1 and RQ1.2 examine DHH ASL users' experience and attitude towards personal assistants, and RQ1.3 and RQ1.4 focus on potential interactions with such devices:

RQ1.1: Have DHH ASL users used devices like this up to now, and what has been their experience?

RQ1.2: If DHH ASL signers could interact with a personal-assistant device in ASL, would they be interested in its use?

RQ1.3: What types of commands would DHH ASL signers imagine using with such a device?

RQ1.4: How would DHH ASL users imagine waking up, interacting, and receiving responses from such a device?

3.4 Research Methods

We conducted a two-phase study, with interviews among 21 DHH adults who identify as ASL signers, guiding the design of a survey with 86 DHH participants. The questions in both the interviews and survey were aligned with the four research questions above. An initial interview phase with open-ended questions, followed by a larger survey study (to enable quantitative prioritization of some interview responses), was selected as the methodology, given the initial exploratory nature of the research questions, which focused on users' experience, interests, and imagined use of a device that understands ASL commands. Both phases of this study were approved by our Institutional Review Board (IRB), and informed consent was obtained prior to participation.

Our methodologies are based on prior work to understand interest in personal assistants among other user groups with disabilities, e.g. [137]. Prior work had used an analogous interview and survey methodology to establish a sub-group's interest in new interactions [43, 56]. We specifically recruited DHH individuals who use ASL on a daily basis. Recruitment on a university campus will not yield a fully diverse sample of the entire ASL signing population. This was a motivation for our mixed methods design, where our online survey reached a more geographically and demographically diverse sample.

3.4.1 Initial Interview Methods and Analysis

The goal of this interview study was to gain an initial understanding of users' views and to support the formative development of questions and answer-choice options for our subsequent survey questionnaire. A total of 21 DHH participants were recruited for interviews through on-campus and social-media ads, with two eligibility criteria: (1) identifying as DHH and (2) using ASL on a daily basis. Participants self-identified as 12 women, 8 men, and 1 as non-binary, with a mean age 24 and standard deviation 3.56. Eleven self-identified as Deaf, 4 as deaf, and 6 as hard of hearing. These interviews were conducted in ASL by a DHH researcher at our laboratory, and participants' responses were transcribed by the researcher during the interview session. Each interview [6] was scheduled for 70 minutes, and participants were compensated $40.

[6] Our participant demographic questionnaire and data, and all interview questions are provided in appendix B.

We conducted an iterative semantic thematic analysis [28] of the transcripts. One researcher recorded notes from all the interview transcripts on 100 post-it notes using Miro software and organized the post-it notes into 42 initial themes across each section of the interviews (e.g. device usage experience, interacting with the device, and usage expectations from the device). The researcher then identified 10 emergent codes that applied to each of these sections (e.g. usage expectations from the device included possible scenarios in which the device will be used, possible commands to give to the device, where would the device be placed, and concerns with using the device). While going through each of the transcripts, the researcher updated the codes as necessary. The researcher then performed another pass of the transcripts and verified the codes and highlighted categories within the codes [7].

To analyze open-ended responses about potential uses of personal-assistant devices, participants responses were categorized into 10 groups as above, e.g. DHH specific sound alerts in homes, weather-related questions, alerts, timers, etc. To select answer-choices for multiple-choice questions on the survey questionnaire, we used these 10 categories, along with an "other" option with a write-in textbox.

3.4.2 Survey Methods and Analysis

Our analysis of the interview responses informed the design of the questionnaire for this survey study. We refined the wording of some interview questions based on any clarification noted during the interviews, and we also selected answer-choices for the questionnaire based on interview responses. For many questions, an "other" option was included so that respondents could offer answers that had not been anticipated based on our analysis of the interview data. The questionnaire for this online survey was created using Google Forms. The survey included a mixture of question types, including multiple-choice, short answer (open-ended text), long answer (open-ended text), and Likert-scale questions.

Since our participants were DHH ASL signers and our focus was on personal assistants that understand ASL, it was necessary to provide questions and answer-choices redundantly in the form of both English text and ASL video. A DHH native ASL signer on our research team recorded videos of ASL versions of all question items, to ensure an accessible survey for ASL-fluent DHH individuals. In the case of open-ended questions on the survey, participants provided responses as English text.

[7] We provide a large image showing our affinity mappings in appendix B.

To ensure that survey participants included demographic and geographic diversity among the U.S. DHH community, the survey was distributed nationally by advertising on locally and nationally-owned social media pages, and national organizations affiliated with the DHH community were contacted for advertising. The survey was advertised from March 2020 to the end of April 2020. From 86 respondents, three raffle winners were selected to receive $100. Out of these 86 respondents, 49 identified as women, 37 identified as men, with mean age 47 with standard deviation 23.5. Seventy identified as Deaf, with the other 16 identifying as deaf or Hard-of-Hearing. The participants were very spread out across the U.S., including individuals from over 20 U.S. states.

The survey consisted of multiple open-ended questions. Answers to the questions on the possible scenarios in which the device will be used and possible commands to give to the device were analyzed using deductive coding, using the codes that we inductively identified during the interview transcript analysis. The outlier scenarios were recorded separately. This approach helped us collect a dataset of possible commands and usage scenarios which the DHH users would be interested in using as discussed in RQ1.3. For the open-ended question about where in the home a device would be placed, we used an affinity mapping approach to organize the locations and rationales mentioned by participants [8].

3.5 Results

To avoid redundancy and to enable triangulation between the findings in this mixed-methods study, the results of the interview and of the survey are *interleaved* below, for each research question.

3.5.1 RQ1.1: Have DHH ASL users used devices like this, and what has been their experience?

As shown in fig. 3.1(a), 80.2% of survey respondents had never used a personal-assistant device, with only 6 participants (7%) using such devices on a daily or weekly basis. This is significantly lower (CHI-squared test, $p < .0001$) than general public usage, as reported in prior research [128], which had found that 72% of their participants reported having used personal-assistant devices.

[8] In appendix B, we provide our survey study demographic data, questions, and affinity mappings.

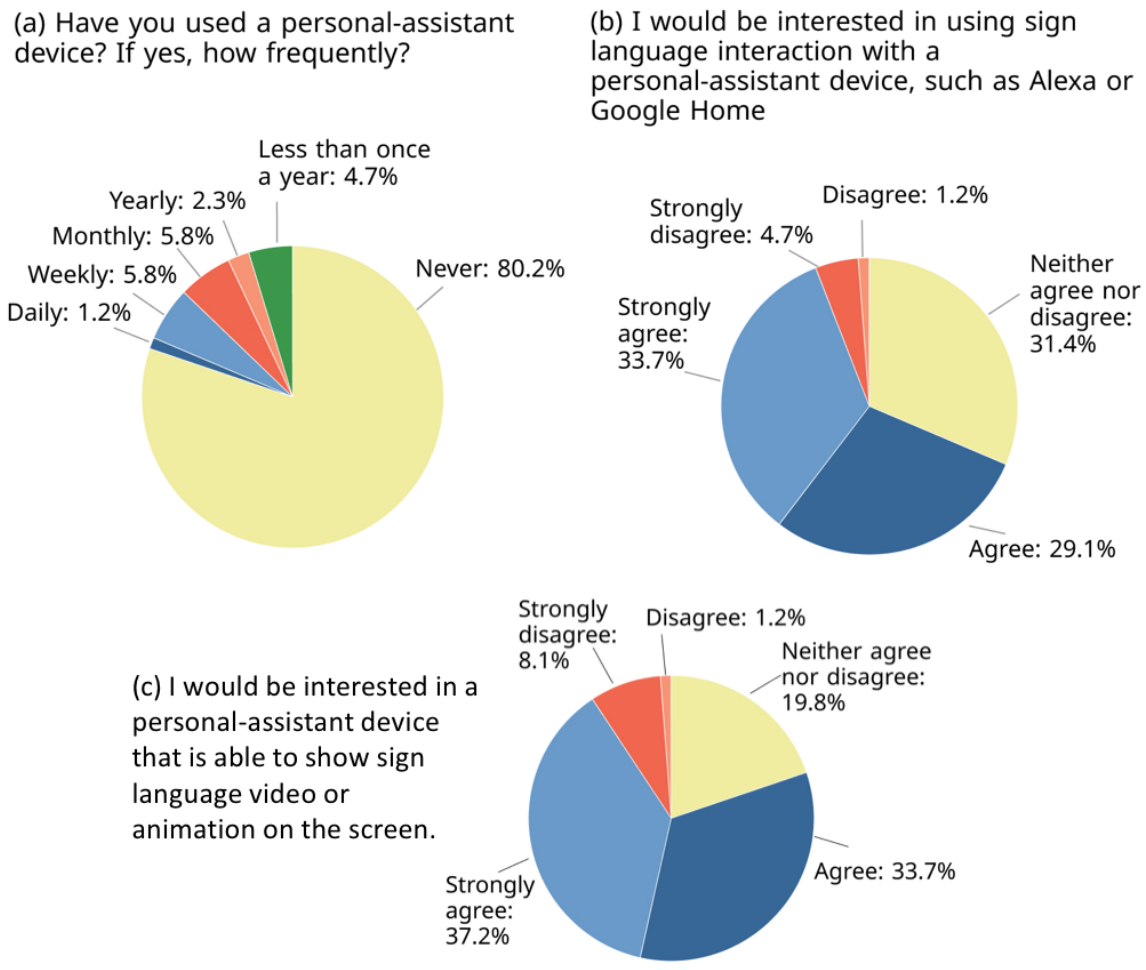


Figure 3.1: Responses from the 86 survey participants about their (a) prior frequency of use of personal assistant devices, (b) interest in using sign-language interaction with a personal assistant, and (c) interest in a personal assistant that can display sign language animation on the screen.

Among survey participants who had previously used personal assistants, 17 responded to a question about how they interacted with the device, with 11 speaking to the device with their own voice. One demographic question asked what percentage of ASL vs. spoken English participants used in their home. All 11 of these respondents reported that they used English at least 50% of the time in their home. Other modes of input mentioned included typing commands as text, using text-to-speech to generate audio commands to the device, or selecting suggested commands on the device's screen. Among the same 17 participants, the most popular usage was for weather-related requests (9 people).

A similar pattern arose in our analysis of responses from the 21 participants in our initial interview study: 9 interviewees had tried to interact with a personal-assistant device using their voice, for various purposes: to learn about the weather, to set timers, to play music, or to stream music. P3 reported using the device to control lights and home security cameras. However, interviewees reported that they had faced difficulties while interacting with the device using their voice: The device did not always understand their spoken commands, and they faced problems understanding the device's response. For instance, P3 discussed how the device did not understand them, saying, "*Sometimes Alexa doesn't pick up my voice, so I have to ask her 3 or 4 times.*" Some participants explained how the device did not seem like it was meant for them, with P8 saying, "*It is designed for people who speak, I don't speak fluently so it is not useful for me.*" or P18 saying, "*I would have to speak to it and I don't so that means I would have to touch the screen.*" P7 called for the device to provide additional input options, saying, "*Be more inclusive, allow me to sign rather than designing product for hearing.*" Overall, our results from both the survey and interviews indicate that DHH users have tried to use personal-assistant devices (at a lower rate than the general population), but in most cases, they have found such devices to be inaccessible.

3.5.2 RQ1.2: If DHH ASL signers could interact with a personal-assistant device in ASL, would they be interested in using it?

Asked if personal-assistant devices were able to understand ASL, whether they would be interested in using such devices, survey participants responded as shown in fig. 3.1(b). 33.7% of respondents indicated "Strongly Agree," and 29.1% indicated "Agree." Compared to our earlier findings about current

usage in fig. 3.1(a), these responses suggest a potential for greater usage of such devices among DHH ASL signers, if devices understood ASL commands. The survey also asked how the device should provide output, with a Likert item: "I would be interested in a personal-assistant device that was able to show me sign language video or animations on the screen." Here, 37.2% "Strongly Agreed" and 33.7% "Agreed" with this statement (fig. 3.1(c)).

We asked the interview participants if they were interested in using a personal-assistant device that would allow someone to communicate with it using American Sign Language (ASL). All 21 participants indicated they were interested in such a technology because it would make it easier for them to communicate with the device. For instance, P10 said, "*Yes because signing is faster than texting,*" and P18 commented "*It would make my life easier, I wouldn't have to control everything by touch, I could just sign.*" P16 added, "*Yes because we both would be able to communicate with each other.*"

The interview participants also indicated different scenarios where the device could be helpful. For instance, P6 mentioned that "*it would be nice to know when baby is crying and other things such as doorbell, mail, etc.*" Other participants also suggested using the device to connect with lights or security cameras, search for recipes, or learn about the weather. For instance, P3 said, "*I would use this at home because I have lights, cameras, music connected. Everything would be connected at home.*" P17 said they would "*look up recipes while in kitchen while I have dirty hands,*" and P21 said, "*Yes it would be nice so I could just be more prepared rather than not knowing what the weather is like without looking at my phone.*"

Accuracy Concerns

While there was interest in this technology, participants also expressed concerns. Some worried whether the device would understand them, due to, e.g. accents or fluency-levels in ASL. P20 said they "*would be interested only if it can accommodate with varied ASL skill-level,*" and P14 believed that the "*tech is not at the point where it can recognize various signs because people sign so differently. You would have to sign clearly.*" To enable the device to adapt to a user's unique signing style, P11 wanted a "*training session during the set up so the system can understand the varied signs.*" Participants were concerned that if the device did not understand their signing, then alternative forms of input would be cumbersome, e.g.

with P21 commenting that if the system did not understand them then "*I would have to fingerspell and I don't want to do that. It should be smooth.*" Other participants were concerned about accuracy when signing far from the camera. For example, P17 indicated interest "*if the camera could be wide enough of pick up signs from the distance,*" and P11 was concerned that "*it would require me to move in front of the device for the camera to see me, and I don't know how I would like that.*"

Privacy Concerns

Participants also expressed concerns relating to privacy, especially since the device would need a camera to understand the ASL commands. Several participants mentioned that they were concerned that the device could pick up on signs that were not meant for it. Notably, this issue of privacy is a focus of recent features of devices like Amazon Alexa, Google Assistant, or others, which include features to disable or block the microphone and camera, or provide other safeguards for privacy [128]. Since some of our initial interview participants had raised concerns about this issue, our survey questionnaire also included a few Likert items to gauge participants' opinions on this issue. Five Likert items were presented:

1. From a privacy perspective, I would be concerned about having a device with a camera.
2. The device might pick up on some signs that were not meant for it.
3. It is important to have an option of turning off the microphone sometimes for privacy.
4. It is important to have the option of turning off the camera sometimes for privacy.
5. It is important to have a physical cover to block the camera sometimes for privacy.

As shown in the diverging stacked bar graph in fig. 3.2(a), a majority of the survey respondents indicated that it was important to have the option of turning off the microphone and camera for privacy, and that it was important to have a physical cover to block the camera. Respondents also agreed that they were concerned about having a device with a camera, and that the device might pick up on some signs that were not meant for it.

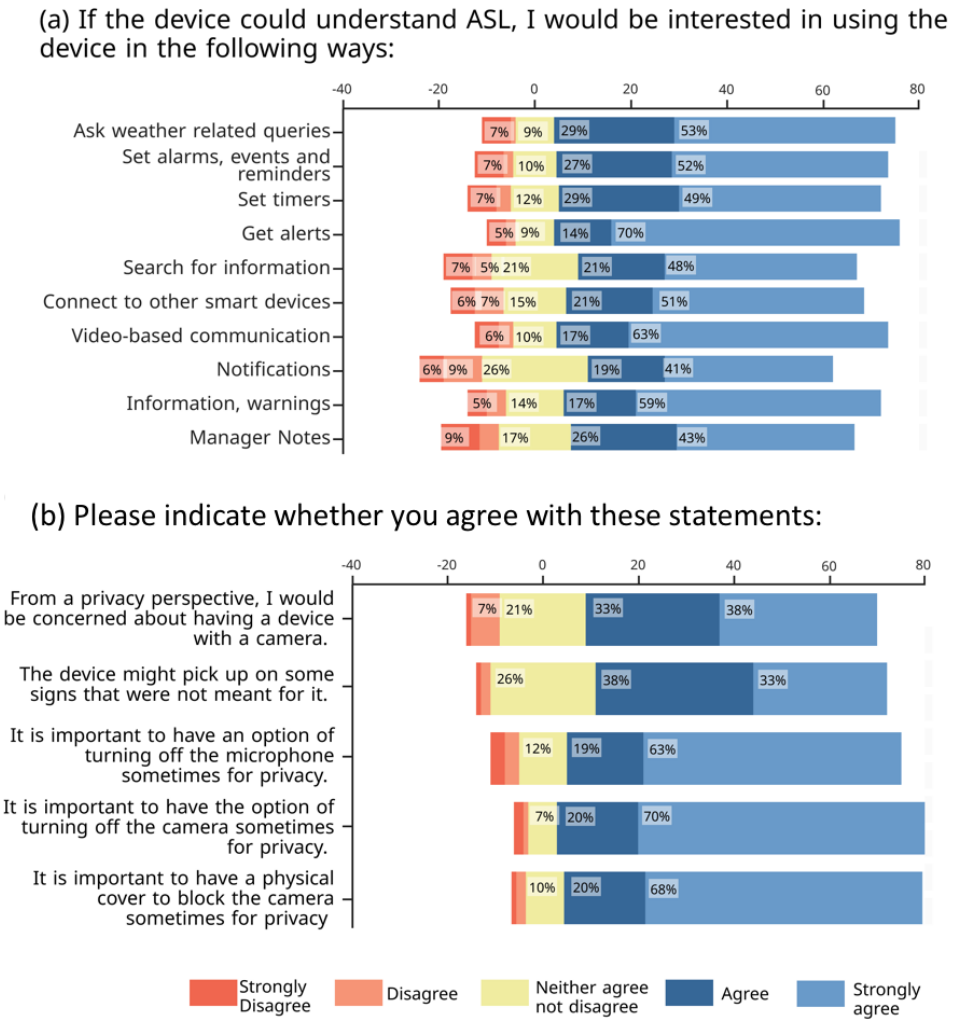


Figure 3.2: Responses from the 86 survey participants on a 5-point Likert scale about their (a) interest in various ways of using a personal-assistant device and (b) privacy concerns

3.5.3 RQ1.3: What commands would DHH ASL signers imagine using with such a device?

To address this research question, we used our initial interview study to gather open-ended responses from participants, in order to identify several popular categories of commands, which were later used to construct a question item for the later online survey, to gather quantitative data about users' interest in each category.

In our initial interviews, we asked participants to think of potential commands they might like to give to a personal-assistant device that could understand ASL. If a participant was uncertain how they would specifically say a command in ASL to the device, they were also invited to just explain the general use or purpose. After several minutes of encouraging the participant to offer such suggestions, as the flow of ideas slowed, the interviewer showed the participants a list of popular commands used by people to interact with current voice based personal-assistant devices, extracted from prior sources [89, 103, 128, 145]. The interviewer asked the participant to select the commands on this list that might be useful, and to suggest how they might issue a command in ASL to a device in this general category. The 10 categories identified through analysis of interview data appear in fig. 3.2(b), e.g. "ask weather-related questions," "set alarms, events, reminders," etc.

In the survey study, we included a question item that asked participants, for each of these 10 categories, whether they would be interested in using a device in this way, if it could understand ASL commands. Figure 3.2(b) presents the results of these Likert-items using a diverging stacked bar graph.

To understand how participants understood each category and provide guidance for future efforts to gather video datasets for training sign-recognition models, we asked participants, for each category, to share specific commands that they would be interested in issuing to an ASL personal-assistant device. Table 3.1 gives an example of the commands mentioned by participants in each of these categories. To avoid the need for users to record themselves with a webcam, participants' suggestions of these commands were collected through the questionnaire in the form of English text. Further research would be necessary to determine how users would specifically convey such commands in ASL.

Table 3.1: List of categories of commands developed from thematic analysis of interview and survey responses, with examples of specific commands suggested by participants in the survey study

Category	Selected Quotes
Ask weather-related questions	P08: What weather will it be on the drive to work? P29: What is the temperature outside at 4pm P33: Hey Google, do I need to bring my coat?
Set alarms, events, and reminders	P06: Set alarm - 7am don't forget to pack lunch. P16: Please add ultimate frisbee game on March 23 at 7pm P40: Alexa, Today Sched, When doctor (social, family, out, back, etc.) Tomorrow Sched, Friday Sched
Set timers	P9: Set a timer for 25 minutes P42: timer ring 10 min
Get alerts (e.g., doorbells, smoke alarms)	P7: Any alert? P9: Notify me of doorbells P21: Turn on the alert for doorbell P32: Is there someone at the door?
Search for information (e.g. recipes, movie times)	P21: What's the best movie recommendations? P26: Amazon delivery package information please
Connect to other smart devices (e.g., lights, TV, cars)	P62: Light on when I come in the kitchen. P70: Lights on please P79: Turn TV on and go to ESPN
Video-based communication (e.g., videophone/VRS)	P32: Call mom P20: Set up videophone for the meeting in 30 minutes. P33: Hey Google, call my mother on SVRS. P51: Who called me while I was in the shower?
Notifications (e.g., read, delete notifications)	P20: Please alert me in case of crisis emergencies. P18: Please add notifications about coronavirus. P42: flight arrive time what?
Information, Warnings (e.g., traffic, weather conditions)	P7: Is there traffic in I95 highway? P16: Is traffic near my work bad or steady? P27: Corona virus recent info P41: What time should I leave home to arrive at [place] by [time]?
Manage notes (e.g., to-do lists, shopping lists)	P9: Make a shopping list, add eggs to it P4: Please add burgers to the shopping list. P41: Show me my Wegmans shopping list.

3.5.4 RQ1.4: How would DHH users imagine waking up, interacting, and receiving responses from a personal-assistant device?

As discussed in section 2.1, interaction with personal-assistant devices consists of a sequence of steps, including, e.g., waking up the device, issuing the command, and receiving a response. To gain insight into how DHH ASL signers may prefer to engage in such interactions with a device in ASL, we discussed these aspects of the interaction with our interview participants. Since this discussion required participants to consider hypothetical technologies, this topic was more suitable to an in-person discussion with a researcher who could clarify participants' comments.

Waking-up the device

To provide the participants with context for the questions about the interaction process, the interviewer described the typical sequence of interaction steps when a user engages with a voice-based personal-assistant device, as summarized in section 2.1. The interviewer also displayed to the participant a captioned video of a hearing person interacting with a voice-based personal-assistant device – pausing the video to emphasize the sequence of interaction steps. After that, the interviewer asked the participants how they would imagine waking up a device capable of sign-language interaction and about how they would like the system to display the results of their command.

A majority of the interview participants (13 of 21) said they wanted to wave their hand in the direction of the device. In Deaf culture, waving your hand in someone's direction is a culturally acceptable method for gaining attention [151]. The remaining interview participants suggested various other ideas for how this wake-up process could occur: Some participants suggested making noise (e.g. clapping or tapping) to wake the device, e.g. with P11 suggesting "clapping or snapping or some noise to alert her [the Alexa device] to wake up." Others suggested touching the device itself to get its attention, with P21 explaining that they would "touch the device, similar to how we tap people's shoulders to get their attention. It will sign and caption 'Yes I am awake now.'"

After participants provided their own initial suggestions for how the wake-up process could occur, the interviewer offered participants a list of options for how a device could be awoken, so that participants

could react to each. This list included: hand-waving in the direction of the device, signing the device's ASL name-sign (an ASL sign that could represents the name of the device), fingerspelling the device's English name, signing a multi-word "wake-up" phrase in ASL, using some a physical gesture that is not an ASL sign, pressing a button on a remote, pressing a button on a smartphone app, pressing a button on a smartwatch app, making noise with their hand (e.g. clapping, tapping, snapping), or sending a text message to the device. While prior to considering this list of options, a majority of participants had suggested hand-waving methods for waking the device, after considering these options, participants' interest shifted toward push-to-talk style methods, which require physically touching the device or pressing a button.

Participants interested in physical-touch methods often mentioned concerns about false-positives or false-negatives in the detection of wake-up requests: In false-positives, the device may wake up when the user had not intended it to do so, perhaps due to the system incorrectly detecting signing or gestures. P15 believed that a touch-based method "*feels more in control, because if you use the wave and someone is near the Alexa and signing, she [Alexa] might hear the conversation and think to look up that conversation.*" Similarly, P17 worried "*if I am waving to get other people's attention then the device will wake up and I don't want that.*" In false-negatives, the device might miss the user's attempt to wake it. For instance, P20 preferred "*touch, [because it] will ensure that the device will wake up. If I wave maybe the camera won't recognize it.*"

A minority of participants were still interested in hand-waving based wake-up, with most explaining the convenience of not having to depend upon finding their phone, smartwatch, or remote in order to start an interaction with the personal-assistant device. Participants explained that they would need to remember to have another device with them, which would not provide a functionally equivalent experience as that of hearing people. For instance, P12 indicated that "*you might not have your phone or what if you are trying to use her to find your phone. Also, you may not be nearby so waving to get her attention will be quicker.*" P21 suggested that "*[Waving] is the easiest for me because then I wouldn't have to look for my devices to use Alexa.*"

Device Response Modality

To understand DHH ASL users' preferences for the response modality of personal-assistant devices, we asked interview participants to suggest how they would want the device to show them results and answer their queries. We collected these ideas and presented them to the survey participants in the form of a multiple-choice question. Figure 3.3(a) displays the number of times each of our 86 survey participants selected each option on this multiple-select question; the text-based result was the most preferred form of response modality. Examples of "other" options written in by survey respondents include: "output sent to printer, like fax or email message" and "colored flashing lights."

The preference among survey participants for text-based output was notable, given the results presented in section 3.5.4, which indicated that participants were interested in receiving output in the form of ASL. To understand why they may still prefer text-based output, we examined data from our interview study, in which we had also asked participants about how they would prefer to receive a response from the system. Similar to the survey results in fig. 3.3(a), a majority of interview participants (13 out of 21) indicated that they would prefer text-based output. Some participants explained that they believed text output would be faster to consume than ASL output, with P14 explaining that "*text would be the best one, because if it's ASL video, then there could be a lag due to many factors. Text would be faster and easier.*" Others preferred text output because they would not need to devote their full attention to the device to the same degree as if it were displaying results in a more dynamic modality, e.g. with P22 saying, "*I feel like text would be faster than reading because I may be busy doing other things. So if it's sign, I would have to fully attend to the screen, whereas reading I can quickly skim.*" A few interview participants specifically mentioned concerns they may have in regard to ASL-animation output modalities, with P9 indicating that a system displaying ASL output should "*show captions too.*" P9 had concerns about the speed of ASL animation output and said, "*If its animations, I would be concerned that it would be too fast or too slow.*"

Placement of the Device in a Home

Since placement of a personal-assistant device [145] affects how the device is used, survey participants were asked to respond to an open-ended question about where they would set up such device in their home and why. Following the analysis method described in section 3.4, responses were categorized into various household locations: The top four places in the home where users indicated they would place such a device were the living room, kitchen, bedroom, and home office. Counts of how many participants mentioned each room are displayed in fig. 3.3(b); notably, respondents could suggest multiple rooms in which they would place the device, with all rooms mentioned by a participant contributing to the data in fig. 3.3(b). When explaining why they would place the device in a particular room, many respondents mentioned the "living room," explaining that they spend the most time there; others mentioned that it was a central location in the house or a larger space. For instance, P40 explained how this location would provide access to key information, saying, "*Living room. I see its value as information such as news, weather, etc.*" P56 indicated they would place the device in the kitchen, since it "*is where there's no tv/video screen I prefer to watch it while I cook.*" Although prior research had not included quantitative survey data enabling a direct comparison, the locations in the home that hearing individuals mentioned in [128] included similar locations e.g. living room, kitchen, etc. Whereas, some participants might have less sound-awareness when they are in their home, possibly due to not using an assistive hearing device (e.g. hearing aids, cochlear implant) while in the personal comfort of their home. Thus, they might want to place the device in specific rooms in the house, e.g. P76: "*in my room because I am often deaf in there.*"

In our initial interview study, we had asked participants about their thoughts and concerns around the placement of the device in their home. When discussing this issue, a majority of the interviewees mentioned concerns regarding the device not being able to see their signing, e.g. with P11 being "*mostly concerned with not having an open space that would block the view of the system.*" This issue of being visible to the device also led some participants to worry they would not have the same experience as hearing users, who are able to use voice-based devices without the need to physically approach the device, e.g. as P12 explained, "*I am concerned that I still wouldn't get the same experience as hearing people. They*

don't have to get up and be near the device to interact with it while I would have to get up and walk to the device for it to even see me sign."

Visibility was also on the minds of other interview participants, who mentioned concerns about lighting in various rooms, e.g. P9 wondered, "*Will I need to have a light on for it to see me or how will I know the room is bright enough?*" Similarly, P33 said they "*would place it under a well-lit spot, where it can clearly see my ASL commands.*" As discussed in section 3.5.2 above, some interview participants raised concerns about privacy with having a camera device in their home, and for some, this influenced their thinking about where they would place the device. For instance, P30 explained that they would put the device in their "*kitchen because that's the least place that I need privacy. I'm not sure how I would feel with a camera in my bedroom at all times who can understand me.*"

Overall, our findings as to where DHH signers would place a device in their home revealed that the most popular locations were the kitchen or living room. The reasons mentioned by participants for this choice were varied: Some participants mentioned issues that might be relevant for any user, e.g. having the device in a room where the user spends a lot of time, but other participants mentioned issues relating to the sign-language modality of interaction, e.g. visibility, lighting, or privacy concerns.

3.6 Discussion

Our findings provide **motivation** for HCI researchers and designers to explore technology for ASL interactions with these personal-assistant devices. Our study investigated DHH individuals' **prior use** of personal-assistant devices, with our survey revealing that over 80% of DHH respondents reported having never used a personal-assistant device before (fig. 3.1(a)). When compared to the general population, as reported by Microsoft Voice report (62% of respondents have used digital assistant and 72% among them used using voice search), our DHH participants' had significantly less experience in using these technologies. These devices pose an accessibility barrier due to their voice interactions. As deaf speech is not clearly recognizable to automatic speech recognition [60], DHH users face challenges. Additionally, survey participants indicated on average they use ASL more than 50% of the time in their interactions with family, friends, and at school or work. We found that despite the limited prior usage

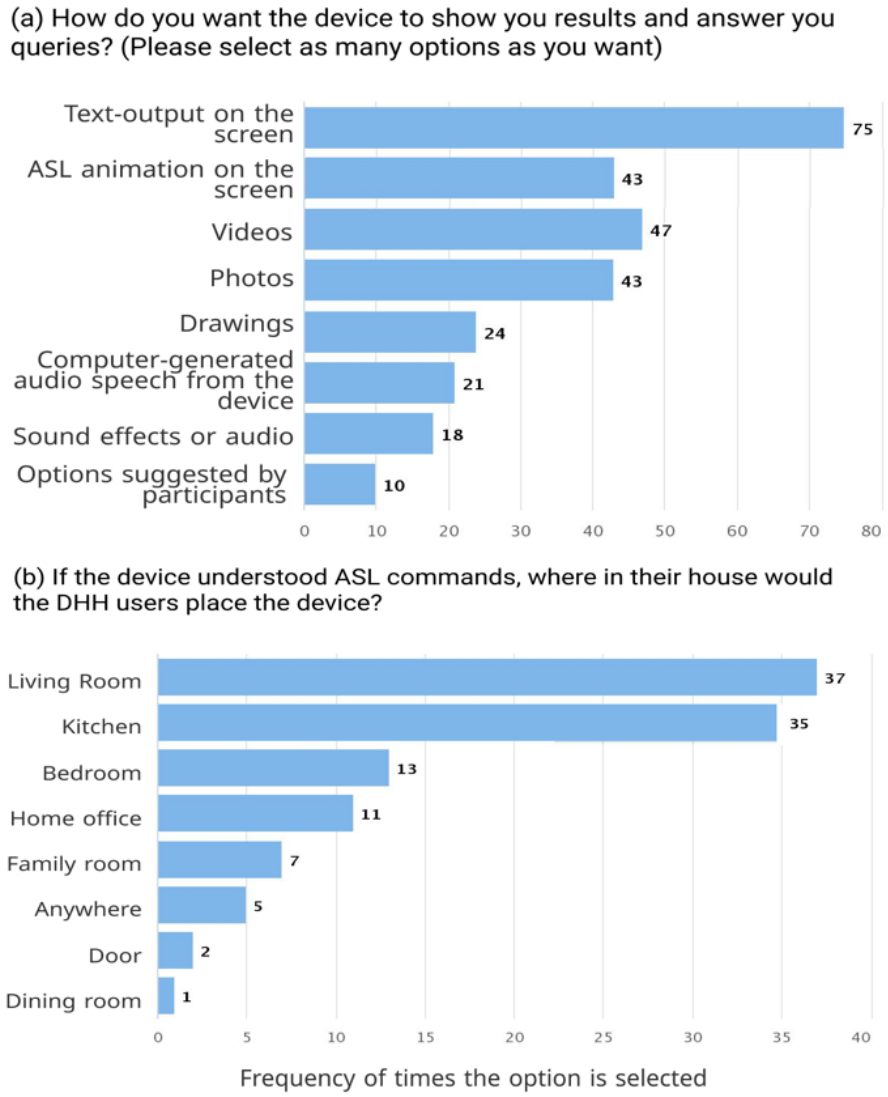


Figure 3.3: Frequency of participants who (a) selected different result mediums and (b) would place the device in the different locations in their houses

experience among DHH individuals, there was **great interest** in using personal assistant technologies if such devices supported sign-language interaction, e.g. with 33.7% of participants indicating "Strongly Agree" and over 29.1% "Agree" (fig. 3.1(b)). A notable aspect of this finding is the comparison between the rates of prior use, as compared to the greater rates of interest if this technology could support sign-language interaction. This suggests a potential benefit if sign-language interaction capability could be added to these systems.

To provide guidance for future developers, it is important to understand whether DHH users may be interested in using these devices for **different purposes than the general population**, thereby motivating new features, rather than only making existing features accessible. As presented in section 3.5, respondents reported that they would be most interested in using the device to get alerts about the surroundings, e.g. doorbells and alarms. This finding is in alignment with prior research on the interests of DHH individuals in smart-home technology for maintaining an awareness of surrounding ambient sounds [56, 112]. Using a personal-assistant device for alerts was of interest to DHH participants, but this may differ for hearing users. Similarly, DHH ASL users often use video-based communication, and expressed great interest in making use of the device for video-based communications. DHH ASL users wanted devices to connect to other video-based communication platforms, including Deaf-specific applications such as SVRS, a video relay application [155]. Participants were also interested in "weather-related queries," "set alarms, events, and reminders," and "set timers." Similar categories of use had been reported in prior studies on personal-assistant devices among the general population [89, 103, 145].

To create sign-recognition technology, a video dataset is needed of DHH users issuing ASL commands. To avoid naive assumptions about **which commands would be important to collect** in such a dataset, this type of foundational research about DHH users' interest in specific commands is needed.

Our study also provided some **preliminary insights for future designers** of ASL devices. For instance, DHH users share their views on how to **wake up** the device. A majority of interview participants suggested using hand-waving, a common method for gaining attention in Deaf Culture. Prior research on voice-based personal-assistant interactions had examined push-to-talk (press a button before speaking a command) or talk-to-talk (begin speaking, potentially with a 'wake word' to issue a command) as

methods of waking up devices. Certainly, the suggestion of tapping the device to gain its attention is similar to push-to-talk approaches, and we also see an analogy between the hand-waving suggestion and the typical talk-to-talk approach, as the use of hand-waving in Deaf culture is similar to how someone may use their voice (e.g. calling someone's name or saying "Hey!") to gain attention.

Our study had also investigated users' preferences for how to **receive output or results**. In fact, we observed a contrast here: When asked if they would be interested in ASL output from devices, users had indicated strong interest (section 3.5.2), but when presented with a list of various output modality options (section 3.5.4), they indicated a preference for text-based results. The interview study shed light on this contrast: A majority of interview participants (13 out of 21) also preferred text-based output, and they explained that they believed text output would be faster to consume than ASL output. Participants also had concerns about whether ASL-animation output would be understandable.

Finally, as part of our investigation of how users would like to interact with a personal-assistant device in ASL, we also asked users about **where they would place the device in their home**. In prior research among the (majority hearing) general population, Sciuto et al. [145] found that current users of personal assistants place devices in various locations, and this placement affected how devices are used. As discussed in section 3.5.4, if personal assistants were capable of understanding ASL commands, the most popular placement in homes among DHH ASL users would be the kitchen or living room; these locations are also popular placements of devices among the general population [145]. When asked to discuss their rationale for this placement, DHH participants mentioned some considerations that would be common to all users, e.g. having the device in a room where the user spends a lot of time. However, they also discussed issues affecting device placement that were specific to the sign-language modality of interaction. Since the camera on the device would need to capture video of the DHH user, a device may need to be placed strategically so that it has a good view of the user. Participants discussed issues of visibility, lighting, and privacy concerns. Notably, some recent work with DHH users has investigated sound-awareness device placement in homes [87], and as many DHH participants in our study expressed interest in their personal-assistant device being used for sound-awareness applications, the findings of that related work may also inform the issue of personal-assistant device placement among DHH users.

More broadly, the findings in our study have **implications for the field HCI**. For instance, this chapter contributes to a broader theme within the field of computing accessibility that designers should not assume that the needs and preferences of a sub-group of users will be identify to those of the general population. In this case, designers may assume that DHH users' preferences or patterns of use would be similar to the general population, and one contribution of this chapter is evidence this is not the case.

3.7 Conclusion

This chapter investigated DHH users' views about ASL interaction with personal-assistant devices. The studies presented included engagement with DHH users to understand how they would be interested in using devices capable of accepting sign-language input commands, i.e. desired features, usage scenarios, or other expectations for such systems. Interviews were conducted with 21 DHH ASL signers, and then the interview questions and thematically-analyzed responses informed an online survey launched nationally, for which we received 86 DHH ASL signers from over 20 U.S. states.

This work provides a basis for future phases of research to investigate DHH user's interactions with these devices, and to construct a dataset of videos of sign-language personal-assistant commands, which may be useful for sign-language recognition researchers. Broadly, this research contributes to improving the accessibility of conversational-interaction user-interfaces, an increasingly ubiquitous mode of interaction for personal-assistant devices.

The main empirical contributions of this work include:

1. Evidence that DHH users use personal assistants significantly less than the general population
2. Evidence of DHH users' interest in ASL interaction with such devices
3. Prioritized list of "categories of commands" DHH users are interested in issuing to a personal assistant, as well as a list of DHH-specific user cases for personal assistants
4. Evidence of DHH users' privacy concerns with camera-based interaction in their homes
5. Initial user reaction to wake-up interaction approaches and response-display for ASL personal-assistant devices

Chapter 4

DHH Users' Preferences Among Wake-Up Approaches during Sign-Language Interaction with Personal Assistant Devices[9]

4.1 Introduction

As briefly explained in chapter 2, personal-assistant devices are becoming popular and ubiquitous. These physical devices, e.g., smart speakers or smart screens, respond to user queries; these devices provide information or/and enable control of other smart devices. To interact with a personal-assistant device, the user needs to get its attention, typically by saying a wake-word, i.e. "Alexa" for Amazon Alexa or "Ok, Google" for Google Assistants. Once the device is ready, the user issues the command.

Interaction with these devices is typically voice-based and poses accessibility barriers for people who are DHH, many of whom would prefer sign-language interaction, rather than text input or non-

[9]The information in this chapter is based on a project where I advised and guided a graduate student at RIT (Vaishnavi Mande), working with my advisor, Dr. Matt Huenerfauth. I collaborated on the study and stimuli design, conducted the data collection, and advised and assisted in data analysis and writing a paper published as a late-breaking work submission at the ACM CHI'21 conference [111].

sign gestural input [139]. Given advances in computer vision technologies [23], HCI researchers are beginning to consider future device-interaction using American Sign Language (ASL) commands [15, 61]. Chapter 3 of this dissertation asked people to imagine how they might wake up a device in a survey-based study, but did not have the ability to clarify what users intended nor to show people mock-ups of how this might work.

In this chapter, using a formative interview study with 21 DHH ASL signers, we identified 6 approaches for wake-up interactions for potential sign-language-enabled personal assistant devices. We evaluated video prototypes of these 6 approaches with 12 DHH ASL signers, and a qualitative analysis revealed key attributes users' considered when selecting their preference of a wake-up technique. The empirical contribution of this study is in identifying the preferences and concerns among DHH users in regard to this new form of interaction, which provides guidance for future designers of these systems.

4.2 Wake-up Interactions

Personal-assistant devices can be thought of as spoken dialogue systems, which typically enable question-answer interaction with users [41]. This interaction requires the user to first obtain the device's attention, before issuing a command, a process that is referred to as "waking up" the device. Generally, the user calls to the device by speaking a wake-word; calling the device by its name is the most commonly used wake-word technique [136], i.e. "Alexa" or "Echo" for Amazon devices or "Ok, Google" for Google devices. This interaction may be categorized as a "talk-to-talk" method. Alternatively, some devices support "push-to-talk," whereby a user may press a button to invoke the personal-assistant device without speaking a wake word [8].

Relatively little prior work has focused on this wake-up interaction. One study identified usability problems with wake-words, e.g., the need to construct a sentence to place the wake-word first, the robotic nature of the wake-word, or even accidental device wake-up due to similar-sounding words [4]. The authors proposed avoiding the use of wake-words and propose alternate approaches; however, this work did not consider DHH users nor sign-language interaction. Another study investigated the effectiveness of wake-up techniques for conversational agents among children, comparing several ap-

proaches: a wake-word, pressing a digital button, pressing a physical button, gazing towards the device, using a mouse pointer on the device screen, and other techniques [32]. That study revealed that among users who are children, a physical button (a push-to-talk technique) was the most appropriate solution.

No prior research has investigated the device wake-up process among DHH users, especially in the context of sign-language interaction. Within the cultural and linguistic context of American Sign Language (ASL) users, it is useful to consider analogs of various push-to-talk and talk-to-talk methods, as well as the typical ASL dialogue structure. In U.S. Deaf culture, it is acceptable for an individual to tap someone gently on the shoulder to get their attention. If beyond the reach to tap, someone may wave their hand in the air, in the direction of the person, until eye contact is established [151]. As ASL is a visual language, individuals must ensure that there is proper lighting and line-of-sight such that their conversational partner may clearly see their manual signs and linguistic facial expressions, e.g., avoiding standing in front of bright light or window [151]. With this context in mind, we conducted studies to explore how DHH users may prefer to wake up personal assistant devices if they were to interact with those devices in ASL.

4.3 Study 1: Formative Interviews

We conducted interviews with 21 DHH ASL signers to collect ideas and recommendations about how users would like to wake up a personal-assistant device, with which they could interact in ASL. Results from these interviews were used to identify potential wake-up interactions that we evaluated in a subsequent study.

4.3.1 Methodology

Participants

We recruited 21 DHH ASL signers (8 female, 12 male, and 1 non-binary) from our university through poster advertisements. Each interview was scheduled for 30 minutes and was conducted face-to-face in ASL by a DHH researcher from our lab. Our participants were between the age of 18 to 25, and all had

some college education. Most participants had very little experience with personal-assistant devices, but all reported having tried a personal-assistant device at least once. There was 1 participant who owned 6 personal-assistant devices and used them regularly. Table 4.1 shows the participant demographics and prior experience with personal-assistant devices in this study.

Table 4.1: Study 1 participant demographics and prior experience with personal-assistant devices

PID	Gender (Age)	Identity	English used at home/ school	Education	Familiar with the device	Ways of interaction	Problems faced in the interaction
1	M(18)	deaf	37.50%	BS	Not used	No	-
2	F(24)	deaf	90%	BS	Hearing friends and family	Not used	-
3	NB(24)	HH	70%	MS	Have a few, 6 or 7; Google home, Alexa	Almost everyday with voice or my phone	Understanding what they're saying. Sometimes Alexa doesn't pick up my voice
4	M(24)	Deaf	37.50%	BS	Friends have device	Not used	-
5	M(24)	HH	35%	BS	No	Not used	-
6	F(21)	Deaf	25%	BS	No	Not used	-
7	F(22)	Deaf	0%	AAS	No	Not used	-
8	M(22)	deaf	50%	BS	Seen it on TV	Not used	-
9	F(21)	deaf	37.50%	BS	Not used	-	-
10	M(23)	HH	25%	BS	Seen it on TV	Not used	-
11	F(23)	deaf	5%	BS	No	-	-
12	F(25)	deaf	75%	BS-MS	Have 1, parents have 1	Everyday with Voice	Doesn't understand me in first try
13	F(36)	HH	0%	BS	Seen it on TV	Sometimes, with Voice	Sometimes it doesn't understand me like 's' words/sounds
14	F(23)	deaf	37.50%	BS	Parents	Once a month maybe using Voice	I can't always understand what it says
15	F(24)	Deaf	0%	BS	Seen it on TV	No	-
16	F(25)	deaf	50%	BS	Friends and family have one	Tried it before using Voice	Couldn't understand my voice and I couldn't hear the device
17	M(28)	HH	20%	NA	Seen it on TV	Once tried with my voice	Maybe my voice isn't clear enough
18	M(24)	deaf	7.50%	BS	No	Not used	-
19	F(29)	deaf	5%	AAS	Seen it on TV	Not used	-
20	F(23)	HOH	70%	BS	No	I tend to talk but if not I will type	Deaf accent sometimes harder to understand
21	M(23)	deaf	37.50%	BS	Family own Alexa	Used it once and it didn't work Voice	Used my voice but Alexa did not understand me

Procedure

In the interview, we asked questions about participants' familiarity and usage experience with current voice-controlled personal assistant devices, their expectations for interacting with these devices in sign-language, their ideas about possible wake-up approaches, and concerns they envision with such interaction.

During the interview, the interviewer demonstrated to participants the typical steps involved in interacting with a voice-based personal-assistant device, by displaying a captioned video of a user engaging with a voice-based device. The purpose of this video was to provide participants with context about

how the wake-up process typically occurs. The researcher paused the video to indicate the initial wake-up phase of the interaction, to clarify the specific portion of the interaction that was the focus of the interview.

Interviews were transcribed into written English for analysis, and an affinity mapping methodology was used to identify users' ideas for waking-up the device. In this process, participant quotes were organized and grouped, and our analysis resulted in identifying major types of wake-up interactions that had been mentioned by participants.

4.3.2 Study 1 Findings

Our analysis revealed that users' envisioned six major types of wake-up interactions. Four can be classified as talk-to-talk approaches, i.e. signing the ASL sign-name of the device, waving in the direction of the device, finger-spelling the device name using the English letters, and clapping to get the devices' attention. (Details of each are discussed below.) The other two approaches are push-to-talk techniques, i.e. using a phone app to trigger the device or using a physical remote control. These six types of wake-up approaches were investigated further in Study 2 (section 4.4).

Talk-to-talk Techniques

A majority of the participants (13 out of 21) suggested using talk-to-talk methods, such as using an ASL sign or waving in the direction of the device to wake it. As mentioned above, waving one's hand in someone's direction is a culturally acceptable method for gaining attention in Deaf culture [151]. Users also suggested waving in a specific pattern to wake-up the device, using the device name in the form of sign-name (a unique ASL sign used to uniquely identify someone), or fingerspelling the English letters of the device name. Users expressed concern that commonly used signs or waving might lead to accidental device wake-ups. For instance, P17 mentioned, "*what if I am waving to get another person's attention then the device will wake up and I don't want that.*" Few participants suggested making noise (e.g. clapping or tapping), for instance, P11 suggested "*clapping or snapping or some noise to alert her [the Alexa device] to wake-up.*"

Push-to-talk Techniques

Other participants (8 out of 21) suggested push-to-talk techniques, i.e. using a physical button to get the device's attention by pressing a button on another device, e.g., a smartphone app or a physical remote control.

For example, P11 said, "*I like [using] the phone app because it is easy to control,*" and P08 suggested using a physical remote control that is paired with the personal-assistant device, commenting "*A remote or something to press.*" Participants interested in push-to-talk methods mentioned how these wake-up approaches were more reliable. Specifically, users mentioned how push-to-talk approaches could avoid false-positives (the device waking up when the user had not intended it to do so, perhaps due to the system incorrectly detecting signing or gestures) or false-negatives (the device missing a user's attempt to wake it). For instance, P20 preferred "*touch, [because it] will ensure that the device will wake up. If I wave maybe the camera won't recognize it.*"

4.4 Study 2: Video Prototype Evaluation

While Study 1 had enabled us to collect some ideas from users who imagined how they might wake-up a personal assistant device that understands sign language, there was a limitation in that study. Specifically, participants had to imagine their interaction. To provide a means for participants to better visualize each type of wake-up interaction without being overwhelmed by the personal-assistant device interaction, we developed video simulations in which a DHH actor demonstrated using each of the six wake-up techniques (see fig. 4.1 for a video storyboard). By displaying these video prototypes in Study 2, we hoped to gain further insight into the factors DHH users had in mind when they considered which wake-up approaches they preferred.

In order to create the 6 video simulations of each wake-up techniques, we filmed a DHH actor interacting with a personal assistant device. This was a Wizard-of-Oz set-up where a hearing person was voicing commands to the device while we recorded a DHH actor pretending to issue commands to an Amazon Echo Show device in ASL. The video recording location, device placement, command given to the device and actor were constant in all the simulations. Only the wake-up technique changed with

each video. Figure 4.1 shows the video storyboard layout and screenshots of the six wake-up techniques.

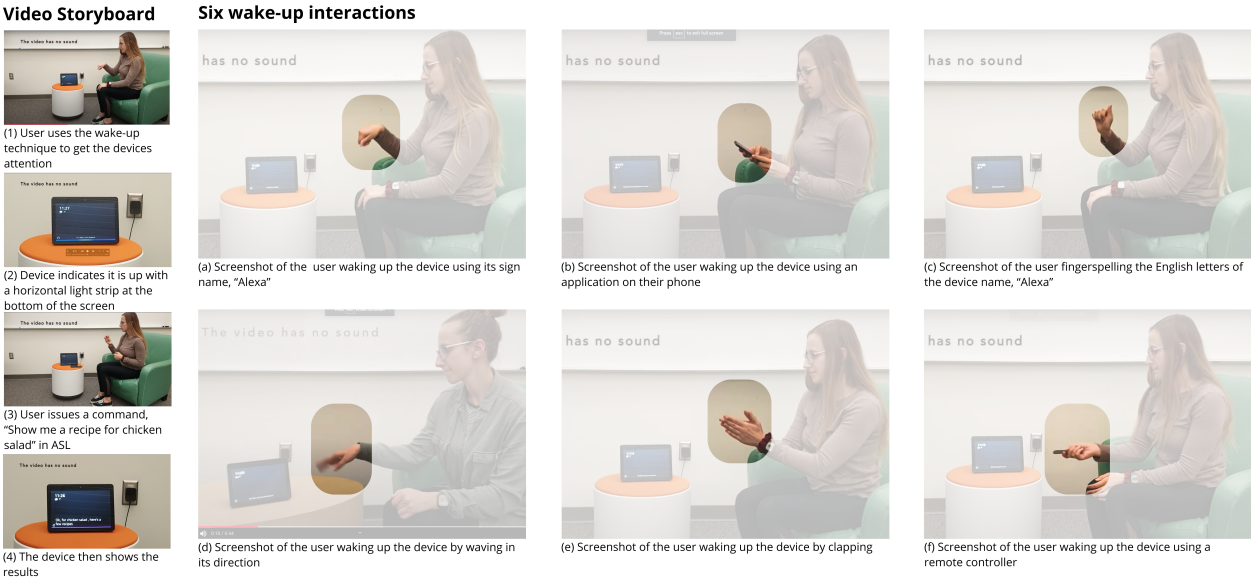


Figure 4.1: Video storyboard with the device-user interaction steps: (1) user uses the wake-up technique (2), device wakes-up, (3) user gives the command, and (4) device responds. To the right are screenshots of the actor using the wake-up techniques: (a) using the sign-name technique, (b) using the phone application technique, (c) using the fingerspelling technique, (d) using the wave towards the device technique, (e) using the clapping technique, and (f) using the remote technique.

4.4.1 Methodology

Participants

We recruited 12 DHH ASL signers (6 Male and 6 Female) who were in the age range of 21 to 29. All participants indicated they were aware of what personal-assistant devices were, but a majority of the participants (10 out of 12) did not have any experience of using such a device. The remaining two had used a device using their voice. Table 4.2 gives participant demographics and prior experience with personal-assistant devices.

Table 4.2: Study 2 participant demographics and prior experience with personal-assistant devices

PID	Gender	Age	Identity	Usage Frequency	Familiar with the device	Ways of interaction
1	Female	24	deaf	Friends	No	-
2	Male	25	HH	I have Google Nest, Amazon Alexa	Yes, daily	Speak using my own voice (90% of the time). Other times I use the touch screen or manually use the phone app
3	Female	36	HH	Commercial	No	-
4	Female	23	Deaf	Friends	No	-
5	Male	25	Deaf	No	No	-
6	Female	24	Deaf	Family	No	-
7	Female	25	Deaf	Grandma	No	-
8	Female	29	deaf	Parents	Yes, once	Texted through phone app
9	Male	23	HH	Friends	No	-
10	Male	24	deaf	Commercial	No	-
11	Male	21	deaf	Home + Bixby on phone	Yes, very often	Speak to the device
12	Male	29	deaf-blind	Store	No	-

Procedure

We conducted a within-subject evaluation of the video prototypes, with the one independent variable being the wake-up technique. The sequence in which the wake-up technique videos were shown to the participants was counterbalanced via a Latin Square schedule. Each session lasted 45 minutes, and participants were compensated with $40. The session was conducted in ASL and later transcribed in English for analysis.

We collected demographic data, including participants' familiarity and experience with personal-assistant devices. Next, we presented and discussed each video-simulation. For each, we asked the participants to share their thoughts on the technique, including any benefits or problems they envision. We analyzed the transcriptions of the sessions using an affinity-mapping methodology, by inductively grouping the participant quotes, based on the various trade-offs or factors discussed.

At the end of the study, we asked participants to rank the six wake-up techniques from least- to most-preferred. We encoded these ranks with integers (1 to 6), and then we summed the responses for each technique for analysis.

4.4.2 Study 2 Findings

Based on participants' ranking of wake-up techniques, using the ASL sign-name of the device was the most preferred. The remaining techniques in descending order of preference were: waving in the direction of the device, clapping, using a remote control, using a phone app, and fingerspelling the English name of the device.

The affinity-mapping analysis of participants' open-ended responses revealed that participants were concerned about various factors when comparing their preference for each wake-up approach. Factors mentioned by participants included: whether interaction success depends upon the surrounding environment (e.g. lighting), whether the interaction depends on availability of another device, the reliability of the technique, how convenient it would be to use that technique, if it was easy to use, the speed of the technique, or the speed with which the device would wake-up. Users' responses are summarized below, presented in the preference-ranking order determined during the study.

Using the device sign-name

As discussed above, a sign name is an ASL sign that is used to uniquely identify a person. Participants preferred the idea of assigning a sign-name to the device and then waking up the device whenever they produce that sign-name. Participants indicated that using a specific ASL sign solely for waking the device may avoid accidental wake-ups, which users believed would be more likely using other wake-up techniques like waving or clapping. P1 said, "*sign name is more specific than the wave... If another person had the same sign-name it could become an issue but I think that is rare.*" Participants liked that this technique would be fast, e.g., P9 indicated that, "*there is not a lot of unnecessary time needed, similar to wave.*" Participants also mentioned how this approach would be more convenient than fingerspelling English letters of the device name. Participants also commented how this technique does not require the user to carry an additional device. However, participants did note that this wake-up technique is dependent upon the surrounding environment, i.e. having sufficient light and being in the camera-range of the device. For example, P2 noted, "*Sometimes the device may not be able to see the sign in dark,*" and P9 stated, "*my concern is how good the device would recognize me signing the name across the room.*"

Waving in the direction of the device

Section 4.2 discussed how, in Deaf culture, it is common to wave in someone's direction to get their attention. Participants indicated this wake-up technique would feel comfortable and natural, like interacting with a person. P7 pointed out that "*(the wave method) keeps your hands in the same spot during waving and then signing.*" Additionally, participants discussed how waving would be more convenient for people who prefer to use sign-language over English. As P1 mentioned, "*I think it could benefit DHH especially those who prefer sign over written English.*" Participants also liked how this technique was not dependent upon the user having an additional device. However, a majority of participants noted that this wake-up method is susceptible to accidental wake-ups; P11 said, "*something in the background may get Alexa's attention like if a cat waves at it and Alexa may get its attention*". Similarly, P7 said, "*I am concerned if I were to wave to someone else if the device would accidentally wake up.*" Similar to the sign-name method, participants noted that this method depended upon the lighting and distance to the camera.

Clapping to wake-up the device

Similar to other talk-to-talk methods (using the device sign-name, waving, and fingerspelling), participants mentioned how they liked that the clapping technique did not require the user to carry another device. Participants commented that they found this method to be simple to execute, fast, and comfortable. While participants mentioned that this approach would work regardless of lighting or camera distance, several participants did mention that they would require training to select the appropriate loudness of clapping. P6 mentioned, "*In the beginning I would have to figure out how loud I would need to clap but eventually would figure that out.*" Also, similar to the waving technique, participants mentioned how clapping is susceptible to accidental device wake-ups because of background noises. P9 noted, "*Alexa can't detect whether people are clapping for fun or clapping for her attention.*" To avoid that, P12 suggested a pattern of claps for waking up of the device, saying "*maybe set up how often you need to clap like 1 or 2 times.*" As clapping by different people may sound alike, some participants were concerned that anyone could access and operate the device.

Using a remote to get the devices' attention

Many personal-assistant devices come with an additional remote-control device, which can be used to trigger the system. Using a remote to wake the device would be classified as a push-to-talk technique. Participants liked that this wake-up method would work regardless of surrounding lighting or distance to the camera. They also liked that this approach would be fast, and it would likely avoid false positives or negatives. P12 suggested that this approach may be preferred by older users, saying "*The remote is a good replacement for those who are senior citizens or people who are annoyed with fingerspelling.*" Participants also noted that this technique may feel familiar, as remote controls are ubiquitous, e.g., P6 said, "*(Remote technique) is easy as we already use remotes and are okay with this concept.*" Despite these advantages, participants did not rank this approach highly. Many noted that this approach would not provide DHH users with a hands-free experience (like talk-to-talk methods). As P8 said, "*I would like the remote options but going through these options, I don't think it is the best option. Plus, one of the biggest appeals about Alexa is that you just have to say her name you don't need a remote or get up.*" Participants were also worried about misplacing the remote or the remote battery dying.

Using a smartphone app to wake the device

When discussing this approach, participants raised several factors similar to those when using a remote control, e.g., commenting on how this may be a fast or reliable method of waking the device. Additionally, participants believed this approach would be easy to use, e.g., as P10 said, "*everyone has their phone with them so I think it would be easier to use that*". Similarly, P12 said, "*I would use it for smart home kits like smart home devices, and [it would have] less errors so I know Alexa would wake up right away.*"

P8 also suggested that there are "*situational benefits like if you can't sign to Alexa, you can text (the command using the phone app).*" Similar to the remote-control wake-up approach, participants noted that they might misplace their phone or its battery may die. Participants suggested that using a phone app might be useful as a reliable backup approach for waking a device, e.g., as P11 said, "*I would use the phone app if Alexa didn't catch my signing.*"

Fingerspelling the device name

Fingerspelling the device's English name was the ranked as the least-preferred method by participants. Although participants noted that this wake-up method would be hands-free (i.e. nothing to hold in the hand or touch) and would not depend upon the user having an additional device, they also noted that it may be slower and more error-prone, e.g., due to spelling mistakes. P10 wondered, "*I am curious how picky Alexa would be like what if I misspelled her name.*" Participants noted that this method may be less convenient for people who prefer sign to English. P1 said, "*It takes a little bit longer to spell name and may not be as efficient for others who may have difficulty with fingerspelling,*" and P8 was concerned, "*Some people have a hard time moving their fingers so fingerspelling would be no good.*" Participants were worried about the device's accuracy in detecting fingerspelling, e.g., with P12 saying, "*sometimes it (fingerspelling) can become sloppy.*" Additionally, participants noted that this wake-up approach would be dependent upon the lighting and distance to the camera.

4.5 Discussion

The findings of our two studies have revealed preferences and concerns of DHH users for how to wake up future personal-assistant technologies that could understand sign language. A key contribution of this work has been identifying a set of six wake-up techniques, as recommended by 21 DHH users who participated in a formative interview study. In addition, our subsequent study, with video prototypes, enabled 12 DHH participants to visualize how these approaches may work. In addition to indicating their overall ranking preference among the wake-up techniques, participants discussed the trade-offs between various wake-up approaches, and they identified key factors that affected their preferences of each.

In this study, we identified the specific trade-offs and factors for each of the wake-up techniques. These factors were based on the convenience and reliability of the wake-up techniques and did not rely on specific brand of device shown in the video. In this section, we discuss the participants' concerns about wake-up interaction aligned with two key underlying factors:

Convenience of using the wake-up technique.

Overall, we found that participants were more inclined towards techniques that were easy for them to use and easy to access. Talk-to-talk techniques provided them with a hands-free experience, requiring no prior setup to interact with the personal-assistant device. Similarly, these techniques also enabled users to keep their hands free, in order to next issue the command in ASL. Broadly, we found that users preferred methods of waking up the device that enabled them to have as equivalent an experience as possible to hearing individuals who use voice-based interaction, e.g., with wake-words. However, participants discussed how talk-to-talk techniques (except for clapping) depended upon the lighting or camera distance in the environment, which could restrict users' access to the device in some situations. Despite push-to-talk techniques being more robust to these environmental factors, participants still did not find them as convenient to use.

The link between privacy and the reliability of the wake-up technique.

Our findings revealed that participants were broadly concerned with the reliability of the wake-up technique. Participants indicated a clear preference for wake-up methods that avoid accidental device wake-ups. In particular, they were concerned that the device not have to access conversations that were not meant for it. In addition to concerns about false wake-ups, participants were also concerned about the privacy implications of a camera-based interaction with the device, which is necessary for ASL interaction.

Although participants noted the reliability benefits of push-to-talk techniques, this did not lead participants to prefer them to talk-to-talk techniques. Although it would be ideal for future designers of sign-language based personal-assistant devices to identify wake-up techniques that are both convenient and secure, our study suggests that DHH users prioritize convenience. This finding is in alignment with prior work on usable security which reveals the importance of any security and privacy approaches to be easy to use [2, 13].

4.6 Conclusion

This chapter investigated wake-up approaches for sign-language-enabled personal assistant devices. Through formative interviews with 21 DHH participants, in Study 1 (section 4.3), we identified six potential wake-up interactions. We created wizard-of-oz video prototypes of a DHH user demonstrating each form of wake-up interaction with a personal-assistant device. In study 2 (section 4.4), 12 DHH participants discussed factors that influenced their preferences among these prototypes. Our findings revealed pros and cons of various wake-up techniques, as well as factors that shaped users' views of these interactions. Our findings provide guidance to future researchers and designers of this technology.

EPILOGUE TO PART I

This is the end of Part I of this dissertation. Chapter 3 presented a mixed-method study which started with interviews to inform the design of a larger online survey. This study provides a basis for future phases of research to investigate DHH user's interactions with these devices, and to construct a dataset of videos of sign-language personal-assistant commands, which may be useful for sign-language recognition researchers. Chapter 4 focuses on the wake-up portion of the interaction with personal assistants, and employs formative interviews followed by video prototype evaluations. Part I of this dissertation explored:

RQ1.1: Have DHH ASL users used devices like this up to now, and what has been their experience? We found evidence that DHH users use personal assistants significantly less than the general population. (section 3.5.1)

RQ1.2: If DHH ASL signers could interact with a personal-assistant device in ASL, would they be interested in its use? Whereas they currently use these devices less than the general population, we found evidence of DHH users' interest in ASL interaction with such devices. (section 3.5.2)

RQ1.3: What types of commands would DHH ASL signers imagine using with such a device? We have formed a prioritized list of "categories of commands" DHH users are interested in issuing to a personal assistant, as well as a list of DHH-specific user cases for personal assistants. (section 3.5.3)

RQ1.4: How would DHH ASL users imagine waking up, interacting, and receiving responses

from such a device? We present evidence of DHH users' privacy concerns with camera-based interaction in their homes, and initial user reaction to wake-up interaction approaches and response-display for ASL personal-assistant devices. (section 3.5.4)

RQ2.1: What are the factors considered by DHH users to judge a wake-up interaction accessible and usable for a personal assistant device that understands ASL? We identified six potential wake-up interactions, and identified factors that influence DHH preferences among these prototypes, providing guidance to future researchers and designers of this technology. (section 4.5)

Part II: Dataset collection

PROLOGUE TO PART II

Part I has shown there is a lack of prior research on the usage of personal assistant devices (which are becoming ubiquitous) by DHH users. The gap in knowledge about DHH users' experience with personal assistant devices is addressed, and further research on this topic is strongly motivated. Chapter 2 has posed an ASL data bottleneck problem for sign language technologies (e.g. automatic recognition of ASL). Additionally, the coronavirus pandemic drove the broader issue of how to best collect data from DHH participants using remote methodologies. Through an internship with Microsoft Research, motivated by prior work which has shown that traditional ASL data collection succeeds more using specific prompts of English words or passages for users, I investigated remote ASL data collection at scale, testing the validity of using a scalable online sign language data collection platform, presented in chapter 5.

A corpus of labelled, isolated signs may help develop individual sign recognition, which would be useful for digital personal assistants that respond to simple signed commands. After exploring a platform to collect these, in chapter 6, I investigated whether this methodology could be extended to also generate a continuous signing dataset, while supporting bilingual content. For more complex commands to personal assistant devices, there needs to be natural conversation with complete sentences, which would need continuous sign language data.

Specifically, Part II of this dissertation investigates these research questions:

> **RQ3:** How can DHH and signing communities be enabled to curate sign language datasets that overcome limitations of traditional in-lab collection (e.g. limited demographics, controlled environments, limited size and quality, expensive post-processing and labeling)?
>
> **RQ4.1:** How can everyday signers efficiently contribute to continuous sign language datasets?

RQ4.2: Ensuring that the DHH community is involved in the process, how would the platform be designed? What are the design criteria?

RQ4.3: How would DHH users respond to crowd-generated content?

RQ4.4: Can the platform incentivize contributors by being a sign language bilingual resource?

After these data collection efforts, I investigated how this might benefit AI researchers in the future who are working on personal assistant technologies. In chapter 7, I employ a remote data collection protocol, using a Wizard-of-Oz prototype device that appears to understand ASL, DHH users were allowed to spontaneously wake-up and interact with a personal assistant device in sign language, albeit in a limited manner. This methodological setup and logistics is explained, and the collected dataset is described. Additionally, I describe the process of analyzing and annotation of this dataset, and share this information publicly.

While this remote Wizard-of-Oz experiment put the knowledge learned in Part I into practice and taught us many things (as discussed in Part III, specifically chapter 9), there were limitations that came with the lack of real-world conditions. The last chapter in this part (chapter 8) conducts an in-person, physical Wizard-of-Oz experiment to enable investigation of aspects that were not possible through the remote protocol, such as allowing users to change their location and reference to objects inside the room while interacting with a personal assistant.

Chapters 5 and 6 in the beginning of Part II focus on the general case of collecting single-word and continuous-utterance ASL signing using remote modalities. Datasets of individual words and longer phrases may be part of the overall training data that could benefit personal assistant technologies – such systems are likely to need to identify individual one-word commands, and some longer phases of continuous signing. Then, chapters 7 and 8 use Wizard-of-Oz data collection strategies (one remotely, and the other using an in-person methodology) to collect ASL personal-assistant commands, to further contribute to training data of such systems.

Chapter 5

Exploring Collection of Sign Language Videos through Crowdsourcing[10]

5.1 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 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) [11] people who primarily use a sign language [127], and a growing number of hearing people who use sign languages socially or in language classes [66].

Developing Artificial Intelligence (AI) and Machine Learning (ML) models that handle sign lan-

[10] The information in this chapter is based on a joint project with Dr. Danielle Bragg, along with Fyodor Minakov, Dr. Naomi Caselli, and Dr. Bill Thies. This work was conducted as part of an internship I had at Microsoft Research, where I led the user study design and analysis. The results were published as a paper at the CSCW'22 conference [22].

[11] 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.

guages 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 [23]. 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 [58] or social media posts [91]) 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.

5.1.1 Motivation for 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 [7]. Some crowdsourcing initiatives exist outside of such platforms, and instead enable people to contribute directly to specific initiatives, for example Wikipedia [176]. 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 [84, 152] is a type of crowdsourcing that seeks to advance scientific research by leveraging small contributions from individual ”citizens.” Citizen science falls within the broader um-

brella term of ”organic crowdsourcing” [95], 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 [153]) 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 [38], amassing common-sense knowledge [173], and labeling tasks [169, 172]). Incentivization can also be provided by revealing information to contributors about themselves (e.g., LabInTheWild [138]). While there has been some preliminary work on designing general platforms to collect data from people with disabilities [131], 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.

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 [55]. Instead, research teams often employ complex tagging manuals and/or video-based controlled vocabularies (e.g., [74, 114]). The labellers need advanced linguistic expertise in both languages and training in specialized annotation software (e.g. ELAN [163]), making the process expensive and time-consuming.

To enable DHH and signing communities to curate sign language datasets that overcome such lim-

itations, 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, 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.

5.1.2 Contributions

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.

5.2 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.

5.2.1 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.

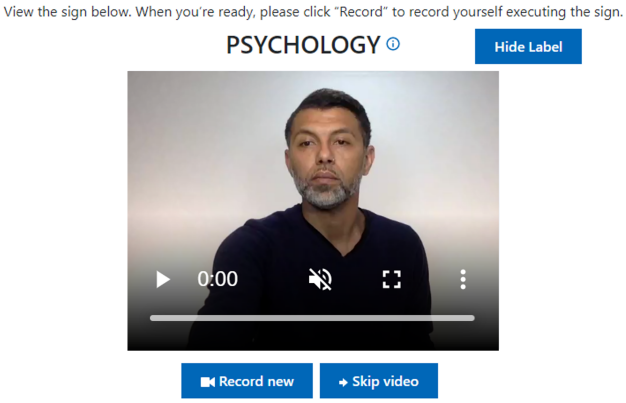
4. Qualitative Feedback: After completing the above tasks on the website, the form asked several

questions about participants' overall experience using the website. In closing, they were asked for basic demographics and compensation information.

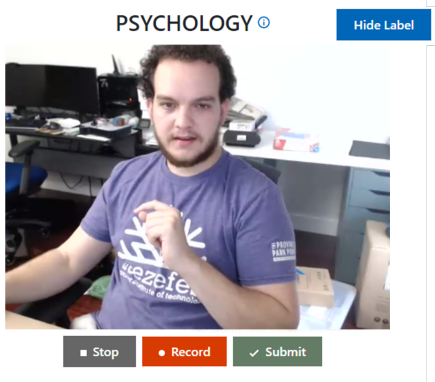
5.2.2 ASL Crowdsourcing Platform Prototype

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., [26, 30, 68]), educational websites and resources (e.g., [34, 39, 75, 104, 110, 180]) and some, though fewer, games (e.g., [20, 185]). Particularly relevant to our work is [20], 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.

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 were designed to scalably solve post-processing difficulties with minimal training of contributors, particularly 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 iterative design and testing with 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. We use a citizen science approach to crowdsourcing, enabling contributions in order to advance sign language research. Its components and implementation are detailed below.



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



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

Figure 5.1: 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 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

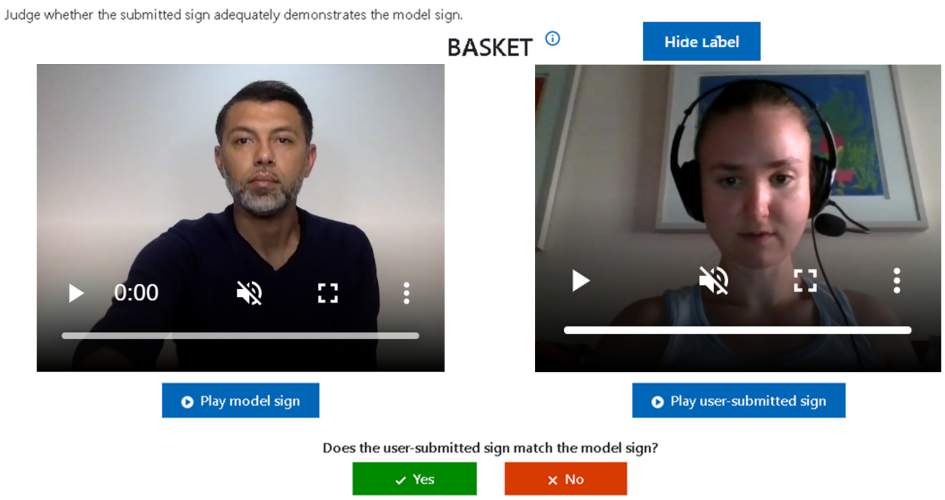


Figure 5.2: Quality control task, where users check whether another user recorded the same content as in the prompt, demonstrated with a recording of BASKET. The purpose of this task is not to rate the signer's execution, but to verify that the user contributed a copy of the specified content. Reviewers view the prompt video and the user-submitted video, and answer a Yes/No question: "Does the user-submitted sign match the model sign?"

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 5.1 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 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.

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).

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 fig. 5.2. 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.)

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.

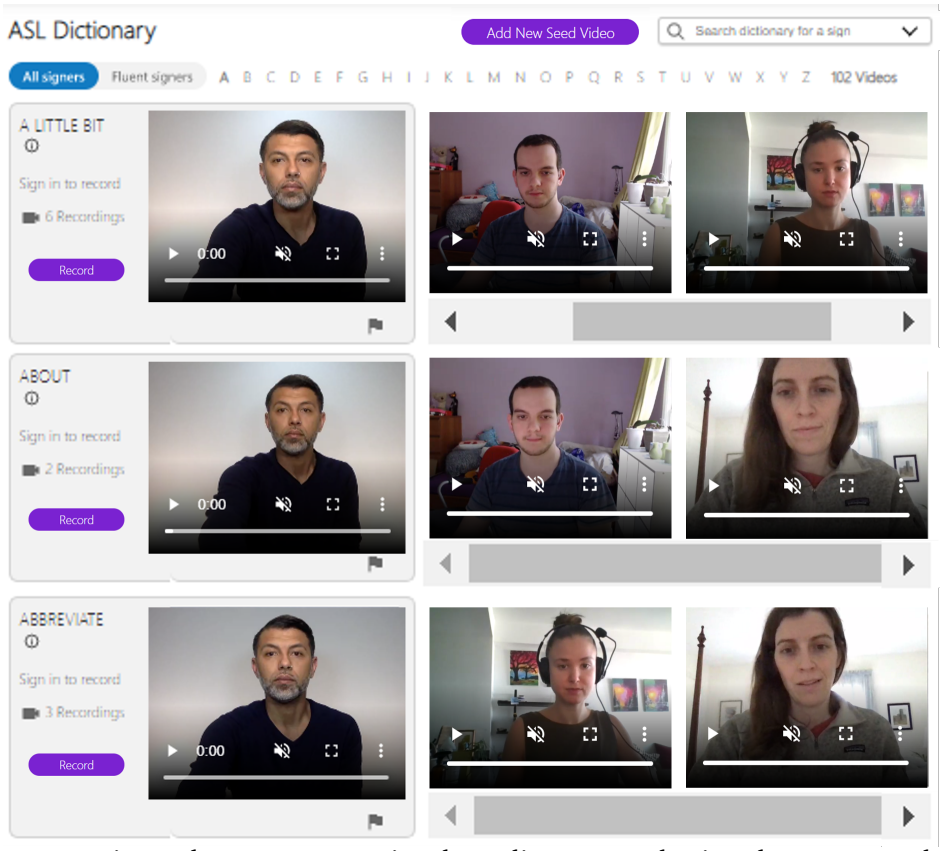


Figure 5.3: Dataset view, where users can view how diverse people sign the same words or concepts. They can search for signs, and also directly add to the dataset by clicking 'Record' for a particular sign, or 'Add New Seed Video' to add a new sign to the database vocabulary.

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.

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 over secure protocols. The website was seeded with model videos from ASL-LEX [31, 146], a large labelled corpus of ASL vocabulary, with permission.

5.2.3 Prompts Used

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

Signs to Record

All participants were asked to record the same set of 60 signs, listed in table 5.1. 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.

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 5.1: 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 [31].

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 table 5.2, 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 [20] (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.

	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 5.2: 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 5.1), this set of 30 was chosen to span the same phonological properties and levels.

5.3 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.

5.3.1 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.

5.3.2 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.

Evaluation Process

To ground our analysis and enable comparison, we adopt the methodology established in [20] 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.

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 table 5.3). 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.

		Crowdsourcing Prototype				Control Mobile App [20]			
		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 5.3: Expert evaluations of the a sample of 180 ($\sim$ 10%) videos collected in our user study. Two experts answered the same set of questions as in [20], 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 [20], for completeness.

	Crowdsourcing Prototype	Control Mobile App [20]
DHH	91.89%	98.00%
Hearing	100%	99.50%
Total	96.67%	98.75%

Table 5.4: Percent of recordings where at least one expert's evaluations indicate the video is appropriate for training real-world recognition models.

Quality for Training Recognition Models

Table 5.4 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.

5.3.3 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 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.

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.

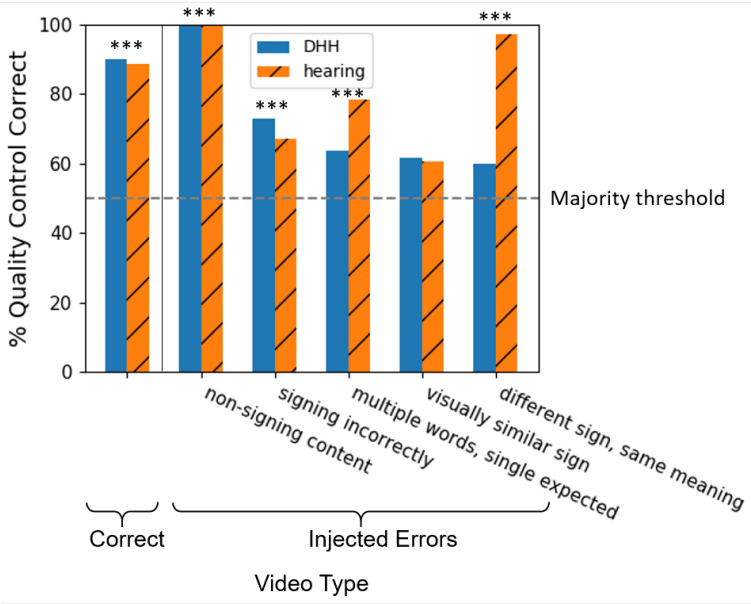


Figure 5.4: Participants' accuracy at performing quality control checks, for various types of videos: correctly signed videos (left), and five types of injected errors (at right). The majority vote was statistically significant (***) for all video types except "visually similar sign" (compared to random). Significance codes: ***$< .0001\bar{6}$, **$< .001\bar{6}$, *$< .008\bar{3}$.

Quality Control Reliability

Figure 5.4 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 with a Bonferonni correction ($p < .001/6 = .0001\bar{6}$), 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 ($t(28) = -3.085, p =$

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

Table 5.5: 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.

.0048) , unlike for any other video type, by t-tests comparing DHH and hearing individuals' accuracy rates on each question type with Bonferonni correction ($p < .05/6 = .008\bar{3}$ 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 5.5 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.

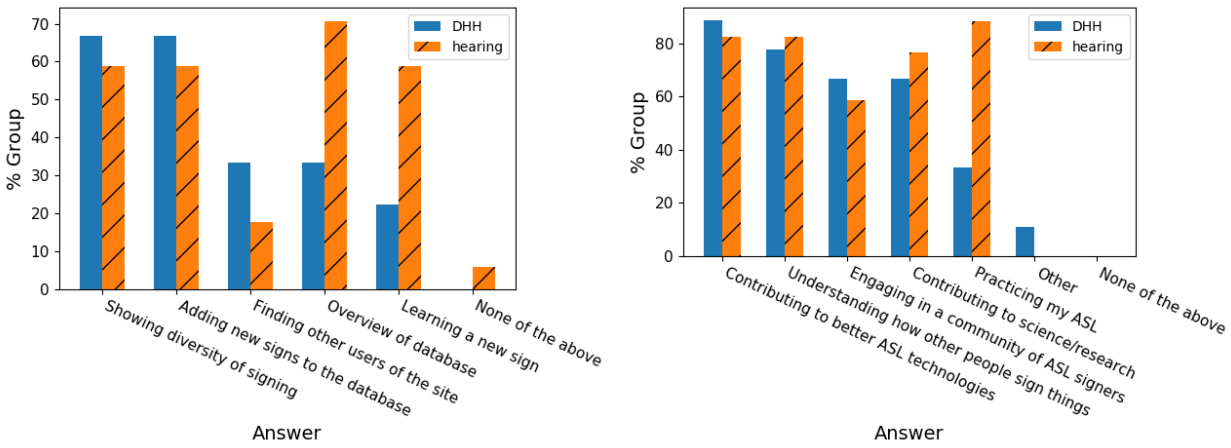
5.3.4 Participant Feedback

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 5.5 shows the benefits they reported for a) the dataset view specifically, and b) the website overall.

For the database view (fig. 5.5a), 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 (fig. 5.5b), 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. fig. 5.6 shows participants' reported concerns. The most common among DHH and



(a) Dataset view

(b) Overall website

Figure 5.5: 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.

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). Participants noted reasons such as learning, viewing different people signing, and ease of use as positives, and also noted 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.

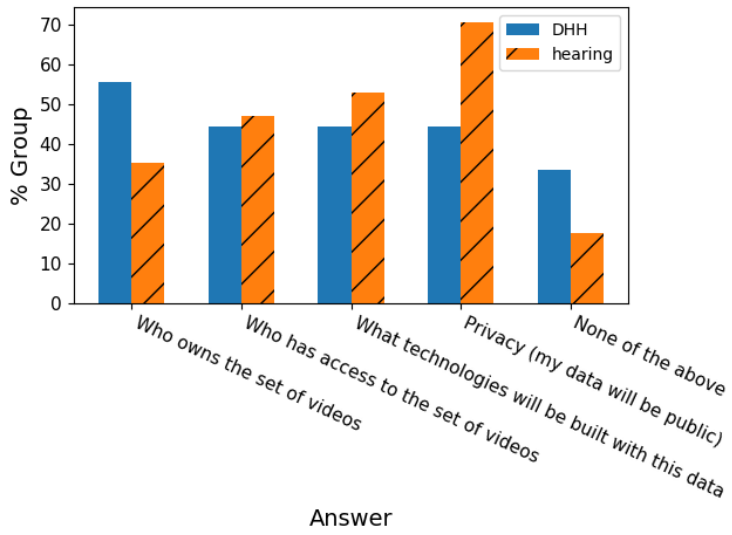


Figure 5.6: Concerns reported by participants, in contributing to the website, separated into DHH and hearing groups.

5.4 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.

5.4.1 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. [161, 162]), sidewalk accessibility mapping [142, 147], image caption creation (e.g. [143, 144]), visual question answering for blind and low-vision users (e.g. [17, 19]), and real-time speech captioning for DHH users (e.g. [98, 99]). While some of these research projects have produced large datasets (e.g. [71]), 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 [52]. Prior research has also shown that people with disabilities themselves, including DHH people, want to contribute to datasets that will benefit their disability communities [131]. Beyond accessibility, there is an even wider array of crowdsourcing projects, many of which have succeeded at scale (overviewed in section 5.1.1). 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.

5.4.2 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 fig. 5.4). 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 [27, 59], 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 fig. 5.5). 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 (fig. 5.6). 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.

5.4.3 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.

5.4.4 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 (fig. 5.6), 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. [156]), 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 [24, 100], 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 [21]. 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.

5.5 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% appro-

priate 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.

Chapter 6

ASL Wiki: An Exploratory Interface for Crowdsourcing ASL Translations[12]

6.1 Introduction

Chapter 5 explored crowdsourcing to collect word-level sign language videos for training ML models. Creating scalable continuous sign language datasets with diverse, real-world signers is a difficult task, and there is a lack of informational resources in ASL (and other signed languages). This is a problem for both the DHH community and sign language processing researchers. Nobody has explored the question of how to enable everyday signers to efficiently contribute to continuous content translation efforts, or how DHH users might respond to crowd-generated content.

Approximately 1 in 6 adults in the U.S. is Deaf or Hard-of-Hearing (DHH), and prior literacy research shows that over 17% of deaf adults have "low literacy" [5]. Signed languages are the primary languages of Deaf communities worldwide, and they are completely distinct from local spoken/written languages. For example, American Sign Language (ASL) is the primary signed language used in the U.S., but it is a completely different language from English – not a one-to-one mapping. As a result, if a person

[12]The information in this chapter is based on a joint project with Dr. Danielle Bragg and Fyodor Minakov. I had started this work during my internship at Microsoft Research, and continued through an ongoing collaboration after that time. I led this effort, e.g., design of the site, user study, analysis, etc. I also led the writing of a paper that was published at the ASSETS'22 conference [63].

is fluent in ASL, they are *not* necessarily fluent in English reading and writing. ASL is often DHH signers' primary language, and they typically prefer ASL over English, are more comfortable with, and understand content better in ASL [79]. Among this bilingual community, there is a wide range of literacy levels (e.g. studies have found fourth-grade reading levels among DHH high-school graduates [167]). Research has found that as a result there are lower educational outcomes among DHH individuals and lower rates of employment (and salaries) among DHH adults [6].

A major obstacle facing DHH signers is a lack of educational resources in sign language. Many educational resources are available in text (e.g. textbooks, literature books, online encyclopedias, etc.), but not in a signed language. As there is no standardized written form of ASL and sign language is typically in video form, these text-based interfaces do not adequately support users who prefer a signed language. Because of this lack of ASL content, DHH users often have to look up individual English words on a separate website or interface (e.g. English-to-ASL dictionaries) [26], and re-read the English content they are trying to consume [16]. Even though individual words can be looked up when necessary, this is not efficient, does not help to understand English grammar, and may be insufficient for DHH signers trying to understand English text. It would be helpful if an ASL version of the target content was available – having the entire sentence/article signed might be preferred by a DHH ASL signer rather than looking up individual words and/or re-reading multiple times.

At the same time, advancing sign language research and technology is currently impeded by lack of sign language data [23]. Existing ASL datasets typically offer a set of individual ASL signs, with their respective English meanings, and/or ASL glosses. They do not have sufficiently representative and diverse signers – they often consist of homogeneous sign language interpreters, small sets of signers, and poorly labeled videos of unverified quality (listed in [20]). In order to more fully understand and model the language, labelled continuous signing (i.e. complete sentences with annotations) from diverse signers is needed. However, creating such a dataset is extremely difficult. It is not only expensive to produce, but it also requires a massive amount of human labeling and annotation, since there is no automated system to do so. It is also hard to enable a large pool of contributors, since most in-person data collection efforts are limited to individuals who live close-by within commuting distance, and

have time in their schedules to contribute. How to enable everyday signers to efficiently contribute labelled continuous content, and how DHH users might respond to crowd-generated content remain open questions.

In this work, we present a novel interface that addresses two needs at once: 1) it provides a bilingual information resource and 2) it simultaneously generates a continuous labelled signing dataset that could be used by artificial intelligence researchers, ASL linguists, ASL learners, DHH ASL signers, and others. Our interface provides a side-by-side ASL (video) and English (text) synchronized interface, where users are able to read/view articles simultaneously in both languages. Users can also use this platform to contribute ASL translations of existing English texts in the communal database. For this exploratory work, we seed the interface with popular English Wikipedia articles, which are translated into ASL, and refer to this prototype system as "ASL Wiki". However, the same interface could be seeded with any long text, and could be used with any pair of written and signed languages. In terms of dataset creation, by enabling contributors to record segments of English text with known contents, the interface eliminates the need for humans to later segment and align the text and video. Such intensive labelling work is commonly done in creating parallel corpora containing signed language and spoken/written text (e.g. [58]).

To help understand the effectiveness of such an interface, we ran two exploratory studies. First, to better understand the user experience with the interface, we conducted an exploratory user study where 19 participants used the interface to consume and generate content, and shared feedback. Our results suggest that DHH individuals find real-world value in our interface, thought it was easy and intuitive to use, and were excited to see further development and identified several target audiences they would recommend the site to. Second, we also conducted an exploration into the quality of translations that can be generated through our interface. Results suggest that the translation quality is comparable to the quality of translations created through state-of-the-art setups for sign language translation. We conclude by discussing future work that this initial exploration introduces.

6.2 Related Work

In this section, we focus on work relevant to our two motivations: supporting bilingual content, and supporting sign language data collection efforts.

6.2.1 Sign Language Educational Resources

Existing resources that make information available in a signed language compromise a small number of dictionaries ([30, 68]), educational materials ([34, 39, 75, 180]), lexical databases ([74, 146]), and mobile vocabulary apps ([104, 110]). Several examples of these are listed in [20]. The landscape of existing sign language resources is very small compared to the resources available for spoken and written language users, who are typically considered by default. There have been limited attempts to create browser tools that provide signed translations of written content, to create signing avatars, and to more generally create recognition and translation systems [20]. These tools and resources are not viable due to the very limited amount of labeled signing videos with diverse, well-representative signers. There are some DHH content creators that strive to support accessibility of information and support the DHH community, such as the Daily Moth – a group who "deliver news in video using American Sign Language" [117]. However, as these efforts are sponsored and often composed of a small group of people, they are limited in the amount of content they can create, and often have to selectively offer a handful of content options – for instance, the Daily Moth says "the deaf host, Alex Abenchuchan, covers trending news stories and deaf topics on new shows Monday-Fridays". Many people in the DHH community praise the Daily Moth due to the level of access it provides, being a bilingual information resource for selected news happening around the world [133, 177].

Our interface would enable crowdsourcing to address the problems of large-scale sign language data collection and diversity, naturalness, all while serving as a bilingual educational resource.

6.2.2 ASL and English Bilingualism

Prior work suggests that bilingual resources are useful for DHH fluent signers, rather than having any negative information-overload effects. Psychology researchers have established that it is not costly to

switch from single to dual lexical retrieval (using two languages at once), and revealed a significant cost to turning off a language, which bilingual DHH users might do while trying to understand English text alone [51]. This suggests that an ASL and English bilingual interface, such as the one we have developed in this work, could be beneficial to DHH fluent signers by providing greater accessibility than English text alone.

The value of bilingualism in ASL and English has been further substantiated by Deaf-led organizations. The National Association of the Deaf (NAD), a nonprofit organization whose mission "is to preserve, protect and promote the civil, human and linguistic rights of deaf and hard of hearing people in the United States of America." Internationally, NAD represents the U.S. to the World Federation of the Deaf (WFD), an international human rights organization. NAD supports bilingualism, using ASL and English, in the home and educational environment for DHH individuals. They advocate that bilingualism is important and effective because it fosters "positive self-esteem, confidence, resilience, and identity, factors necessary for lifelong learning and success" [126].

Despite the value of bilingual ASL/English resources, few exist. The Deaf Studies Digital Journal (DSDJ) "is the world's first peer-reviewed journal dedicated to advancing the cultural, creative, and critical output of published work in and about sign languages and Deaf culture" [90]. It is a bilingual and bimodal publication primarily presented in both ASL and English. It features academic work in other sign languages, and offers scholarly articles, commentary, literature, visual arts, film/video, interviews, reviews, and archival history footage and commentary. To date, there have been 5 issues (spanning 2009-2020) with about 150 articles total. In the most recent issue, each article has a split side-by-side view showing ASL (or other sign language) on the left, and English text on the right. The content is synchronized so that the English sentence being signed is highlighted. The video has controls so that the user can control playback of the signed video. Our interface builds on this, similarly providing side-by-side views in both languages.

To the best of our knowledge, there has been only one past attempt to systematically provide sign language translations of existing text. Signly[13] is a recent commercial effort to add "synchronous, invision, sign language translations on any webpage for any deaf sign language user". They enable website

[13]https://signly.co/

visitors to select English text they would like translated into British Sign Language (BSL), which is sent to a professional interpreter for translation. Once the translation video is created, website visitors can click on the English text to trigger a pop-up translation video at the bottom-right corner. While this company helps make English texts online accessible, users have to request translations, and website creators have to contact and pay Signly to incorporate and maintain their services. Scale is also limited, as the translations are done by the Signly team. Our interface is similarly motivated to provide access to English text online. However, we enable crowdsourcing translations to streamline and scale data collection, and to enable a more diverse and representative group to contribute. We also display the text and video side-by-side in a more bilingual manner.

The lack of bilingual resources and lack of data (2.3.2) motivated our "ASL Wiki" interface.

6.3 ASL Wiki Prototype

We have created "ASL Wiki" – a prototype site where people can crowdsource ASL translations of English articles, providing a community resource that supports accessibility as a bilingual information resource, while also tackling the lack of continuous ASL datasets with English labels. In this section, we describe our prototype and design process.

6.3.1 Design Process and Criteria

We engaged in an iterative design process to arrive at our "ASL Wiki" website design. We first identified design criteria the platform needed to meet (e.g. that the text used is available for use on the platform and in a dataset, that participants can contribute remotely without specialized hardware, and that translations are segmented and labelled). With these identified, we started with drawn designs which were iteratively refined and implemented. Throughout the process, we continued to meet with stakeholders consisting of a group of interdisciplinary Deaf and hearing individuals who have deep ties with the DHH community and incorporate their input. These stakeholders tried out the evolving prototype, and also discussed the project and provided guidance.

Through our meetings, we chose to explore creating and reading bilingual versions of Wikipedia

articles, rather than play scripts, books, or other resources. We decided on Wikipedia articles because they are generally neutral, publicly available, and popular informational resources. There also exist other parallel corpora of Wikipedia content which have been useful for natural language processing and artificial intelligence/machine learning.

Our iterative design process uncovered specific user requirements of our interface. We found that the interface needed to show ASL and English at the same time, so that users could see both and easily look at one or the other as they wished. Our interface also needed to show which English portion is being signed in the current ASL video, so that users can keep track of their position in both the video timeline and the English article. Users who are recording their videos should have an efficient, streamlined way to record sentences, meaning that the interface should not pose unnecessary overhead. It should be allowed for multiple people to submit recordings for the same English sentence, as different people might sign differently (e.g. regional accents or varied interpretations), or have preferred signs for specific English words.

6.3.2 "ASL Wiki" Design

Homepage

We took inspiration from the idea that Wikipedia is a "free content, multilingual online encyclopedia written and maintained by a community of volunteers" [176]. On the homepage of the ASL Wiki site, on the left hand side, is a checkbox list of featured categories. Users can use these checkboxes to bring up relevant articles, which appear in the middle with fractions indicating how many sentences there are in the article, and how many of these sentences have been recorded by at least one user. Being clickable, the rows of article titles also display a "Record" button that takes you to the reading/recording interface (discussed further in subsequent section).

On the top of the homepage is an introductory title and paragraph, along with an ASL video of someone signing this text. Once you are logged in, on the top right of the page is a button that allows you to view and edit your profile, or sign out. Next to this button is a gamifying trophy icon displaying the number of sentences the logged-in user has recorded. This was added as it is a common element of

social media sites that display the number of "posts" an user has submitted. It potentially incentivizes the user by showing them how much sentences they have recorded.

In the middle of the page, between the top banner and article table of contents, is a numbered instruction summary to remind users how to navigate and use the interface. Especially as users are able to leave the site and come back later, and since they navigate into and out of specific articles, they may need a persistent reminder of how to use the site, which is why we added this. A screenshot of the homepage is shown in fig. 6.1. Once the user selects an article, they are taken to the reading/recording view. This view has a toggle on the top to switch between the recording view and the reading view.

Recording and Reading View

In both recording view and reading view, the main layout is the same: it is a split, side-by-side bilingual interface. On the left is a placeholder for an ASL video. On the right is the article in English.

If in record view, once the user selects a sentence on the article, the ASL video placeholder becomes a self-view of the user's webcam, so that they can see themselves. Their self-view is overlaid with a head and body guide to encourage users to center themselves in the recording. A 3-second countdown commences, and then the user would sign the English sentence in ASL. While they are recording, the according English sentence is highlighted, to mark and keep track of their place in the article. When they are finished, clicking on a stop button underneath their self-view stops the recording, and displays their recorded video for playback. If the user approves, clicking "Keep" will submit the video, and auto-progress to the next sentence in the article, or clicking "Redo" will prompt them to redo the recording. The English sentences that have been recorded will show a video camera icon. There is also a guide on the top, above the English article, to remind users how to use the interface. Also, underneath the ASL video placeholder is a picture demonstrating a good recording setup and a bad example, to remind users that they should be sure to position their webcams so it captures their upper body and that their arms/hands do not go out of frame while signing. There is also an upvote/downvote button where users can give feedback on the ASL video.

In reading view, the site enables users to access parallel content in ASL and English. After toggling

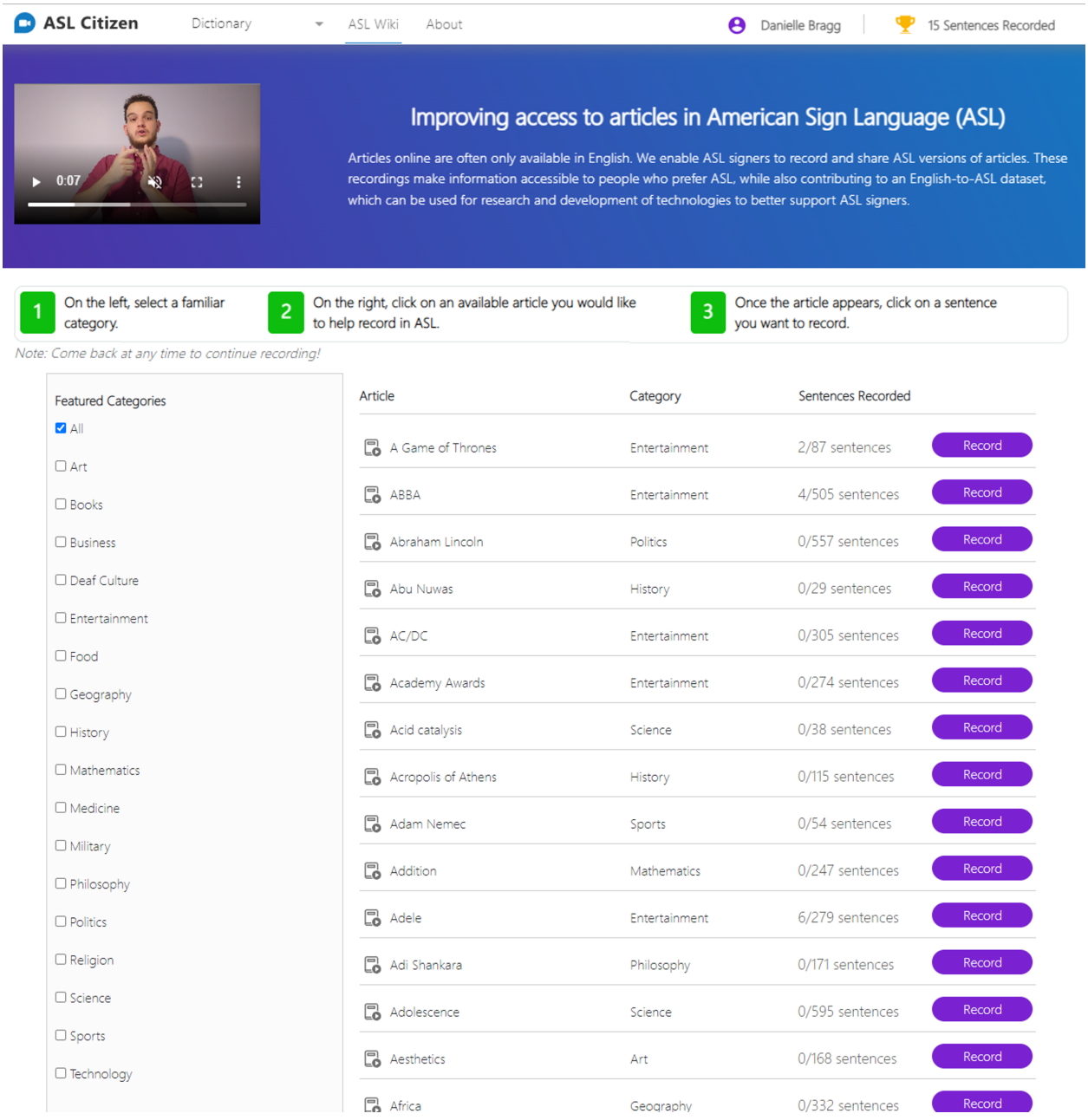


Figure 6.1: Screenshot of ASL Wiki homepage.

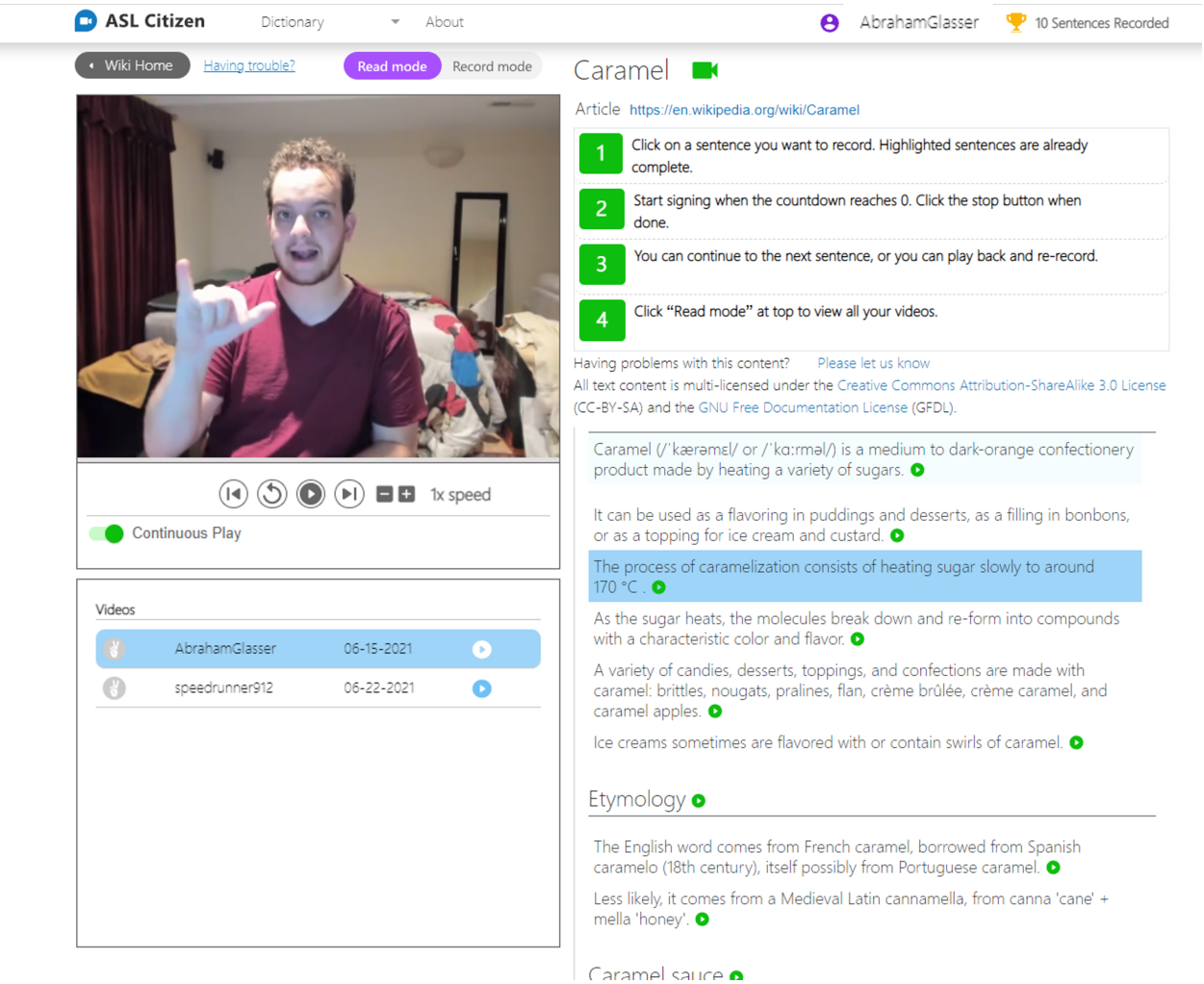


Figure 6.2: Screenshot of reading view of article "Caramel".

to reading view, the same English article is kept, and now shows a "play" icon next to the sentences that have been recorded. Clicking on a sentence will highlight that sentence, and play the respective ASL video. Once the video completes, it auto-progresses to the next sentence. There is a playback control underneath the video so that the user can go back, forward, redo, pause/play, and control the playback speed. There is also a toggle to turn on or off the auto-progression. It is possible that multiple users would sign the same English sentence, so underneath the ASL video is a list of the users who submitted videos for the currently selected English sentence. The user has the option to switch between signers if they desire. A screenshot of a sample reading view is shown in fig. 6.2.

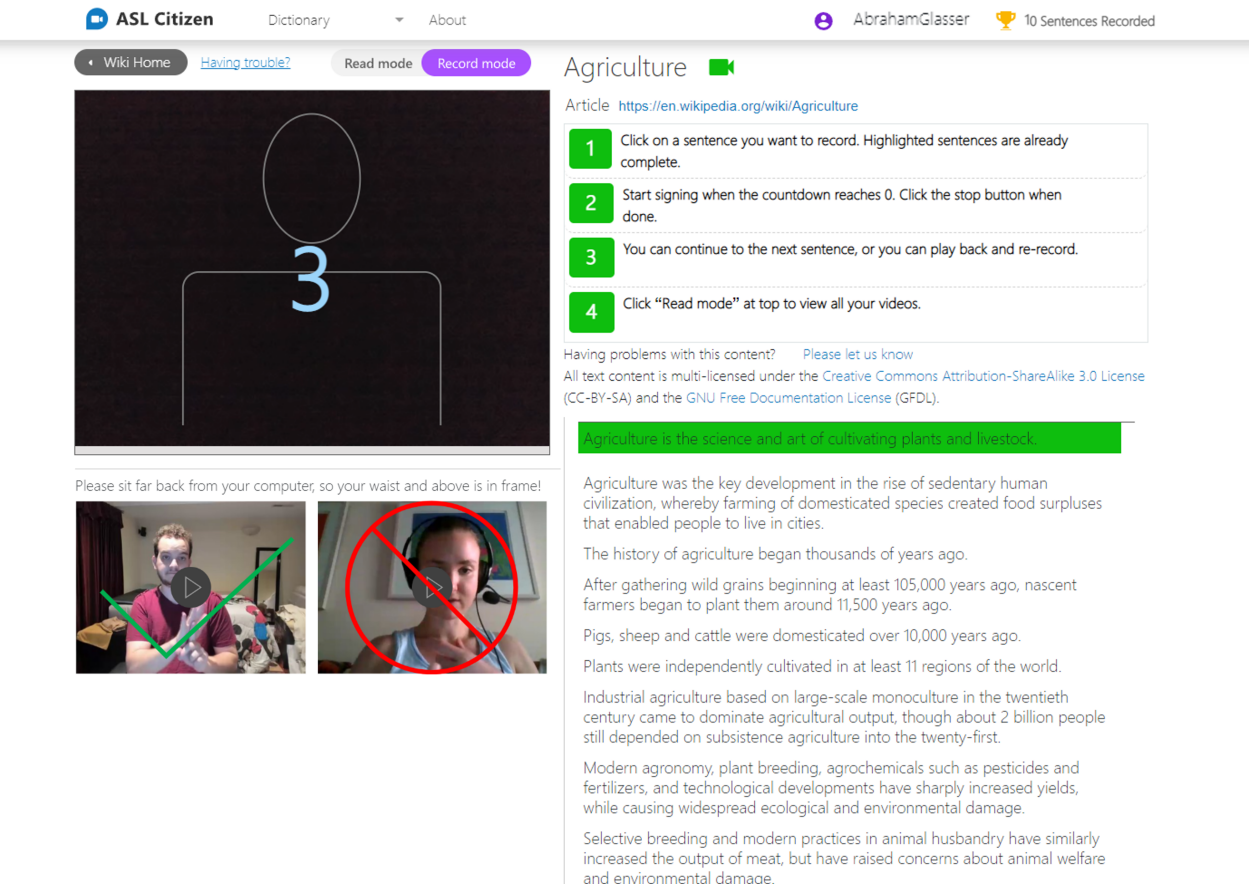


Figure 6.3: Screenshot of recording view of article "Agriculture".

6.4 User Study

To explore the usability of our ASL Wiki site design, we ran a remote user study, with Institutional Review Board (IRB) approval. In this user study, participants answered survey questions, tried out the reading and recording views, and discussed interview questions about their experience.

6.4.1 Participants

Recruitment

Participants were recruited via mailing lists, social media posts, and snowball sampling. The recruitment criteria was that they use ASL, are 18 or above years of age, and have a computer with a webcam. 19 participants were recruited in total. The sessions ran for about 1 hour, and participants were given a $30 (USD) Amazon gift card for their participation.

Demographics

Out of the 19 participants, 15 identified as Deaf, 3 deaf, and 1 Hard-of-Hearing. 11 identified as female, and 8 male. The average age of all participants was 26.1 with standard deviation 2.2.

Participants self-reported their ASL fluency on a scale from 1 (I do not use ASL) to 7 (I am fluent). The average fluency was 6.4 with a standard deviation of 1. Generally, all participants were educated, with only 3 out of 19 not having a bachelor's degree yet at the time of participation. Participants were diverse, with 12 self-identifying as White (e.g. German, Irish, English, Italian, Polish, French, etc), 5 as Asian (e.g. Chinese, Filipino, Asian Indian, Vietnamese, Korean, Japanese, etc.), 1 as Black or African American (e.g. African American, Jamaican, Haitian, Nigerian, Ethiopian, Somalian, etc), and 1 as Middle Eastern or North African (e.g. Lebanese, Iranian, Egyptian, Syrian, Moroccan, Algerian, etc). The 19 participants came from 8 different U.S. states.

Prior Experience with ASL and English

All participants reported that they read English text online daily (n=5) or multiple times a day (n=14). It was reported in the demographics survey that participants read English text via websites, books, articles, video transcripts, and social media posts. Along with these 5 options, we had also listed podcasts (and "other") as the answer-choices on this survey question, but nobody selected that.

Participants were asked the question "How often have you encountered websites you wish provided ASL videos instead of or in addition to English text?" 3 answered "multiple times a day", 5 "daily", 5 "weekly", 2 "monthly", 2 "less than once a year", and 2 "never".

All participants except one said that they watch ASL videos online frequently (1 said "yearly", 3 "monthly", 5 "weekly", 5 "daily", and 5 "multiple times a day"), typically through video blog (vlog) posts, YouTube videos and other social media videos. Participants commented that they have seen content on various social media platforms where someone is signing in ASL, and there are English captions visible, so they have seen bilingual/bimodal content before, and are comfortable with it. 10 out of 19 participants said that they have at least once created content like this that had both ASL and English, and that they created the ASL video first and added English subtitles afterwards. 9 out of these 10 did this to post on social media, where they have both DHH and hearing friends, and 1 said they only did it for a homework assignment or class project in college.

6.4.2 Procedure

An online form walked participants through the study procedures while a DHH fluent ASL signer was on a video call with the participant. Each participant was scheduled for their own session, and the entire procedures took approximately 1 hour. The procedures were as follows:

1. **Consent:** Participants engaged in a consent process with IRB-approved language through the online form. The researcher on the video call checked whether the participant needed any portion of the consent language signed in ASL so that it was fully understood.

2. **Background:** Through the online form, participants were asked multiple-choice questions about

their prior experience with using English and ASL online.

3. **Reading:** After this, they followed instructions on how to access the ASL Wiki site and sign in, and were directed to the "Caramel" article which had been entirely pre-recorded by a DHH ASL signer from our research team. They engaged with the interface to read this article until they were satisfied.

4. **Recording:** Next, they were instructed to select any article of their choice and record themselves signing. Since we wanted to closely match a real-world experience of using our site, participants were given the flexibility to record as much (or as little) as they wanted to, but were told to use the recording interface until they were confident that they got and understood the full experience of recording and contributing to the site, and were told they would discuss their experience afterwards.

5. **Semi-structured interview:** While the fluent DHH ASL signer continued to be on a video call with them, they engaged in a semi-structured interview with guiding questions spanning short answer, long answer, and Likert-scaled question items. The interview focused first on the reading view, asking about their experience and understandability using the interface, and then were asked questions about their experience and challenges (if any) while recording. Lastly, questions were asked about the overall concept of the site, what they liked and disliked, and whether they would recommend the site to others. Section A.1 provides our interview questions.

6. **Demographics:** After the interview portion, participants returned to the online form where they filled out demographics and compensation information.

6.5 User Study Results

We discussed with each of the 19 DHH participants during the experiment to gauge their reactions and experiences with the reading and recording views of the interface. We evaluated how they used the site, to understand their motivations, challenges and strategies, and the benefits they took away from

the site. We thematically analyzed the interview responses and performed statistical analysis of their responses to the questionnaire. We also collected feedback and identified several target audiences who the users would recommend the interface to.

6.5.1 Reading View

ASL vs. English

Participants valued having both English and ASL versions of articles available for consumption. On average, participants self-reported that, while reading the "Caramel" article that our research team had entirely pre-recorded, they looked at the English part 65% of the time, and the ASL video 35% of the time. Participants explained that the English part is faster to read, with P1 saying "*...it's faster to read and skim through. It's more of a habit because I'm used to reading English articles*". P8, who reported looking at the English part 50% of the time, said "*I like ASL. [It] is more visual and I can visualize it better, but for English I can read it faster. If I just want to consume the content and save time, I would look at the English 100%. If I wanted to fully understand, learn, visualize, maybe 50/50 – I'd also be curious what it looks like in ASL*".

Participants were asked to indicate how understandable the ASL content and the English content in the "Caramel" article they viewed was, on a scale from 1 (very difficult) to 5. For the ASL content, the average was 4.6 (s.d. .7), and 4.8 (s.d. .4) for the English part. Even though participants said that the English part was very easy to understand, all 19 participants answered "Yes" to the question "Was it helpful to view the content in both English and ASL?" P11 explained "*Yes, I can imagine how it would be helpful for the general. It's a nice tool for me to use, and I would like having it even if I don't use it much*".

Interface Usability

On a scale from 1 (very difficult) to 5 (very easy), participants said that the interface was very easy to use, giving it an average of 4.5. Most of the difficulty came from not having prior experience and not knowing what to expect with the interface, e.g. P10 saying "*I didn't think there was any information overload – in the beginning I wasn't fully sure what to do. Maybe the first sentence could be highlighted*

with the video, that would make it more clear there is ASL there". P2 commented "*at first when I opened it, I wasn't sure what to do – my eyes caught the English part first, and I ignored the left half – and then it took me a while to realize that the left side was empty until I clicked on some text, and the video player showed up. [...] I think there should have been some kind of tutor/illustrations with directions of how to use this site before I went ahead and looked at an actual article*".

Most participants (12 out of 19) did not use the upvote/downvote button that was available to them while viewing the "Caramel" article. Some participants said that they did not see it, while some did but decided not to use it. P12 said that they do not use it in general, such as upvoting/downvoting on Reddit, liking/disliking on Facebook or YouTube. P9 said "*I didn't know about the feature until I arrived at this question. I normally skim through contents*", and P14 said "*I wasn't focusing on providing feedback on performance*". Those who did use it generally said that they wanted to give feedback, with P4 saying "*I wanted to give feedback on the video, so I clicked yes – I noticed the signing was clear and matched the English so I went ahead and clicked yes*". Some participants such as P6 emphasized it was important: "*I think it's important to use, yes I would use it, it gives feedback to other people and I can help this website advance and develop in the future and make sure it has good content*", and P16 suggested it would prevent misunderstandings, saying "*I don't want some signers to use wrong signs or say it in the wrong way which will make viewers misunderstand. We want to avoid that*", and P11 made an analogy to real-world applications they've seen, reminiscing "*Yes, it's the same as FAQs or other articles that say "was this article helpful?" – this is the same situation*".

We collected some feedback about the interface, to understand how our interface could be improved and help inform future work on such bilingual interfaces. These feedbacks typically consisted of user interface preferences and suggestions, such as coloring and layout styling. There were also some suggestions about the fundamental system. P7 suggested a different layout: "*For me, I would prefer top and bottom rather than side by side, so it's kind of like captions. It was a little challenging for me to have it side by side*". P13 suggested making the recorded videos easier to find: "*One suggestion I have is that it might be nice to have a separate scrolling bar other than the browser one where it'll indicate the recorded statement bits. E.g. code changes in a code review*". Besides these feedbacks, users also complimented the interface,

P4 said "*I liked the clarity, green highlight, follow each other, I liked the time/playback, matching*", P13 "*What I liked about the interface is that each statement and section is reasonably spaced out which makes it easier to read and I like how there's a clear indicator whether if there's recordings for it or not*".

6.5.2 Recording View

A total of 202 sentences were recorded from our 19 participants. On average, participants recorded 11 sentences. Participants recorded in 25 total articles from the Entertainment, Deaf Culture, Sports, Books, Mathematics, Technology, Food, Geography, Art, and Politics categories.

Challenges and Strategies

Participants were asked if they found any content challenging to record. They reported that they generally selected articles from topics they thought they were the most familiar and comfortable with. For instance, P11 said "*I picked the content I was most comfortable with, and it was straightforward and just facts, so it wasn't challenging. I can imagine if I picked a STEM article or something complicated it would be challenging*". Some participants commented that they felt it was challenging to actually translate the English content into ASL, because they were not sure how to sign some words, or were not sure how to make it so it wasn't a word-for-word English-to-ASL translation, but rather a concept-to-concept translation – P18 said "*it can be a bit challenging to keep it simple and brief yet informative*", P16 reflected that there were "*some words that I'm not sure if they have signs for them*", and P3 summarized "*sometimes I have to reread and think about how I will sign it to try not to be too English*".

We asked participants if they had any strategies they employed while recording content. Most said they did not – they commented about trying not to be too "English" in their signing, with P1 saying "*I would read first, and then think about my understanding of it, and try my best to explain it in ASL. I wanted to avoid one to one or exact English translations*" and P7 commenting "*I tried to find simpler sentences, but most sentences required a lot of fingerspelling. It was challenging to use it, I didn't really think through it, I just read the English part a couple times and then tried my best*". We noticed that not all participants started at the top of their selected articles. It seems that some participants selected sentences throughout

their articles and did not always record consecutively.

Interface Usability

To ensure whether the interface itself caused any significant issues for participants trying to record themselves, we asked them to rate, on a scale from 1 (very difficult) to 5 (very easy), how easy the interface was to use. For the 19 participants, the average was 4.6. It did not appear that the interface caused any further challenges to the recording experience, with P1 saying "*I thought it was straightforward and simple*", P3 "*...liked Redo/keep, add playback/review to watch it before deciding*". Some participants had suggestions about the interface to make their experience better, such as the seemingly abrupt countdown that started as soon as they clicked on an English sentence to record, but there were conflicting responses as some participants said they disliked it, e.g. P7 suggesting "*maybe instead of auto countdown, I felt more pressure, I would rather click on the sentence and then have a record button*" and P13 who said "*I dislike that I can't manually start recording*", while some liked it, e.g. P11 "... *I liked the countdown, 3-2-1*". Participants also made some suggestions for extra features, such as being able to trim the video before submitting, moving the placement of the self-view, an explicit way to "skip" (rather than "redo" or leaving the page).

6.5.3 General Experience

Enthusiasm

After participants had tried out our bilingual interface, we asked whether they "wish more content online provided both English and ASL?" from a scale from 1 (strongly disagree) to 5 (strongly agree). The average response was 4.6 (s.d. .76). Participants gave examples of where they have wished they had access to both English and ASL. These examples included but were not limited to news, podcasts, articles, social media and entertainment. Participants mentioned the Daily Moth, where they have seen both ASL and English captions or transcripts, but they mentioned that these are selected specific news, and they wish they had access to a more broad, general selection of news around the world. Some participants mentioned that they wish they had this kind of bilingual resource when they were learning

about things for their classes, projects, and homework. These findings suggest that people may want to use a tool, similar to our novel interface, in the real-world.

Despite the 19 participants' desire of having more content online provided in both English and ASL, they were not as interested in generating this sort of content themselves. When they were asked "Would you be interested in generating content available in both English and ASL? (1-5: Strongly disagree–Strongly agree)", their average response was 3.6 (s.d. 1.6). Some of their rationale included not wanting to record themselves and/or posting publicly, with P1 saying "*I personally would not, because I personally don't like recording myself and posting online publicly*", P10 "*No, because I feel like I'm signing wrong, or feel that people would judge my signing for being English, etc*", P17 "*No, I'm a camera shy*". Participants who indicated that they are interested were inclined to do so because they felt they would be giving back to the community, and supporting this concept of accessibility, e.g. P2 "*Yes, I wouldn't mind – because I feel like there is a lot of ASL content out there that is not neutral, where the people who are signing are biased, or give biased information. This would be nice and I would like to help increase access while still keeping neutral and spreading information in a neutral manner*" and P3 "*Yes, because if I can get access like this, why not I give back, I don't want others to miss out*".

Personal benefits

One participant mentioned that the site would benefit them because they can use it while teaching, to make sure their students have access and can understand the content fully. Many participants mentioned the site would help them understand content better and go through content faster, since they wouldn't have to spend time looking up specific English words in a separate interface and/or re-reading the English text multiple times. Some participants said that this would also help them improve their signing and presentation skills, since they could benefit from watching their own videos, or pick up new signs for unfamiliar words. For instance, P14 said "*I can improve how best I can interpret English in ASL*", P4 similarly saying "*If I record, I could benefit from watching my own videos, I will see if I signed it clear and understood it well. I would also benefit from reading myself, and others would benefit by reading my videos that I contribute*". When we asked them what kind of content they would like to see on the website,

they mentioned things they were studying, e.g. P2 "*related to my major, tutorials on 3d design software, art, technology, art terminology, for example gothic art history, etc.*", things they were interested in, as P11 brought up "*nutrition, diets, women's health, for example there's a lot of things that are related to hormones, specific foods affecting things, having ASL there would be nice*", general news, information and topics, with P8 saying "*news, health, podcasts, could be a safe place for community involvement, like an area for people to post news around the world, gaming area, etc. Make subcommunities for gamers, etc, same concept as Facebook groups, Reddit subreddits, etc. But everyone is deaf and uses ASL*", among several others. Many participants were very supportive of the idea, and did not care what kind of content is available, as long as a lot is available, with P19 saying "*... every site should have this option, all kind of topics are welcomed*", P8 agreeing "*as much as possible, no limits*", P5 "*there's so many topics to choose from, I would just pick the best and most informative articles for education*", and P7 "*not that I can think of, general Wikipedia articles would be good*". This shows that participants were very supportive of the site.

Concerns

We asked the participants whether they had any concerns while using a site like this. 7 out of 19 participants explicitly brought up the concern that there wasn't control over the quality of users' submitted videos. For instance, many commented that people may not have professional backgrounds, or that they may have something inappropriate or unintentional (such as other people) in the backgrounds of the videos they submit. People also mentioned that users have varying devices and webcam technologies, so that the quality of the videos themselves may not be as good as they'd like – perhaps the lighting would be bad, the video would be choppy or blurry, etc. A few participants also mentioned that the site may find users who misinterpret or inaccurately translate content. P18 said "*It can be misinterpreted easily if the translator is not professional or a novice*", P11 brought up that "*not everyone can translate well, so that would be my concern – there might be some bad videos. I recommend having STEM topics assigned to people who are specialists in that field*". A participant also brought up the issue of privacy, stating that they are concerned about the privacy of their data, and who would "own" it and who would be able to

access it, especially if it was public.

Participant Impression

Overall, participants said that they enjoyed using the website, and that they thought it was "cool to use". When asked "Would you want to use a website like this to read content in the future?", 14 participants said "Yes", and 5 said "No". The participants who said "No" said that they are already comfortable with reading English text alone, and do not require ASL for reading comprehension. Despite this, the participants, on average, said 4.5 to the question "How likely are you to recommend this website to others? (1-5: very unlikely – very likely)". Participants suggested many different groups of people who they would recommend this site to. They would recommend it to DHH individuals, because of the communication barriers they face, as P2 said "*I feel for DHH people, and others who are not good at English, and have communication barriers and have a lack of education, they can learn well through this site*", P7 said "*I would recommend it to people who I know grew up signing and struggle with English, they could improve their reading skills and understand content better*", P16 "*pretty much everyone with ASL especially for people who have weak English skills*", and P4 bringing up "*international friends who don't know English very well, it would help understand English and ASL, or other people whose first language isn't English*".

Many participants mentioned they would recommend the site to people who are learning ASL, since the site is bilingual and has synchronized English and ASL content, as P3 says "*friends who are interpreters, on their own time learn ASL/translation, receptive skills, signing skills*", P15 "ASL students for learning and people who are thriving to learn ASL", P17 "*If the website has enough recordings or gains popularly among users, I would recommend to a friend who isn't fluent in ASL*", and P19 summarizing both ASL learners and the DHH community "*this would be great for people learning ASL. They can practice their receptive skills, and learn how to follow the ASL grammar structure and sign placements. This is also good for every person in the deaf community who may prefer reading captions only some days and ASL other days, or anyone who has a preference in how they absorb information*". Since the site has been seeded with articles from Wikipedia, which are normally informative and neutral, participants suggested peo-

ple who often look up information, or use information in their profession, such as researchers, with P6 saying "*school educational use, like for students to do research, or college students/professionals to record videos, k-12, community college, ...*", P5 suggesting teachers "*I would recommend it to teachers. I think the website would be best for education and is very educational rather than recreational, so teachers could use it to record content and provide information online*", and P8 recommending learners: "*Maybe people who want to learn more things, learners, people who typically look stuff up and read things*".

6.6 Translation Quality Exploration

While our ASL Wiki site was designed to facilitate translation contributions, how the interface design may impact translation quality is unclear. Interpreters typically generate ASL translations of English texts in large sections (e.g. paragraphs). In contrast, our interface elicits of text segmented into sentences to enable readers to access spot-translations within long texts. Our interface also provides built-in mechanisms to facilitate the translation process (e.g. marking completion progress within the text, providing the text and recording interface in the same tool).

To explore the potential impact of the interface on translation quality, we ran a small experiment comparing a set of recordings generated through our interface to a comparable set generated through state-of-the-art recording setups. Specifically, we paid four professional Deaf interpreters to record 20 articles in both setups, and then paid two fluent ASL experts to evaluate all the recordings, and compared the results. Our results suggest that the quality of translations created through ASL Wiki are comparable to those created through standard state-of-the art setups, with potential slight improvements to translation accuracy and recording quality.

6.6.1 Procedure

Video Generation

We paid four Certified Deaf Interpreters (CDIs)[14] to translate a set of 20 Wikipedia articles twice – with both our interface and with their standard translation setup. We chose to work with professional

[14]https://rid.org/rid-certification-overview/available-certification/cdi-certification/

Deaf interpreters in order to enable comparison to state-of-the-art translations. Each CDI was assigned 5 articles to record twice. and we counterbalanced the procedure, so that two CDIs started with our interface and then used their standard setup, and the other two CDIs did the reverse.

In the standard recording procedure, the interpreters were given access to the plain text, and asked to record a translation of the text in sections. They were instructed to use their typical setup and procedures for such jobs – for example, referencing the text and/or personal notes and recording through a video camera app on their laptop or smartphone. This is a standard type of translation job taken on by professional ASL interpreters (e.g. to translate written questions in a survey, or to translate consent form language).

Each CDI translated their own set of 5 Wikipedia articles. Each set spanned a variety of topics, including both technical and non-technical topics. In total, 17 topics were covered in these 20 articles (identified through topic modeling on the most popular 810 English Wikipedia articles): Geography, Entertainment, Sports, Deaf Culture, History, Science, Mathematics, Medicine, Business, Politics, Technology, Military, Philosophy, Food, Books, Religion, Art. Article length ranged from 105-627 words (avg 309), and from 4-29 sentences (avg 15). In total, we collected 308 recordings through our ASL Wiki interface (corresponding to individual sentences), and 111 recordings through state-of-the-art interpreter setups (corresponding to sections).

Video Evaluation

To compare the quality of the two recording sets, we paid two fluent ASL linguists to evaluate each video along five dimensions. These dimensions capture the accuracy of the translation from English to ASL (Q1), the quality of the ASL independent of the English (Q2-Q3), and the completeness of the data captured (Q4-Q5). The dimensions and exact questions that the experts answered for each video are listed below. In addition, the experts had the opportunity to enter additional notes for each video, and we also engaged in a debrief meeting to gather their feedback and observations about the video sets as a whole.

Q1) Translation accuracy: How well does the ASL recording convey the meaning of the English? (Scale

of 1-5)

Q2) Linguistic correctness: How correct is the ASL execution (e.g. were there many mistakes with handshape, movement, grammar, etc.)? (Scale of 1-5)

Q3) Signing naturalness: How natural is the ASL (i.e. how similar is it to ASL you might run into in real life)? (Scale of 1-5)

Q4) Recording quality: How good is the recording quality (e.g. is it blurry, is the lighting good, etc.)? (Scale of 1-5)

Q5) Signing captured: Is the full signing space captured in the video (i.e. hands, torso, surrounding are)? (Yes/No)

6.6.2 Results

The expert evaluations of the recordings generated through our interface and through the CDIs' standard setups were comparable, across all five explored dimensions. Figure 6.4 shows the overall results – average score and standard error for Q1-Q4, and the percent of videos that were evaluated as having captured the full signing space for Q5. We ran two-sided Wilcoxon-Mann-Whitney tests with Bonferonni correction to compare evaluations of the two interfaces for Q1-4. For Q1 (Translation accuracy), Q2 (Linguistic correctness) and Q3 (Signing naturalness), there was no statistically significant difference ($p>.0125$). For Q4 (Recording quality), the test showed statistical significance Q4 ($U=83175.5$, $p<.005$). We also ran a χ^2 test to compare Q5-Signing captured, which was not statistically significant ($p>.05$).

During our debriefing, the expert evaluators identified some patterns in the data. They noted that the recordings, in particular those created through ASL Wiki, contained straight translations rather than interpretations. For example, the interpreters did not tend to elaborate on concepts from the text to ensure that the meaning in ASL is clear, or to provide additional context not provided in the text. Instead, they tended to stick to the exact text. They also noted some examples where it seemed that the interpreters had not done the full prep work to understand the content they were translating. For example,

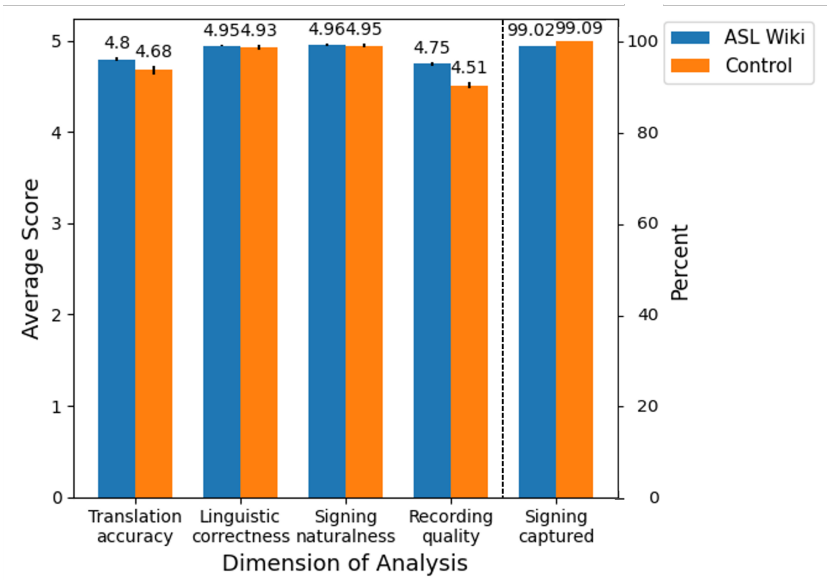


Figure 6.4: Comparison of expert evaluations of ASL translations of 20 Wikipedia articles, recorded by CDIs through ASL Wiki and a control state-of-the-art setup. For Q1-4, (Translation accuracy, Linguistic correctness, Signing naturalness, and Recording quality), the bar chart shows the average and standard error of expert evaluation. For Q5 (Signing captured), the bar chart shows the percent of recordings evaluated as having captured the full signing space.

this was evident to one expert in a translation of some plant anatomy, which lacked the appropriate visual representation. One expert also noted that they had expected to see a larger difference in quality between the recording sets, in particular due to the difference in text segmentation lengths. They were surprised that there was not a larger difference in translation accuracy and quality for the longer and shorter excerpts.

6.7 Discussion

Generally, the results of our exploratory studies suggest that it may be possible to use specialized interfaces to crowdsource ASL translations of English text, to provide valuable bilingual resources to the community and to curate ASL data. Our user study results suggest that users would find value in such a bilingual ASL and English platform, and would be willing to contribute, especially if incentivized. At the same time, our translation quality exploration suggests that the interface enables high-quality translations. In this section, we provide further discussion on our exploratory work, the limitations of

this initial work in this space, and related future work.

6.7.1 User Experience

Because participants were not incentivized further for contributing more videos during our user study, the majority of participants only contributed until they figured out and were satisfied with the user interface and experience for the recording view, with an average of 11 sentences per user. It seemed that participants generally chose to contribute to topics that were personally meaningful to them, especially those who contributed a larger number of recordings. It is possible that an expanded range of topics that interest more people would thus incentivize contributions from the community. Further incentivization, such as credit for class or monetary payment could also be beneficial to deployment at scale.

Participants indicated that the reading and recording interfaces of our website design were easy to use. Even though participants all thought the site was easy and intuitive to use, several would rather only use it for reading bilingual content, rather than contributing ASL videos. They thought it was helpful to view content in both English and ASL, and mentioned several cases where they wish they had this level of accessibility in media. They talked about some of the challenges and strategies used while recording. The website was strongly supported and all participants identified populations that they would recommend the site to. Participants also suggested many different topics that could be added to the interface that would benefit them and others in mind.

Even participants who commented that they were fluent in English and ASL still indicated that seeing content in a bilingual, bimodal form was useful. Even if it did not help them understand the content itself better, some participants still mentioned that they could pick up new signs or improve their signing and presentation skills. Overall, participants enjoyed using the website, and identified several use cases and target audiences who they would highly recommend the interface to.

During the interview portion of the user study, we collected feedback from participants so that we could further iterate upon our design. These feedbacks would also be useful for future researchers who want to generalize our interface, and potentially use it for other signed or written languages. While our exploratory user study serves as a proof-of-concept, several research questions have arised. We have

identified several research avenues and next steps as a result of this work.

6.7.2 Translation Quality

In our translation quality exploration, it is possible that linguistic correctness was slightly more reliable with our interface due to reduced cognitive load. Our interface provided required shorter excerpts of text to be translated. It also simplified the recording task by keeping track of where the user was within the text, auto-progressing to the next excerpt, and providing the text and video feedback in a single interface rather than requiring the interpreter to manage two separate interfaces for these components. It is also possible that the recording quality was slightly better on average with our interface because the quality of the recording was less dependent on the quality of apps that the interpreter has available to them. While we did not provide hardware, we provided built-in recording software in our website, unlike state-of-the-art setups that are dependent on the recording software that interpreters have access to and know how to use.

While our exploration suggests comparable translation quality with ASL Wiki compared to state-of-the-art translation setups, it still leaves open questions about the impact of isolated interface components. For example, it would be interesting to examine the effect of different text segmentations within our interface, possibly ranging from individual words, to sentences, to paragraphs or sections. Similarly, it would be interesting to experiment with the effect of different types of visual cues for orienting the translator within a page of text. It is also possible that the impact of the interface on the translation quality may vary depending on the experience or fluency of the user.

6.7.3 Limitations and Future Work

The website was switched off after the study, so users could not return to it if they wanted to. There is a need for a larger, more longitudinal study to see how users use the site over a period of time in the real world, rather than a short 1 hour session where they use the site for the first time and answer survey and interview questions with a researcher. Leaving the website on, and having a longitudinal study would enable investigation of motivation behind user participation, showing if users desire further

incentivization beyond the scientific and accessibility contributions of the site.

Additionally, most of our participants already had a Bachelor's degree, which may have biased our results; as a result, it is important for future studies to capture more diverse participants from the DHH community. Such studies would allow for deeper insight in user participation and behavior, and the additional data collection would enable deeper linguistic analysis and open up several research questions.

Since some participants in our user study skipped sentences, selecting nonconsecutive sentences to contribute, there are gaps in the articles. Our user study participants supported the idea of the website, said it was easy to use, but many of them said they would personally not contribute themselves. To encourage users to contribute in completeness, further research is needed to investigate different incentivization methods. There are several ways we can imagine this happening, such as strengthening the gamification inside the website (emphasizing the experience points they earn as they contribute, displaying a leaderboard of the top contributors), or monetary compensation for some arbitrary milestone of amount of ASL videos an user contributes. Another possible avenue to investigate is educational tasks, e.g. ASL interpreting students could contribute to gain credits for certification or program requirements.

For this user study, we chose to implement a stand-alone website pre-populated with a sample of Wikipedia articles, limiting the type of content available for participants. We chose this implementation, rather than a web plug-in or other setup with broader content for several reasons: ability to choose English texts that are open for public use, utility to users in having a complete translation as opposed to sparse translations across more content, and implementation feasibility. Still, user study participants brought up many different types of content and explained their experiences with other real-world content. Consequently, other types of content, and expansive interface designs including web plug-ins, should be explored. The utility of our interface with other signed/written language pairs, or exploring other potential user groups (e.g. those recommended by participants, such as K-12 students) could also be investigated. Different use cases may or may not require other interface changes, which would be explored in this research avenue.

There were two major concerns brought up during our user study. Users were concerned about

the level of control over data quality – since this is a crowdsourced approach, it is the contributors' responsibility to have a good background in their signed videos, ensure there is good lighting, and that the video is not choppy or blurry. The other major concern was privacy. This is a very complicated topic ([21, 24, 100]), and more research is needed about privacy concerns when it comes to crowdsourced ASL datasets. Another data quality research question is whether the crowd would be able to control the data quality at a bigger scale. We included an upvote/downvote button where participants could give feedback, but we did not study this further, since 12 out of 19 participants did not use it. We also had a small number of sentences from each participant, but if a larger and more longitudinal study was conducted, it could be investigated how users use this feature.

We have also run a small experiment comparing the quality of translation recordings made through our interface and through a state-of-the-art setup. This exploratory study suggested that the quality of translations created through ASL Wiki are comparable to those created through state-of-the-art setups, and potentially might enable slight improvements. While this is promising, we have not evaluated the crowd-generated dataset from our participants (as we did not have a control dataset to compare to, since general community members do not normally engage in translation). It would be useful for future researchers to investigate this, as well as to conduct in-depth linguistic analysis. For instance, it is possible that our interface reduced the cognitive load of the signer, as well as the technical requirements, which may have elicited more natural and linguistically correct translations.

As mentioned above, privacy is another issue that may impact the design and use of ASL Wiki and future work. The research community has only recently begun to explore privacy concerns related to sign language videos and thought about how they can be addressed [24, 100]. This initial work began to explore the impact of filtering videos, for example by blurring the video or anonymizing facial features. However, acceptability of these approaches is poorly understood, and their technical implementation is limited. Indeed, it is possible that the community might prefer different approaches altogether, for example protective licensing or enhanced security and transparency of data use. Once a better understanding of the privacy needs and appropriate solutions have been developed, such techniques could be incorporated into ASL Wiki and similar applications, and make a ripe area for future work.

6.8 Conclusion

The lack of sign language bilingual resources and the lack of sign language datasets are difficult problems to solve, mainly due to the cost, resources needed, and amount of human effort required to label and annotate data. In this work we have addressed both of these problems by presenting a novel interface. Our interface provides a side-by-side ASL and English synchronized interface, streamlines pre-labeled data collection, and enables a crowd to contribute to piecemeal translation. We pioneer exploration of the question of how to enable everyday signers to contribute to continuous content translation efforts, and how DHH users would respond to crowd-generated content.

Chapter 7

Virtual Prototype Implementation and Remote Data Collection[15]

Prior chapters had identified a viable platform to scale a dataset of individual signs without scaling prohibitively expensive labelling problems. They had also extended that work by creating a novel interface that can be used to collect sentence-level signs. While those efforts supported collecting data necessary for general sign language recognition technologies, they did not focus on the specific domain of signing directly to a personal assistant device.

7.1 Introduction

This chapter returns the focus of the dissertation research to the context of personal assistants. Specifically, this chapter discusses a methodology to enable DHH users to interact with an actual personal assistant that appears to understand ASL. This chapter describes how this was done using a remote modality, and in a later chapter, an in-person modality is discussed. Since Part II of this dissertation focuses on the issue of data collection, this chapter will describe the setup used for collection and the overall dataset that was acquired. The detailed analysis of what had been collected will be presented

[15]The information in this chapter is based on a joint project with my advisor (Dr. Matt Huenerfauth), and graduate students I supervised at RIT (Matthew Watkins and Kira Hart) whom assisted me with data collection and qualitative data analysis. The results were published as a paper at the CHI'22 conference [64].

later (in chapter 9 in Part III of this dissertation).

7.2 Wizard-of-Oz Methodology

Up until now, prior work has asked DHH users to imagine interacting with a personal assistant device that can understand ASL, but have not yet had the opportunity to actually do so. When DHH users are given this opportunity, they might spontaneously query the device in ways they had not imagined previously, and it is important to capture this behavior. Thus, we are motivated to setup a prototype device that would allow DHH users to freely issue commands to a personal assistant in ASL, in order to start establishing guidance for future designers of such devices, as well as sign language recognition researchers, e.g., specific commands to support, ASL terminology to use for command and control of the device, how the device should respond when there is a potential error, and other insights.

Due to conditions during the COVID-19 pandemic at the time of this study, our collection of ASL signing data had to occur in a remote manner, using the Zoom videoconferencing platform. In our sessions, DHH users interacted with a device using sign-language commands which were “voiced” into spoken English by an interpreter who had remained unseen by the participant, to make it appear as if the Alexa device itself had understood the sign-language command. We chose the Amazon Alexa Show 10-inch (2020 model) for this study because it is a popular consumer personal assistant device with a large screen that could display results of commands and had captioning features to display what the device said. While a similar Wizard-of-Oz methodology was employed in [139, 179], these studies did not give participants the opportunity to spontaneously interact or engage in back-and-forth dialogue with Alexa. Additionally, these prior studies did not investigate the other phases of the device interaction or conduct in-depth interviews with the participants discussing their user experience and opinions.

Before the participant had joined the scheduled Zoom meeting, an Amazon Echo Show device was set up with its own video stream on the Zoom call, with video and audio initially turned off, named "Alexa." The sign language interpreter also joined the Zoom meeting as a separate meeting participant, with their video and audio turned off, and with their "name" shown in the Zoom meeting as "software" (to hide their identity and to convince the participant that a software program was translating their

commands, rather than a live person). While the interpreter was muted on the Zoom meeting, they were, in reality, sitting physically in the same room as the Amazon Echo Show device and thus were able to speak commands to it. A member of our research team, the moderator, was also on the Zoom call, visible and ready to meet with the DHH participant. Figure 7.1 gives a diagram to demonstrate this setup.

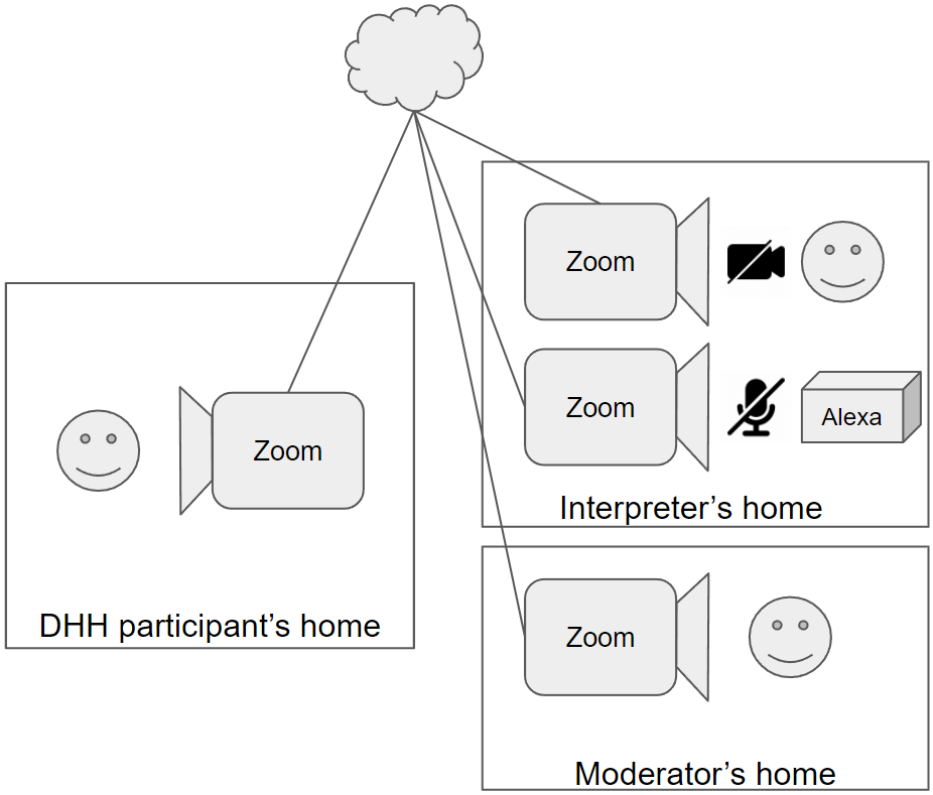


Figure 7.1: Diagram illustrating remote Wizard-of-Oz setup

After the initial interview questions, when it was time to commence the Wizard-of-Oz interaction, the moderator used ASL to ask Alexa to join the Zoom meeting and to turn on its stream. The moderator explained to the participant that Alexa was here and ready to start, and that the DHH user could try out any command they would like (with the exception of connecting to smart devices, such as their phone, TV, lights, doorbell, etc., due to the virtual nature of the experiment on Zoom due to COVID-19). The participant was told that they would have to get Alexa's attention before it would be ready to listen for a command, and that Alexa would only be watching the DHH participant and not the moderator. The

DHH participant then proceeded to try out different ASL commands and queries.

When the interpreter voiced "Alexa," to wake-up the device, the device displayed a horizontal blue line across the bottom of the screen to show it was ready for a query. When the command was processed, the device displayed a result (e.g. pictures, video, search results, etc.), and it had caption text displayed on-screen for its reply. Figure 7.2 shows a screenshot from a sample video where the Echo Show device displays this blue line, and it can be seen the device is also displaying a picture with text.

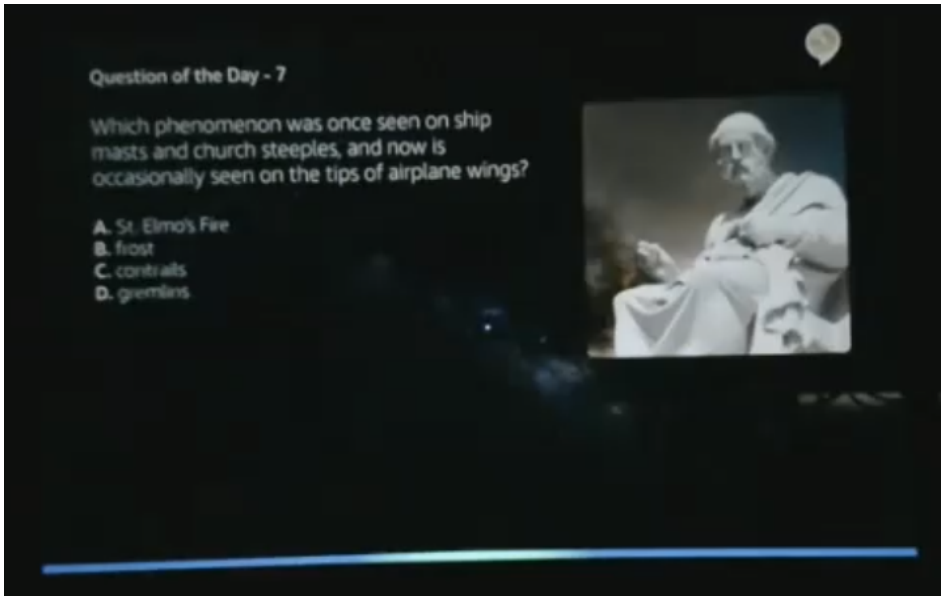


Figure 7.2: Screenshot from sample video showing blue line to inform the user that it is ready for a query

When the time allotted for the recording session elapsed, the moderator told the participant that time was up, asked Alexa to leave, and then continued with the aforementioned final interview questions.

7.3 Dataset Collection and Annotation

There were 21 participants in this study. For each participant, the Zoom meetings were recorded using the built-in recording feature, and also using a screen-recorder software on the moderator's computer. One recording focused on capturing the DHH participant alone, while the other recording focused on capturing the entire Zoom meeting, including Alexa's stream, so that it could be seen how Alexa re-

sponded.

Three members from our research team watched the entire set of recordings (approximately 21 hours total) to transcribe everything the participant had signed, and all of Alexa's responses. ASL commands were annotated using glosses, which are English words that are conventionally used to refer to transcribe ASL signs. In total, there were over 1,400 utterances transcribed. Table 7.1 illustrates the layers of annotations added to this dataset, and table 7.2 shows the first 10 rows of the annotation file.

As part of this transcription, researchers described every "wake-up" command they saw. After all data was written into a spreadsheet, the research team agreed on a set of codes to describe the unique wake-up methods they saw. Using these codes, each member went through and assigned each transcribed wake-up utterance to a code.

Table 7.1: Virtual Wizard-of-Oz data annotation descriptions (actual data sample shared in table 7.2)

Column	**Description**
Participant ID	ID for the participant
Video Filename	Name of the .mp4 file associated with the annotation
Command number within Video ID	Number of command inside the video (each video has multiple commands)
Wake up Method	Which wake-up method was employed by the participant (see table 9.1)
Command in English	Transcription of the ASL command in English
Error Type (If it happens)	Error type, which will be one of the following: 1) **Silence** (Alexa ignored the command) 2) **Confusion** (Alexa heard but didn't understand the command, Alexa says something like "I don't know that") 3) **Suggested** (#2 but Alexa suggested something at the end) 4) **Undesired** (Alexa understood the command but didn't give the desired result) 5) **Failure** (Alexa crashed, hardware error, software error (e.g. captions stuck), etc.) 6) **Question** (Alexa understood but missed key information thus asking the participant for confirmation) 7) **N/A** (No error, Alexa responded as expected)
How participant followed up on the error	Follow-up type, which will be one of the following: 1)**Repeated** (Self-explanatory – same signing and wording) 2) **Reworded** (Self-explanatory – changed signing or wording, including only changing the greeting) 3) **Ignored** (Ignored, moved on) 4) **Question** (The participant asked either Alexa or the researcher about the error they're seeing – the participant basically saying "what can i do?" or "what do i do?") 5) **Played Along** (When Alexa suggested something, sometimes they'll accept the suggestion even though it wasn't what they were expected. Otherwise, the error type was #4 but they accepted and asked about that result) 6) **N/A**

Table 7.2: Sample annotations from section 7.3

Participant ID	**Video Filename**	**Command number within Video ID**	**Wake up Method**	**Command in English**	**Error Type**	**Error follow-up**
P01	P01.mp4	1	Hello	Hello can you schedule a plan with my friend at noon to 2P with my friend Koby?	Silence	Repeated
P01	P01.mp4	2	Hello	Hello can you schedule a plan with my friend at noon to 2P with my friend Koby?	Silence	Reworded
P01	P01.mp4	3	Hello	Hello can you schedule a plan with my friend at noon to 2P?	Silence	Reworded
P01	P01.mp4	4	Hey	Hey I want to schedule noon to 2 with my friend	Silence	Repeated
P01	P01.mp4	5	A-l-e-x-a	Alexa I want to schedule noon to 2 with my friend	Confusion	Ignored
P01	P01.mp4	6	Hey-A-l-e-x-a	Hey Alexa give me tips for cleaning.	Silence	Repeated
P01	P01.mp4	7	Hey-A-l-e-x-a	Hey Alexa give me t-i-p-s for cleaning.	Suggested	Played Along
P01	P01.mp4	8	None	Basketball	N/A	N/A
P01	P01.mp4	9	None	Sure	Silence	Ignored
P01	P01.mp4	10	A-l-e-x-a	Alexa what is the weather today in Foxboro?	Undesired	Reworded

For the main phase of issuing commands and requests, after the spreadsheet of transcriptions was complete, BERTopic, a topic modeling tool, was used to automatically identify some common topics in the list of commands [69]. Considering this initial analysis as inspiration, the research team went through this list and re-examined the set of transcribed commands, prior to holding a meeting among the team to discuss the set of codes to use when categorizing the commands that had been transcribed. After an initial list of topic categories was composed, the research team iteratively examined each of the commands and labeled them with the respective categories. For example, the most frequent category collected was "command and control," where users adjusted device settings and navigated through its interface, including issuing "Yes"/"No" commands. The next top 4 categories were "entertainment," "lifestyle," "shopping," and "trivia, calculations." The full results of this analysis appears in a later chapter, chapter 9.

During the original transcription process, our team also made note of whether any error, breakdown, or unexpected response had occurred. After transcribing every error, the research team used a similar method above to identify a set of labels to use in categorizing the errors. After coding a subset of the data and holding a further discussion, all three team members labeled all of the errors in the data set to categorize them into six categories. The team iteratively labeled the errors in the dataset and met to discuss any disagreements in the labeling. Since we were also interested in how the user behaved after an error happens, the team also transcribed what the user had done to follow-up on these errors. Using a similar procedure as above, the team iterated until arriving at five labels for the user behaviors that were observed. Most of the time, participants "ignored" (where they ignore the erroneous response and move on to the next command), or "repeated" (repeated the command with the same signs and wording) when there was an error. The full results appear in chapter 9.

7.4 Dataset Release

We collected over 1,400 user utterances across many different topics, many of which reflect current events or news, e.g., the COVID-19 pandemic. The composition of this dataset may be useful for designers of personal assistant devices to understand different ways DHH users would use this technology,

and the dataset itself may be useful for sign-language-recognition researchers as training or testing data for their AI models. In comparison to use among the general population, the "DHH-specific" category of commands revealed some ways the DHH population may make special use of such devices. Some commands in this category related to accessibility features (e.g., captioning on the device), and some were questions about services and technology (e.g., finding deaf professionals, assistive technology, deaf education and studies). There were also requests to launch video calling platforms, such as FaceTime and Video Relay Services (e.g., Convo [36]). In a few cases, participants asked Alexa if it could sign or tell deaf-related jokes.

We share the video and annotation dataset publicly, at this URL: http://doi.org/10.17910/b7.1392, so the actual ASL recording of these commands is available to the research community – something that no prior work with personal assistants has done. Designers of personal-assistant devices may also benefit from noting the specific ASL vocabulary items and phrases that DHH participants used spontaneously for various "command and control" commands, e.g., for navigating the user interface of the device or responding to device prompts. Figure 7.3 shows screenshots from two sample videos for demonstration.

At http://doi.org/10.17910/b7.1392, there are 21 "session" folders – 1 for each participant in the user study. Each folder is named "P01", "P02", and so on, until "P21". In each respective folder, is a .mp4 file named "P01", "P02", and so on. There are also two supplementary materials. They are both `.csv` files. The first one is basic demographics for the 21 participants in this study. This file is named `demographics_for_sharing_chi2022` and descriptions of the content are given in table 7.3. The second is named `data_annotations_chi2022` which contains the annotations for the 21 videos, as described earlier in this section. Descriptions for this are given in table 7.1. Figure 7.4 illustrates the folder structure and contents of this dataset.

The descriptions of the columns in our demographics file are given here in table 7.3, and the demographics data itself is given in table 7.4, showing the basic demographics of our participants in the experiment as described earlier in this section.

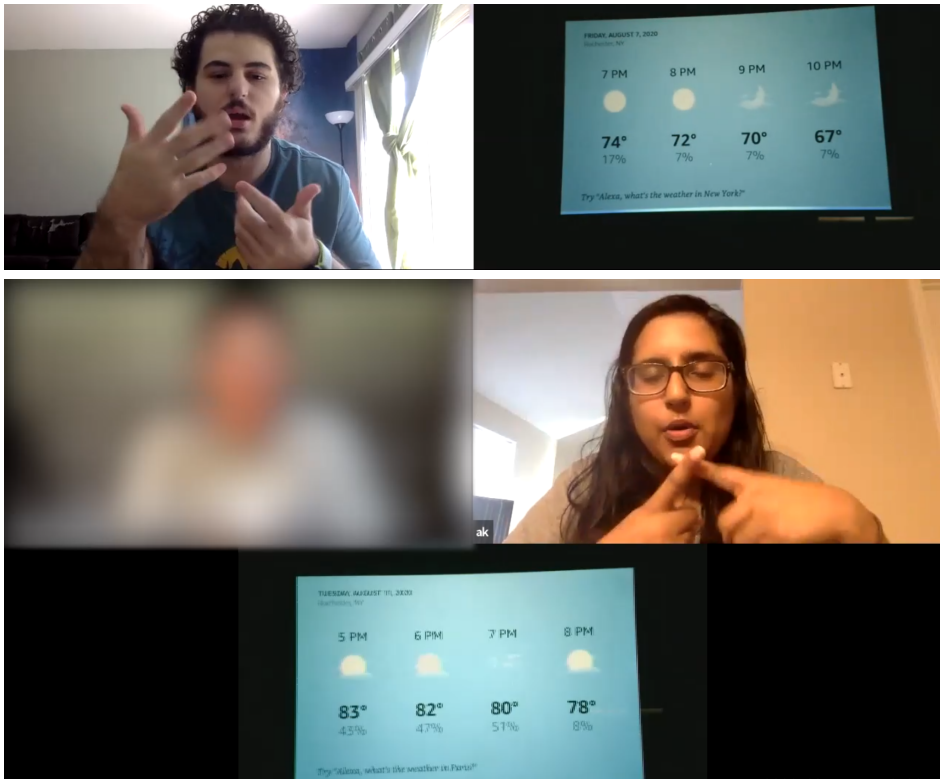


Figure 7.3: Screenshots from selected remote Wizard-of-Oz videos

Table 7.3: Basic participant demographic data columns and descriptions (actual data shared in table 7.4)

Column	**Description**
ID	Arbitrary consecutive integers to label a row
Gender	Gender of the participant
Birth Year	Birth year of the participant
How would you describe yourself? [Deaf/deaf/Hard of Hearing/Hearing]	Participant answer to this question.
At what age did you become DHH?	Participant answer to this question.
At what age did you begin to learn ASL?	Participant answer to this question.
Are your parents DHH?	Participant answer to this question.
Did your parents use ASL at home?	Participant answer to this question.
In elementary school, did you use ASL?	Participant answer to this question.
What language do you use at home? [0= 100% English, 10= 100% ASL]	Participant answer to this question.
What language do you use at work/school? [0 to 10]	Participant answer to this question.
What language do you use with friends/family? [0 to 10]	Participant answer to this question.

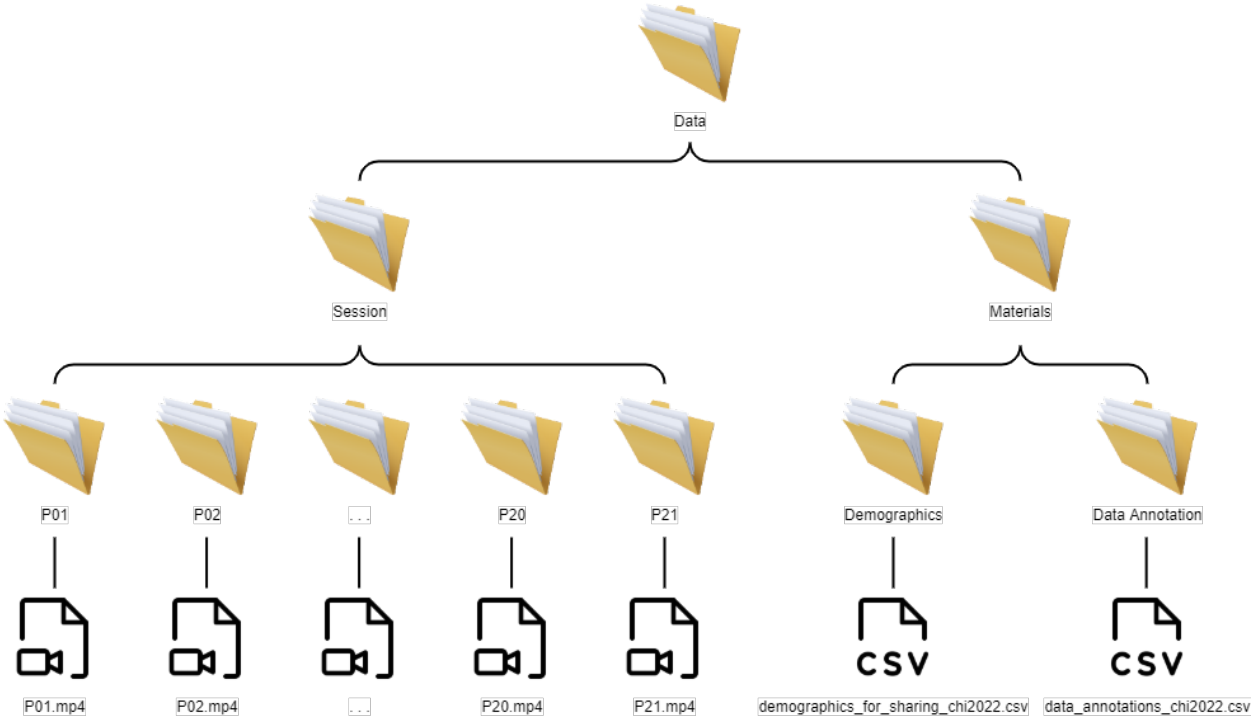


Figure 7.4: Diagram illustrating folder structure and contents of the dataset

7.5 Contributions

Sign-language-recognition researchers may benefit from the videos in our dataset, which demonstrate the variety of linguistic structures and word-order of these commands. Considering this data, researchers can ensure that sign-recognition models for use in personal assistant devices are able to work with a variety of vocabulary, people, and signing styles. For instance, while several participants signed in fluent ASL, some used a more English-like word-order or structure to their signing. This type of code-switching between fluent ASL and English-like signing is common among DHH signers, e.g., when they are signing with someone whom they believe may be a novice signer [105]. We speculate that the experience of Alexa sometimes giving inappropriate responses to commands may have led some participants to naturally engage in such code-switching behavior.

Our findings and discussions (presented in chapter 9) from analysis of this data contribute to improving the accessibility of conversational-interaction user-interfaces through sign-language interaction, to help mitigate the emerging accessibility barrier that the proliferation of voice-controlled interfaces

are posing for DHH users. In addition to these empirical contributions, the disseminated video dataset, which is the first of its kind, will be useful for computer-vision sign-language recognition researchers.

Table 7.4: Basic demographics of participants from section 7.3

ID	Gender	Birth Year	How would you describe yourself?	At what age did you become DHH?	At what age did you begin to learn ASL?	Are your parents DHH?	Did your parents use ASL at home?	In elementary school did you use ASL?	What language do you use at home? [0=100% English, 10=100% ASL]	What language do you use at work or school? [0 to 10]	What language do you use with friends or family? [0 to 10]
1	Male	1997	Deaf	0	1	No	Yes	Yes	8	5	6
2	Female	1996	Deaf	0	0	Yes	Yes	Yes	10	5	8
3	Female	1994	deaf	0	16	No	Yes	No	10	5	5
4	Male	1995	Deaf	0	4	No	Yes	Yes	10	10	10
5	Female	1997	Deaf	2	11	No	Yes	Yes	10	10	7
6	Female	1994	deaf	0	2	No	No	No	10	5	5
7	Male	1983	Deaf	0	3	No	Yes	Yes	10	9	7
8	Female	1983	deaf	0	0	No	No	Yes	10	9	5
9	Male	1997	Deaf	0.5	0.5	No	Yes	Yes	4	9	5
10	Male	1995	Hard of Hearing	3	3	Yes	Yes	Yes	7	5	9
11	Male	1994	Deaf	1	5	No	No	Yes	2	10	7
12	Female	1996	Deaf	0	3	No	Yes	Yes	10	6	10
13	Female	1996	Deaf	4	5	No	No	Yes	9	10	8
14	Male	1998	Deaf	0	0.5	Yes	Yes	Yes	10	10	7
15	Male	1994	Deaf	2	3	No	Yes	Yes	10	5	5
16	Female	2002	Deaf	0	0	Yes	Yes	Yes	9	6	8
17	Female	2001	Deaf	0	0	Yes	Yes	Yes	10	5	10
18	Female	1990	Deaf	0	0	No	Yes	Yes	9	10	10
19	Male	2000	Deaf	2	2	No	Yes	Yes	8	10	10
20	Male	1999	Deaf	0	0	Yes	Yes	Yes	8	6	9
21	Male	2000	Hard of Hearing	0	16	No	No	No	0	0	1

Chapter 8

In-Person Wizard-of-Oz Data Collection

8.1 Introduction

While the remote Wizard-of-Oz methodology described in chapter 7 directly engaged with the DHH community and enabled us to learn several valuable insights, it had limitations due to the lack of real-world conditions.

As is normal with general users of personal assistants, a personal assistant could essentially be placed anywhere in a residential setting and the user would be free to locate themselves wherever they'd like as long as their voiced commands are picked up by the device [128]. If automatic sign recognition was embedded in current personal assistant devices, they would have to capture users through an integrated webcam, meaning the user could still freely move around in a home, as long as they are within line of sight. During the Zoom videoconference calls in chapter 7, participants had to position themselves in front of a computer with a webcam, had to sign in a limited manner so that their ASL was visible in their video input for the Zoom meeting, and they were limited to the personal-assistant video stream for output from the device. Additionally, during the remote protocol, users were not able to issue commands or requests that use other smart devices, such as lights or a TV, which is a popular feature of personal assistant technologies, especially with modern smart TVs being commercially released with integrated voice-controlled personal assistants.

In this chapter, an in-person Wizard-of-Oz protocol was developed, allowing DHH users to interact

with a personal assistant using ASL in a natural and instinctive manner while in a home-like living room and kitchen environment. This enabled us to investigate several observational research questions, such as some linguistic properties of the in-person interaction where DHH users can use their full signing space, referencing to objects in the room, and where they naturally position themselves in proximity to the device in a room. While this chapter describes the employed data collection methodology, chapter 10 focuses on the data analysis, including a formal list of research questions and results from the observational, quantitative, and qualitative analysis.

8.2 Room Setup

Considering that personal assistants are placed in the homes of users with a variety of residential styles, to emulate a residential living room and kitchen setting, we set up common living room furniture, and placed a kitchen area next to it. Inside these areas, different tasks were setup for participants so that they replicated typical behaviors of personal assistant users inside a household.

For the "living room" setup, the following materials were used:

- **Echo Show:** The Amazon Echo Show is a smart speaker part of the Amazon Echo line of products, and is a popular consumer personal assistant device (using Amazon Alexa). It has a screen for displaying output to the user, and has an integrated camera. Figure 8.1 shows a picture of the device.
- **Fire TV:** The Amazon Fire TV stick (fig. 8.2) is a digital media player and microconsole developed by Amazon. It is compatible with and can be controlled through Alexa. The Fire TV stick was plugged into a TV via HDMI, for displaying streamed media.
- **Table:** A table was used to prop the Echo Show and TV, and was put in the back end of the living room area.
- **Three-seat couch:** A couch is a common living room furnishing, and provides different seating positions while remaining on the same couch. This couch was positioned near the Echo Show, in a way so that the different seating positions impact the view of the user from the device.

- **Side Chair:** A side chair is also a common living room furnishing, and is a fixed seating position. This chair was placed in the front of the living room area, opposite of the Echo Show device for greatest visibility and convenient proximity to the couch and kitchen area.
- **Coffee table:** A coffee table was put in front of the couch, giving space for storage of materials, e.g., those included in participant tasks.
- **Table lamp:** A side table with a lamp was placed in between the couch and side chair, to serve two purposes. First, it acted as a boundary of the living room area and guided participants to stay within the experimental setup. Secondly, it served as a smart device as part of the experiment tasks, using a WiFi and bluetooth-enabled lightbulb.

Figure 8.1: Picture of Amazon Echo Show device
Source: https://www.techhive.com/article/583487/amazon-echo-show-2nd-generation-review.html

Figure 8.2: Picture of Amazon Fire TV stick
Source: https://www.mytrendyphone.eu/shop/amazon-fire-tv-stick-4k-alexa-voice-remote-269586p.html#gallery-4

The kitchen area, which is opposite of the couch, had:

- **Counter:** The counter was a long table that spanned the majority of the kitchen area.
- **Equipment:** There were things here on the counter for the experiment tasks, such as a water dispenser, tea and a powdered drink mixture, prop food, and a smoke alarm.
- **Cabinet:** There was a cabinet for storage of the supplies used in the kitchen area.
- **Floor lamp:** This served as the primary "light" used for the kitchen, and it also served as a smart device, as well as a room boundary.

The front (the direction in which the user would face) of the living room and kitchen areas had a fabric backdrop to further enhance the feeling of a natural residential setting. A layout of all the materials listed above is provided in fig. 8.3.

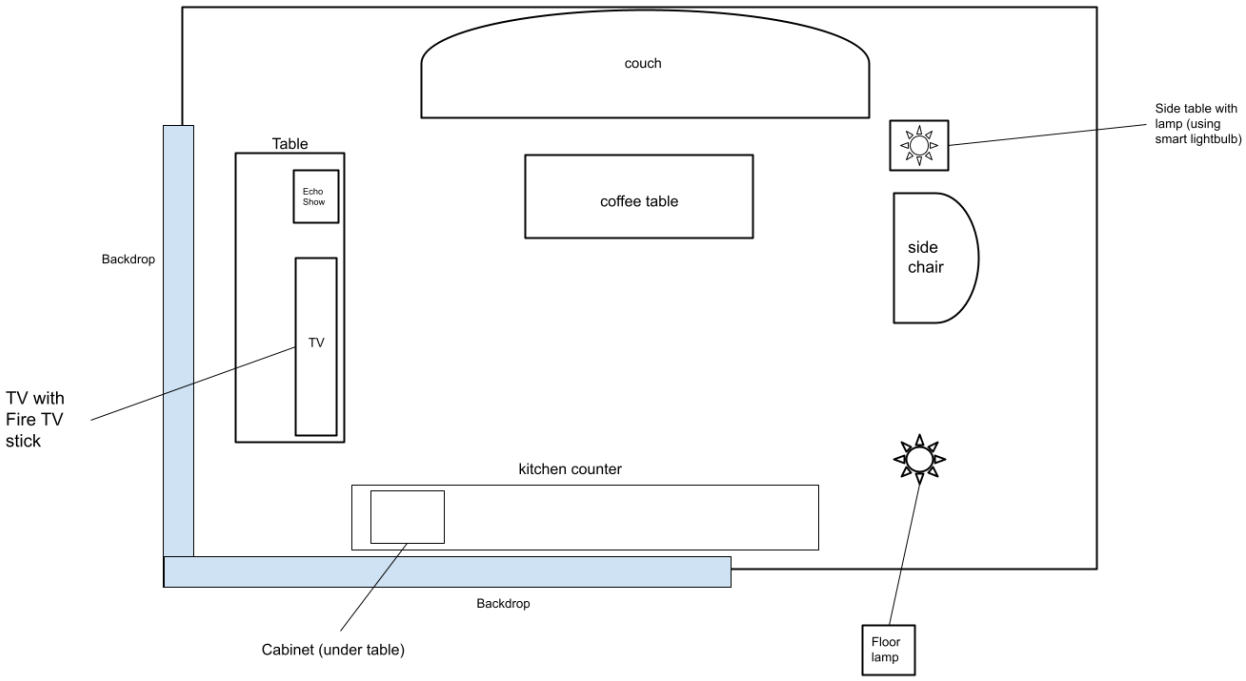


Figure 8.3: Room layout for in-person experiment (without Wizard-of-Oz related equipment)

8.3 Wizard-of-Oz Methodology

To enable the Wizard-of-Oz, an ASL interpreter needed to be hidden and still be able to see the DHH user, in order to voice their ASL signs into spoken English for the Echo Show Alexa device. A wide webcam was placed on top of the Echo Show and plugged into a laptop behind the device. This laptop initiated a Zoom videoconference meeting, allowing the ASL interpreter to join from a different room. When the interpreter voiced the commands in English, the audio coming from the laptop speaker was be picked up by the Echo Show device.

During the virtual Wizard-of-Oz experiment, a researcher who moderated the session, was always visible on the participant's Zoom view. In order to give participants the most comfort and naturalness during this in-person setup, the moderator (an ASL signer) was also in a separate room (simultaneously complying with university-mandated COVID-19 policies that were present at the time of writing). They joined the same Zoom meeting so that they could see the participant, and their Zoom stream was broadcast on a separate TV that was between the Echo Show device and the couch. When it was the participant's time to go through the experiment tasks and interact with the personal assistant, the moderator had their video stream turned off and only turned it on to answer any questions that arose, or when they needed to direct the participant. Since it was be useful for both the moderator and interpreter to be able to see the Echo Show device screen output to know what it was displaying, another webcam was be setup to capture this and also joined the same Zoom meeting.

8.4 Dataset collection and Annotation

The Zoom meeting was recorded, and included a variety of different perspectives. The device-view camera saw the personal assistant's field-of-view in a residential living room and kitchen area, capturing the user's location and signed commands to the device. This camera was also capturing the audio from the ASL interpreter who was speaking commands to the device in English. Since the device-screen camera was also streaming in the same Zoom meeting, it was recorded and automatically synchronized with the timing of the device-view recording. Additionally, a separate camera was placed outside of

the experiment boundary and propped high on a tripod to capture a "bird's eye view" of the entire setup. This was an offline recording and served as an alternate view of the participant in the room, akin to a security camera an user may place in their home, which could be used in future applications of sign language recognition technology. These recordings were used for analysis, which is presented in chapter 10. Figure 8.4 provides an updated room layout diagram showing these equipment additions.

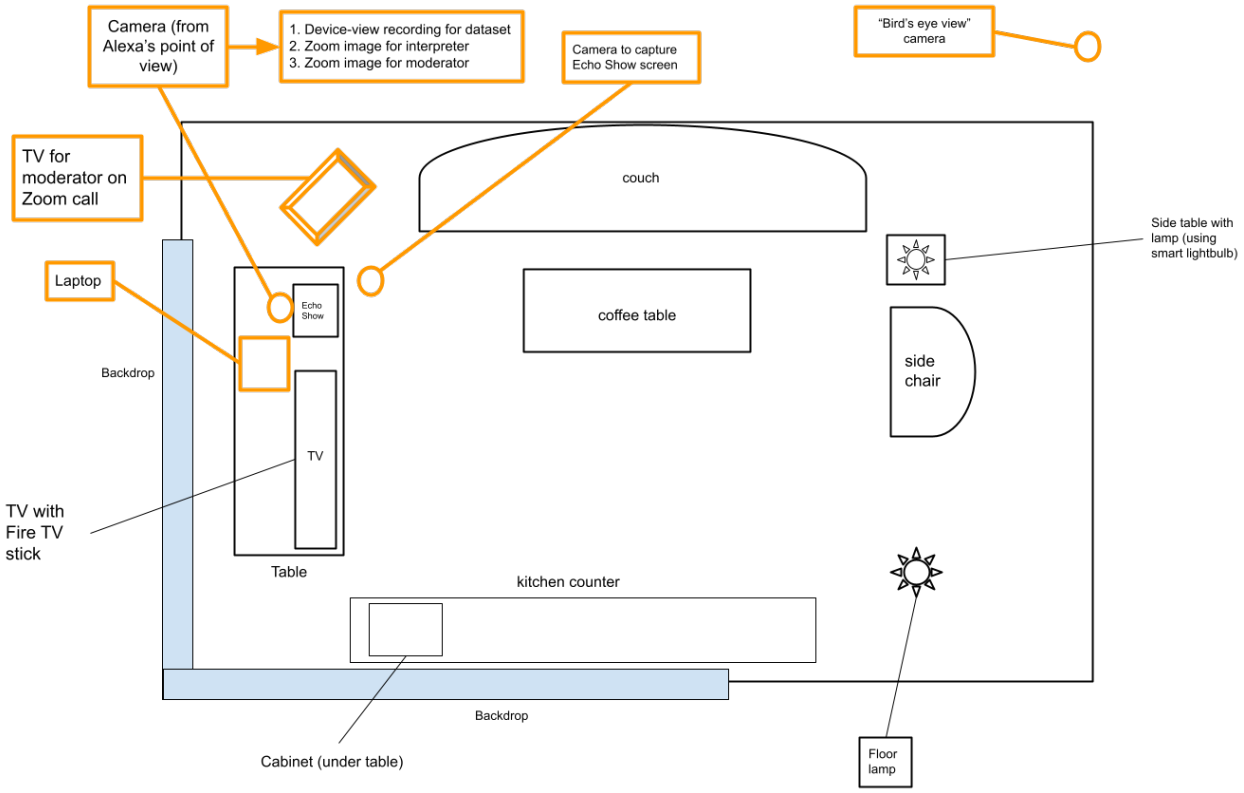


Figure 8.4: Room layout for in-person experiment with Wizard-of-Oz additions in bold orange

Three members of our research team who use ASL watched the "device-view" and "bird's eye view" video recordings for each of the 12 participants and transcribed every single ASL command-phrase, totaling 531. Then, they iteratively re-watched the video recordings to add several layers of annotation metadata. Table 8.1 describes the labels that were added, and table 8.2 gives the top 10 rows of the resulting annotation file.

Table 8.1: In-person Wizard-of-Oz data annotation descriptions (samples shared in table 8.2)

Column	Description
Participant ID	ID for the participant
Video Filename	Name of the .mp4 file associated with the annotation
Start timestamp	Video timestamp marking the beginning of the ASL command
End timestamp	Video timestamp marking the end of the ASL command
Command in English	Transcription of the ASL command in English
Location	Participant location inside the room when the command was issued. There are 10 different position labels; 4 are seated and 6 are standing. See fig. 8.5.
Looking at Alexa	Whether the participant was making direct eye-contact with the Alexa device during this command, which will be one of the following: 1)**Repeated** (Self-explanatory – same signing and wording) 2) **Reworded** (Self-explanatory – changed signing or wording, including only changing the greeting) 3) **Ignored** (Ignored, moved on)

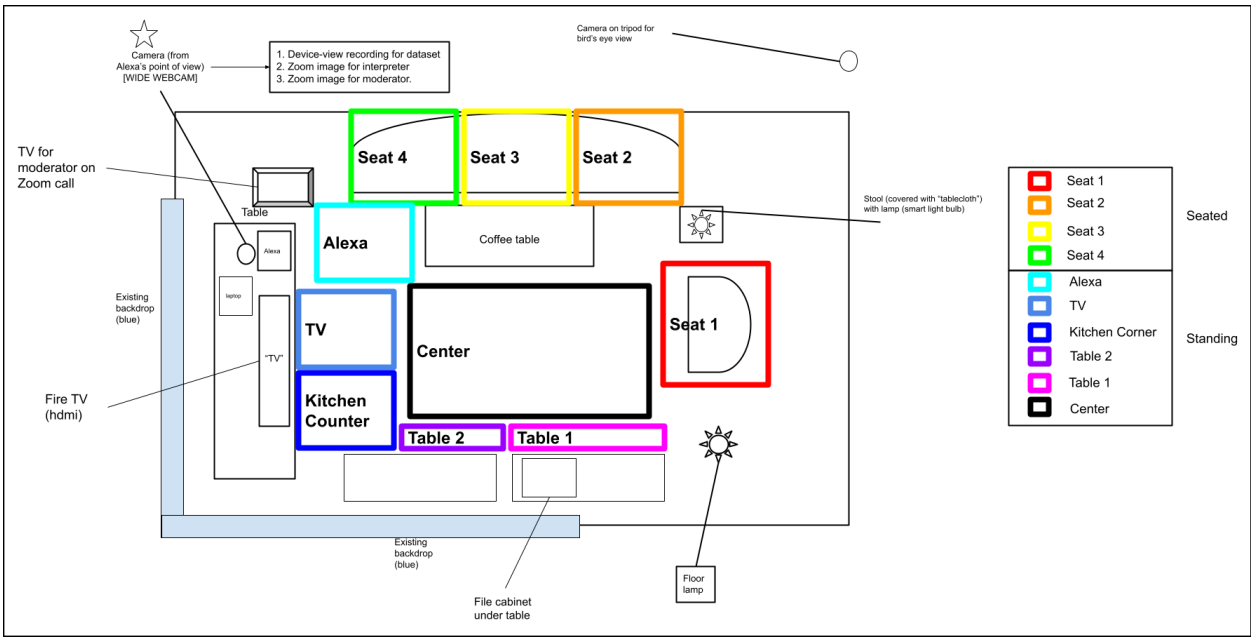


Figure 8.5: Room layout with labels for different positions where an in-person Wizard-of-Oz participant may be located

Table 8.2: Sample annotations described in section 8.4

Participant ID	Video Filename	Start Timestamp	End Timestamp	Command in English	Location	Looking at Alexa
P01	P1zoom.mp4	0:31:00	0:35:00	"Alexa please turn the lights on"	Center	Entire
P01	P1zoom.mp4	0:50:00	0:54:00	"Alexa please turn the lights off"	Center	Entire
P01	P1zoom.mp4	1:55:00	2:01:00	"Alexa how many cups is 30 grams?"	Center	Entire
P01	P1zoom.mp4	2:36:00	2:42:00	"Alexa can I watch YouTube on the TV?"	Center	Partial
P01	P1zoom.mp4	3:10:00	3:17:00	"Alexa can you dim the TV brightness?"	Center	Entire
P01	P1zoom.mp4	4:08:00	4:13:00	"Alexa can you turn the lamp on?"	Center	Entire
P01	P1zoom.mp4	4:44:00	4:48:00	"Alexa can you turn the floor light on?"	Center	Entire
P01	P1zoom.mp4	4:59:00	5:04:00	"Alexa can you turn the living room light off?"	Center	Entire
P01	P1zoom.mp4	5:12:00	5:17:00	"Alexa can you turn the TV off?"	Center	Entire
P01	P1zoom.mp4	5:44:00	5:50:00	"Alexa how much caffiene should I drink in a day?"	Center	Entire

EPILOGUE TO PART II

This is the end of Part II of this dissertation. Chapter 5 explored the potential for a crowdsourced, scalable online sign language data collection platform and showed its viability for collecting a corpus of labelled, isolated signs, which help develop technology involving individual sign recognition, e.g., digital personal assistants that respond to simple signed commands. Chapter 6 explored whether this methodology could be extended to also generate a continuous signing dataset, necessary for natural conversation with complete sentences, supporting complex commands for personal assistant devices.

Chapter 7 conducted a remote Wizard-of-Oz study to allow DHH users to spontaneously wake-up and interact with a personal assistant device in sign language, and the resulting dataset was described and shared. Chapter 8 then conducted an in-person Wizard-of-Oz experiment to investigate aspects that were not possible through the remote protocol, such as how users change their location inside the room and utilize three-dimensional signing space. Part II of this dissertation explored:

> **RQ3: How can DHH and signing communities be enabled to curate sign language datasets that overcome limitations of traditional in-lab collection (e.g. limited demographics, controlled environments, limited size and quality, expensive post-processing and labeling)?** A sign language crowdsourcing platform was built that allowed users to record themselves signing particular signs, and perform quality control checks on other contributors' videos. The platform enabled automatic labelling of all user-contributed videos, and was able to scale the dataset without scaling labelling problems. Results suggested that the crowd can generate high-quality recordings appropriate for training models and can perform quality control checks on others' videos with high reliability. (section 5.3)

RQ4.1: How can everyday signers efficiently contribute to continuous sign language datasets? A novel interface was developed that provides a side-by-side ASL and English synchronized interface, streamlining pre-labeled data collection. The platform was presented as a piecemeal translation, where articles were broken down into sentences that were used as contribution prompts. (section 6.3.2)

RQ4.2: Ensuring that the DHH community is involved in the process, how would the platform be designed? What are the design criteria? The interface shows English and ASL at the same time, and shows which English portion is being signed in the ASL video, so users can keep track of their position via both the English sentence and the video timeline. The platform allows for efficient recording and did not pose unnecessary overhead to the recording operation. It was allowed for multiple users to contribute to the same English text, to account for different signing styles (e.g. regional accents or varied interpretation) or preferred signs for specific words. (section 6.3.1)

RQ4.3: How would DHH users respond to crowd-generated content? Overall participants enjoyed using the site, and commented on several potential use groups and use cases for the platform. Participants showed concerns about controlling the quality of contributors' videos (e.g. professional backgrounds, lighting, etc.). (section 6.5.3)

RQ4.4: Can the platform incentivize contributors by being a sign language bilingual resource? Despite citing several personal benefits and strongly agreeing that they wish more content like this was available, the participants were not significantly interested in generating content for the platform themselves. A major rationale described by the participants was that they did not want to publicly post a recording of themselves and not having control of who can see and use it. (section 6.5.3)

Part III: DHH behavior with a Personal Assistant that appears to understand American Sign Language

PROLOGUE TO PART III

Part I addressed the gap in knowledge about DHH users' experiences with personal assistant devices, Part II developed and tested remote data collection methodologies, and lastly devised an in-person Wizard-of-Oz method to investigate aspects that were not possible with the aforementioned remote protocol. Now, here in Part III, I present the data analysis and discussion from the Wizard-of-Oz experiments:

Analysis of the remote Wizard-of-Oz data collected previously (chapter 7) is presented in chapter 9, which investigates each stage of interaction with a personal assistant. For instance, it is important for designers of personal assistants and developers of sign-language recognition technologies to understand how a DHH user might wake-up the device or initiate a command. Next, to ensure that personal assistants will function to DHH users' satisfaction, it is necessary to know what categories of commands and requests need to be supported. Further, it is important for AI researchers to know what these commands look like in ASL to create automatic sign-language recognition models. More complex, common real-world scenarios may require dialogue and conversation with personal assistants in order to achieve a desired result, but these interactions often have errors or breakdowns. This chapter also describes how DHH users recover or respond if such disruptions occur.

Chapter 8 (from Part II) conducted an in-person, physical data-collection setup that emulated a residential living room and kitchen, in order to address the limitations that came with the lack of real-world conditions during the remote Wizard-of-Oz experiment. In this part, chapter 10 describes how the data that was collected during this study was analyzed to obtain insights about DHH users interacting with a personal assistant that appears to understand ASL while in a home-like environment. For instance; how users may change their location inside the room and utilize their full signing space (e.g. point-

ing at items in the room to refer to them in ASL), both of which had not been possible via a Zoom videoconference call.

Specifically, Part III of this dissertation investigates:

RQ5.1: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users instinctively "wake-up" the device or initiate a command?

RQ5.2: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, what categories of commands/requests do users produce?

RQ5.3: What do these commands look like in ASL?

RQ5.4: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users recover or respond when there is an error or breakdown?

RQ5.5: After DHH ASL signers had the opportunity to interact with a personal assistant device in sign language, did their interest in such interaction increase, decrease, or stay the same?

RQ6.1: What are some linguistic properties of in-person interaction with full signing space, such as referencing to other objects inside the room?

RQ6.2: Based on observation, how would a DHH user naturally position themselves in proximity to a personal assistant device in a residential-like living room and kitchen area?

RQ6.3: When given the opportunity to interact with a personal assistant device that is physically in the same room as them in a residential setting, what are DHH users' experience and opinion on this?

Chapter 9

DHH Users' Behavior, Usage, and Interaction with a Prototype Personal Assistant Device that Understands Sign-Language Input[16]

9.1 Introduction

The previous survey-based study (chapter 3) asked DHH participants to imagine commands they may use with personal assistant technologies. That study had also gathered some commands in English, participants had been asked to imagine how they might interact with a device. In addition, our previous interview-based research (chapter 4) began to examine how DHH users might like to "wake up" a personal assistant device before giving it a command in sign language, but that study did not observe how DHH users would spontaneously attempt to wake-up a device when given the opportunity to do so. When faced with the opportunity to actually interact with a personal assistant device in sign language,

[16]The information in this chapter is based on a joint project with my advisor (Dr. Matt Huenerfauth), and graduate students I supervised at RIT (Matthew Watkins and Kira Hart) whom assisted me with data collection and qualitative data analysis. The results were published as a paper at the CHI'22 conference [64].

the way in which users may behave and how they would construct commands in sign language may differ, as compared to how they might imagine doing so hypothetically.

To provide guidance for designers of this technology, as well as shed light on the specific types of sign-language commands DHH users may wish to perform with this technology, this chapter analyzes the data that was collected using a video-conferencing Wizard-of-Oz methodology (described in chapter 7). DHH signers tried ASL interaction with a personal-assistant device with a hidden human ASL interpreter who translated commands into spoken English for the device. About 21 hours of video recordings of these interactions were transcribed and annotated for analysis, yielding a dataset of over 1,400 individual ASL commands and interactions with the technology. We observed the variety of ASL commands that users performed, the linguistic structure of these commands, the way in which participants "woke up" the device to initiate each command, and how users responded when errors occurred during the interaction. As compared to prior work, this observational study revealed new ways in which DHH users chose to wake-up their device, ASL-specific terminology signers chose to use for device command-and-control interactions, and ways in which participants shifted between ASL and English-like ways of interacting with their devices, especially after errors.

This chapter empirically contributes the first observational study of the behavior and interaction of DHH individuals engaging with a personal assistant device that appears to understand ASL, to answer research questions about the types of ASL commands that users performed, ASL-specific vocabulary and structures used in those commands, how users initiated commands to wake the device, and how users responded or reformulated commands when there was an error in the interaction. While prior studies had investigated some of these issues by asking DHH users to imagine using such devices, this is the first study to give users the opportunity to experience an interaction. These findings provide guidance for future designers of this technology, e.g., specific commands to support, ASL terminology to use for command and control of the device, how the device should respond when there is a potential error, and other insights. These findings also provide guidance to creators of sign-language recognition technology, in prioritizing vocabulary or structures that must be recognized in order to support natural ASL interaction.

9.2 Related Work

Recent research has drawn attention to issues of artificial-intelligence (AI) fairness in regard to sign-language technologies, e.g., [21, 93], motivating greater inclusion and leadership among DHH individuals in the design and dataset creation for such technologies. Other recent work has drawn attention to fairness and accessibility issues specifically in regard to personal-assistant technologies: In a survey-based research study about the usage of personal-assistant devices by people with various disabilities [137], researchers found that few DHH individuals currently used these devices. Those same researchers found that despite marketing that would suggest personal-assistant systems are accessible, many current tools continue to have accessibility issues, and there is a need for research on the usage of these devices by people who are DHH.

Personal assistant devices typically use spoken-modality input and output, which is inaccessible for DHH users. Some devices have a visual interface that can display captioning [17] and features for typing or selecting commands on a screen [18]. However, both of these accessibility features require the user to have written-language reading or writing skills, and many DHH users prefer interaction in ASL or have lower English literacy skills. Given these challenges, other DHH researchers have called for HCI researchers to explore interaction methods for DHH users with personal assistants *before* they become ubiquitous in daily life [139].

Prior work [111] has discussed how users' interaction with personal-assistants devices is typically described in terms of several stages, including, e.g., activation, command, and response (potentially followed by a reattempt of the command by the user in case of error). Typically, speaker-based personal-assistant devices are voice-activated and will process a user's command if prefixed by an activation word, sometimes called a "wake-word" [77]. For example, a user may need to say, "Alexa," to get the attention of an Amazon Echo device, and then the user would speak the command. After processing the command, the device typically provides an audio and/or visual response to the user. In the subsections that follow, we discuss related work for each of these three phases, with a focus on any prior work among DHH participants or which has considered sign-language interaction with such devices.

[17] https://www.amazon.com/gp/help/customer/display.html?nodeId=GK2BSY9F55EM56YL
[18] https://www.amazon.com/gp/help/customer/display.html?nodeId=GBUJQF9ZX3TV7MK6

9.2.1 Device Activation

To activate a personal assistant device, traditionally a "wake word" or phrase is spoken prior to the command, e.g., "Hey Siri," "OK Google," "Alexa," or "Hey Microsoft" [181]. Some personal assistants may also be activated through physical interaction with the device or an accompanying smartphones app [67]. Our prior research (chapter 3 and chapter 4) on wake-up modalities has categorized such approaches as: "push-to-talk" (touching a device physically to wake it prior to a command) and "talk-to-talk" (speaking or signing a "wake word" to gain the device's attention prior to a command).

Chapter 4 investigated preferences and concerns among DHH users for waking personal-assistant technologies that could understand sign language (in the hypothetical future). The authors identified six wake-up techniques through formative interviews, and then created a video prototype demonstrating each. Using these videos, in a subsequent study, participants discussed the trade-offs between various wake-up approaches, and identified key factors that affected their preference for each. The authors called for future HCI researchers to implement a working prototype or interactive Wizard-of-Oz set-up so that users would have first-hand experience trying various wake-up interactions.

Chapter 3 included interviews with DHH ASL-signing participants, asking them to imagine interacting with a personal assistant that could understand ASL. Participants suggested hand-waving (in Deaf culture, waving your hand in someone's direction is a culturally acceptable method for gaining attention [151]), making noise (e.g., clapping or tapping), signing the device's ASL name-sign (an ASL sign that could represent the name of the device), fingerspelling the device's English name, among other approaches.

Although some prior research had begun to investigate potential wake-up approaches among DHH participants, there was a limitation: None of that work had provided participants the opportunity to actually interact with a device in sign-language. Rather than depending upon participants to imagine such interaction, new insights may be revealed when DHH participants spontaneously interact with a device; capturing and understanding this behavior is important for future designers of this technology.

9.2.2 Issuing Commands and Requests

There has been substantial research and development on improving personal assistant technologies. For example, large companies such as Google, IBM, and Microsoft have been making progress in automatic speech recognition (ASR), the underlying technology personal assistants use to recognize voice commands [78, 82, 148]. They have also been examining voice recordings of queries that led to errors, to see how they might need to expand the syntax of the types of spoken commands they can process or to expand the feature set of the system over time [101, 158]. The popularity of these technologies has increased: In 2019, 72% of respondents of a global survey [128] reported using a digital assistant, and 45% reported owning one, with an additional 26% planning to purchase one soon. The report also covered popular use of digital assistants for music (63%), lighting (57%), security cameras (38%), and thermostats (37%).

With accessibility barriers for DHH users amid this growing popularity among the wider population, research is needed into potential ASL interactions, to understand both what DHH users want to do with the device, and how they would express it in ASL. Chapter 3 began to to address this knowledge gap through a survey of DHH ASL signers who had been asked to imagine how they would use a personal-assistant in ASL; the analysis of those survey responses led to a set of "categories of commands" DHH users would be likely to give to a personal assistant, which included: "Ask weather-related questions (e.g., temperature, need umbrella)," "Set alarms, events, and reminders," "Set timers," "Get alerts (e.g., doorbells, smoke alarms)," "Search for information (e.g., recipes, movie times)," "Connect to other smart devices (e.g., lights, TV, cars)," "Video-based communication (e.g., videophone/VRS)," "Notifications (e.g., read, delete notifications)," "Information, Warnings (e.g., traffic, weather conditions)," and "Manage notes (e.g., to-do lists, shopping lists)." That study revealed potential DHH-specific use cases, e.g., notification about sounds in the home. In addition, our survey had asked participants to indicate their interest in using a personal assistant device that could understand ASL commands.

While that prior work revealed that DHH users were interested in ASL interaction with a personal assistant device, the limitation was that participants never actually experienced sign-language interaction with these devices.

While categories of potential commands were imagined by participants in that prior study (chapter 3), participants had typed their potential command suggestions in English. Research is needed in which DHH users have the experience of actually producing commands in ASL to a device, not only to increase the ecological validity of their suggestions, but also to enable the video-recording of such ASL commands. Such videos can be analyzed to understand the ASL linguistic structure and vocabulary of the commands, and to serve as potential training data for AI researchers building sign-language recognition systems.

9.2.3 Device Response and Command Reattempt

Personal-assistant devices are not perfect, and errors are common when users interact with these systems. For instance, one study revealed that smart speakers respond to approximately 1 in 3 user requests with an error message [184]. To improve the experience of users, some recent work, e.g., [184], has focused on how conversational user interfaces should reply to the user if an error occurred. However, to our knowledge, no prior work has examined how devices would or should respond to users who have submitted commands in sign-language.

As personal-assistant devices provide a conversational user interface, users often follow-up on the device's reply. While companies developing personal-assistant technologies may examine recording logs from users or conduct internal usability testing on how users respond to errors, within the published HCI research literature, how users behave and respond when errors arise during personal-assistant interaction has been an under-studied research area.

In the most closely related prior study [118], researchers performed sequence analysis of users' interactions with a specific user-interface that accepted voice commands, for a more narrow domain than modern personal assistants: The system allowed users to add, edit, and delete events from their calendar. In that prior work, the authors described a set of user-behavior codes they employed when analyzing their recordings, such as: repeating the same utterance, removing a word, using a new keyword, adding more detail, or trying again from scratch.

Among DHH users or through sign-language interaction, no prior research study has examined how

users would behave or respond when there are errors during a personal-assistant interaction. Despite this gap in the literature, it would be valuable for designers of these devices to know how DHH users might respond to an error or breakdown, so that they understand what users want the device to do in such situations. Further, for designers of sign-language recognition technology, understanding the types of vocabulary or linguistic structures DHH users might use when reacting or responding to an error would inform the development of their technology.

9.3 Research Questions

As in our section 9.2 above, our four research questions have been sequenced below according to the stages of interaction with a personal assistant. However, despite this temporal ordering, we view our primary contribution from this study to be our findings regarding research questions RQ5.2 to RQ5.4, while the contribution from research questions RQ5.1 and RQ5.5 is secondary. These secondary questions enable comparison to prior work, in which DHH users had been asked to imagine interacting with devices, to explore device activation (RQ5.1) [111] and interest (RQ5.5) [62].

For the device activation stage, designers of personal assistants and developers of sign-language recognition models may benefit from understanding how a DHH user might wake-up the device or initiate a command, yet no prior work had examined how DHH users would behave when given the opportunity to actually interact with a personal assistant device using sign language. Hence, our first research question is:

- **RQ5.1**: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users instinctively "wake-up" the device or initiate a command?

When a user is issuing a command to a device, some prior survey-based research had identified commands of interest to DHH users. However, that study had asked DHH ASL signers to merely imagine how they would use a device and to indicate, in English, some commands they might perform.

This motivates our second research question:

- **RQ5.2**: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, what categories of commands/requests do users produce?
- **RQ5.3**: What do these commands look like in ASL?

Our Related Work analysis revealed a lack of prior research on how DHH users might react or follow-up when an error occurs during a personal-assistant interaction in sign-language – motivating our third research question:

- **RQ5.4**: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users recover or respond when there is an error or breakdown?

In prior survey-based work in which DHH users imagined how they might interact with a personal assistant in sign language (chapter 3), participants reported that they would be interested in such interaction. However, it is unknown, after actually having the opportunity to experience interacting with a prototype, whether users' interest may change. Thus, our fourth research question is:

- **RQ5.5**: After DHH ASL signers had the opportunity to interact with a personal assistant device in sign language, did their interest in such interaction increase, decrease, or stay the same?

9.4 Research Methodology

The data collection step has already been presented in chapter 7. To remind the reader, DHH participants were enabled to interact with a personal-assistant device in sign language, through a Wizard-of-Oz (WoZ) set-up to prototype this technology. The recordings from these interactions were analyzed redundantly by three researchers (two native ASL signers and one English-ASL interpreter) to address research questions 1 to 3. In addition, an identical pre- and post-survey was conducted to assess participants' interest in such devices, to address research question 4.

9.4.1 Recruitment and Participant Demographics

A total of 21 DHH participants were recruited through on-campus and social-media ads. The eligibility criteria for participation was being at least 18 years old and having used sign-language at home when they were a child and/or attended an elementary or middle school where they used sign language every day. Our rationale for recruiting ASL signers with higher fluency from early language exposure is that we wanted to investigate the linguistic structure of ASL commands produced by participants, and given this focus, we wanted to collect data from signers who would produce signing with ASL linguistic structure, rather than more English-like signing. Furthermore, we did not specify whether participants were required to have experience with personal assistant devices. In prior work [62, 111, 137], it was discovered that very few DHH users have and use personal assistant devices, due to their accessibility barriers, despite there being interest in the use of these devices. That prior work had also revealed that many DHH individuals who had not used such devices had still seen this technology in media, and some had witnessed household members who are hearing using such devices.

Participants self-identified as 11 men and 10 women, mean age 25 with standard deviation 4.9, all living in the U.S., across 12 different states. Fourteen self-identified as Deaf, 3 as deaf, and 2 as hard of hearing. While 6 reporting having DHH parents, 17 of the participants reported using sign language at home. Eighteen of the participants used ASL during elementary school, with 9 describing their elementary school as a daytime school (where the students commute to), 1 as a residential deaf institute (where the students sleep at the dorms), 8 as mainstream (where both hearing and DHH students attend), and 3 as both (spending some years at one and then transferring to another). Among participants' education level, 6 had a high school diploma, 1 had some college, 11 had a Bachelor's degree, and 3 had a graduate degree. The average number of members in each participant's household was 3.5 (standard deviation 1.2), while the average number of household members that use ASL was 2.9 (s.d. 1.5). On a question about the percentage of use of ASL in various contexts (100% meant using ASL the entire time and no English, and 0% meant using spoken/written English all the time with no ASL), participants' average response was: 83% (s.d. 29%) at home, 71% (s.d. 28%) at work or school, and 72% (s.d. 24%) with friends and family.

9.4.2 Questionnaires and Consent

This study was approved by our institutional review board (IRB) for the protection of human subjects, and participants signed an informed consent and video-recording release at the beginning of the study, which granted permission for their video recordings to be included in a public dataset (described in chapter 7). Each session was scheduled for 70 minutes, and participants were compensated $40. Sessions were moderated by a DHH researcher in ASL, using Zoom for video conferencing and for making a video recording of the participants' responses and signing.

We made use of a set of interview questions shared in prior work (appendix B), asking a subset of questions redundantly both before and after participants' experience with the prototype. We began with asking about participants' familiarity with personal-assistant devices, whether they had one in their household, and how they may have interacted with it, if applicable. A question asked right before participants started their interactions with the prototype was whether they agreed with the following statement (5-point, 1=Strongly Disagree, 5=Strongly Agree): "I would be interested in using sign language interaction with a personal assistant device." After the prototype experience, we asked this same question again, to enable a before-and-after comparison, for our fourth research question.

Participants were told that they would have an opportunity to try out an Amazon Alexa Echo Show device. After the moderator invited Alexa to join the meeting, a new Zoom-meeting participant video appeared during the video conferencing meeting, showing the screen of the device. The moderator explained to the participant that they could ask Alexa anything they wanted to, and that Alexa would watch and understand their ASL commands. After ensuring that the recording began, the moderator invited the participant to start interacting with Alexa.

After the device-interaction session and after asking the Likert item about their interest, the moderator asked the participant to respond to the demographic questions.

9.4.3 Details of Wizard-of-Oz Setup and Recording

Chapter 7 describes our Wizard-of-Oz setup that we used to collect data for this study. We had some intial interview questions and, when it was time to commence the Wizard-of-Oz interaction, the moderator

used ASL to ask Alexa to join the Zoom meeting and to turn on its video stream. When the time allotted for the recording session elapsed, the moderator told the participant that time was up, asked Alexa to leave, and then continued with the aforementioned final interview questions (appendix B).

For each of the 21 participants, the Zoom meetings were recorded using the built-in recording feature, and also using a screen-recorder software on the moderator's computer. One recording focused on capturing the DHH participant alone, while the other recording focused on capturing the entire Zoom meeting, including Alexa's stream, so that it could be seen how Alexa responded.

9.4.4 Analysis of Responses and Recordings

Section 7.3 describes our procedure for analysis and annotation of the data collected during this experiment. Three members from our research team watched the entire set of recordings (approximately 21 hours total, over 1,400 utterances) to transcribe ASL commands, Alexa's responses, "wake-up" commands, command categories, errors, and error follow-ups.

9.5 Findings

At the beginning of the study, participants had been asked "Have you ever seen smart personal assistant devices like Amazon Alexa or Google Home, which allow someone to give commands or to ask questions?" While 14 out of 21 participants answered "Yes," only 5 out of 21 lived in a household with at least one personal assistant device. Only 2 out of 21 participants personally owned a personal assistant device. For these two people, one said that they use it on a monthly basis by speaking to it with their own voice. The other person said that they use it on a daily basis by speaking to it with their voice and also by typing commands on a phone app.

9.5.1 RQ5.1: How do people instinctively perform a "wake up" command in this interactive setting?

Table 9.1 displays the set of wake-up codes used to label all of the participants' videos, along with descriptions, ASL gloss, and frequency for each one. As mentioned above, ASL glosses are a set of

Table 9.1: Device wake-up codes, descriptions, ASL glosses, and frequency

Wake-up Method	Description	ASL Gloss	Frequency
None	No wake-up method (only the command/query)		1081
A-l-e-x-a	Fingerspelling "Alexa"	fs-ALEXA	233
Hey-A-l-e-x-a	ASL sign ATTENTION-WAVE (waving hand with palm facing down) followed by fingerspelling "Alexa"	ATTENTION-WAVE fs-ALEXA	57
A-l-e-x-a-Ending	Fingerspelling "Alexa" at the end of a command	... fs-ALEXA	14
Hey	ASL sign ATTENTION-WAVE (waving hand with palm facing down)	ATTENTION-WAVE	10
Hello-A-l-e-x-a	ASL sign HELLO (salute from head) followed by fingerspelling "Alexa"	HELLO fs-ALEXA	7
Hello	ASL sign HELLO (salute from head)	HELLO	5
Curious	ASL sign CURIOUS (hand near neck)	CURIOUS	4
Do-do	ASL sign DO-DO (repeated pinch, palms facing up)	DO-DO	4
Hi-A-l-e-x-a	ASL sign ATTENTION-WAVE_2 (waving hand with palm facing forward), then fingerspelling "Alexa"	ATTENTION-WAVE_2 fs-ALEXA	3
H-e-y-A-l-e-x-a	Fingerspelling "Hey Alexa"	fs-HEY fs-ALEXA	2
A-l-e-x-a-do-do	Fingerspelling "Alexa" followed by ASL sign DO-DO	fs-ALEXA DO-DO	1

English-like labels that can be used to transcribe specific ASL signs; for consistency and replicability, we made use of the set of gloss labels used within the American Sign Language Lexicon Video Dataset (ASLLVD) [123]. In some cases, signers made use of fingerspelling, which is a method for producing a sequence of alphabet letters on the hands to spell an English word.

In 1,081 cases, participants did not use any wake-up method before issuing a command; many of these were follow-ups to a question, command-and-control input from the participant, or responses to yes/no questions from the device. Among those cases in which a wake-up method was used, the most popular was A-l-e-x-a, in which the participant began the command by fingerspelling "Alexa." Overall, we observed 11 different wake-up methods, including, for example: A-l-e-x-a, Hello, Hey, H-e-y, Hi, Curious, and Do-do. Table 9.1 explains the differences between these signs, gives their formal ASL glosses and frequency, and fig. 9.1 has screenshots of these different signs.

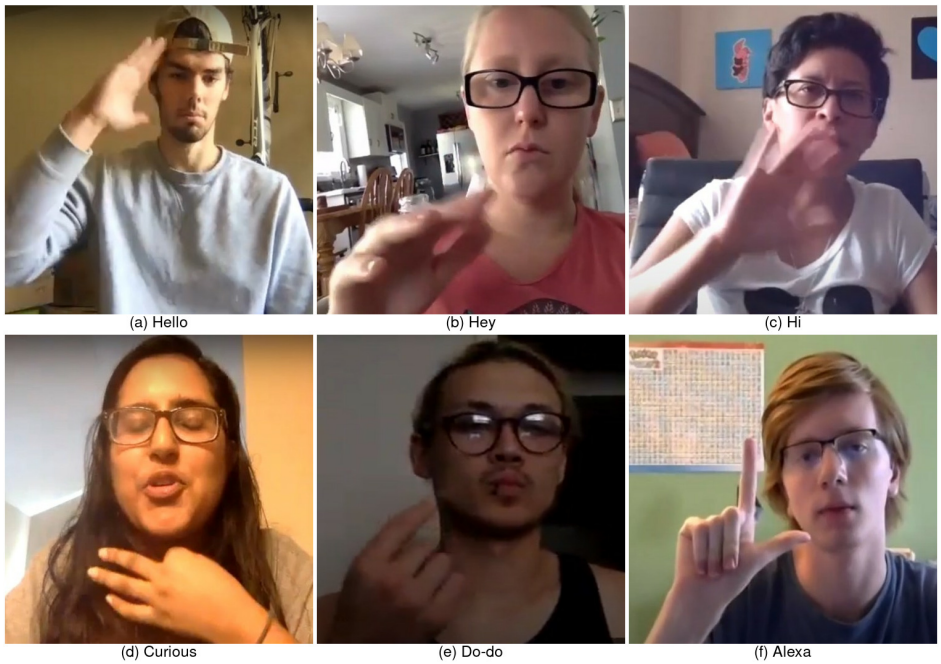


Figure 9.1: Screenshots of various wake signs, coded: (a) Hello, (b) Hey, (c) Hi, (d) Curious, (e) DO-DO, and (f) A-l-e-x-a

9.5.2 RQ5.2: What categories of commands/requests do people make in ASL with an Alexa?

Table 9.2 shows command categories that resulted from our analysis of the videos, a short description of each, some sample commands of each category, and the number of times a participant performed a command of that category in our dataset (chapter 7). Our analysis of the videos led us to label commands using 15 different categories; the most popular category was "Command and control" (where people change the device settings, navigate the device, say Yes/No), and the next 4 top categories were "Entertainment," "Shopping," "Lifestyle," and "Trivia, calculations."

9.5.3 RQ5.3: What do these commands look like in ASL?

Table 9.2 lists categories of commands we observed, explanations and examples of each category, and the number of occurrences of each in our dataset (chapter 7). Since ASL and English are different languages, the grammar, syntax, and word-order between sentences in each may differ. For this reason, we also

Table 9.2: Command-topic categories, descriptions, sample command and ASL gloss, and frequency

Category	Description	Sample command	ASL gloss of sample	Count
Command and control	Personal assistant device settings, navigation, Yes/No	Can you turn up the brightness?	CAN YOU TURN UP BRIGHT	352
Entertainment	Videos, riddles, jokes, games	Can you google funny dog videos?	CAN YOU fs-GOOGLE FUNNY DOG VIDEO	162
Lifestyle	Health, pets, travel	Where can I buy pet food?	WHERE CAN I BUY PET FOOD QMWG	127
Shopping	Miscellaneous shopping, finding items for purchase	Where can I buy a bike?	WHERE CAN I BUY BIKE QMWG	126
Trivia, calculations	Looking up facts, miscellaneous information, conversions	Tell me the most interesting fact about Earth.	TELL ME ABOUT MOST INTERESTING ABOUT EARTH	124
Sports	Sports related schedules, updates, people	Alexa, how did the Bucks do last night?	fs-ALEXA HOW DID fs-BUCKS DO LAST NIGHT	96
Food/restaurant	Recipes, nutrition, restaurants	I want to eat vegetarian dinner tonight.	I WANT EAT VEGETABLE DINNER TONIGHT WHAT	83
News	News updates, general and specific	What's happening in news today?	WHAT HAPPENING IN fs-NEWS TODAY	55
Weather	Weather forecasts	Alexa, what's the weather here in <Redacted City Name>?	fs-ALEXA WHAT WEATHER HERE <Redacted Signs>	54
Alarms/Events	Setting alarms, reminders, scheduling events	Alexa, I want to schedule noon to 2 with my friend.	fs-ALEXA I WANT TO SCHEDULE NOON TO TWO WITH MY FRIEND	51
DHH-specific	Accessibility, DHH related topics	Please give me caption for this video.	GIVE CAPTION IX VIDEO QMWG	47
COVID-19	Information about Coronavirus disease (COVID-19),	Alexa, are there any cases of COVID here in <Redacted City Name>?	fs-ALEXA ANY fs-CASES OF COVID HERE <Redacted Signs>	43
TV/Movies	Finding movies and TV shows, looking up information	Alexa, Which movie has the highest gross pay?	fs-ALEXA WHAT MOVIE fs-IS HAS HIGHEST PAY fs-GROSS	36
Colleges/Universities	Information, schedules relating to college/university	Alexa, what's the schedule for fall 2020 at <Redacted University Name>?	WHAT SCHEDULE fs-FALL TWENTY TWENTY <Redacted Sign>	32
Websites	Navigating, using miscellaneous websites	Okay, open up macrumors.com	FINE OPEN fs-MACRUMORS DOT fs-COM	19

provide some examples below of ASL-gloss transcriptions for selected commands, to demonstrate the variety of structures used, across different users.

In ASL, typically words are expressed using ASL signs, but at times signers may use fingerspelling to articulate the letters of an English word. In fluent ASL signing, this is typically used only for proper names, movie/book titles, or specialized jargon for which a specific ASL sign might not exist. As expected, we observed that signers used fingerspelling for certain proper names in their commands (indicated in transcriptions as items with prefix "fs-"), e.g., "WHO BETTER fs-LEBRONJAMES OR fs-MICHAELJORDON." However, we also observed cases in which participants used fingerspelling for words for which there did exist an ASL sign, sometimes for emphasis or when the sign may be less commonly known, e.g., "WHO INVENT APPLE fs-APPLE," "HOW OPEN fs-COCONUT," "PLEASE TELL ME HOW-MUCH COST LIVE fs-DUBAI." In these examples, the signer spelled "Apple" after using the ASL sign for it, and in the case of "coconut" and "Dubai" while ASL signs exist for these items, they are less commonly known.

In ASL questions, the WH-word (e.g., "who," "what," "where") often appears at the end of the sentence, rather than at the beginning in English [119]. For instance, participants' questions included the following: "WHO fs-JOHNDROCKFELLER WHO," "fs-ALEXA I LOOK GOOD CHINESE FOOD WHERE," "MOVIE HIGHEST PAY MOVIE HIT WHAT," "CHICAGO BLACK fs-HAWKS AND fs-LASVEGAS GOLDEN KNIGHTS, THEY PLAYED LAST NIGHT. fs-SCORE WHAT." However, we found that once participants noticed that the system was not perfect in understanding or responding to their questions, e.g., after the device gave a response that did not seem accurate, participants transitioned to a more English-like word-order for their WH-questions, e.g., "WHERE CAN I BUY TABLE COVER OVER. WHERE CAN I BUY."

In ASL, yes-or-no questions sometimes end with the QMWG "question mark wiggle" sign at the end, which is performed by the signer holding their dominant hand in space with their index finger extended and rapidly switching between a hooked and extended configuration [12]. In fluent ASL signing, this is often done at the end of sentences in case of emphasis, or in settings in which the signer may worry that their facial expression, which indicates that the sentence is a question, cannot be clearly seen. We noted in our dataset (chapter 7) that participants commonly used this sign not only at the end of yes-or-no

questions, but they also inserted it at the end of some wh-questions, e.g., "WHERE I GO fs-OIL CHANGE CAR QMWG," "WHAT ABOUT DEAF PROFESSIONAL PHOTOGRAPHER DEAF QMWG."

During ASL, signers can associate people, things, or ideas under discussion with locations around their body; to refer to these locations again during the conversation, they can point to these locations in space [113]. This referential use of space corresponds to use of pronouns, e.g., "she" or "it," in spoken languages. Such pointing to space during ASL signing is glossed as "IX" (corresponding to the signer "indexing" or pointing to this spatial location that represents that entity). Participants made use of the space around their body in this manner during commands and questions during our study, e.g., "MY NEPHEW BIRTHDAY TOMORROW. IX FOUR. WHAT SHOULD I GET IX."

Not all input to the device was in the form of questions. Participants also gave imperative commands to the device in ASL, e.g.: "PLAY GAMES WITH ME," "CLOSE-UP PICTURE ZOOM," "CONVO VRS OPEN." This final example is a command that may be of particular interest among DHH participants, as "Convo" is a video-relay service (VRS), which enables ASL signers to place telephone calls to hearing individuals using an ASL interpreter as an intermediary.

Participants used ASL to ask about accessibility features of the device. For instance, if the device played a video in response to a command or request, many participants asked if captions could be enabled, e.g., "CAN CAPTION VIDEO."

We also observed how participants used ASL to respond to questions from the device in ASL, e.g., "YES," "SURE," "NO," or "DON'T-KNOW," or to issue command-and-control inputs to the device, including return to a previous/home screen of the device (e.g., "fs-BACK HOME PAGE," "HOME," "GO HOME," "CLOSE fs-APP," "GO BACK," "QUIT"), navigating lists of choices (e.g., CAN SEE MORE PICTURES QMWG," "SHOW ME MORE," "MOVE NEXT STEP," "DOWN," "MORE," "NEXT," "NEXT STEP"), or selecting items on lists (e.g., "OPEN," "SHOW-ME NUMBER THREE," "CLICK FIRST"). At times, participants used polite language, e.g., appending "PLEASE" or "THANK-YOU" to their commands.

9.5.4 RQ5.4: How do users recover or respond when there is an error/breakdown?

Device errors were noted during our analysis and were categorized into six types. The most common was when Alexa received the command, but did not show a desired result, e.g., when it misinterpreted part of the query or showed something that didn't match what the user had been looking for. Another error type was when Alexa remained silent, even though it had been activated and should have received a command. In this case, Alexa stayed on the home page and did not show any results nor reply to the user. Other times, Alexa heard but did not understand or was unable to process the command, and replied with something along the lines of "Sorry, I don't know that." In other cases, Alexa followed up with a suggestion, e.g., "Sorry, I don't know that. Do you want to try <different query>?" There were times when Alexa understood the command, but needed some additional information; in such cases, Alexa asked the user for clarification, e.g., "Sorry, I didn't get that location" or "What do you want me to play again?" Sometimes an error was caused by a hardware or software breakdown, e.g., Alexa crashing or when the closed captions froze on the screen after Alexa had finished replying.

Table 9.3 shows the set of codes that emerged during our analysis of how participants responded or followed-up after errors; the table includes a short description and frequency for each of the five types of user behaviors that we observed. Most often, participants simply ignored the error and moved on to a different query. The second most popular behavior was to repeat the query, with the same exact wording. Third, participants reworded their commands, changing the specific signs that they used in their command. For example, one participant changed an initial question of "fs-VEGAN OR MEAT EATER BEST WHICH" to "fs-VEGAN ITSELF HEALTHY QMWG."

In a few cases, the participants instead commented to or asked the researcher/moderator on the Zoom call about the error, for example, asking whether Alexa went off topic, "REALLY IX fs-ALEXA DEVIATE WHAT". Such behavior might also be typical in settings in which a device is used in the home while a bystander is present.

Table 9.3: User behaviors following personal-assistant errors

Error-follow-up code	Description	Count
Ignored	Ignored the erroneous response and simply moved on to the next command	229
Repeated	Repeated the command with the same signs and wording	205
Reworded	Repeated, changing some signs and wording of the command	129
Played Along	Accepted and went along with what Alexa responded, despite being undesirable	25
Question	Asked or commended to Alexa or the researcher about the error they are seeing	18

9.5.5 RQ5.5: Did users' interest in sign language interaction with a personal assistant device increase, decrease, or stay the same, after having the opportunity to experience a prototype?

Before participants started their interactions with the prototype, we asked whether they agreed with the following statement (5-point, 1=Strongly Disagree, 5=Strongly Agree) "I would be interested in using sign language interaction with a personal assistant device." After the experience, we asked this question again. The mean of participants' (N=21) responses was 4.29 before the experience and 4.19 afterward, suggesting that participants were interested in personal assistants that could understand ASL. We performed a Two One-Sided Test (TOST) for statistical equivalence, with a margin of 0.5 on the 1-to-5 Likert response scale, and found that participants' responses before the experience were statistically equivalent to those after the experience.

9.5.6 General Observations

During the recording sessions, participants wore casual clothing and typically stayed in their living room or bedroom. Most participants only had their upper body in the screen. Generally, the interaction between Alexa and the participants was somewhat slower than would be typical for a spoken interaction, due to the ASL-interpretation latency delay between the end of the command and the beginning of Alexa's response; we observed this interval to be roughly ten seconds. In addition to interpreting, this latency was due in part by time taken by Alexa to process and prepare its response to the user.

Some participants started off their interaction with Alexa while signing ASL at full speed and then transitioned to slower signing or more English-like signing over time, especially after noticing some

erroneous responses. There were also participants who appeared more cautious in the beginning, focusing more on spelling out words and being clear, and then they started to increase their signing speed when they realized that Alexa could handle more than they thought. The rest maintained their ASL signing speed and fluency during of the session, with only occasional switch to fingerspelling of specific words when errors occurred repeatedly.

Based on facial expressions and observable demeanor, participants' response to interacting with Alexa appeared positive at the beginning of the study. For some participants, when they encountered some errors or miscommunications during the interaction, their apparent enthusiasm with interacting with the system seemed to dip after a few minutes of interaction. Most participants seemed to learn what types of commands or phrasing worked well with the device over time, leading to greater success and apparent enthusiasm with using the device by the end of the session. In some cases, participants laughed when Alexa made an absurd error, but other participants became visibly frustrated when they need to repeat their command more than a few times, e.g., when they were simply trying to request Alexa to return to the home page of its user interface. In spontaneous comments during and at the end of the session, several participants noted that while there were some technical limitations in what the device could do, they were fascinated in being able to interact with the device in ASL.

While Alexa displayed its response to the command or query, participants generally nodded and/or produced an eyebrows-raised facial expression indicating acknowledgement or wonder/surprise at the response. On the other hand, when an error occurred, we observed that participants generally shook their head negatively and/or produced a facial expression indicating aversion. In some cases, participants reacted to Alexa's responses using an ASL sign; for instance, some participants used the aforementioned QMWG sign on its own as a reaction, meaning of "Really?" or "Seriously?"

Although participants spent the full recording session giving commands and queries to the device, many participants eventually ran out of ideas of things to say to the device. In such cases, participants sometimes looked around their room for inspiration of other questions or commands.

Participants indicated that they were generally happy that captions were displayed on the Alexa screen. However, when the device was displaying a video in response to a command, participants

complained when a specific video did not automatically have captions displayed. Participants also mentioned that when Alexa was displaying information on its screen in response to a command or query, e.g., displaying a news article, sometimes the captions were blocking their view when reading articles.

During our review of videos of the participants' signing and the corresponding English commands that had been voiced by the interpreter, no significant errors were observed from the ASL-to-English voice interpretation. When voicing English commands to the device, the interpreter used standard English grammar, rather than voicing a word-for-word transliteration of the ASL-syntax command; this fluent English translation was provided because Alexa was expecting input commands in standard English word order. In cases in which participants modified their signing to use more English-like word-order, e.g., when the device did not seem to understand their original command, the interpreter continued to voice English-word order commands to the device, albeit now with more direct guidance from the participant about the English wording and structure to use.

To understand the interpreter's spoken English input, Alexa uses a state-of-the-art automatic speech recognition (ASR) engine. The interpreter who voiced the commands for the DHH users was a native English speaker born in North America, and ASR systems generally work well for such speakers. We observed that the device even understood proper nouns, e.g., city names, sports teams, or celebrity names, that were voiced by the interpreter. We observed that the errors in the interaction mainly arose not from speech-recognition errors, but rather from misunderstanding of the query itself or the user's intent.

Sometimes, the Alexa device displayed English text on its screen corresponding to the spoken words it had heard voiced by the interpreter. It would typically display this on the top of the screen, with the rest of the screen containing the response from Alexa. Most participants recognized immediately that the text appearing on the screen in such cases was the English translation of what they had signed, but some participants were uncertain of the meaning of this text, e.g., wondering whether it was Alexa's response to their command or a question for them to answer.

9.6 Discussion

In this section, we comment and speculate about the findings, discuss some implications for designers, and motivate future research on personal assistants with ASL signers.

9.6.1 Discovering New Approaches to Device Wake-Up

As described in the RQ5.1 findings (section 9.5.1), there were different ways that DHH users activated the device. In Deaf culture, it is typical to wave your hand in a person's field of view to get their attention. There were several signs that were used to wave hello – coded "Hello" (hand from head into air like a salute), "Hi" (waving hand with palm facing out), "Hey" (waving hand with palm facing down), and "HEY" (fingerspelling the letters H-E-Y). There were also some signs used for device activation that are common in ASL conversation – coded "Curious" (using the ASL sign CURIOUS) and "Do-do" (looks like pinching the thumbs and index fingers with palms facing up). These signs are colloquialisms in ASL and are typically used by culturally Deaf individuals, e.g., prior to asking a question. The set of wake-up approaches in our dataset (chapter 7) differed from those discussed in a prior interview study that had asked DHH participants to imagine ways of waking up a device (chapter 4); that study reported talk-to-talk methods such as waving the hand, spelling or signing the device name, or clapping. We did not observe any participants using clapping for device wake-up in our dataset, and our dataset included more variations of hello-like signs than had been discussed in that prior work. This demonstrates that, when participants are actually given the opportunity to interact with a prototype, they may spontaneously try things they had not imagined before.

9.6.2 Use Cases and Commands of Interest to DHH Users

Table 9.2 presented 15 categories covering the wide variety of commands in our dataset shared in chapter 7. Over 1,400 user utterances were collected and, as described in that chapter, ways the DHH population may make special use of such devices were revealed. These include accessibility features, questions about services and technology, requests to launch video calling platforms, and sign or tell deaf-related jokes.

9.6.3 ASL Linguistic Aspects of Commands

In section 9.5.3, we listed examples of ASL-gloss transcriptions of commands, demonstrating variation in how DHH ASL signers structure and sign their queries. In chapter 7, we also share the video and annotation dataset [19], so the actual ASL recording of these commands is available to the research community – something that no prior work with personal assistants has done. Designers of personal-assistant devices may also benefit from noting the specific ASL vocabulary items and phrases that DHH participants used spontaneously for various "command and control" commands, e.g., for navigating the user interface of the device or responding to device prompts.

Sign-language-recognition researchers may benefit from the videos in our dataset, which demonstrate the variety of linguistic structures and word-order of these commands. Considering this data, researchers can ensure that sign-recognition models for use in personal assistant devices are able to work with a variety of vocabulary, people, and signing styles. For instance, while several participants signed in fluent ASL, some used a more English-like word-order or structure to their signing. This type of code-switching between fluent ASL and English-like signing is common among DHH signers, e.g., when they are signing with someone whom they believe may be a novice signer [105]. We speculate that the experience of Alexa sometimes giving inappropriate responses to commands may have led some participants to naturally engage in such code-switching behavior.

9.6.4 DHH Users Responding to Errors

Intrinsic to personal assistant devices, there were several possible sources of error during this Wizard-of-Oz study. For instance, the interpreter voicing the commands behind-the-scenes may have mistranslated something, possibly because they could not see the person well. It is also possible that Alexa misunderstood the spoken English of the interpreter, or even misunderstood during the processing of the English words (e.g. semantic or natural language processing issues). Our observations, as discussed in section 9.5.6, did not capture any significant errors in the ASL-to-English voicing or the English words understood by the device. There were still several errors that occured, mainly due to the processing of

[19] http://doi.org/10.17910/b7.1392

the English words and Alexa not responding with what the participants were expecting.

Our general observations of the sessions discussed how some participants began signing to Alexa in full speed with fluent ASL linguistic structure, perhaps initially thinking or hoping they would be able to communicate with Alexa very naturally, as if it was like another deaf person. However, after some erroneous responses, their signing speed slowed down. Table 9.3 had presented the various ways in which participants responded after Alexa made an error during an interaction. Some of these behaviors were analogous to those that may occur during spoken interaction with a device. For instance, the "reworded" cases in which the participant repeated their command with a slight change to its wording – or only repeated or emphasized a subset of their original command that Alexa seemed to have misunderstood. While this may also be a typical conversational behavior among spoken-language users, the propensity for our participants to clarify or emphasize a misunderstood ASL word by switching to fingerspelling its English equivalent may be unique to ASL context of this interaction.

9.6.5 DHH Users' Imagination vs. Experience

Our findings in section 9.5.5 described how participants' interest in sign-language interaction with a personal assistant device remained the same after engaging in this Wizard-of-Oz based prototype experience. This finding suggests that our Wizard-of-Oz implementation in this study was sufficient to simulate DHH users' imagined success and experience in interacting with such devices.

9.7 Conclusions

Through an observational study of DHH signers interacting in ASL with a personal assistant, via a Wizard-of-Oz approach, we recorded a dataset of over 1,400 ASL utterances. An analysis of the recordings has provided insights and motivation for personal-assistant designers and sign-language-recognition researchers. Specifically, this study has addressed various phases of personal-assistant interaction, including the device activation, issuing a command, and responding to errors.

We empirically contributed novel ways that DHH signers wake up the device, beyond those which had been discussed in prior work. We revealed a set of topic categories covering the wide variety of com-

mands that users produced during our recording sessions, including some commands of specific interest to DHH users. We also describe various ASL vocabulary and linguistic structures that our participants employed throughout their commands. Our analysis also revealed some unique ways users responded when the device produced erroneous output, e.g., switching to English fingerspelling of words.

Chapter 10

Evaluation of In-Person Interaction in ASL with a Personal Assistant

10.1 Introduction

As a reminder to the reader, chapter 7 set up a remote Wizard-of-Oz study to directly engage with the DHH community using an ASL personal assistant prototype, and chapter 9 analyzed this data, investigating each stage of the ASL interaction. While this work revealed several valuable insights, there were inherent limitations due to the lack of real-world conditions. Chapter 8 set up an in-person Wizard-of-Oz user study, and in this chapter we present the research questions and data analysis.

10.2 Research Questions

During the remote protocol, users were asked to try out the personal assistant and continue to invent different commands and requests to try. So, while we were able to observe many different types of commands and curate a list of command-categories, we were unable to capture important linguistic features that would arise during in-person interaction. Additionally, participants could not try out things that involved other things in the room, e.g. smart devices like lights or a TV.

It is important for personal assistant designers and developers of sign recognition technology to

understand different linguistic characteristics of ASL that may influence the way they use and interact with a personal assistant. For instance, a DHH user may "point" to an object in the room to refer to it within a command issued to the personal assistant. The in-person study provided an opportunity for users to interact with a personal assistant device while being in a natural environment, focused on typical activities they may do while at home.

> **RQ6.1:** What are some linguistic properties of in-person interaction with full signing space, such as referencing to other objects inside the room?

In a home, a DHH user would not be limited to one position; they would be freely moving around the room and could spontaneously interact with a personal assistant at any time, so it is important for future designers of personal assistants to make sure that they are able to capture the user wherever they are in the room.

> **RQ6.2:** Based on observation, how would a DHH user naturally position themselves in proximity to a personal assistant device in a residential-like living room and kitchen area?

Through the virtual Wizard-of-Oz setup, users were not physically in the same room as the personal assistant device, as they would be if they owned one in their home. It is unknown what their general experience and reactions to this would be.

> **RQ6.3:** When given the opportunity to interact with a personal assistant device that is physically in the same room as them in a residential setting, what are DHH users' experience and opinion on this?

10.3 Research Methodology

The user study setup has already been presented in chapter 8. To remind the reader, a room was decorated to emulate a residential living room and kitchen area. TVs and webcams were strategically used, along with the Echo Show Alexa device, to create the Wizard-of-Oz setup, so that participants could interact with the device naturally in ASL.

10.3.1 Recruitment and Participant Demographics

A total of 12 DHH participants were recruited through social media postings and advertising on university campus. Like the remote Wizard-of-Oz study; the eligibility criteria for participation was being at least 18 years old and having used sign-language at home when they were a child and/or attended an elementary or middle school where they used sign language every day. Participants met two members from our research team in a research studio on-campus, and were compensated $40.

Participants self-identified as 7 men and 5 women, mean age 29 with standard deviation 10.5. 10 self-identified as Deaf, and 2 as Hard-of-hearing. 7 were born DHH, one became DHH at age 2, two at age 3, and two at age 4. While 3 participants reported having DHH parent(s), 8 reported that their parent(s) use at least some sign at home, and 6 reported using sign in elementary school. 9 described their elementary school as mainstream (where both hearing and DHH students attend), and 3 as a daytime school for the Deaf (where the students commute to). Among participants' education level, 5 had a high school diploma, 2 had some college, 3 had a Bachelor's degree, and 2 had a Master's degree. The average number of members in each participant's household was 5 (standard deviation 1.3), while the average number of household members that use ASL was 2.7 (s.d. 1.6).

10.3.2 Study Procedure

The room layout (described in chapter 8) was checked to ensure that everything was in place. Once the participant arrived, they signed a informed consent form (this study was approved by the IRB, and included a video-recording release). Then, participants were given an overview of the user study setup, and had an opportunity to ask any questions they may have before the research team initiated a Zoom meeting and left the room, allowing the participant to unmask (to remind the reader, there were university-mandated COVID-19 policies in place at the time of this study).

After moving to a separate room, the research team joined the Zoom call, and asked the participant to start going through a task list. The task list was designed to reproduce a typical "day" at home, giving the participant a few different goals to achieve while interacting with the personal assistant device. The tasks were specifically designed to evoke participants to interact with the device while in different

locations throughout the living room and kitchen area. The task list also included a variety of expected results and output modalities from the personal assistant, such as information on the screen, media on the TV, and controlling the lights. While the participant was going through the task list and issuing commands to the device, the interpreter was watching through the Zoom call and voicing any commands to spoken English for the Echo Show. The moderator had their video turned off, but was watching and taking notes about what the participant was doing, noting anything interesting to bring up and ask about during the post-experiment interview portion of the study. The task list given to the participant was as follows:

- **START:** You get home, make yourself familiarized with the space, interact with Alexa
  - Ask Alexa to turn on the light
  - Try out a few commands
- **Task 1: Making drink**
  - [See drink task]
  - Ask any questions related to making drink
- **Task 2: Watching YouTube video on living room TV**
  - You want to watch a YouTube video on the TV, but the TV is off. Ask Alexa to help you.
  - Now you are watching a video, but you feel that the lights are too bright, ask Alexa to do something about it.
  - Ask Alexa about something from the video you are watching.
  - While you are watching the TV, it is possible that the smoke alarm will go off and you would not know. Ask Alexa to help you out.
- **Task 3: Getting snack - Food and making tea**
  - [See snack task]

  - You need the light on in the kitchen area, ask Alexa to help you out.
  - Since you are not in the living room area, you do not need the TV and table lamp on, use Alexa to address this.
  - Some teas contain caffeine. Ask Alexa a question you have about caffeine.
  - You come across an unfamiliar word on the box of tea – use Alexa to help you understand.

- **Task 4: Setting up video chat**
  - Ask Alexa to set up a video chat with a friend

- **Task 5: Ordering food**
  - Ask Alexa to find best pizza places
  - Ask Alexa to show the website of the most interested pizza place
  - Or Ask Alexa to recommend the most popular pizza
  - Tell Alexa to place an order of pizza recommended
  - Tell Alexa to notify when the pizza was delivered

- **Task 6: Reading a book**
  - Tell Alexa to adjust the brightness of the room light
  - Ask Alexa a question about the book for getting information
  - Tell Alexa to notify when doorbell is ringing or fire alarm is on, etc.

- **END** the tasks

"Drink task" is as follows:

- There are teaspoon and cup measures available.
- Add approximately 30 grams of powder to 12 ounces of water and stir completely.
  - Remember, Alexa can help you out with conversions!

"Snack task" is as follows:

- Ask Alexa to turn on the floor lamp for you. When you are in the kitchen, use Alexa to turn off the table lamp and the TV.
- There is prop food, a plate, and a napkin available.
- Make cold tea: Put the tea bag in a mug and dispense water in it.
  - You realize this is your 5th cup of tea today! Use Alexa to figure out how much caffeine you should drink maximum per day?

After the participant finished going through these tasks, the researcher moved back to the main room with the participant, and started the interview portion of the study. The entire interview was conducted in ASL, and the researchers wrote down the participants' responses in English. The interview questions started with asking the participant for general comments about their initial thoughts and reactions during the user study. Then, the participant is reminded of the tasks they did, with questions interleaved. The researcher then asks about any of the interesting observations from their notes while watching the participant conduct the tasks, and also asks about their locations inside the room while interacting with the device. Near the end, questions are asked about the line of sight between the Echo Show and the participant, and whether the participant had any opinions or feedback about this. Questions from the system usability scale are presented through ASL videos from [81]. Lastly, participants are asked if they had any negative comments about the experience, such as something that was not comfortable or suggestions for changes and future improvements. The researcher's copy of the interview questions was as follows:

- Before we go ahead with the interview questions, do you want to share anything about your general initial thoughts/reactions about your experience? Anything at all!
- Please think back to your first interaction with the device at the beginning – what was your first impression? How close to the device did you get when you were interacting with it?

- If you recall, next you made a drink, and you asked Alexa for some measurement conversions. What was your experience interacting with the device while you were doing this task? Was it comfortable? Did it go well?
- Next, how was your experience using Alexa to watch YouTube on the TV? Would this be something you would do if you owned a personal assistant device that could understand your ASL commands?
- While you were getting a snack and making tea in the kitchen area, you interacted with Alexa. How did you feel about this interaction, since you were doing a task in the kitchen area?
- When you "ordered food" using Alexa, you had a dialogue with the device as you completed the task. Did this interaction go smoothly?
- <Bring up and ask about anything that was interesting during the session> E.g. I noticed you did XYZ, could you tell me more about that? Why did you decide to do that?
- What were your thoughts when you did XYZ?
- <Bring up their positions during the study, e.g. "while you were doing X, you were sitting/standing at Y"> Can you tell me about why you decided to sit/stand where you were? Were you thinking about where to place yourself? Did these different placements impact the way you interacted with Alexa?
- <If they used pointing to point at lights or other things in the room> When you did X, you pointed at Y, did you expect Alexa to know what you were pointing at?
- Did you think about whether Alexa could "see" you? How did you judge whether Alexa can "see" you and is ready for a command?
- Do you think Alexa should also be able to "see" the other items in the room? (e.g. the items you referred to)
- (demo opportunity) Since we are still here in the room, could you show us where you think, if any, Alexa could see you the best and where you are most comfortable, and where you think

Alexa cannot see you. [make sure to capture this via drawing or camera – can use living room for interview]

- Would you have interacted differently if there were other people in the room with you?
- If you were in a different environment, such as a bedroom, how would you use the device differently than you did here?
- Go through ASL-SUS questions about their experience (ASL videos and answer sheet available via http://latlab.ist.rit.edu/assets2017sus/).

After the interview questions, participants responded to a demographic questionnaire, and were paid and offboarded.

10.4 Data Analysis and Findings

As chapter 8 described, the video recordings from the user study were iteratively annotated (see tables 8.1 and 8.2). The 531 annotations were analyzed, and the video recordings were observed to identify different aspects of the in-person personal assistant device ASL interaction. This section is broken down into the research questions posed earlier in this chapter, and has the findings interleaved.

10.4.1 RQ6.1: What are some linguistic properties of in-person interaction with full signing space, such as referencing to other objects inside the room?

Throughout the participants' ASL interaction with the Alexa device, it was evident that there were a lot of linguistic features similar to person-to-person ASL interaction. These features were apparent throughout all the stages of interacting with the device; from wake-up to responding to device output. To start, eye-contact is critical for Deaf ASL users – and participants carried this to personal assistant device interaction, even though the device is inanimate. For 403 (76%) of the 531 commands, the participants kept eye-contact with the Alexa device the entire time. 114 times there was partial eye-contact, and for only 8 out of 531, there was no eye-contact. For 6 of 531, the participant's head was

off-screen on the device-view camera (but it is estimated that they are looking at the Alexa device), as shown in fig. 10.1.

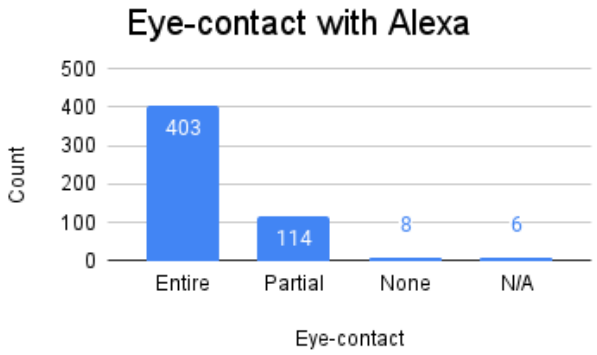


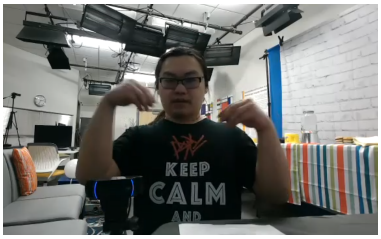
Figure 10.1: Count of in-person Wizard-of-Oz Alexa eye-contact

During the commands themselves, many different examples of sign language space utilization, both linguistic and non-linguistic were captured. In sign language, signing space is the three-dimensional space in front of the signer, from the waist to the forehead, and from one side of the body to the other, used to represent physical space and conceptual structures (e.g. syntactic, temporal, and topographic placement). For some tasks, participants were encouraged to use Alexa to turn on/off or change the brightness of various lights inside the room. A task also asked participants to use Alexa to display media on a TV. When participants issued such commands, it was common for them to look at the light or TV during or after the command. Participants also used many directional verbs (signs that include the subject, verb, and object in one movement) in their commands, such as signing "BACK" in the direction of the TV, "INFORM-ME", and "LET-ME-KNOW" (moving from the Alexa device to the user). Often, when participants were fingerspelling a word, they would use their other hand to point at the fingerspelling hand. As is normal in ASL, participants used classifiers (signs using handshapes that are associated with specific categories/classes such as size, shape, usage, or meaning) throughout their commands to refer to movements, light brightness, distance, etc. Participants also pointed at objects they referenced in their commands.

After making commands, participants' language use continued to include linguistic characteristics typical of human-to-human interaction when responding to Alexa's output processes. As Alexa re-

(a) P1 looking at the floor lamp

(b) P3 pointing at the TV



(c) P6 using classifiers to represent the lights

(d) P10 using classifiers to represent smoke

Figure 10.2: In-person Wizard-of-Oz screenshots showing four examples of linguistic features: (a) shows P1 looking at the floor lamp after they had issued a command asking Alexa to turn it off. (b) shows P3 pointing at the TV during a command related to it. (c) shows P6 using classifier signs to show the lights turning off. (d) shows P10 using classifiers to sign smoke before spelling "SMOKE" for a command related to the smoke alarm.

sponded, participants would often backchannel (indicating that they were paying attention and understanding incoming information) and/or talk to themselves in ASL. For instance, similar to how hearing people would say "mmhm" and nod their head to show they are following a speaker, the DHH participants often nodded their head and signed things such as "OH-I-SEE", "VEE", "INTERESTING", "OK". Participants also talked to themselves, especially when the interaction did not go smoothly. For example, participants signed "FAIL, OK OK", "THAT THAT", "FINE", and other things that may not be an explicit response addressed to Alexa. Some participants talked to Alexa after the response, such as "THANK YOU ALEXA" and "THAT THAT THANK YOU fs-ALEXA".

Participants also changed their language with Alexa as they progressed through the study tasks. We observed that participants changed their word or sign-choices, and fingerspelled things more, and/or changed their grammatical structure to be more English-like, especially as Alexa failed to satisfy their queries. For instance, P3 fingerspelled the word "Pekoe" very slowly and repeatedly to try to help Alexa understand the question. They also changed the angle of their hand so that the handshapes of the

letters were more clear for the device's camera view. P3 also changed the wording of this question from "A-L-E-X-A PLEASE EXPLAIN WORD P-E-K-O-E" to "A-L-E-X-A EXPLAIN P-E-K-O-E", making their question simpler and use less signs. Eventually, Alexa responded with an explanation and satisfied the participant before moving on to the next task.

10.4.2 RQ6.2: Based on observation, how would a DHH user naturally position themselves in proximity to a personal assistant device in a residential-like living room and kitchen area?

Figure 8.5 shows the Wizard-of-Oz study room layout with 10 position labels added. After each of the 531 commands were annotated, the frequency for each user-location was counted, and fig. 10.3 shows the totals. For 225 (43.1%) out of 531 commands, the participant was located in the center, folowed by Alexa (16.9%), Seat 1 (16.9%), TV (14.5%), and then the rest were below 6%.

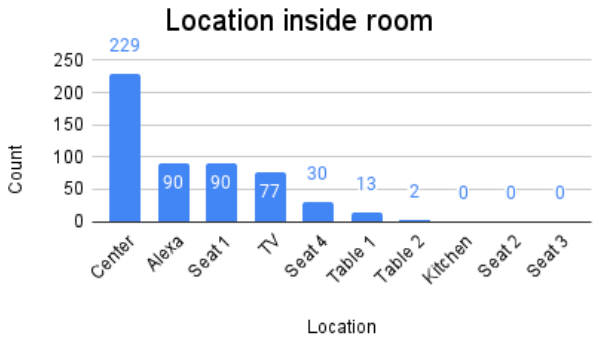


Figure 10.3: In-person Wizard-of-Oz participant location frequency

Participants were asked about the different tasks they conducted throughout the user study (10.3.2), and gave mixed answers; some said that a task went smoothly with Alexa giving the correct response, while others said the task failed and they moved on. While participants' levels of satisfaction about their experience varied, all participants mentioned ways in which they had to adjust their behavior to ensure that Alexa was able to understand them. P6 said "*I thought I needed to sit close to have Alexa understand me. I did not think Alexa would pick up on my signs from far away.*" P5 said that "*I thought that if I was farther, then Alexa would not be able to see me because of the quality of the camera. I chose my location*

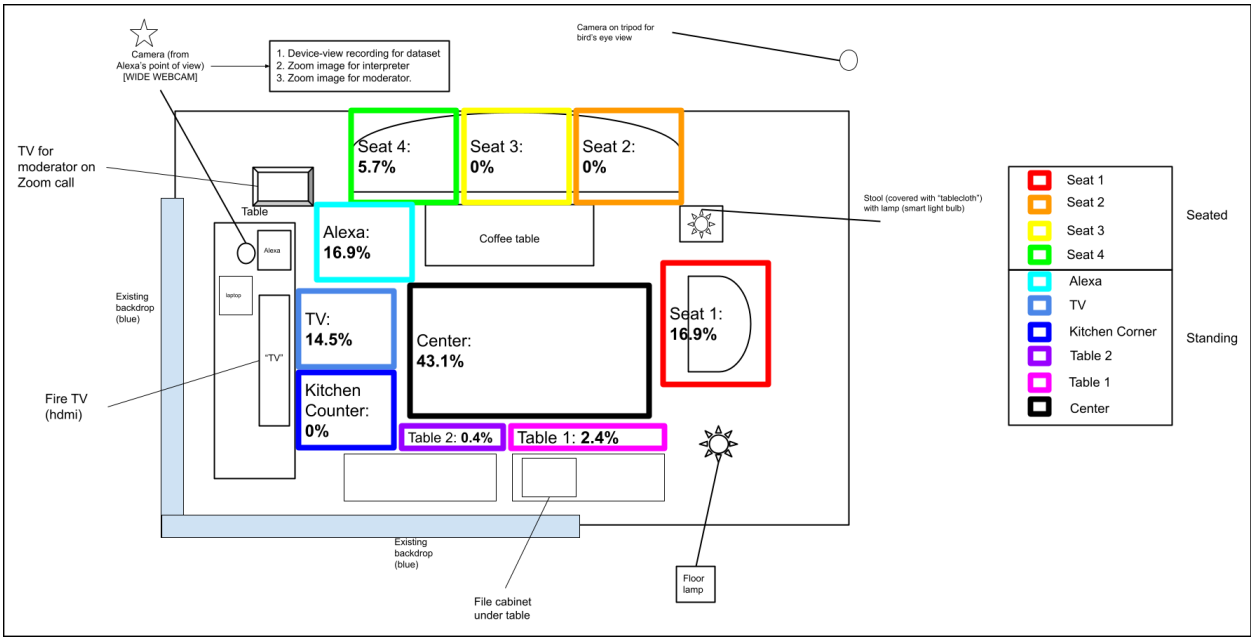
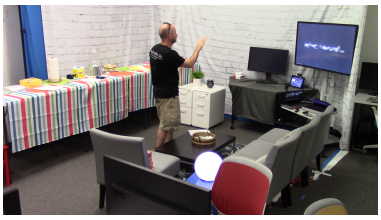
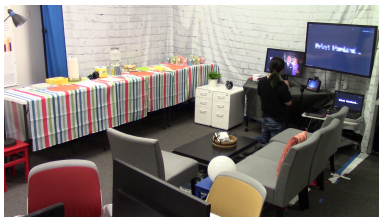


Figure 10.4: Room layout of different participant placements during in-person Wizard-of-Oz, with percentages added for each position

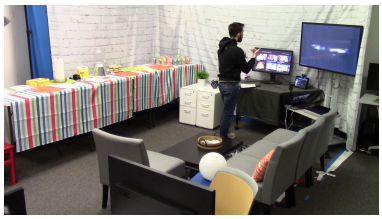


(a) P11 in "Center"



(b) P6 in "Alexa"

(c) P12 in "Seat 1"



(d) P8 in "TV"

Figure 10.5: In-person Wizard-of-Oz screenshots showing the bird's eye camera view of the top four locations inside the room that participants placed themselves at.

based on the camera I saw and what I thought the optimal distance would be. I enjoyed moving around because I don't like staying in one place. I suggest having two Alexa devices, one in the living room and one in the kitchen so I could move around and change positions easily." P11 said "*I did not move much during the experiment because I felt like the camera was angled and stayed in one place. If the camera moved, maybe I would have tried different places for the Alexa to see me.*" During the interview, participants were asked where they thought Alexa could see them best. Several participants said that location "Center", "Alexa", "TV", and "Kitchen Corner" (see fig. 8.5) were closest to Alexa and therefore were the best. Several participants also commented that you have to face Alexa, otherwise it would not be able to see the person signing. One person suggested a 360-degree camera be placed at the center of the ceiling in order to see the participant anywhere they were.

10.4.3 RQ6.3: When given the opportunity to interact with a personal assistant device that is physically in the same room as them in a residential setting, what are DHH users' experience and opinion on this?

Initially, participants were excited about the technology, with P2 commenting "*It was interesting to use sign to communicate with Alexa and it seemed easy*" in the post-experiment interview. P4 said that "*it was cool, and I think a lot of Deaf people would enjoy interacting with Alexa in this setting.*" Participants said they were surprised that Alexa was able to understand their signing, i.e. P10 saying "... *very interesting and I thought it was impressive when Alexa can understand ASL ...*" Some participants, especially if they did not have prior familiarity with Alexa, were not used to the interaction, e.g. P3, who said "...*I don't have a lot of experience using Alexa so it was a little bit awkward but it was an exciting experience.*" P6 said "*it was odd to talk to Alexa when there was no one else around. I felt like I was talking to myself.*"

During the interview, participants mentioned that, as the study went on and they had some time interacting with the Wizard-of-Oz Alexa, they became frustrated and had several criticisms. For instance, participants had comments about Alexa's limitations for processing queries and giving straightforward answers. For example, P3 said "*Alexa struggled with answering some of the pizza questions and looked up Lucy's instead of Luccis...*" After performing then kitchen recipe task requiring measurement conversion,

P1 said that "*the Alexa did not do the correct conversion so the task was not completed.*" P5 said that "*Alexa would not give me a straightforward answer and gave way more information than necessary, and that I would prefer just the number to appear on the screen.*" P11 summarized: "*First, I was very excited. But then I realized that it takes a long time and it doesn't always understand things. With more specific questions it can't quite do the task. The responses from Alexa tend to be off from what I want them to be*".

Participants were asked if they think they may have interacted differently if there were other people present in the room with them. All 12 participants commented that Alexa may have difficulty recognizing signs meant for it, since there may be other people signing in the background. Due to this, several participants suggested that users would have to move closer to Alexa, since it would be more difficult to get Alexa's attention with multiple people in the room. A few participants ideated that Alexa could utilize facial recognition technology to detect the correct signer trying to get its attention. Participants were also asked about whether they would have interacted or used Alexa differently if they were in different environments or rooms inside a home. A few scenarios mentioned were using it as an alarm (vibrating and flashing lights) in the bedroom (P1), controlling a computer or TV in an office (P2), and making phone calls and helping with work tasks (P4). Other participants said it was best used in a common room such as a kitchen or living room (P3,6,7,9,10). These participants rationalized that they would not want it in a bedroom or some other rooms due to privacy concerns.

10.5 Discussion

The study previously presented in Chapter 9 was virtual – participants were not physically in the same room as the Alexa device, and interacted with it through a Zoom video conference. From the physical nature of this study, we were able to capture properties of the personal assistant device interaction beyond the device wake-up and ASL commands. The findings from the "Zoom study" allowed participants to think of their own queries to issue the device and enabled investigation on how users instinctively wake-up devices and initiate their commands, as well as capture what different commands look like in ASL and how users responded after a breakdown or error during the interaction. Here, we provided a task list that participants followed, rather than freely interacting with Alexa, and focused on other

properties that contribute to the overall interaction experience. Together, these studies provide guidance for future technology development, and break down the overall experience into actionable goals, both short and long-term.

The linguistic features that were captured in the video recordings give important insight comparing human-human interaction to human-device interaction. For instance, it was shown that eye-contact was essential for all the participants interacting with Alexa, even though it is an inanimate device. In addition to this eye-contact for getting attention and starting a command, the eye-gaze during the commands themselves also contributes to the ASL command; it was shown that participants would gaze to the lights or TV when they were issuing a command related to them. Various forms of referential and spatially depictive signing were also important during this interaction. Participants pointed to objects inside the room when referencing to them and also used classifiers to describe what they wanted to do with them. We also observed participants back-channeling to the Alexa device, similar to how they might indicate to a human conversational partner that they were paying attention to their signing. Participants also engaged in self-talk in ASL, e.g., whether they were satisfied with the device output. As participants saw device errors or undesired answers, they often changed their language by choosing different words to use, fingerspelling words, and/or changing the structure of their sentences to be more English-like. Future sign-recognition technology embedded in a personal-assistant device would not only need to understand the vocabulary and linguistic structure of commands being issued to the device, but it would also need to understand these various forms of linguistic phenomena that we have observed during these interactions, many of which are typical during human-to-human ASL conversation.

Recently, after the work for this dissertation was completed, Google added a new feature to their Assistant that aligns with the aforementioned importance of eye-contact for ASL interaction. Google added "Look and Talk" to their Nest Hub Max [20]; where users can now, while they are satisfactorily close to the device, look at the screen and speak their command without the device-activation phrase "Hey Google." As virtual assistants are increasingly embedded into different form factors, such as smart phones and watches, future work is needed to figure out how such properties of the interaction would translate.

[20] https://support.google.com/googlenest/answer/11410414?hl=en

531 commands were captured and annotated from our 12 participants. As described above, these commands were issued by participants while they were located in various areas of the room, with those locations being at a variety of distances and angles from the device's camera view.. It is essential for personal assistant devices to be able to understand users if they move around the device's field of view, within reason. Trying out ASL commands from different distances and positions inside a room would be important for developers of sign-recognition technologies. It would also be important for designers of personal assistants to make hardware decisions as to what type of integrated camera would be necessary to capture an ASL signer who might be located close to or far away from the device, and might be positioned at an angle where their body is not directly facing the camera. Participants explained that their position inside the room also depended on whether they could see the output on the device screen. This informs designers that the device size and output modality have an impact on the user behavior and interaction. For instance, small text shown in a subregion of the device's screen may require the user to be closer to the device, but full-screen large text or display of videos/pictures might allow for users to be at a greater distance from the device. As different device form factors become common, it is necessary to investigate this, and how user needs and preferences would vary with different devices.

Participants were initially excited, surprised, and amazed that the Alexa device appeared to understand their signing. However, they commented that they became frustrated over time and had several criticisms about the experience. This shows that appropriately setting users' expectations is important for applications of future sign recognition technologies, such as this context of personal assistant devices. Developers of sign recognition models and designers of personal assistants need to empirically investigate how variations in latency and perceived accuracy of their technology influence users' satisfaction – as well as whether there exist minimum levels of latency or accuracy needed for this application to be sufficiently usable. In this research, participants also commented about other people being in the room, and how that may impact the ability of the personal assistant device to understand their ASL commands, since the device could incorrectly decide that something that had been signed to a human in the room had been intended as a command to the device. Future researchers and developers of sign-language-based personal assistants will need to consider how to determine when users are addressing

the device or other people in the room.

10.6 Conclusions and Future Work

The analysis of the video recordings and the interview questions contribute toward addressing the gap in knowledge about how DHH users react to interacting with a personal assistant in-person, physically in the same room as them. We have gained insight from DHH ASL signers' perspectives, crucial for developers of personal assistant technologies who want to make their devices accessible for these users.

We have discussed linguistic properties of the in-person interaction not investigable during the previous remote Wizard-of-Oz, and explained how these properties have an impact on the personal assistant device interaction. Also, we have discussed other things impacting the interaction such as the location of the user inside the room and the device output modality. We also discussed the general user experience and opinions on the interaction through interview questions. Through all of this, we have motivated several avenues of future work for developers of sign recognition models and designers of personal assistants.

EPILOGUE TO PART III

This is the end of Part III of this dissertation. We analyzed the data collected in the virtual Wizard-of-Oz protocol (from chapter 7), discovering insights about each stage of ASL interaction with a personal assistant device. In this part, I also followed up via an in-person, physical Wizard-of-Oz experiment (chapter 8), and collected more data. Analysis of this data (chapter 10) provided further insights about ASL interaction with personal assistants, including several topics that could not be explored via the virtual experiment. Part III of this dissertation, through analysis of the video recordings and interview questions from these Wizard-of-Oz studies, addressed a gap in knowledge and informed designers of personal assistants and sign-language researchers:

RQ5.1: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users instinctively "wake-up" the device or initiate a command? Through analysis of over 1400 utterances, 12 different wake-ups were identified, and information is provided about each one. (section 9.5.1)

RQ5.2: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, what categories of commands/requests do users produce? A list of categories was compiled, and descriptions along with samples are provided for each, as well as frequency. (section 9.5.2)

RQ5.3: What do these commands look like in ASL? Several linguistic properties are mentioned, with examples, and variety in signing style is described. Also, despite the remote, virtual nature of the experiment, some participants utilized the space around their body during commands and

questions in the study. (section 9.5.3)

RQ5.4: Based on observation of the behavior of DHH ASL signers interacting with a personal assistant device in sign language, how do users recover or respond when there is an error or breakdown? Recordings of the entire interaction for every participant was analyzed, and 5 different behaviors were identified when there was an error or breakdown. (section 9.5.4)

RQ5.5: After DHH ASL signers had the opportunity to interact with a personal assistant device in sign language, did their interest in such interaction increase, decrease, or stay the same? Participants' interest was statistically equivalent before and after the experience of interacting (albeit remotely) with a personal assistant device that appeared to understand ASL. (section 9.5.5)

RQ6.1: What are some linguistic properties of in-person interaction with full signing space, such as referencing to other objects inside the room? Several interesting and important linguistic features were captured and discussed. Clear evidence of space utilization and usage of eye-contact/gaze is captured (section 10.4.1).

RQ6.2: Based on observation, how would a DHH user naturally position themselves in proximity to a personal assistant device in a residential-like living room and kitchen area? Video recordings of 531 commands from 12 participants resulted in 10 location-labels. A frequency chart is shown, and participants commented on their positioning rationale in interview questions (section 10.4.2). We found that participants tended to position themselves directly in front of the device, generally standing in the middle of the room or relatively close to the device screen so that they could see the output.

RQ6.3: When given the opportunity to interact with a personal assistant device that is physically in the same room as them in a residential setting, what are DHH users' experience and opinion on this? Interviews with participants were thematically analyzed, capturing reactions and opinions which changed throughout their interaction (section 10.4.3). Participants indicated initial excitement which tended to diminish when the device did not meet their expectations of

accuracy, and they discussed potential concerns about how the device would behave if there had been multiple people in the room.

Part IV: Privacy concerns

PROLOGUE TO PART IV

Thus far, Part I has used a mixed-method study to provide a basis for the research to investigate DHH users' interaction with personal assistant devices (discussed in Part III), and to construct a dataset of videos of sign-language personal-assistant commands (discussed in Part II).

During the previous parts of this dissertation, the issue of privacy concerns has come up, e.g., a camera-based personal assistant would follow the user wherever they naturally position themselves in a home. Potential DHH users of personal assistants were concerned about the device picking up on signs that were not meant for it, as well as capturing other people in their homes, citing the importance of being able to manually control whether the device can see/hear them, e.g., with a physical cover that blocks the camera section 3.5.2 (chapter 3). Participants in the ASL data collection studies were concerned about the privacy of their data, e.g., who would "own" it and who would be able to access it, especially if it was shared publicly (sections 5.4.4 and 6.5.3 (chapters 5 and 6)). The issue of privacy is a focus of recent consumer devices, such as the Google Nest Hub and the Amazon Echo Show, which have added features that can disable the microphone and camera.

Parallel to how Part II of this dissertation was structured, where I explored a more general ASL data collection methodology before focusing on the narrow domain of personal-assistant interactions, in this part, I will consider the more broader situation of DHH individuals sharing videos online, before focusing on the case of personal assistants. To understand how modern, state-of-the-art face-transformation technology can be used to preserve anonymity, and to learn what factors impact DHH users' acceptability of this technology, I investigate the following research questions:

- **RQ7.1:** Is state-of-the-art face-disguise technology capable of perserving facial expressions and

natural human appearance for sign language video?

- **RQ7.2:** What are DHH users' interest in and impressions of this technology for protecting anonymity, including users' views of various dimensions of system performance?

Future personal assistant technology, combined with developments in sign-language recognition, potentially would use integrated cameras to capture an ASL signer and process their video to understand the signed command. Since these videos would be captured and transmitted across the Internet to servers for processing, there are privacy concerns in the context of using a personal assistant. In chapter 12 I conducted a study to interview participants (who have just interacted with a personal assistant device in ASL using a Wizard-of-Oz approach) in order to evaluate whether they would find this face-disguise technology to be suitable for helping to protect their privacy for the videos collected and transmitted by the device:

- **RQ8:** Would using a state-of-the-art face-disguise technology to anonymize DHH users' ASL recordings (that are used for device processing) before they are processed by a personal assistant device alleviate privacy concerns?

Chapter 11

American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users[21]

11.1 Introduction and Motivation

In this chapter, we focus on the more general case of people sharing sign language videos online, before focusing on the case of personal assistants in the next chapter.

While there are many sign languages, our work focuses on American Sign Language (ASL), used by over 500,000 people in the U.S. [116]. Although often used in countries in which English is spoken, ASL is a language distinct from English, produced by movements of the face, head, hands, and torso [11, 40, 119, 159, 170]. Being able to communicate anonymously in one's preferred language is essential for participating in a variety of social, professional, and societal contexts. Some prior work

[21] The information in this chapter is based on a joint project with Sooyeon Lee, Becca Dingman, Zhaoyang Xia, Dr. Dimitris Metaxas, Dr. Carol Neidle, and my advisor, Dr. Matt Huenerfauth. I collaborated on the study design and creation of study stimuli, and significantly contributed to data collection, analysis, and writing of a paper published at ACM ASSETS'21 [100].

[10, 109] has focused on techniques to hide the face of a user for privacy protection in circumstances where this may be important. For instance, Internet users may visit discussion boards to ask questions about sensitive topics; individuals may express dissenting political or religious views that could subject them to persecution; or essential professional activities like academic peer-review may require anonymity. While it is relatively straightforward for users of written languages to engage in anonymous written communications online, such options have not been available for users of sign languages. These languages generally lack a written form in common use among the language community, and therefore video-based communication, which reveals the face, is necessary.

While users of spoken language can hide their face on online video-sharing platforms [53, 83, 154, 166, 183], this option is not available to ASL users, as the face conveys essential linguistic information [11, 40, 92, 119, 170]. Barriers to private communication in one's primary language limit online debate or enquiries, e.g., in relation to sensitive topics, such as reproductive health, domestic abuse, or substance abuse, which prior research has revealed to have higher prevalence in the DHH community [14, 135]. Anonymizing the face, while retaining the key linguistic information it conveys, would also enable peer review of academic publications in sign language, conformity in appearance when multiple individuals contribute to a composite video or collection (e.g., entries in a video ASL dictionary), and privacy protection when users contribute videos to ASL datasets for AI research – applications discussed in [24, 108].

Over the past decade, real-time tools for face transformations have become popular among consumers, e.g., to make someone appear to be wearing makeup [86] or overlay a virtual cute animal mask [188]. More recently, AI technologies for real-time face transformation (sophisticated technologies that preserve facial expressions) have matured and become available to non-technical users for producing realistic videos in which a synthetically generated human face in a video is driven by the face of another person. As compared to earlier face-filter technologies (simplistic technologies that do not preserving facial expressions), these advancements enable new applications for DHH ASL users, as it is now possible to replace the face while preserving detailed facial expressions and head movements.

In this research, we conducted an interview study to evaluate prototype face-disguise technology (a

generic term for technologies that obscure the face) applied to videos of human ASL signers, influenced by recent image-to-video technology [149, 150, 165, 187], for replacing the face in a video with a new face from a given photograph, preserving facial expressions and head movements. In one prototype variation, the torso of the human remains in the video, and in another, the torso is hidden to disguise the clothing and body for further obscuring the identity of the signer. For comparison, we also evaluated a simpler face-filter with a virtual cartoon-like Tiger mask, previously evaluated in [24]. In a 70-minute appointment, participants: (1) viewed disguised videos and attempted to identify the person in the original video from a line-up of photos, (2) viewed original and disguised videos processed by prototype variations, and provided subjective feedback about each, and (3) viewed videos of themselves transformed by this technology. In a semi-structured interview, participants discussed their views of the technology, preferences among appearance options, factors affecting acceptability, potential uses, and concerns.

The contributions of this work are empirical and include: (1) The first evaluation with DHH ASL users of modern face-transformation technology, capable of preserving ASL linguistic facial expressions, revealing its effectiveness at preserving anonymity; (2) Quantitative and qualitative evaluation of understandability, naturalness, and anonymity-preservation, to compare prototypes varying in their appearance transformations; (3) Evidence of users' views on the acceptability of this technology, its potential uses, and their concerns; (4) Identification of users' perceived trade-offs among understandability, naturalness, and anonymity protection, with design considerations from our analysis; (5) Evidence of ways in which preservation and transformation of identity relate to users' acceptance of this technology.

11.2 Prior work

11.2.1 Existing Methods of Conveying ASL Anonymously

While researchers have acknowledged the importance of enabling deaf signers to communicate anonymously online [48, 49], most prior efforts to address this problem have aimed to produce artificial writing systems for sign language or to create tools to allow deaf signers to create their own animations

of a virtual human signing their message. Despite efforts to invent sign language writing systems, e.g., [9, 124, 160] or related technologies [25], no writing system has yet gained widespread popularity within the DHH community. Thus, written communication in ASL is not practical for enabling signers to communicate without revealing their identity.

Other work seeks to enable users to create synthetic animations of sign languages, which could, in principle, produce anonymous messages. Prior sign language animation research has largely focused on machine-translation contexts [23], but some work examines how to enable users to script the movements of virtual humans to perform sign language, e.g., [46, 73]. Unfortunately, existing tools are not yet sufficiently expressive to produce clear virtual animation, nor are the tools and techniques for building novel animated messages likely to become simple enough for use by non-experts, despite recent efforts [3, 171]. In summary, despite work on writing systems and avatar technologies, no existing approaches yet provide a satisfactory solution to the challenge of anonymous communication in sign language.

11.2.2 Accessibility of Written/Spoken and Sign Language Online Content Creation

Prior work has examined DHH users' interests, current practices, and barriers, in relation to producing content to share online, e.g., [33, 47, 48, 79] or in the context of social media interaction [108]. When privacy is a concern, DHH users must currently use written English to prepare online messages or content. Given the diversity in written-language literacy levels among DHH individuals [167] and the preference of many DHH users for communication in ASL, DHH users face barriers to online participation [108], if they wish to preserve their anonymity during interactions. This is an inequitable situation, as hearing individuals can express themselves online much more easily, in written or spoken form (assuming that their voice is not recognizable and their face is disguised).

Prior work has revealed particular challenges for users who prefer to produce content in sign language, as they must create and post a video of themselves, with their faces and physical appearance visible to whoever watches the video. Recent research [108] has highlighted challenges that DHH ASL signers face in participating in social media sites by recording and sharing ASL video. As reported in

[108], the need to hold the phone with one hand (e.g., while standing) in order to record themselves leaves only one hand for signing, which is not ideal, because signing in ASL normally requires two hands. Adding text captions to videos to enable them to be understood by individuals who do not know ASL is also time-consuming. The authors provided potential solutions for these challenges, such as incorporating automatic captioning into social media platforms. While [108] focused on barriers to communication on social media platforms, our work focuses on preserving DHH individuals' privacy in video communication. In summary, prior work has revealed that there is strong interest among DHH users for technologies that could facilitate ASL-based communication online, especially in a manner that is privacy preserving; yet existing technologies are not providing an adequate solution to this challenge.

11.2.3 Video de-identification for privacy in video sharing sites

Some recent work has investigated face-disguise technology for motivating ASL signers to feel comfortable sharing videos in public ASL datasets for research [24]. This study is of particular interest, in that participants were asked questions about their interest in and impressions of face-disguise technology – albeit within this specific context of contributing to a dataset. Participants were able to see their own video transformed through some simple face-filter technology, including a filter that overlaid a cartoon tiger face on top of the signer's face without preservation of any facial expressions, aside from the degree to which the mouth opens. Participants were more willing to share their video publicly with filters mitigating privacy concerns, yet they were dissatisfied with the fact that the filters did not preserve facial expressions.

In the video/photo sharing context, trade-offs between the utility of the anonymized video/photo and privacy protection have also been investigated [54, 70, 72, 102]. Prior work has studied how the level of obfuscation from various image filtering techniques (e.g., blurring, pixelization, masking) affects the viewer's experience and the utility of the video/image for specific tasks, e.g., patient training video in a clinical setting [54]. As found in prior work [24], obfuscation from some common privacy enhancing techniques does not satisfy ASL signers because facial expressions are not preserved. Prior research suggests that providing adequate privacy protection for various contexts and uses requires careful se-

lection of the relationship between the level (ranging from no recognition to full recognition) and the types (e.g., blurred, masking, face disguise) of anonymization. Focusing specifically on DHH signers, our study differs from prior work in two ways: (a) We investigate more advanced face-transformation technologies capable of preserving facial expressions; and (b) We investigate these technologies for preserving privacy in ASL videos for a wider variety of uses and contexts, e.g., participation on social media platforms.

In recent years, there has been tremendous progress in technologies for analyzing and synthesizing video of human faces, e.g., [10, 125, 164, 165, 187], with new applications in smart home technologies [168], health [42, 76], and other fields. Another key application of this technology has been for de-identifying videos in order to preserve privacy, e.g., [10, 109]. While most work has focused on technical details and performance of this technology, some researchers have conducted research with human participants to understand their interests in or concerns about this technology. Advances in this technology have led to recent public awareness of “deep fake" technologies for producing seemingly realistic videos of humans, in which the movement of the face is based on the performance of a human in an original video. The ease of creating videos that impersonate someone, making it appear that they are saying or doing things that they had never said or done, has raised significant ethical concerns [96, 141].

Given the complex face and head movements used in ASL for a variety of linguistic purposes, e.g., involving subtle movements of the eyebrows or head [11, 40, 119], there has been a question as to whether the resulting video would sufficiently preserve these key linguistic elements of the performance. Some researchers have begun to design face-disguise technology with a particular focus on preserving such elements of the performance [149, 150, 165, 187], necessary for applying this technology to sign language videos. However, there is a need for empirical research with DHH ASL signers, to understand the performance of this technology, as well as users’ impressions and judgments of its suitability for the task of anonymizing ASL videos to be shared online.

11.3 Research Goals and Methods

Emerging face-transformation technology has the potential to create realistic videos with new faces; yet prior work has revealed ethical concerns with the use of such technology. While some research has examined DHH users' interest in simple face-filter technologies for specific contexts, no prior study with DHH users has investigated state-of-the-art face-disguise technology capable of preserving facial expressions and natural human appearance for sign language video. As these new technological capabilities emerge, it is important to understand DHH users' interest in and impressions of this technology for protecting anonymity, including users' views of various dimensions of system performance, e.g., understandability and naturalness of appearance. The goal of this research is to guide the development of ASL-optimized face technology and inform designers of future applications for these users.

We conducted an interview-based study with 16 DHH individuals who reported using ASL on a daily basis; each participated in a 70-minute Zoom teleconference meeting with a DHH ASL-signing researcher. In this IRB-approved study, the participants were shown examples of videos of ASL signing processed by prototype face-transformation technology (section 11.3.1). Prior to transformation, some of these videos had been of the participant, submitted to us in advance of the appointment, and some were of other ASL signers from a public research dataset of ASL signing. The interview was conducted entirely in ASL, while the researchers typed notes in English. Participants were asked a mixture of open- and closed-ended questions about their subjective impression of the videos, especially in regard to how well they preserve anonymity, their understandability, and other factors, as described in section 11.3.2.

11.3.1 Anonymization Technology Prototypes

In this study, we compare multiple prototype technologies for disguising the face of an ASL signer. We refer to our first prototype as **tiger-face**, a simple video filter technology, similar to those used in SnapChat, in which a 3D mask is virtually overlaid on the face in the video. **Our rationale** for selecting this prototype is three-fold: (a) It reflects the state-of-the-art of consumer-grade face technologies popular during the 2010s; (b) The specific filter was used in a prior study that had examined DHH users' interests in using filters to hide their identity [24], the open-source tiger-face filter from Jeeliz

[88]; and (c) It also provides a baseline point-of-comparison for participants, to determine whether the more computationally intensive facial-expression-preserving transformations were useful. The filter detects the human's face and overlays an animated tiger avatar head, which emits blue bubbles from its mouth, triggered whenever the human's mouth opens. Participants in that prior study commented on the limitations of this filter, which does not preserve any other facial expression details, e.g., eyebrow movements, despite this being linguistically important in ASL. We included tiger-face in our study as a baseline for comparison, reflective of the prior state-of-the-art for available face-disguise technologies.

Our prototype, **with-torso**, is based on recent work on image-to-video transformation and video editing, to enable the replacement of the underlying facial geometry, while preserving the linguistically significant facial expressions [149, 150, 165, 187]. **The rationale** for including this transformation in our study was that it reflects a state-of-the-art facial image animation and transformation technology. This specific technology was selected because of its ability to animate face images based on image-to-video transformation, to enable the replacement of the underlying facial geometry by editing the latent facial representations [165, 174]. The torso and background of the signer are not touched or modified in any way. Colloquially, we may refer to the face of the signer being "swapped" with a different human face, based on an input photograph of the desired "target face." However, the resulting output video actually appears as a blend of the facial structure of the original signer and those of the individual pictured in the "target face," resulting in a novel composite face that mimics the head movements and facial expressions of the original signer. Sample images of the output of this transformation are shown in fig. 11.2.

The third prototype, **without-torso**, is identical to the with-torso prototype, except that the signer's torso and the background are both replaced by a flat gray color, as shown in fig. 11.2. **The rationale** for including this transformation is that identity may be revealed not only by the face, but also by body appearance, clothing, or background, especially if the person viewing the video is familiar with the person in the video.

For both with-torso and without-torso, the resulting output can be varied, by selection of different "target faces," and throughout our study we displayed videos based on a variety of target faces, selected

from the Chicago Faces Dataset [106, 107]. We took into account the gender and race/ethnicity of the person in the original video, and we selected target faces of other people with corresponding demographic characteristics – with variation in age, hair style, and hair color. **The rationale** for selecting these variations was that they reflect common options for the selection of video-game avatars or personalized emojis on social media, and several pilot interviews with DHH ASL signers prior to our study revealed their interest in such options. More details about the transformations used in the separate phases of this study are described below. Figure 11.2 shows screenshots from a few videos and their transformations.

11.3.2 Study Design

The 70-minute appointment was temporally partitioned into three phases, for participation in three different activities. During each phase, the participant viewed the videos and then answered semi-structured interview questions. In the first phase, we evaluated face disguise technology from the perspective of participants' seeing a disguised video of other people. In the next phase, the understandability, naturalness, and anonymity protection of the transformed videos were assessed, with participants viewing a variety of face-disguise options. (Prior to the main study, we had conducted pilot interview studies with DHH participants to ask them about their interest in technologies for disguising the face, and this had suggested that understandability, naturalness, and anonymity may be key issues for users, which helped us in finalizing the design of our interview questions for this phase.) In the final phase, participants saw themselves disguised, and they commented on the acceptability of the transformed videos and shared other concerns.

11.3.3 Phase 1 of the Study

The first phase focused on evaluating **how effectively videos had been disguised** by the with-torso and without-torso software; participants were asked to attempt to identify the original person in the video. The source videos used in this phase of the study were from the Boston University American Sign Language Linguistic Research Project [120, 121, 122]. To produce a variety of videos, we selected two

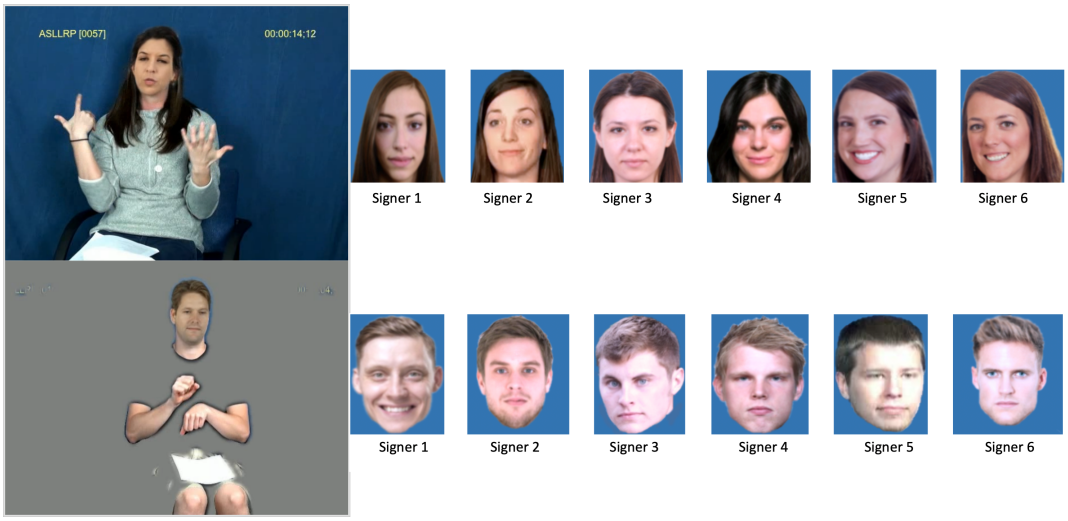


Figure 11.1: Disguised videos shown in phase 1, along with line-up photos including the actual signer and other face images selected with similar hair and skin color; to measure the effectiveness of the disguise, participants were asked to guess the correct face.

videos of a male signer and two of a female signer from this dataset; in each video, the signer produces 1-2 ASL sentences. Next, we processed the videos using each of the two prototypes, with-torso and without-torso, using two different “target faces" for each (two male target faces for the male signer, and two target faces for the female signer). Overall, this yielded 16 disguised output videos.

Each participant viewed one disguised video of the male signer, and one disguised video of the female signer. One video was processed using the with-torso prototype, and the other, using the without-torso prototype. The order in which these stimuli were shown to participants, and the assignment of prototype-condition to each gender, were counterbalanced via Latin square. After viewing each video, participants were shown a line-up of six different faces, one of which was the true face of the ASL signer in the anonymized videos. The order in which these line-up faces were shown to the participants was also counterbalanced via Latin square. Figure 11.1 shows example line-up photos for both the male and female faces. After participants guessed which face was the original person in the video, they indicated their agreement with the Likert item: “It was very difficult to guess the original signer." Phase 1 concluded with questions about participants’ opinions of the videos and their difficulty in guessing the signer, including whether seeing the original signer’s body and background made it easier to guess the original signer’s face.

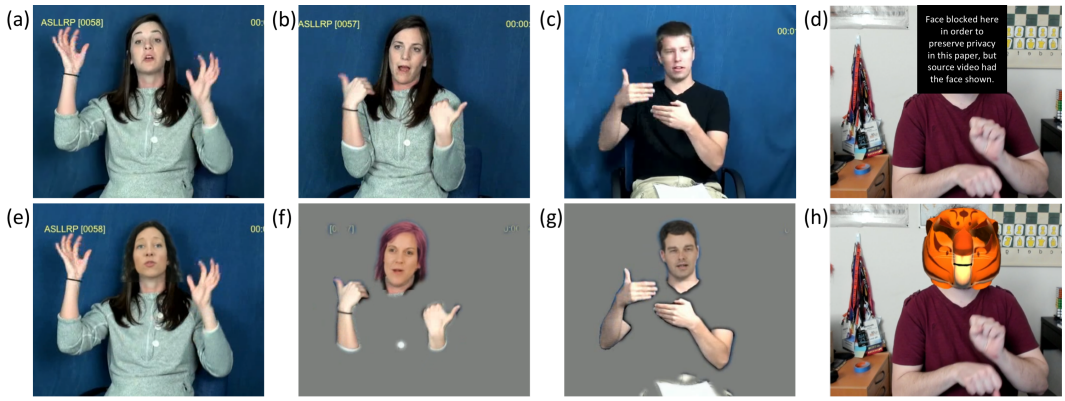


Figure 11.2: Sample of videos shown in phase 2: (a-d) source videos and (e-h) transformed videos below corresponding source, e.g., (a) transformed to (e). Samples include: (e) with-torso, (f-g) without-torso, and (h) tiger-face. Source videos (a-c) from [120, 122] and (d) illustrates the type of videos participants submitted (blocked here for anonymity).

11.3.4 Phase 2 of the Study

The second phase focused on the **understandability, naturalness, and anonymity-protection** of videos from all three prototypes, including with-torso and without-torso videos based on a variety of target faces, as well as the tiger-face prototype. In this phase, each participant viewed a total of 34 videos, half based on a source video from a male signer from [120, 121, 122], and half from a female signer from the same dataset. For each signer, participants were shown an original, unmodified video, followed by 16 transformed videos associated with that source video. The 16 transformed videos consisted of several sets, each of which focused on one appearance characteristic that varied within each set:

- age (3 videos; based on a young, middle, and older-aged target face),
- artificially colored hair (3 videos; blue, pink, and green colored hair),
- natural-colored hair (3 videos; light, medium, and dark shades),
- with-torso (2 videos with the torso visible – all the others had the torso removed), and
- tiger-face (1 video shown with an animated cartoon tiger face, as used in [24]).

The order of these sets was counterbalanced between participants, and whether male or female videos were shown first was also counterbalanced. After the first with-torso video was shown, the re-

searcher on the video call interrupted the participant to ask the participant to indicate agreement with each of three Likert items, "This video was completely understandable," "This video was very natural in appearance," and "This video disguised the identity of the original signer completely." Similarly, as soon as the first without-torso video was shown, and immediately after the tiger-face was shown, the participant was asked these same three questions. After the participant viewed all videos in this phase, semi-structured open-ended interview questions were asked about the overall understandability, naturalness, and anonymity-protection of the transformations.

11.3.5 Phase 3 of the Study

In the third phase, participants saw a video of themselves transformed using all three prototypes so that we could evaluate their view of **how acceptable this technology is for disguising their own videos**. While the tiger-face prototype could run in real time, the with-torso and without-torso prototypes required additional processing time. Thus, prior to the appointment, we asked participants to submit a video of themselves signing a short ASL passage. Because of limitations in the anonymization prototype and in order to ensure good-quality output, participants were instructed to make sure they had good lighting and a plain background, and they were asked to pull shoulder-length or longer hair back in a ponytail. Participants were also asked to remove any glasses, headgear, and hand jewelry. Lastly, participants were asked to sign in a manner that avoids having their hands obstruct their face, as the prototype system is not robust to face occlusions. For this reason, signers were given an ASL gloss script for a specific passage to perform that excluded signs in which the hands would come close to the face, while also requiring the grammatical use of several facial expressions in ASL: "BOOK, I BUY. TODAY, YOU BORROW. BOOK, READ YOU? BOOK WHERE?"

During phase 3 of the appointment, participants viewed 13 transformed videos, based on the video they had submitted. Six were with-torso, with another six without-torso, using the same set of target faces. The target-face set was matched to the participants' self-reported gender and apparent race/ethnicity in their submitted videos. The 13th video was a live demo website with the tiger-face effect, which participants were instructed on how to use.

After viewing all videos, participants responded to open-ended questions about their perception of and preference among the videos, whether they thought the quality of these videos was good enough for them to consider using software like this, and whether it would be helpful for them to have software that could anonymize videos. Finally, participants were asked what situations they would or would not use this software for, whether they thought it would be acceptable for other people to use software like this, and whether they had concerns about software like this.

11.3.6 Participants

Via social-media postings, we recruited 16 DHH adults who use ASL on a daily basis; 12 indicated that ASL was their primary language. Four participants had used ASL since birth, 6 learned ASL by age 5, and 6 learned ASL during their late teens (with all in this latter group having used ASL for at least 8 years). Participants' ages ranged from 19 to 47 years old (median 27.5). Eight self-identified as male, 1 as non-binary, and 7 as female. Participants' education levels varied: 1 had some undergraduate education, 1 had an associate's degree, 10 had a bachelor's degree, and 4 had a master's degree. Eight self-identified as Caucasian, 1 as Black, 3 as Asian, 1 as Vietnamese, 1 as Latino, 1 as Asian & Hispanic, and 1 as Spanish & Native American. Table 11.1 shows basic participant demographics.

11.3.7 Data Analysis

All data collected from the three phases of studies were analyzed with both quantitative and qualitative approaches. We conducted statistical analysis with Friedman tests on the quantitative data, and we performed an iterative thematic analysis [28] on our qualitative data, employing both deductive and inductive approaches. We manually developed a deductive coding framework with the main topics of our interview questions. In the framework, we aggregated all the data and iteratively performed open coding using colors. Then codes were generated with the color-coded data and organized with categorization. Finally, main and sub-themes were identified and developed using a bottom-up approach. We went through the same process with the data from all three phases of the study.

ID	Gender	Age	Ethnicity	ASL Primary Language?	Age learn sign?	Occupation	Education	Area of Study
1	F	23	Spanish, White, Native American	No	5	Student	BS	Science
2	M	28	Latino	No	17	University Disability Office	BFA	Design
3	M	23	White/Caucasian	Yes	5	Student	MS	Technology
4	F	30	Black	Yes	3	Academic Counsellor	MS	Social Science
5	M	47	White	Yes	0	University Admissions	MS	Social Science
6	M	26	Asian	Yes	7	None	BS	Computing
7	F	25	White/Caucasian	Yes	0	Engineer	BS	Technology
8	M	24	Vietnamese	Yes	12	Student	BS	Design
9	NB	26	Asian	No	18	Student	MS	Business
10	F	26	White/Caucasian	Yes	0	Administrative Assistant	BS	Computing
11	F	27	Asian	Yes	5-7	Student	AS,BS	Science
12	M	25	White/Caucasian	No	19	Tool Maker	AS	Technology
13	M	19	White/Caucasian	Yes	1-1.5	Student	BS	Business
14	F	23	Asian, Hispanic	Yes	0	Student	BS	Social Science
15	F	26	White/Caucasian	Yes	17	Lab Technician	BS	Social Science
16	M	27	White/Caucasian	Yes	3-5	None	BFA	Design

Table 11.1: Demographics for ASL Anonymization Study

11.4 Findings

To investigate the usefulness of the anonymized ASL video, we compared three prototypes (with-torso, without-torso, tiger-face) along three evaluation dimensions: understandability, naturalness, and anonymity. During the study we had collected some quantitative data, e.g., participants' Likert response to questions in phase 2 about each of these dimensions. Our quantitative analysis consisted of conducting Friedman tests, which indicated a statistical significance in understandability and naturalness among the three types of transformations, but no statistical significance for anonymity-protection. Following up with pairwise Wilcoxon signed-rank tests, significant differences among the types of transformations were identified. In our qualitative analysis, we found that the participants overall perceive the video transformation as interesting and useful. However we observed differing perspectives among participants in regard to how they compare these three prototypes along the three dimensions, as well as how this affects their overall views on the ASL video anonymization and its value. We present the details of the findings in the following sections.

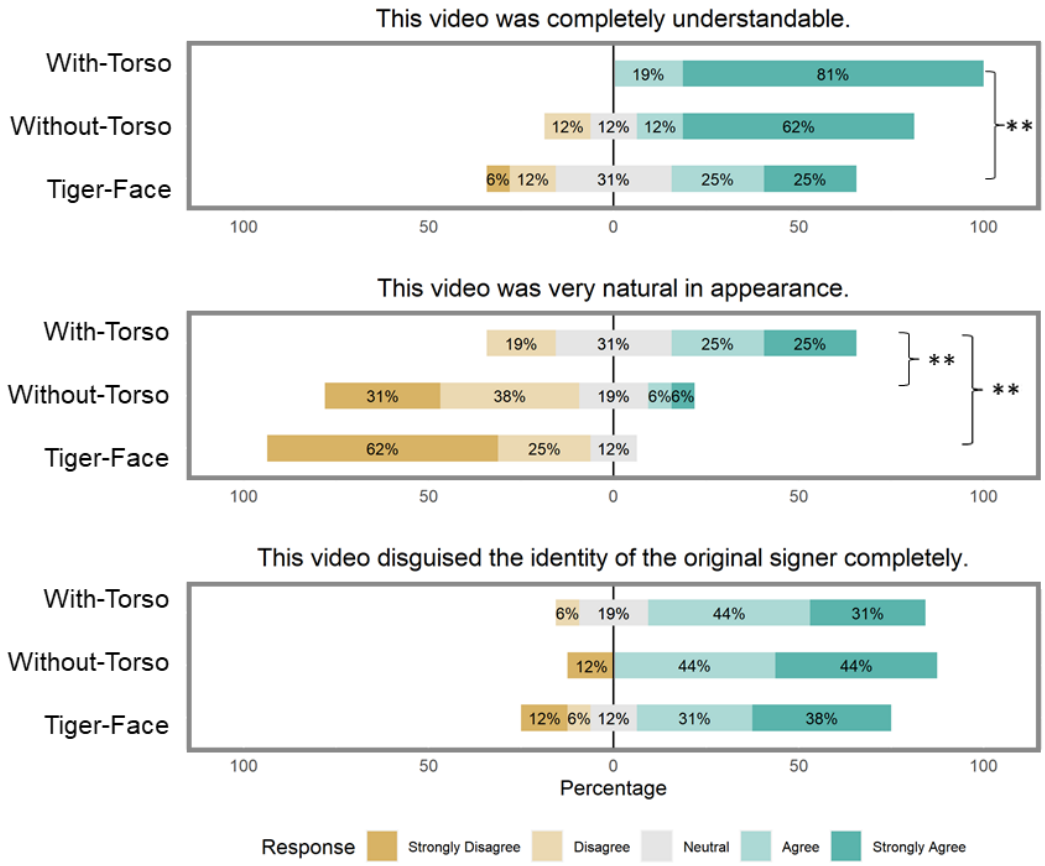


Figure 11.3: Participants' agreement with Likert items in phase 2 of the study, for each of the 3 prototypes.

11.4.1 Understandability

Quantitative Analysis

Figure 11.3 displays participants' responses during phase 2 of the study to the Likert item "This video was completely understandable." Analysis with a Friedman test revealed that the type of video transformation had a significant main effect on understandability ($p < .05$). Overall, 81% of respondents strongly agreed with this statement in regard to the with-torso videos, 62% of respondents strongly agreed in regard to without-torso videos, but only 25% agreed in regard to tiger-face videos. Post hoc pairwise analysis with a Wilcoxon signed-rank test and Bonferroni correction revealed that **participants believed the with-torso videos were more understandable than the tiger-face videos** ($p<0.01$). However, no

significant difference was observed between with-torso and without-torso, nor between without-torso and tiger-face. Overall, these quantitative findings indicate that ASL signers believed that the modern 3D face transformation videos with a torso displayed (with-torso) were more understandable than the simple mask-overlay videos (tiger-face), when viewing videos of ASL.

Qualitative Analysis

The overall feedback in regard to the understandability of the anonymized videos of all three prototypes was **generally positive**, which aligned with the quantitative findings presented above. Most participants indicated that the transformed videos were clear and conveyed the same information as the original videos. Among the three transformations, participants agreed that the with-torso version was easiest to understand, but they believed the without-torso version was still relatively understandable, although somewhat less so.

Participants commented that even though the hand movements were intact in both the with-torso and without-torso videos, they felt that **understandability was reduced when the torso was cut out**. P9 said *"When you take out the torso, it was harder to understand."* P13 provided the reason: *"Without the body it was hard to detect the body language."* P14 implied that the removal of the torso may have interfered with their ability to focus on the message, and consequently to understand: *"It was distracting to have no body."*

Although participants said that seeing the body was useful, they emphasized that **facial expression was most important for understandability**. While both the tiger-face and with-torso videos retained the signer's original body appearance, all but one participant indicated that the tiger-face was the least understandable, because of the absence of the facial expression. P6 and P8 said respectively: *"Tiger face did not really bring the same information, the facial expressions were lost in that video"* and *"For tiger face, there are no facial expressions, feels weird."* Some participants described how the tiger-face itself was distracting. P13 said *"with tiger face, it was not very clear because it kind of blocked the signing because the face was big."* P5 expressed that while for a short video, the tiger-face animation could be understood with great effort, that *"if the tiger face was longer video then I might not understand as much."* Beyond

the ASL linguistic information on the face, P5 described how the face conveys other information: *"The message is the same but because I see different faces - one person looks stoned and other person looks like messy hair - so I interpret things differently - same messenger but different feeling."*

11.4.2 Naturalness

Quantitative Analysis

Participants' responses during phase 2 of the study to the Likert item "This video was very natural in appearance" are shown in fig. 11.3. Quantitative analysis with a Friedman test revealed a significant ($p<0.01$) main effect of the type of video transformation on participants' rating of naturalness. Post hoc pairwise testing with a Wilcoxon signed-rank test, with Bonferroni correction, revealed that participants believed that the with-torso videos were significantly more natural ($p < .01$) than both the without-torso or tiger-face videos. Post hoc testing did not reveal any pairwise significant difference between without-torso and tiger-face videos. Overall, these findings indicate that **participants believed the with-torso videos were the most natural in appearance**.

Qualitative Analysis

Participants commented that **none of the three prototypes was completely natural**. In alignment with the quantitative findings above, most agreed that without-torso videos were less natural than with-torso videos, e.g., P14 said *"With torso was better because it looks more natural."* P13 agreed: *"With body was better because you can see the whole body as natural."* Participants indicated that they disliked the gray background color of the without-torso videos, and they also commented on there being visual "noise" at times, e.g., a flickering effect due to video-transformation artifacts. A few participants expressed concerns about insufficient facial expression in the disguised video, inadequate skin color match (between the face and the neck/arms), or unnatural hair color.

All but two participants commented that the **tiger-face videos were most unnatural** and explained that this was because no human face was visible. Among those with the minority opinion that tiger-face videos were more natural than without-torso videos, P9 explained that *"with the tiger face, you still see*

the body, body language, you can see the body shape. The face does not look natural but the rest of body was natural." P1 was alone in believing that the tiger-face was the most natural of the three prototypes, saying *"because it was slightly believable while the others weren't."*

11.4.3 Anonymity

Quantitative Analysis

An analysis of the Likert response data in phase 2 in regard to the statement "This video disguised the identity of the original signer completely" did not reveal any significant effect of the type of video transformation on participants' response to this item, as shown in fig. 11.3. However, participants' high level of agreement on this item for all three prototypes suggests that they were **all seen as effective at disguising identity**. The vast majority of responses to this question were strongly agree or agree: 75% of respondents for the with-torso videos, 88% of respondents for the without-torso videos, and 69% of participants for the tiger-face videos.

After viewing a disguised video in phase 1, participants were asked to guess the original human face from a photo line-up, including other faces we had selected with similar race, gender, and hair style. From the male line-up, 9 of 16 participants were able to guess the original signer correctly. From the female line-up, 4 of 16 participants guessed the original signer correctly. This finding provides some evidence for **the potential of face swapping for anonymizing ASL videos**.

Qualitative Analysis

In addition to being asked in phase 1 to guess the identity of the signer in the line-up of photos, participants were also asked to comment on how difficult it was for them to do so. Almost all participants agreed that it was hard to guess the original signer from the transformed videos of both with-torso and without-torso versions. Participants explained that they could not use the face as a clue, but **they made use of used other appearance details, such as head shapes or skin color**. P14 described how he tried to approach this task, explaining that he tried to *"remember the skin color based on the arms."* P13 used a similar strategy, explaining that they made their selection *"based on the color of skin except for*

face."

There was consensus among participants that the with-torso version would make it easier to identify someone from their clothes, background, or body shape, **especially if the signer was a friend or family member**. Overall, participants believed that the without-torso transformation would be most effective for anonymization. As P4 explained, *"Knowing the person and seeing their torso and background would make it easier to identify them because the more you hang out with the person you know their body language and how they sign."* P14 agreed that the without-torso videos had the greatest anonymity protection: *"Without torso is the best, sometimes you can identify people by the body shape, etc, but without seeing the body it is very difficult to guess despite that it might be harder to understand or not natural."*

Some participants believed that the tiger-face videos were most effective at disguising the face, **which is simply blocked, without any facial expressions revealed**. However, P13 explained that there are trade-offs between the ability of some prototypes to disguise the face or to disguise the body. As P13 explained, *"Without torso is the best. It covers the face and also hides the body language. You can't look at the body shape, size, etc. For tiger face, it hides the face the best but it doesn't hide the body at all. Without-torso has the best balance at hiding body but keeping facial expression."*

While participants agreed that without-torso videos were most effective at preserving anonymity, all participants commented that they would prefer to view a video with a torso – because of naturalness and understandability, as discussed previously. Several commented that it would be useful if this technology could make **modifications to the body of the signer instead of removing it**, e.g., suggesting that the tool could change the signer's clothing.

11.4.4 Preferences for Transforming Specific Characteristics

Throughout the study, participants viewed disguised videos of both themselves and other people, with a variety of characteristics transformed, e.g., age, hair color. Many participants indicated their preferences for video transformations **that closely matched their own traits**, such as race, age, hair, and skin color. P6's comment conveys this clearly: *"What I liked was all the faces using the same race or traits as me...I liked the faces that looked similar to my face.*" Participants also emphasized the importance of having the

transformed video match their own age. For instance, in response to a question about which of their own transformed videos of was their favorite, P7 answered: *"4th video. It was similar age, and looked the most natural to me."* P7 went on to emphasize the importance of the age feature for the natural appearance of the video: *"I would use a similar age, but I don't care about the other features as long as it doesn't look way off or too unnatural.*" In the same vein, P1 expressed unhappiness with a transformed video with an older looking face, indicating that it was the least favorite video, and commenting, *"I didn't like that you made me old, I didn't like the age change."*

What mattered most to participants was whether **changing specific characteristics reduced the naturalness** of the resulting video; participants generally disliked transformations that resulted in artificial-looking hair color or the tiger face. In fact, all but one participant disliked having their hair transformed into bright colors. For instance, P2 commented, *"I didn't like using the different hair colors like purple hair was strange.*, and P9 added, *"it was funny to see my hair color look different."* Similarly, all but one participant disliked having their video transformed into the tiger face – with participants commenting on its artificial appearance and oversized head.

To a lesser degree, participants preferred transformations of characteristics that **supported understandability**. For instance, P12 noticed that some transformations preserved facial expression more clearly than others: *"it was easier to understand the younger faces than the older faces because I could see their mouth move."* Participants also mentioned that some transformations led to a **distracting result**, which interfered with their visual focus and thus their understanding. For example, P13 said, *"with tiger face, it was not very clear because it kind of blocked the signing, the face was big, and was distracting."* For the without-torso version, P14 commented, *"It was distracting to have no body".* P4 disliked brightly colored hair, explaining: *"it seemed distracting for me.*"

Participants' preference among most transformations **did not depend upon whether it was applied to videos of themselves or of others**, with one exception: the removal of the torso from a video. Before participants viewed their own transformed videos, all of participants favored the with-torso videos, commenting on the natural appearance. Participants tended to retain this preference until they viewed their own transformed videos during phase 3, at which time half of the participants switched their

preference from with-torso to without-torso, because they were worried that identifiable characteristics were visible in their with-torso videos.

In fact, upon seeing transformed videos of themselves, some participants not only became interested in the without-torso feature, but they also wondered how they could strategically transform **as many demographic and appearance characteristics as possible**, to protect their identity. For instance, P13 suggested: *"for improving anonymization, I would use a neutral color skin on the arms, neck, etc. And doesn't have to match gender, you could use neutral gender or opposite gender instead of having to match."* However, some participants noted that using this technology to change the skin color of one's face could produce offensive or insensitive results, with P9 musing, *"There could be a few issues with race..."*

Finally, participants believed that the appropriateness of specific appearance transformations would **depend upon the context of use**, as some situations required more anonymity or seriousness. P8 said, *"Doesn't matter to me which appearance, it's more about how serious I want to be when hiding my anonymity. If I wanted to hide, as is, I would pick without torso, doesn't really matter what hair color/age.*" P5 indicated that *"If its formal, then it needs to look real/natural. Suppose Biden was presenting with a funny tiger face then I would be more resistant to watching while if it was comedian using it then I would understand. I think context is important.*"

11.4.5 Potential Uses

Participants identified a variety of possible use cases for ASL video anonymization technology. In particular, nearly all participants agreed that the technology would be useful for safely expressing **personal views on sensitive or confidential topics**. P3 was interested in using it *"to avoid being targeted, want it to be anonymous. Some people might want to share important information but don't want to tie it to their identity.*" P7 wanted to us this technology to *"post videos where I say things that I don't want associated with my identity. For example, political, abuse reports, protests, etc.*" P10 was interested in using this tool to *"share my personal experience or feelings and I didn't want people to know who I am.*" P2 explained they would use if for a "sticky question. If I was telling a powerful, heavy topic but wanted my identity hidden then I would use this. Mostly for sensitive topics." The ASL sign STICKY, used by P2 in their

response, translates to the English concepts of awkward or embarrassing.

Participants also identified uses of this technology on social media, especially when they needed to share information that may be re-shared **beyond their own immediate personal network**, especially when ASL video would be more effective than text. For instance, P13 discussed sharing ASL lessons anonymously: *"I would use it for posting videos that strangers have access to, teaching ASL without revealing my face.*" P11 discussed social media contexts in which protection of privacy is especially important, e.g., *"social media, OnlyFans, anonymous groups, etc.*." Participants also discussed uses for this technology on personal social media contexts during **fun or casual interaction with people they know**. P8 was interested in *"entertainment with friends and family, like the gaming community.*" Similarly, P9 indicated that the transformed videos themselves may be entertaining or fun to share, explaining, *"I would also use it for entertainment...with friends, assuming they would not share it publicly.*"

Finally, participants described contexts in which they would **not use this technology**, at times disagreeing with uses suggested by other participants. For example, several participants saw no use for this technology when interacting online with family or friends. As P13 explained, *"I would not want to use this if I was just talking with friends or people I know and trust.*" P14 agreed and extended this to fellow students: *"If I was signing on my social media or with friends or schoolwork, I wouldn't use it.*" Most broadly, P2 felt strongly that they *"wouldn't use this software for any other purpose that I am not trying to hide my identity like having fun, etc.*"

11.4.6 Concerns

Participants indicated that they would find it acceptable to watch a video from someone that has been disguised by this technology, **as long as there was an ethical purpose**, e.g., if anonymity was needed in order to share important information or ask sensitive questions. However, participants expressed concerns about use of this technology for unethical purposes, such as harassment, trolling, or degrading someone online – as well as someone using this technology to steal information or to impersonate someone for fraudulent purposes. P5 expressed this trade-off: *"That's ethics. I wish it was safe for everyone to express their thoughts and concerns without being identified; however, this could be misused*

[or] abused so there needs to be a set of rules."

Participants were especially concerned about the potential for this technology to enable someone to **impersonate another real person**. Although participants liked the concept of having a natural face, different from their own, appear in a disguised video, a majority of participants shared concerns about using another real person's face. P8 was worried about this technology producing videos using the faces of their family or friends, explaining, *"I don't want another person to intentionally disguise themselves as another person that I know. It could also be used for a scam or something. The faces should be fake and not from real people."* P1 suggested that a computer-generated virtual avatar face could be used instead of a photo of a real person, explaining *"also [I] don't understand why not using avatar - I would use my own avatar.*" P1 wondered whose faces had been used as the basis of transformation, and she was worried about someone impersonating her, e.g., asking *"Who are the faces they are using? ... Are you using my face to hide other?*" The researcher explained that no face images of participants in the study were used to disguise the face of others. The faces were from a public research dataset [106, 107].

Finally, participants indicated that seeing someone use this technology may lead to **feelings of distrust**, as they may wonder about someone's motives for hiding their identity. For instance P1 indicated that upon seeing a video of someone that has been disguised, she would want to ask that person *"Why you feel the need to hide you face that much?*"

11.5 Discussion

Our findings revealed trade-offs between key dimensions of importance to users, including: understandability, naturalness, and anonymity. In this section, we discuss pairs of these dimensions to inform the design of ASL video de-identification technology, and we examine factors that affect the acceptability of this technology for users.

11.5.1 Understandability vs. Anonymity: Design Considerations

Participants mentioned that facial expression and the movement and location of the body were important for preserving meaning in ASL, and transformations in which the signer's original torso was retained

were rated as most understandable. However, there was a **tension** between greater understandability and participants' perceived degree of anonymity protection. Participants noted that with-torso videos revealed visual clues about the identity of the signer: their individual style of body movement, their clothing appearance, and other physiological traits, such as body size. Given that participants liked the understandability of with-torso videos, they were interested in improving anonymity protection without removing the torso completely. For instance, participants suggested that it would be valuable to extend this transformation so that, rather than hiding the body of the signer, the technology could apply some disguise to the body, while preserving its location and movement. As previously mentioned, some participants suggested virtually changing the body appearance or the clothes of the signer.

Our findings revealed that participants viewed the without-torso and tiger-face prototypes as being relatively similar in their degree of anonymity protection, which was striking given that these two tended to occlude or omit opposite portions of the signer's body. That is, the tiger-face blocked the signer's face—whereas, the without-torso videos omitted the signer's torso while conveying the facial expression information on a transformed face. Our qualitative findings revealed that participants judged the without-torso videos as more understandable; thus, **occlusion of the face led to a relatively greater reduction in understandability, for a similar anonymity improvement**. Recent work by Bragg et al. [24] had investigated ASL video anonymization within the context of motivating users to contribute videos to public research datasets; their participants had used the same tiger-face filter and had similar concerns about the negative effect on understandability of the absence of facial expressions.

For designers creating face-transformation applications, sensitivity to this understandability vs. anonymity trade-off is essential. While it would be ideal for the underlying transformation technology to achieve both high understandability and high anonymity (perhaps as further advances in face and body modification technology are created), in the meantime designers might consider **offering users choices** in transformation options that vary along this trade-off axis. For evaluation of these applications in studies, it is important for **both properties to be measured**, in relation to intended use cases, to avoid optimizing for one at the expense of the other.

11.5.2 Naturalness vs. Anonymity: Design Considerations

Participants indicated that it was important for videos to appear natural; however, our analysis revealed that there was a **trade-off** between naturalness and anonymity protection. Unanimously, our participants indicated that the with-torso videos were the most natural, yet these videos had weaker anonymity protection, as details of the signer's body and background were visible. In contrast, our qualitative analysis revealed that participants believed the without-torso and tiger-face videos were better at protecting anonymity, yet both of these had much lower levels of naturalness, due to the unfamiliar appearance of the torso being cut out of the video or the artificial animal face.

For individuals interested in disguising themselves, a decision must be made about where on this naturalness vs. anonymity trade-off the user would prefer for their video to be. This decision may depend upon the context of use, and designers creating face-disguise applications may wish to **provide users with options** that vary along this axis.

11.5.3 Understandability vs. Naturalness: Design Considerations

Whereas the discussion above identified trade-offs between naturalness vs. anonymity and understandability vs. anonymity, our findings revealed a **complementary relationship** between understandability and naturalness. Participants discussed how improvements in naturalness led to increased understandability, explaining that unnatural appearance could be distracting, which would draw attention away from the message. For designers of transformation technologies for face disguise applications, this relationship is important to consider when making improvements to the technology. In efforts to achieve increases in the understandability of the resulting video, it is important to ensure a baseline level of naturalness, to avoid interfering with the viewer's ability to focus on the message.

While there are relationships among these factors, the signer's intended usage of this technology is likely to influence how these factors are prioritized. Before seeing transformed videos of themselves, participants focused on the **perspective of people viewing videos** of other people who have been disguised, and understandability was seen as being of greater importance so that the message could to be understood. After seeing videos of themselves transformed, they shifted to the **perspective of someone**

who is transforming their own video and became concerned with how they present themselves online, prioritizing naturalness to a greater degree. For instance, we reported that P1 disliked having been transformed into an older face, and this participant later explained: "*I would use this for situations when I want to look nice... In Snapchat you have the filters and you look better than normal, while this technology makes you look worse than normal. If it helps people look better than normal then it would be accepted.*"

Our findings also inform how future designers or researchers investigating this technology should **design studies** to gather requirements from both perspectives. We found that simply asking participants to imagine using this technology to transform their own video was ineffective. Actually seeing their own videos transformed was what had sparked participants to re-prioritize their preferences and requirements for this technology.

11.5.4 Role of Identity in Acceptability: Design Consideration

Given the degree to which face appearance is considered a unique identifying characteristic of individuals, when participants first saw transformed videos of themselves, many expressed **mixed feelings**. Participants were struck with how well the technology had preserved their anonymity, to the point that many did not realize that they were viewing a transformed video of themselves. As P5 said, *"That was me? I didn't realize it was me. It was really interesting because I was watching I was looking for something... I recognizing the shirt...Now that I know it was me, I don't like it.*" In addition, many participants expressed discomfort when first seeing another person's face on their body, as P16 explained, *"I was a little shocked to see the faces changed, huge difference.*"

Our findings also revealed that the acceptability of specific transformations was dependent on a participant's concept of the **individual characteristics of their own identity**—and whether the technology had preserved, hidden, or transformed each. For instance, we mentioned earlier that half of our participants changed their preference from with-torso to without-torso videos upon seeing the first transformed video of themselves. They noticed personal traits on their body that were visible, e.g., ring on finger, clothes, nail polish. Beyond this risk to their anonymity, participants disliked partially transformed videos, as the unnatural appearance made the result appear fake. They were more comfortable

when **all or none** of their identity characteristics were disguised or hidden.

When viewing a transformed video of another person, participants preferred for the video to retain as many characteristics of the original signer as possible. However, when considering how to transform a video of themselves, participants saw two sides to this issue: If the characteristics of the disguised face were similar to their real appearance, then they could **convey individual elements of their identity** when transmitting their message. Knowing the gender, age, or race of the person who had produced a message may be important context. On the other hand, selecting characteristics that differ from their real identity could provide a better disguise, thereby protecting anonymity better.

While several participants expressed interest in being able to transform their face into that of someone of a different **race or gender**, we did not enable this option in our study. Our rationale was that the current version of our prototype was limited to changing the face of the signer—not the skin color on the neck, arms, or body. Because of ethical concerns, we did not display videos in this study in which a face was overlaid on a body of someone of a different race, to avoid producing videos that may be insensitive or offensive. Future designers of face-transformation technology may need to address this desire for users to be able to replace their face with characteristics unlike their own, while providing guidance for users about ethical use of this technology.

The concept of identity was at the heart of many participants' ethical concerns over potential misuse of this technology described in section 4.6. Using real people's faces could lead to identity theft or impersonation that damages someone's reputation. Participants were concerned about someone using their face in this manner. To avoid such misuse, future designers of this technology could display a a disclaimer on the video to indicate that it has been transformed—or rather than using real faces as the target result, videos could be a composite/hybrid of the source and target face, and through this combination thereby producing a **novel identity** for the individual appearing in the resulting video.

11.6 Conclusion

DHH ASL signers who disprefer communication through written English must use video of ASL to communicate online, thereby revealing their face, which conveys essential linguistic information. These

users currently lack effective options for communicating anonymously online, which prevents them from discussing sensitive topics or other activities. New advances in face transformation technologies enable replacing faces in video at a level of quality that preserves linguistic facial expressions and head movements essential for ASL. We conducted an **interview study with 16 DHH ASL signers**, who viewed ASL videos of themselves and others transformed by prototypes for disguising the face.

Our study evaluated three key dimensions of acceptability (understandability, naturalness, and anonymity protection), and quantitative and qualitative analysis of our findings revealed relationships among these dimensions. Our findings revealed that a prototype based on modern face-transformation technology was effective for preserving anonymity, and we **contribute empirical knowledge** about participants' assessment of this technology, preferences among various appearance transformation options, factors affecting the acceptability of this technology, uses of interest, and potential ethical concerns with this technology. Our study provides guidance for both designers of face-disguise applications and creators of anonymization technology for providing DHH ASL signers with new options for participating online.

11.7 Limitations and Future Work

Future users of this technology may apply it to videos intended for sharing online, which they may record under various camera set-ups or environments. A limitation of our study is that participants generally produced videos at home in front of a computer, in a setting typical of a video-conference, because of the ongoing COVID-19 pandemic. Future research should investigate a **more diverse range of videos**, produced under a variety of real use-cases, to determine both the performance (understandability, naturalness, and anonymity) of this technology, and whether users' requirements or preferences are influenced by these factors. Further, while participants in our study had the experience of seeing their own video transformed, they did not have the experience of **actually posting that video online** to share with various audiences. A future study could investigate this technology as used in a more realistic context, which could reveal social factors that affect users' acceptance of such technology or preferences for it should be designed.

Another limitation is that the participants we recruited do not reflect the **full diversity of potential users** of this technology, which may include ASL signers who vary in age, technology experience, cultural background, or ASL fluency—as compared to the specific participants in our study, which disproportionately included recent university graduates, with a narrower range of demographic characteristics and life experience. Future work should investigate a wider range of potential users' interests and requirements in relation to this technology.

Section 3.1 explained our rationale for selecting the prototypes and transformations examined in our study, but future research is needed to explore a **wider range of design alternatives**, to understand more of this design space. In addition, we had selected the specific set of transformations applied to each participant's video in phase 3 of our study, but future work could investigate which transformation options participants would choose for themselves, e.g., if they were provided with an interface that enabled them to select among such options. Such a study may provide further insight as to how DHH signers may balance trade-offs among anonymity, understandability, and naturalness.

Recent work has investigated applications of body-swap illusions in virtual reality [132], with users' new appearance leading to **changes in behavior** [182]. Our study did not examine whether signers might change their signing content or style if they were to see their own face transformed in real time; future work is needed to investigate this.

The with-torso and without-torso prototypes in our study were based on modern face transformation technologies, of which the state of the art is rapidly advancing. Future research is needed to understand users' perspectives of these technologies as they **improve over time**. In fact, our work should inform the work of future designers of such technology and of researchers creating the underlying disguise technologies, as we discussed in section 5. In particular, our research has motivated future work on technology for disguising not only the face of a signer but also their body—to better protect anonymity—while also preserving body location and movement, which contribute to the understandability and naturalness of the resulting video.

Finally, there is a need for future research to consider the use of this technology in specific contexts such as with personal assistant technologies that collect video for transmission over the Internet to a

company that processes the data for sign language recognition. For instance, would DHH users be more comfortable with using a personal assistant if it had such anonymization technology available? It is unknown if the existence of this anonymization technology would inspire novel uses of personal assistants by DHH users now that they can use the device to anonymize ASL videos to share.

Chapter 12

Privacy Concerns During ASL Interaction with Personal Assistants

12.1 Introduction

As discussed throughout this dissertation research, future personal assistant devices may use their integrated cameras to capture an ASL signer and process the video for sign recognition. Since the user may be moving around inside their home, the device would be capturing them in various places, potentially "following" the user around or capturing them with a wide-angle camera system. Since these videos would be transmitted across the Internet to a server for processing, there are privacy concerns in the context of using a personal assistant. In the case that there are other people in the household, e.g., guests who are unaware that there is a personal assistant device watching them with a camera, this raises even more privacy concerns. This need to consider users' privacy concerns becomes increasingly important, as digital personal assistants are becoming more ubiquitous.

As section 10.6 discussed, the issue of privacy has become apparent through previous work presented in this dissertation, and we have learned about the various concerns users have around personal assistant technologies. The previous chapter (chapter 11) has shown potential for existing state-of-the-art face transformation technology to effectively preserve anonymity, and described different factors

affecting the acceptability of this technology. However, within the context of personal assistants that can understand ASL, prior work has not investigated whether users would be interested in modern face-disguise technology to be integrated into their interactions with personal assistants to preserve privacy.

12.2 Research Question

In this chapter, I appended a small interview study to the in-person Wizard-of-Oz experiment described in chapter 10, with the following research question:

> **RQ8:** Would using a state-of-the-art face-disguise technology to anonymize DHH users' ASL recordings (that are used for device processing) before they are processed by a personal assistant device alleviate privacy concerns?

12.3 Materials and Procedure

Focusing on the context of interacting with a personal assistant in-person, chapter 10 describes an experiment in which a DHH user would be conducting typical activities in a home-like living room and kitchen area while interacting with a personal assistant device that appears to understand ASL. After this phase, the main interview questions asked about the user's thoughts, reactions, and general experience during the session. At the end of this interview, I appended questions to focus on the issue of privacy; the researcher-copy is as follows:

- Do you have any privacy concerns with a device like this? Since it uses the camera to understand ASL, would you be OK with it "watching" you the whole time? How would you solve this problem? What would you say or do to Alexa if you were concerned?
- How would you resolve this? (demo opportunity – participants can demonstrate while they are in the living room or kitchen area, as the researcher takes notes)

- Are there any specific types of conversations or ASL usage in the home that you really don't want a device to capture?
- Here is an example of a technology that post-processed an ASL video to try to protect the anonymity of the user while keeping the ASL video understandable and natural in appearance. [Showing participant videos from the ASL anonymization study (chapter 11)] If software like this disguised your face before it was sent over the Internet to the company for processing, does this change your level of comfort with such a device in your home?
- If you could use this anonymization technology with the personal assistant, are there any new commands or uses you would do with the device? (For example, using the device to post an anonymous video on social media platforms online)
- Please rate how strongly you agree or disagree with the following statements (1=strongly disagree, and 5=strongly agree):

    1. From a privacy perspective, I would be concerned about having a device with a camera.
    2. The device might pick up on some signs that were not meant for it.
    3. It is important to have an option of turning off the microphone sometimes for privacy.
    4. It is important to have the option of turning off the camera sometimes for privacy.
    5. It is important to have a physical cover to block the camera sometimes for privacy.

12.4 Results

The interview responses were qualitatively analyzed by coding and labeling points, which were used to generate themes. These themes were then iteratively compared and clustered to yield key patterns and themes. The first interview question asked participants if they had any privacy concerns about a device like this that used a camera to understand ASL. Participants gave mixed responses, some citing that they were concerned because they don't know what kind of data Alexa collects, and that they don't know what happens to the data once it is captured. P7 commented "*Yes I definitely have concerns with privacy and Alexa. Alexa is watching me the whole time which is a privacy issue.*" P1 said "*I might be uncomfortable*

to have people watching, knowing that this was connected with Amazon means many people can watch." Others said that they were not too concerned, and suggested that they would simply cover the camera if need be. P2 brought up "privacy laws": "*I'm not too concerned with Alexa because there are privacy laws that should help protect my rights.*" P6 didn't have any concern, saying "*I do not care about having the Alexa video me, I don't feel like its 'watching me'.*" P8 brought up having a double-checking confirmation: "*No, I am not too concerned. It's important that the camera can have a cover. I think it should be fine and then there could be a setting to turn the camera on as long as you click 'OK' and are ready for Alexa to see the commands.*" Some went on to comment that a camera cover would be inconvenient because they would have to uncover it every time they wanted to use the device. P5 said that "*it's weird that it is watching all the time, I am concerned with privacy and would consider using a cover. On a lot of new laptops they have a cover that can lock over the camera, so that could be an option. However, having to move the camera cover each time you want to sign would be annoying. Maybe have the camera angle/view limited so you have to walk into a certain area to be seen. You also need a high security firewall. I don't know what data Alexa collects and am concerned about that aspect.*"

Participants explained that, if they are talking about sensitive information such as banking, work, personal details, social security, etc. (P7), then they would not want Alexa to capture any of that. Others generalized that if they are not talking directly to Alexa, then it should not be watching (P5,10). Some brought up specific rooms where they have concerns, with P11 saying "*It depends. If it's in the living room it is fine. But if it is in the bathroom or bedroom, then maybe not. But in common areas I am not concerned.*" Several participants explicitly suggested covering the camera or turning Alexa off as a reliable (but not necessarily convenient) solution (P1,2,3,4,8,11). There were also participants who were not concerned, such as P6, who said: "*I don't feel like I'd care if Alexa 'saw' any conversations in my home. I am comfortable with Alexa and the camera feature.*"

When participants were shown the ASL anonymization prototypes from chapter 11, with explanation about how that approach could be used to disguise their face before transmitting their video to the remote server that processes their command, several participants said that it would not impact their comfort level with the product (P3,4,5,10). Some others said that they would use it, but gave feedback

on appearance factors such as skin, background color, contrast, and lighting (P1,5,7). When asked if they would personally use the anonymization technology, most participants responded "maybe" and said it depends on the situation and the quality of the prototype (e.g. whether the signing would be clear and understandable).

The responses for the five likert-scale statements were aggregated and a summary is shown in fig. 12.1. In chapter 3, 86 survey participants responded to the same five statements, and the results are shown in fig. 3.2. These survey participants were asked to imagine having such a personal assistant device that can understand ASL, whereas the participants in this study actually interacted with a Wizard-of-Oz prototype (however the sample size was much smaller in this experiment). In both studies, a large majority agreed or strongly agreed with four of the statements, showing they believe it is important. For the statements "It is important to have the option of turning off the camera sometimes for privacy" and "It is important to have an option of turning off the microphone sometimes for privacy" had, respectively, 91% and 82% strongly agree responses. For the same statements, 70% and 63% of the online survey participants strongly agreed.

For the statement "From a privacy perspective, I would be concerned about having a device with a camera", 45% of the in-person participants were neutral, while 71% of the online survey participants either agreed or strongly agreed. A Mann-Whitney U Test was conducted to compare the likert responses from the online survey and the in-person Wizard-of-Oz experiment, and yielded an insignificant p-value of above .3 for the other four statements (.50, .31, .35, and .92). However, for this statement about being "concerned about having a device with a camera", the p value was .059, much closer to being a significant difference.

12.5 Discussion and Future Work

We received mixed responses about the application of the anonymization prototypes developed in chapter 11. Participants did not seem to have strong feelings about the ASL anonymization prototypes, as they mostly said it would not impact their comfort level in terms of privacy. They also gave feedback on the appearance of it (e.g. noise in the video effects and hair/skin/body color) and noted it was

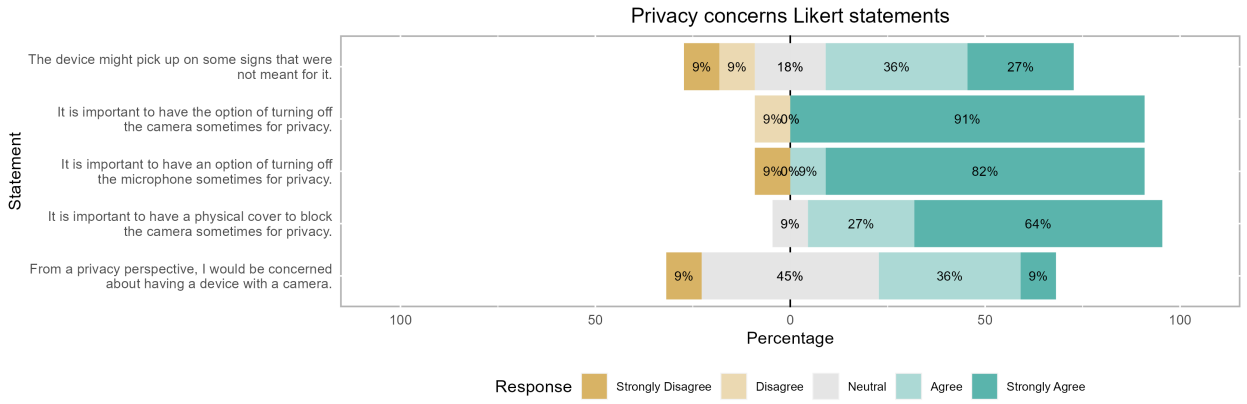


Figure 12.1: Results for likert-scale privacy concern statements

important for the signer to still be clearly understandable. Participants were explained that software like this would pre-process their video, disguising their face before it was sent over the Internet to servers for processing, so it is interesting that participants continued to care about how they appear in anonymization technology in such cases. Regardless, people emphasized that they need to still be clearly understandable after being anonymized, which presents an open challenge on creating technology that can anonymize a signer's identity without compromising the conveyance of their messages, e.g. to be used in adding anonymization-protection to the data such a device captures. Further, it is an open question whether this anonymization technology can be better evaluated with DHH users to ensure they are satisfied with it in different contexts (e.g. could a prototype be good enough for using Alexa, but not be good enough for posting videos online?).

If personal assistant devices can understand ASL in the future, there is more work needed, in the context of DHH ASL users, to understand how privacy concerns impact their comfort level before they decide to purchase such a device, and how this can be addressed. Recent technologies from companies such as Google and Apple have advertised features using data that "never leaves the device". For example, Google advertises security and privacy with their Pixel phones [22]: the "Now Playing" feature that recognizes and identifies music playing around the phone says "Unlike other song-identification services, all the processing happens right on your Pixel. [...] without any audio leaving your phone. Now Playing works fast and is private." Many users who have Apple iPhones use Face ID, where the

[22] https://safety.google/pixel/

phone can be unlocked using the users' face biometrics. Apple says [23] that "Face ID data doesn't leave your device and is never backed up to iCloud or anywhere else." It is an open question whether it would be enough if the devices used an on-board language model to recognize ASL (meaning it would not need to transmit their ASL video across the Internet to a cloud server for processing), and whether that would alleviate privacy concerns about placing such a device in different rooms in users' homes.

During the survey-based study and through interviews with participants who imagined having a personal assistant device that could understand ASL (chapter 3), notifications about sound happening in a household was mentioned as a potential use-case for personal assistants. In the in-person Wizard-of-Oz experiment, an item in a task asked participants about using Alexa to help them out so that they would know if a smoke alarm went off (section 10.3.2). In a prior work, Jain et al. designed and evaluated tablet-based sound awareness prototypes that informed users of sounds happening around their home [87], mentioning privacy concerns about devices that would pick up sounds inside shared spaces. In this small privacy study, we have received similar comments from participants about potential unwanted observation from the device camera. This interview study has also informed developers of sign language recognition technologies to ensure that their models capture the signing that was intended for it, and not pick up unwanted communication in the background.

For the statement "From a privacy perspective, I would be concerned about having a device with a camera", the likert responses from the survey based study (where people answered privacy questions when only imagining using a system) was higher (showing more concern) than that of the in-person study participants. Even though participants commented they felt like moving closer to Alexa so that it could see them better, and making constant eye-contact with the device during the interactions, the likert responses show that they were more neutral about having a device with a camera. It is possible that the experience helped make this more real, helping people realize that they would need to be near cameras to utilize ASL-aware camera-based technologies. Despite that, participants continued to comment and respond that it was very important to have the option to turn off or have a physical block for the device camera and/or microphone.

Several participants commented that they don't want the device to pick up on background commu-

[23] https://support.apple.com/en-us/HT208108

nication not meant for it, and do not want the device to watch them all the time if they are not using it. However, during the main in-person Wizard-of-Oz experiment, participants constantly placed themselves inside the room where they believed Alexa could see them best. As presented in section 10.4, a participant commented that if the device camera moved, they would have tried different locations inside the room for issuing commands. Another participant suggested having a 360-degree camera on the ceiling so that it could see the user anywhere they were inside a room. These suggestions seem to be at odds with the privacy concerns that are evident throughout this interview study that was appended after the main in-person Wizard-of-Oz experience. This echos participants' diverse opinions and reactions to this future technology, and motivates future work.

Privacy has been a challenging topic, especially as technology proliferates, and it has been brought up in many parts of this dissertation work, as well as the current work of other people investigating ASL data collection and automatic sign language recognition. Prior initial work asked participants to imagine the experience of having a personal assistant device that can understand ASL (chapter 3), and this work has further confirmed that DHH individuals would like to be able to turn on/off the camera and/or have a physical cover for the camera. They are also interested in having the same for a microphone, in case that the device uses that as well. In this work, participants note that, while it is important to have, it may be inconvenient to have to turn on or uncover the camera every time an user would like to issue a command. So far, this evidence seems to point towards giving people individual options, but motivates a need for more design work and research on this issue; manifesting an open challenge for future researchers and designers.

EPILOGUE TO PART IV

This is the end of Part IV of this dissertation. We investigated state-of-the-art image processing technology in the context of "anonymizing" DHH users' ASL videos to share online. An interview study evaluated prototype face-disguise technology that preserved facial expressions and head movements (essential characteristics and properties of ASL). The following research questions were posed:

RQ7.1: Is state-of-the-art face-disguise technology capable of perserving facial expressions and natural human appearance for sign language video? When participants first saw their post-processed ASL videos, many did not realize that they were actually viewing a transformed video of themselves. We found that new advances in face transformation technologies enabled replacing faces in videos at a level of quality that preserved their linguistic facial expressions and head movements, essential for ASL. However, participants commented that none of the prototypes presented were *completely* natural (albeit effective at disguising identity). (section 11.4)

RQ7.2: What are DHH users' interest in and impressions of this technology for protecting anonymity, including users' views of various dimensions of system performance? Three key dimensions of importance were identified: understandability, naturalness, and anonymity. The tradeoffs between each possible pair is discussed in depth, and the implications for future designers of face anonymization is described. Overall, participants strongly agreed that the technology disguised the identity of the original signer and identified potential use cases, such as safely expressing personal views on sensitive or confidential topics, and also expressed some ethical concerns, such as being able to use this technology to impersonate another real person. (section 11.5)

This study focused on the general usage of anonymization technology to communicate online, and found that it was effective for preserving anonymity, identified uses of interest, and contributed empirical knowledge about DHH users' assessment of this technology. Chapter 12 conducted an interview study to investigate the potential application of this technology to the specific context of ASL personal assistant interaction:

> **RQ8: Would using a state-of-the-art face-disguise technology to anonymize DHH users' ASL recordings (that are used for device processing) before they are processed by a personal assistant device alleviate privacy concerns?** In an interview study conducted with DHH participants after they had an in-person experience with a device that appeared to understand ASL, participants' views on privacy issues relating to a camera on a device in their home were diverse. Some participants had strong concerns and expressed the desire for a physical cover for the camera, while others were unconcerned with this issue. Such diversity motivates giving users choices in individual privacy settings and options, yet the inconvenience of physical covers motivates future research on more usable privacy controls for an in-home camera. Participants were uncertain whether the face-disguise technologies from the earlier study would be helpful for protecting their privacy in this context.

Chapter 13

Conclusion

Voice-controlled personal assistants are increasingly ubiquitous, and pose urgent accessibility challenges and barriers for DHH users. This dissertation consisted of parallel research efforts investigating issues surrounding this technology, and provides a basis for future design and development of personal assistants that would be able to understand ASL input.

Part I: First, an initial interview study was conducted to inform the design of a larger, online survey that directly engaged with the national DHH community about their interest and requirements for this technology. Next, formative interviews inspired the creation of video prototypes focused on the "wake-up" portion of the interaction of DHH ASL signers with personal assistant devices, and these prototypes were critically discussed with more participants.

Part II: Next, an online ASL data collection platform was designed, and its viability was tested through a crowdsourced approach where participants contributed data and conducted quality-control. This platform was extended into a new platform to encourage contribution of sentence-level ASL data. Then, focusing on the narrow domain of personal assistant commands, a remote Wizard-of-Oz experiment was implemented and collected data. This remote Wizard-of-Oz study gave users the opportunity to spontaneously interact back and forth with a device that appeared to understand ASL, rather than having to imagine the experience, putting the findings from the previous part into practice and building upon them. For instance, while Part I curated expected command-categories and discussions of

wake-up approaches; novel ways users "wake-up" such devices were discovered, ASL commands were captured and annotated (yielding a publicly-released dataset), and linguistic features of the commands were discussed. Since this experiment was limited due to the lack of real-world conditions, an in-person Wizard-of-Oz experiment was conducted, focusing on investigating the linguistic features of the in-person interaction.

Part III: Analyzing the data collected from the virtual Wizard-of-Oz experiment, thematic analysis (e.g. affinity mappings) was done to learn valuable insights about how DHH users behave and command a personal assistant (albeit in a remote manner). Next, analysis of the in-person Wizard-of-Oz experiment revealed several interesting and important linguistic properties (e.g. pointing to other objects inside the room as part of their commands) of the interaction, informing designers and researchers of future personal assistant and sign language technologies.

Part IV: Finally, since privacy concerns had emerged across several of the earlier studies, it was important to consider this in the context of this dissertation work. First, we investigated new advances in state-of-the-art image processing technology, and found it is possible to transform the face of an ASL signer in a video while preserving the facial expressions and head movements essential for ASL. Using the prototypes from this study, we appended a small interview study to the in-person experiment, asking participants about the application of this face-disguise technology after they had some experience interacting with a Wizard-of-Oz personal assistant.

13.1 Contributions

The key contributions of this dissertation are as follows:

Part I Contributions:

- We engaged with DHH community (86 DHH ASL-signing survey respondents from over 20 U.S. states) to address lack of knowledge about DHH user interest and requirements for this technology, such as desired features, usage scenarios, and other expectations for such systems. Through this, we established evidence that DHH users use personal assistants significantly less than the general population and evidence of their interest in using such devices if they understood ASL. Further,

we compiled a prioritized list of commands DHH users are interested in using, as well as a list of DHH-specific user cases. Additionally, we present evidence of privacy concerns, and initial user reaction to wake-up interaction approaches and response-display for ASL personal assistant devices.

- We revealed the preferences and concerns of DHH users for how to "wake up" future personal-assistant technologies that can understand sign language. Building on this, we created a set of six wake-up techniques, and discussed trade-offs between these approaches, identifying key factors affecting DHH user preferences of each.

Part II Contributions:

- We identified a methodology for streamlined collection of ASL data at scale, with automatic labelling of user-contributed videos. Testing its viability, we showed that a crowd of contributors can generate high-quality recordings and can perform quality control checks on one another's videos with high reliability.

- We developed a novel, bilingual interface that provides a side-by-side ASL and English synchronized interface that streamlines pre-labeled data collection, and enables a crowd to contribute to piecemeal translation as motivation for contributing to the dataset.

- We developed a remote Wizard-of-Oz methodology to collect training data of ASL personal assistant commands, and allowed DHH users to spontaneously interact with such a device in sign language. Through this methodology, we collected data and describe the characteristics of the dataset, as well as its properties and annotations (holding the analysis for a later chapter). This collected data and accompanying annotation is released publicly to support future HCI research on the behavior of DHH users of personal assistant systems, as well as serving as potential data for sign-language recognition researchers who are training artificial-intelligence models.

- We conducted an in-person Wizard-of-Oz protocol where participants performed household tasks in a home-like living room and kitchen area experimental setup using a personal assistant that

appeared to understand ASL. Through this experiment, we recorded 531 commands from 12 participants, and iteratively annotated the videos, allowing for analysis (which is also presented in a later chapter).

Part III Contributions:

- Through the first observational study of the behavior and interaction of DHH individuals engaging with a personal assistant device that appears to understand ASL, we analyze the remote Wizard-of-Oz data and provide guidance for future designers of this technology, e.g., specific commands to support, ASL terminology to use for command and control of the device, how the device should respond if there is an error, among other insights. This also provided guidance to creators of sign-language recognition technology, in prioritizing vocabulary that must be recognized in order to support natural ASL interaction.

- Analysis of the in-person Wizard-of-Oz experiment addressed research questions that were not possible due to the inherent lack of real-world conditions of the virtual Wizard-of-Oz setup. With the opportunity to interact with a personal assistant device that is physically in the same room as them in a residential living room and kitchen area, DHH users were able to utilize their full signing space, "pointing" to objects in the room to refer to them within commands issued to the personal assistant. Further, the DHH users were free to change their location inside the room, and analysis and discussion of the variation in user proximity to the device can serve as useful guidance for future sign-language recognition researchers, who must ensure that their technology works for the types of camera distances and angles in this context. Additionally, the feedback and recommendations of participants in the study suggested new avenues for design and research on personal assistant devices, e.g., the potential need for additional in-room cameras to provide greater flexibility in where DHH users can interact with the device.

Part IV Contributions:

- We evaluate state-of-the-art face transformation technologies within the context of ASL video anonymization through a three-phase study with semi-structured interviews. This study revealed

that this technology was effective for anonymization-protection, and we revealed factors impacting DHH users' acceptance of this technology.

- We conducted an interview study showing that this face-disguise technology did not have a significant impact on DHH users level of comfort with ASL-based personal assistant devices. Participants' diverse views on the severity of privacy concerns for personal assistant devices with cameras in their home are discussed, and motivate future work for designers of personal assistants that use an integrated camera to capture ASL input.

13.2 Overall Limitations and Future Work

Throughout this dissertation, we have engaged with the DHH community and addressed the lack of knowledge about DHH user familiarity and prior experience, interest, requirements, and concerns about a hypothetical personal assistant device that can understand ASL. We have also implemented and run two Wizard-of-Oz experiments (one virtual and one in-person), to investigate and estimate user reactions, behavior, and experience with the interaction. However, the extent of our sample sizes is somewhat limited. While our survey study reached 86 participants from over 20 U.S. states, our Wizard-of-Oz experiments were limited to participants found through university social media, mailing lists, and on-campus flyer postings (N=21 for the virtual study, and N=12 for in-person). If such a device was created and sold commercially, it is possible that many DHH individuals from all states, and even internationally, would purchase and use it. While future work can start with user testing locally, researchers and developers need to conduct testing nationally and potentially internationally to ensure good feedback from all potential consumers.

Also, our ASL personal assistant device prototypes, while effective for our research questions, were not real; they used a Wizard-of-Oz setup to appear as such. As sign recognition models are developed and employed, they will need to be intensively tested to check for accuracy and latency with diverse signers in various environments. There are many factors in play here, and there was a recent interdisciplinary review on current efforts, as discussed several times throughout this dissertation [23].

Our laboratory was set up to emulate a living room and kitchen area in a residential setting, but it

still had limitations and potentially influenced the user experience as they were not actually in their own home, which they are accustomed to. Future work should test out new personal assistant devices in the actual homes of many different users, i.e. through a pilot program, to test real-world performance. This may also reveal novel situations or use-cases that occur if personal assistant devices could understand ASL.

This dissertation also focused on DHH ASL users. While this is a large population, there are many people globally who use other sign languages (there are over 300 different ones), and even different forms of ASL (e.g. PSE). There is still a lot of work to be done for ASL technologies, but future researchers and developers need to think about generalizing their procedures and applying it to other languages to ensure inclusivity. It is also possible that there are cultural differences, even intranationally, that contribute to the design of personal assistants, which may impact the user experience. Further, there are various identities and backgrounds within the DHH community; individuals may identify as Deaf, deaf, or hard-of-hearing. [Capitalized "Deaf" is typically used to refer to a cultural and linguistic minority and lowercase "deaf" to refer to audiological status. This cultural identity is complex, deeply personal, and varies globally.] In this dissertation, it was not investigated whether someone's identity affected their views or behavior with these devices in ASL. Throughout the studies, the recruitment criteria generally checked that participants use ASL daily/primarily, and directed that they use ASL during the experiments. Interviews were also conducted in ASL, and ASL videos or interpretations of the study materials (e.g. forms) were available. In reality, the DHH community is diverse, ranging from some users who do not use sign language at all and rely on hearing and speaking, while others rely completely on sign language. Future work could gather a significant amount of participants who self-identify with different statuses and come from different backgrounds, to investigate whether they have different preferences and requirements of future technologies that may utilize sign recognition, such as camera-based personal assistant devices.

Throughout the studies in this dissertation, this work has focused on designs and form-factors of current and near-future personal assistant devices. That is, this work has used an "Alexa Echo Show" device; a home-based, fixed-placement smart display that utilizes a camera and microphone for input,

and a screen and speakers for output. I chose to operate within the likely form factor of personal assistant devices in the near future, striving to make my work as useful as possible. However, this brings a limitation that I did not explore creative designs or invent a new technology. Future work should explore a greater diversity in the design and form factor of technology with voice-control interfaces. For instance, the experience of using one-handed signing to a smartwatch would be greatly different and would uncover several interesting research avenues. There were some comments throughout my Wizard-of-Oz studies where participants suggested alternate designs for the personal assistants, such as using 360-degree cameras or having multiple access points (i.e. via a second device in a convenient location away from the first) for input and output.

13.3 Final Thoughts

As a Deaf ASL signer, I have firsthand experience with accessibility barriers in technology. As voice-based personal assistant technologies proliferate, e.g. smart speakers in homes, and more generally as voice-control becomes an increasingly ubiquitous interface to technology, new accessibility barriers are emerging for many DHH users. Progress in sign-language recognition may enable these devices to respond to sign-language commands and potentially mitigate these barriers, but research is needed to understand how DHH users would interact with these devices and what commands they would issue.

Broadly, as voice-control is becoming a standard feature of smart technologies, there is a risk that a new technology accessibility barrier will be erected that will disadvantage DHH individuals. This dissertation contributes to improving the accessibility of conversational-interaction user-interfaces through sign-language interaction, to help mitigate this risk. This dissertation has directly engaged with the DHH community and established DHH users' interest in personal-assistant technologies, and insights into how they would like to use or interact with these devices. Rather than scientists arbitrarily deciding when technology should be deployed, I investigate factors impacting DHH user preferences, satisfaction, and comfort. I also contribute datasets valuable for computer-vision researchers creating sign-recognition technology, rather than current and existing datasets, which are too diffuse, include non-native signers, and are too expensive to produce.

I am very hopeful that the future will continue to bring us many exciting technologies that help improve our quality of life, as well as to entertain, and that these technologies are fully accessible for DHH sign language users such as myself.

Bibliography

[1] LREC 2020. 2020. 9th workshop on the representation and processing of sign languages: sign language resources in the service of the language community, technological challenges and application perspectives. https://www.sign-lang.uni-hamburg.de/lrec2020/cfp.html (Cited on page 16).

[2] Ali Abdolrahmani, Ravi Kuber, and Stacy M. Branham. 2018. "Siri Talks at You": An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 249–258. https://doi.org/10.1145/3234695.3236344 (Cited on pages 14, 54).

[3] Nicoletta Adamo-Villani and Ronnie B. Wilbur. 2015. ASL-Pro: American Sign Language Animation with Prosodic Elements. In *Universal Access in Human-Computer Interaction. Access to Interaction*. Springer International Publishing, 307–318. https://doi.org/10.1007/978-3-319-20681-3_29 (Cited on page 197).

[4] Shashank Ahire and Michael Rohs. 2020. Tired of Wake Words? Moving Towards Seamless Conversations with Intelligent Personal Assistants. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 20, 3 pages. https://doi.org/10.1145/3405755.3406141 (Cited on page 43).

[5] Oliver Alonzo, Matthew Seita, Abraham Glasser, and Matt Huenerfauth. 2020. Automatic Text

Simplification Tools for Deaf and Hard of Hearing Adults: Benefits of Lexical Simplification and Providing Users with Autonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376563 (Cited on page 91).

[6] Oliver Alonzo, Jessica Trussell, Becca Dingman, and Matt Huenerfauth. 2021. Comparison of Methods for Evaluating Complexity of Simplified Texts among Deaf and Hard-of-Hearing Adults at Different Literacy Levels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 279, 12 pages. https://doi.org/10.1145/3411764.3445038 (Cited on page 92).

[7] Amazon. 2005. *Amazon Mechanical Turk*. https://www.mturk.com/ Accessed 2021-08-27. (Cited on page 62).

[8] Amazon. 2020. Invoking Alexa. https://developer.amazon.com/en-US/docs/alexa/alexa-auto/invoking-alexa.html (Cited on page 43).

[9] Robert W. Arnold. 2007. A proposal for a written system of American Sign Language. (Cited on page 197).

[10] Werner Bailer and Martin Winter. 2019. On Improving Face Generation for Privacy Preservation. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–6. https://doi.org/10.1109/CBMI.2019.8877442 (Cited on pages 195, 199).

[11] Charlotte Baker-Shenk. 1985. The Facial Behavior of Deaf Signers: Evidence of a Complex Language. *American Annals of the Deaf* 130, 4 (1985), 297–304. (Cited on pages 194, 195, 199).

[12] Charlotte Baker-Shenk and Dennis Cokley. 2002. *American Sign Language: A Teachers Resource Text on Grammar and Culture*. Gallaudet University Press. (Cited on page 161).

[13] D. Balfanz, G. Durfee, D. K. Smetters, and R. E. Grinter. 2004. In search of usable security: five lessons from the field. *IEEE Security Privacy* 2, 5, 19–24. https://doi.org/10.1109/MSP.2004.71 (Cited on page 54).

[14] Steven Barnett, Jonathan D. Klein, Robert Q. Pollard, Vincent Samar, Deirdre Schlehofer, Matthew Starr, Erika Sutter, Hongmei Yang, and Thomas A. Pearson. 2011. Community Participatory Research With Deaf Sign Language Users to Identify Health Inequities. *American Journal of Public Health* 101, 12 (2011). https://doi.org/10.2105/ajph.2011.300247 (Cited on page 195).

[15] BBC. 2018. Sign-language hack lets Amazon Alexa respond to gestures. https://www.bbc.com/news/technology-44891054 (Cited on pages 16, 43).

[16] Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. Methods for Evaluation of Imperfect Captioning Tools by Deaf or Hard-of-Hearing Users at Different Reading Literacy Levels. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173665 (Cited on page 92).

[17] 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. (Cited on page 85).

[18] Jeffrey P. Bigham, Raja Kushalnagar, Ting-Hao Kenneth Huang, Juan Pablo Flores, and Saiph Savage. 2017. On How Deaf People Might Use Speech to Control Devices. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 383–384. https://doi.org/10.1145/3132525.3134821 (Cited on page 10).

[19] 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. (Cited on page 85).

[20] Danielle Bragg, Naomi Caselli, John W. Gallagher, Miriam Goldberg, Courtney J. Oka, and William Thies. 2021. ASL Sea Battle: Gamifying Sign Language Data Collection. In *Proceedings of*

the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 271, 13 pages. https://doi.org/10.1145/3411764.3445416 (Cited on pages xxiii, 67, 73, 76, 77, 78, 92, 94).

[21] Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. 2021. The FATE Landscape of Sign Language AI Datasets: An Interdisciplinary Perspective. *ACM Trans. Access. Comput.* 14, 2, Article 7 (jul 2021), 45 pages. https://doi.org/10.1145/3436996 (Cited on pages 89, 119, 149).

[22] Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. 2022. Exploring Collection of Sign Language Videos through Crowdsourcing. In *PACM on Human-Compter Interaction 6, CSCW2*, Vol. 6. Association for Computing Machinery. https://doi.org/10.1145/3555627 (Cited on page 61).

[23] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, 16–31. https://doi.org/10.1145/3308561.3353774 (Cited on pages 16, 17, 43, 62, 92, 197, 238).

[24] 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* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, Article 33, 14 pages. https://doi.org/10.1145/3373625.3417024 (Cited on pages 89, 119, 195, 196, 198, 200, 204, 217).

[25] Danielle Bragg, Raja Kushalnagar, and Richard Ladner. 2018. Designing an Animated Character System for American Sign Language. In *Proceedings of the 20th International ACM SIGACCESS*

Conference on Computers and Accessibility. ACM. https://doi.org/10.1145/3234695.3236338 (Cited on page 197).

[26] 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. (Cited on pages 67, 92).

[27] 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. (Cited on page 86).

[28] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa (Cited on pages 24, 206).

[29] 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. (Cited on page 15).

[30] B Cartwright. 2017. Signing Savvy. https://www.signingsavvy.com/ (Cited on pages 67, 94).

[31] 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. (Cited on pages xxii, 72, 73).

[32] Fabio Catania, Micol Spitale, Giulia Cosentino, and Franca Garzotto. 2020. What is the Best Action for Children to "Wake Up" and "Put to Sleep" a Conversational Agent? A Multi-Criteria Decision Analysis Approach *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3405755.3406129 (Cited on page 44).

[33] Anna Cavender, Richard E Ladner, and Eve A Riskin. 2006. MobileASL: Intelligibility of sign

language video as constrained by mobile phone technology. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. 71–78. (Cited on page 197).

[34] ASL Clear. 2022. https://aslclear.org/ (Cited on pages 67, 94).

[35] Devin Coldewey. 2018. SignAll is slowly but surely building a sign language translation platform. https://techcrunch.com/2018/02/14/signall-is-slowly-but-surely-building-a-sign-language-translation-platform/ (Cited on page 16).

[36] Convo. 2021. Convo VRS. https://www.convorelay.com/ (Cited on page 129).

[37] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2012. Sign Language Recognition Using Sub-Units. *J. Mach. Learn. Res.* 13, 1 (jul 2012), 2205–2231. (Cited on page 17).

[38] 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. (Cited on page 63).

[39] ASL Core. 2022. https://aslcore.org/ (Cited on pages 67, 94).

[40] G.R. Coulter. 1979. American Sign Language Typology. (Cited on pages 194, 195, 199).

[41] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can i Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539 (Cited on page 43).

[42] Xianghua Ding, Yanqi Jiang, Xiankang Qin, Yunan Chen, Wenqiang Zhang, and Lizhe Qi. 2019. Reading Face, Reading Health: Exploring Face Reading Technologies for Everyday Health. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland

Uk) *(CHI '19)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3290605.3300435 (Cited on page 199).

[43] Julie Doyle, Emma Murphy, Janneke Kuiper, Suzanne Smith, Caoimhe Hannigan, An Jacobs, and John Dinsmore. 2019. Managing Multimorbidity: Identifying Design Requirements for a Digital Self-Management Tool to Support Older Adults with Multiple Chronic Conditions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300629 (Cited on page 23).

[44] 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. (Cited on page 15).

[45] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. (Cited on page 17).

[46] Eleni Efthimiou, Stavroula-Evita Fotinea, Theodore Goulas, and Panos Kakoulidis. 2015. User Friendly Interfaces for Sign Retrieval and Sign Synthesis. In *Universal Access in Human-Computer Interaction. Access to Interaction*. Springer International Publishing, 351–361. https://doi.org/10.1007/978-3-319-20681-3_33 (Cited on page 197).

[47] Eleni Efthimiou, Stavroula-Evita Fotinea, Theodore Goulas, Anna Vacalopoulou, Kiki Vasilaki, and Athanasia-Lida Dimou. 2018. Sign Language technologies in view of Future Internet accessibility services. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. 495–501. (Cited on page 197).

[48] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. The

Dicta-Sign Wiki: Enabling Web Communication for the Deaf. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 205–212. https://doi.org/10.1007/978-3-642-31534-3_32 (Cited on pages 196, 197).

[49] Eleni Efthimiou, Stavroula-Evita Fotinea, Christian Vogler, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and Jérémie Segouat. 2009. Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication. In *Universal Access in Human-Computer Interaction. Addressing Diversity*. Springer Berlin Heidelberg, 21–30. https://doi.org/10.1007/978-3-642-02707-9_3 (Cited on page 196).

[50] 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. (Cited on page 15).

[51] Karen Emmorey, Chuchu Li, Jennifer Petrich, and Tamar H. Gollan. 2020. Turning languages on and off: Switching into and out of code-blends reveals the nature of bilingual language control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46, 3 (2020), 443–454. https://doi.org/10.1037/xlm0000734 (Cited on page 95).

[52] Be My Eyes. 2020. Be My Eyes. https://www.bemyeyes.com/ Accessed 2022-04-22. (Cited on page 85).

[53] Facebook. 2021. Facebook Homepage. https://www.facebook.com/ (Cited on page 195).

[54] Jianping Fan, Hangzai Luo, Mohand-Said Hacid, and Elisa Bertino. 2005. A novel approach for privacy-preserving video sharing. In *Proceedings of the 14th ACM international conference on Information and knowledge management*. 609–616. (Cited on page 198).

[55] Jordan Fenlon, Kearsy Cormier, and Adam Schembri. 2015. Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography* 28, 2 (2015), 169–206. (Cited on page 63).

[56] Leah Findlater, Bonnie Chinh, Dhruv Jain, Jon Froehlich, Raja Kushalnagar, and Angela Carey Lin. 2019. Deaf and Hard-of-Hearing Individuals' Preferences for Wearable and Mobile Sound Awareness Technologies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300276 (Cited on pages 23, 39).

[57] 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* (Madeira, Portugal) *(Lecture Notes in Computer Science 7887)*. Springer, 89–99. (Cited on page 15).

[58] 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. (Cited on pages 62, 93).

[59] 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. (Cited on page 86).

[60] Abraham Glasser. 2019. Automatic Speech Recognition Services: Deaf and Hard-of-Hearing Usability. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3308461 (Cited on pages 10, 11, 20, 37).

[61] Abraham Glasser, Vaishnavi Mande, and Matt Huenerfauth. 2020. Accessibility for Deaf and Hard of Hearing Users: Sign Language Conversational User Interfaces. In *CUI '20: Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 55, 3 pages. https://doi.org/10.1145/3405755.3406158 (Cited on page 43).

[62] Abraham Glasser, Vaishnavi Mande, and Matt Huenerfauth. 2021. Understanding Deaf and Hard-of-Hearing Users' Interest in Sign-Language Interaction with Personal-Assistant Devices. In *Pro-*

ceedings of the 18th International Web for All Conference (W4A '21). Association for Computing Machinery, New York, NY, USA, Article 24, 11 pages. https://doi.org/10.1145/3430263.3452428 (Cited on pages 20, 153, 155).

[63] Abraham Glasser, Fyodor Minakov, and Danielle Bragg. 2022. ASL Wiki: An Exploratory Interface for Crowdsourcing ASL Translations. In *The 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, 22 pages. https://doi.org/10.1145/3517428.3544827 (Cited on page 91).

[64] Abraham Glasser, Matthew Watkins, Kira Hart, Sooyeon Lee, and Matt Huenerfauth. 2022. Analyzing Deaf and Hard-of-Hearing Users' Behavior, Usage, and Interaction with a Personal Assistant Device that Understands Sign-Language Input. In *In Proceedings of the CHI Conference on Human Factors in Computing Systems Proceedings* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3491102.3501987 (Cited on pages 121, 147).

[65] Abraham T. Glasser, Kesavan R. Kushalnagar, and Raja S. Kushalnagar. 2017. Feasibility of Using Automatic Speech Recognition with Voices of Deaf and Hard-of-Hearing Individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 373–374. https://doi.org/10.1145/3132525.3134819 (Cited on page 10).

[66] 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. (Cited on page 61).

[67] Google. 2021. Use gestures on your Pixel phone. https://support.google.com/pixelphone/answer/7443425?hl=en (Cited on page 150).

[68] Ann Grafstein. 2002. HandSpeak: A Sign Language Dictionary Online. https://www.handspeak.com/ (Cited on pages 67, 94).

[69] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785 (Cited on page 128).

[70] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. 2005. Integrating utility into face de-identification. In *International Workshop on Privacy Enhancing Technologies*. Springer, 227–242. (Cited on page 198).

[71] 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. (Cited on page 85).

[72] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. 2018. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13. (Cited on page 198).

[73] Alexis Heloir and Fabrizio Nunnari. 2015. Toward an intuitive sign language animation authoring system for the deaf. *Universal Access in the Information Society* 15, 4 (May 2015), 513–523. https://doi.org/10.1007/s10209-015-0409-0 (Cited on page 197).

[74] Julie Hochgesang, Onno Crasborn, and Diane Lillo-Martin. 2022. Sign Bank. https://aslsignbank.haskins.yale.edu/ (Cited on pages 63, 94).

[75] Leala Holcomb and Jonathan McMillan. 2022. http://www.handsland.com/ (Cited on pages 67, 94).

[76] M. Shamim Hossain and Ghulam Muhammad. 2015. Cloud-Assisted Speech and Face Recognition Framework for Health Monitoring. *Mob. Netw. Appl.* 20, 3 (June 2015), 391–399. https://doi.org/10.1007/s11036-015-0586-3 (Cited on page 199).

[77] Matthew B. Hoy. 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medi-*

cal Reference Services Quarterly 37, 1 (2018), 81–88. https://doi.org/10.1080/02763869.2018.1404391 (Cited on pages 9, 149).

[78] Xuedong Huang. 2017. Microsoft researchers achieve new conversational speech recognition milestone. https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/ (Cited on page 151).

[79] Matt Huenerfauth and Vicki Hanson. 2009. Sign language in the interface: access for deaf signers. *Universal Access Handbook. NJ: Erlbaum* 38 (2009), 14. (Cited on pages 92, 197).

[80] 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. (Cited on page 15).

[81] Matt Huenerfauth, Kasmira Patel, and Larwan Berke. 2017. Design and Psychometric Evaluation of an American Sign Language Translation of the System Usability Scale. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 175–184. https://doi.org/10.1145/3132525.3132540 (Cited on page 176).

[82] IBM. 2017. Reaching new records in speech recognition. https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/ (Cited on page 151).

[83] Instagram. 2021. Instagram website. https://www.instagram.com/ (Cited on page 195).

[84] Alan Irwin. 1995. *Citizen science: A study of people, expertise and sustainable development*. Psychology Press. (Cited on page 62).

[85] David Isbitski. 2017. How to Build Alexa Skills for Echo Show. https://developer.amazon.com/blogs/alexa/post/12826e9e-e06a-4ab4-a583-8e074709a9f3/how-to-build-alexa-skills-for-echo-show (Cited on page 9).

[86] Brielle Jaeke. 2016. Sephora boosts augmented reality shopping with real-time facial recognition. http://t.cn/EGjuPZK (Cited on page 195).

[87] Dhruv Jain, Angela Lin, Rose Guttman, Marcus Amalachandran, Aileen Zeng, Leah Findlater, and Jon Froehlich. 2019. Exploring Sound Awareness in the Home for People Who Are Deaf or Hard of Hearing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300324 (Cited on pages 40, 230).

[88] Jeeliz. 2021. Jeeliz website. https://jeeliz.com/demos/faceFilter/demos/threejs/tiger/ (Cited on page 201).

[89] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 506–516. https://doi.org/10.1145/2736277.2741669 (Cited on pages 31, 39).

[90] Deaf Studies Digital Journal. 2009. https://www.deafstudiesdigitaljournal.org/ (Cited on page 95).

[91] 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). (Cited on page 62).

[92] Hernisa Kacorri and Matt Huenerfauth. 2016. Continuous profile models in asl syntactic facial expression synthesis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2084–2093. (Cited on page 195).

[93] Sushant Kafle, Abraham Glasser, Sedeeq Al-khazraji, Larwan Berke, Matthew Seita, and Matt Huenerfauth. 2020. Artificial Intelligence Fairness in the Context of Accessibility Research on

Intelligent Systems for People Who Are Deaf or Hard of Hearing. *SIGACCESS Access. Comput.* 125, Article 4 (March 2020), 1 pages. https://doi.org/10.1145/3386296.3386300 (Cited on page 149).

[94] 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 (Cited on page 15).

[95] Steven Komarov and Krzysztof Z Gajos. 2014. Organic peer assessment. In *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*. (Cited on page 63).

[96] Pavel Korshunov and Sébastien Marcel. 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685* (2018). (Cited on page 199).

[97] Federica Laricchia. 2022. Number of voice assistants in use worldwide 2019-2024. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/ (Cited on pages iv, 1).

[98] 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. (Cited on page 85).

[99] 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. (Cited on page 85).

[100] 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 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*. Association for Computing Machinery,

New York, NY, USA, Article 22, 13 pages. https://doi.org/10.1145/3441852.3471200 (Cited on pages 89, 119, 194).

[101] Abner Li. 2020. 'Hey Google' hotword training updated to boost Voice Match accuracy. https://9to5google.com/2020/04/23/hey-google-voice-match/ (Cited on page 151).

[102] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–24. (Cited on page 198).

[103] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, Adrianna Martinez, and et al. 2018. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2018), 984–997. https://doi.org/10.1177/0961000618759414 (Cited on pages 31, 39).

[104] Colin Lualdi. 2022. Sign School. https://www.signschool.com/ (Cited on pages 67, 94).

[105] Ceil Lucas and Clayton Valli. 1992. *Language Contact in the American Deaf Community*. Brill. (Cited on pages 131, 168).

[106] Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47, 4 (Jan. 2015), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5 (Cited on pages 202, 216).

[107] Debbie S. Ma, Justin Kantner, and Bernd Wittenbrink. 2020. Chicago Face Database: Multiracial expansion. *Behavior Research Methods* (Oct. 2020). https://doi.org/10.3758/s13428-020-01482-5 (Cited on pages 202, 216).

[108] Kelly Mack, Danielle Bragg, Meredith Ringel Morris, Maarten W. Bos, Isabelle Albi, and Andrés Monroy-Hernández. 2020. Social App Accessibility for Deaf Signers. *Proceedings of the ACM*

on Human-Computer Interaction 4, CSCW2 (Oct. 2020), 1–31. https://doi.org/10.1145/3415196 (Cited on pages 195, 197, 198).

[109] Sachit Mahajan, Ling-Jyh Chen, and Tzu-Chieh Tsai. 2017. SwapItUp: A Face Swap Application for Privacy Protection. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*. 46–50. https://doi.org/10.1109/AINA.2017.53 (Cited on pages 195, 199).

[110] Matt Malzkuhn and Melissa Malzkuhn. 2022. The ASL App. https://theaslapp.com/ (Cited on pages 67, 94).

[111] Vaishnavi Mande, Abraham Glasser, Becca Dingman, and Matt Huenerfauth. 2021. *Deaf Users' Preferences Among Wake-Up Approaches during Sign-Language Interaction with Personal Assistant Devices*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451592 (Cited on pages 42, 149, 153, 155).

[112] Tara Matthews, Janette Fong, F. Wai-Ling Ho-Ching, and Jennifer Mankoff. 2006. Evaluating non-speech sound visualizations for the deaf. *Behaviour & Information Technology* 25, 4 (2006), 333–351. https://doi.org/10.1080/01449290600636488 (Cited on page 39).

[113] Richard Meier. 1990. Person deixis in American sign language. *Theoretical issues in sign language research* 1 (1990), 175–190. (Cited on page 162).

[114] Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language corpus. *International Journal of Corpus Linguistics* 20, 1 (2015), 102–120. (Cited on page 63).

[115] Microsoft. 2019. AI for Accessibility Hackathon 2019. https://blogs.partner.microsoft.com/mpn-apac/ai-for-accessibility-hackathon-2019/ (Cited on page 16).

[116] Ross E. Mitchell. 2005. How Many Deaf People Are There in the United States? Estimates From the Survey of Income and Program Participation. *The Journal of Deaf Studies and Deaf Education*

11, 1 (09 2005), 112–119. https://doi.org/10.1093/deafed/enj004 (Cited on page 194).

[117] The Daily Moth. 2022. The Daily Moth. https://www.dailymoth.com/ (Cited on page 94).

[118] Chelsea M. Myers, Luis Fernando Laris Pardo, Ana Acosta-Ruiz, Alessandro Canossa, and Jichen Zhu. 2021. "Try, Try, Try Again:" Sequence Analysis of User Interaction Data with a Voice User Interface. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 18, 8 pages. https://doi.org/10.1145/3469595.3469613 (Cited on page 152).

[119] Carol Neidle, Judy Kegl, Dawn MacLaughlin, Benjamin Bahan, and Robert G. Lee. 2008. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT Press. (Cited on pages 161, 194, 195, 199).

[120] Carol Neidle and Augustine Opoku. 2020. A User's Guide to the American Sign Language Linguistic Research Project (ASLLRP) Data Access Interface (DAI) 2 — Version 2. American Sign Language Linguistic Research Project Report No. 18, Boston University.. http://www.bu.edu/asllrp/rpt18/asllrpr18.pdf (Cited on pages xix, 202, 204).

[121] Carol Neidle and Augustine Opoku. 2021. Update on Linguistically Annotated ASL Video Data Available through the American Sign Language Linguistic Research Project (ASLLRP). American Sign Language Linguistic Research Project Report No. 19, Boston University.. http://www.bu.edu/asllrp/rpt18/asllrpr18.pdf (Cited on pages 202, 204).

[122] Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. 2018. NEW Shared & Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* (Miyagawa, Japan) *(LREC 2018)*. 147–154. (Cited on pages xix, 202, 204).

[123] Carol Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. Challenges in development of the American Sign Language Lexicon Video Dataset (ASLLVD) corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon* (Instabul, Turkey) *(LREC 2012)*. http://www.bu.edu/linguistics/UG/LREC2012/LREC-asllvd-final.pdf (Cited on page 158).

[124] Don Newkirk. 1987. SignFont Handbook. (1987). (Cited on page 197).

[125] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. 2018. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 98–105. (Cited on page 199).

[126] National Association of the Deaf. 2022. Position Statement On ASL and English Bilingual Education. https://www.nad.org/about-us/position-statements/position-statement-on-asl-and-english-bilingual-education/ (Cited on page 95).

[127] World Federation of the Deaf. 2018. *Our Work*. http://wfdeaf.org/our-work/ Accessed 2019-03-26. (Cited on page 61).

[128] Christi Olson and Kelli Kemery. 2019. 2019 Microsoft Voice report. https://about.ads.microsoft.com/en-us/insights/2019-voice-report (Cited on pages 9, 25, 29, 31, 36, 134, 151).

[129] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and R. Bowden. 2012. Sign Language Recognition using Sequential Pattern Trees. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 2200–2207. (Cited on page 17).

[130] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. 2014. Sign Spotting Using Hierarchical Sequential Patterns with Temporal Intervals. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1931–1938. https://doi.org/10.1109/CVPR.2014.248 (Cited on page 17).

[131] 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. (Cited on pages 63, 85).

[132] Tabitha C. Peck, Jessica J. Good, and Kimberly A. Bourne. 2020. Inducing and Mitigating Stereotype Threat Through Gendered Virtual Body-Swap Illusions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3313831.3376419 (Cited on page 222).

[133] Jingnan Peng. 2020. Bringing light to the news, for those who can't hear it (video). https://www.csmonitor.com/The-Culture/2020/0731/Bringing-light-to-the-news-for-those-who-can-t-hear-it-video (Cited on page 94).

[134] Victoria Petrock. 2020. Voice assistant and smart speaker users 2020. https://www.emarketer.com/content/voice-assistant-and-smart-speaker-users-2020 (Cited on pages iv, 1).

[135] Robert Q Pollard, Erika Sutter, and Catherine Cerulli. 2013. Intimate Partner Violence Reported by Two Samples of Deaf Adults via a Computerized American Sign Language Survey. *Journal of Interpersonal Violence* 29, 5 (2013), 948–965. https://doi.org/10.1177/0886260513505703 (Cited on page 195).

[136] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214 (Cited on page 43).

[137] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of*

the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174033 (Cited on pages 12, 14, 23, 149, 155).

[138] 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. (Cited on page 63).

[139] Jason Rodolitz, Evan Gambill, Brittany Willis, Christian Vogler, and Raja Kushalnagar. 2019. Accessibility of Voice-Activated Agents for People who are Deaf or Hard of Hearing. *Journal on Technology and Persons with Disabilities* 7 (2019), 144–156. http://hdl.handle.net/10211.3/210397 (Cited on pages 10, 12, 13, 14, 16, 43, 122, 149).

[140] Kaleigh Rogers. 2018. Augmented Reality App Can Translate Sign Language Into Spoken English, and Vice Versa. https://www.vice.com/en_us/article/zmgnd9/app-to-translate-sign-language (Cited on page 16).

[141] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1–11. (Cited on page 199).

[142] 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. (Cited on page 85).

[143] 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. (Cited on page 85).

[144] 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. (Cited on page 85).

[145] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772 (Cited on pages 10, 31, 36, 39, 40).

[146] 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 (02 2021), 263–277. https://doi.org/10.1093/deafed/enaa038 arXiv:https://academic.oup.com/jdsde/article-pdf/26/2/263/36643382/enaa038.pdf (Cited on pages 72, 94).

[147] 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. (Cited on page 85).

[148] John Shinal. 2017. Making sense of Google CEO Sundar Pichai's plan to move every direction at once. https://www.cnbc.com/2017/05/18/google-ceo-sundar-pichai-machine-learning-big-data.html (Cited on page 151).

[149] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*. (Cited on pages 196, 199, 201).

[150] Aliaksandr Siarohin, Subhankar Roy, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu

Sebe. 2020. Motion Supervised co-part Segmentation. *arXiv preprint* (2020). (Cited on pages 196, 199, 201).

[151] SignGenius. 2020. Do's & Don'ts - Getting Attention in the Deaf Community. https://www.signgenius.com/info-do's&don'ts.shtml (Cited on pages 33, 44, 46, 150).

[152] Jonathan Silvertown. 2009. A new dawn for citizen science. *Trends in ecology & evolution* 24, 9 (2009), 467–471. (Cited on page 62).

[153] 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. (Cited on page 63).

[154] Snapchat. 2021. Snapchat website. https://www.snapchat.com/ (Cited on page 195).

[155] Sorenson. 2020. Sorenson VRS. https://www.sorensonvrs.com/ (Cited on page 39).

[156] 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. (Cited on page 88).

[157] 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. (Cited on page 15).

[158] Greg Sterling. 2018. Study: Google Assistant most accurate, Alexa most improved virtual assistant. https://searchengineland.com/study-google-assistant-most-accurate-alexa-most-improved-virtual-assistant-296936 (Cited on page 151).

[159] William C Stokoe Jr. 2005. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of deaf studies and deaf education* 10, 1 (2005), 3–37. (Cited on page 194).

[160] Valerie Sutton. 1998. The Signwriting Literacy Project. *Impact of Deafness on Cognition AERA Conference* (1998). (Cited on page 197).

[161] 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. (Cited on page 85).

[162] 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. (Cited on page 85).

[163] 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. (Cited on page 63).

[164] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395. (Cited on page 199).

[165] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N. Metaxas. 2018. CR-GAN: Learning Complete Representations for Multi-view Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 942–948. https://doi.org/10.24963/ijcai.2018/131 (Cited on pages 196, 199, 201).

[166] TikTok. 2021. TikTok website. https://www.tiktok.com/ (Cited on page 195).

[167] Carol Bloomquist Traxler. 2000. The Stanford Achievement Test, 9th Edition: National Norming and Performance Standards for Deaf and Hard-of-Hearing Students. *The Journal of Deaf Studies and Deaf Education* 5, 4 (09 2000), 337–348. https://doi.org/10.1093/deafed/5.4.337

arXiv:https://academic.oup.com/jdsde/article-pdf/5/4/337/9835826/337.pdf (Cited on pages 92, 197).

[168] Hitomi Tsujita and Jun Rekimoto. 2011. Smiling Makes Us Happier: Enhancing Positive Mood and Communication with Smile-Encouraging Digital Appliances. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, 1–10. https://doi.org/10.1145/2030112.2030114 (Cited on page 199).

[169] 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. (Cited on page 63).

[170] Clayton Valli and Ceil Lucas. 2000. *Linguistics of American sign language: An introduction*. Gallaudet University Press. (Cited on pages 194, 195).

[171] Vcom3D. 2015. Sign Smith Studio website. http://www.vcom3d.com/ (Cited on page 197).

[172] 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. (Cited on page 63).

[173] 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. (Cited on page 63).

[174] Lezi Wang, Chongyang Bai, Maksim Bolonkin, Judee Burgoon, Norah Dunbar, V. S. Subrahmanian, and Dimitris N. Metaxas. 2019. Attention-based facial behavior analytics in social communication. *30th British Machine Vision Conference (BMVC'19)*. (Cited on page 201).

[175] Jason Ward. 2018. Why Microsoft must bring sign language recognition to Windows and Cortana. https://www.windowscentral.com/microsoft-must-bring-sign-language-recognition-windows-and-cortana (Cited on page 16).

[176] Wikimedia. 2001. *Wikipedia: The Free Encyclopedia*. https://www.wikipedia.org/ (Cited on pages 62, 97).

[177] Wikipedia. 2022. Deaf News. https://en.wikipedia.org/wiki/Deaf_News (Cited on page 94).

[178] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1083–1092. https://doi.org/10.1145/1518701.1518866 (Cited on page 14).

[179] Gabriella Wojtanowski, Colleen Gilmore, Barbra Seravalli, Kristen Fargas, Christian Vogler, and Raja Kushalnagar. 2020. "Alexa, Can You See Me?" Making Individual Personal Assistants for the Home Accessible to Deaf Consumers. *Journal on Technology and Persons with Disabilities* 8 (2020). http://hdl.handle.net/10211.3/215984 (Cited on pages 10, 13, 14, 122).

[180] Alicia Wooten and Barbara Spiecker. 2022. https://www.atomichands.com/ (Cited on pages 67, 94).

[181] Xuchen Yao, Guoguo Chen, and Yuan Cao. 2017. Developing Your Own Wake Word Engine Just Like 'Alexa' and 'OK Google'. https://gputechconf2017.smarteventscloud.com/connect/sessionDetail.ww?SESSION_ID=112905 (Cited on page 150).

[182] Nick Yee and Jeremy Bailenson. 2007. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research* 33, 3 (July 2007), 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x (Cited on page 222).

[183] YouTube. 2021. YouTube website website. https://www.youtube.com/ (Cited on page 195).

[184] Sihan Yuan, Birgit Brüggemeier, Stefan Hillmann, and Thilo Michael. 2020. User Preference and Categories for Error Responses in Conversational User Interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '20)*. Association for

Computing Machinery, New York, NY, USA, Article 5, 8 pages. https://doi.org/10.1145/3405755.3406126 (Cited on page 152).

[185] 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. (Cited on page 67).

[186] 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. (Cited on page 15).

[187] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. 2020. Towards Image-to-Video Translation: A Structure-Aware Approach via Multi-stage Generative Adversarial Networks. *International Journal of Computer Vision* 128, 10-11 (April 2020), 2514–2533. https://doi.org/10.1007/s11263-020-01328-9 (Cited on pages 196, 199, 201).

[188] New Zoogle. 2021. Best Animal Face Changer Apps for Android. https://newzoogle.com/best-animal-face-changer-apps-android/ (Cited on page 195).

Appendices

Appendix A

Supplemental Materials for ASL Wiki

In this appendix, we provide supplemental materials for *ASL Wiki: An Exploratory Interface for Crowdsourcing ASL Translations* (chapter 6).

A.1 Semi-structured user study interview questions

Below is the semi-structured interview questions that were discussed with participants as part of the user study:

- Role/relation to ASL: What's your role/relationship with ASL? (e.g. native speaker, primary language, ASL teacher, use ASL at work, etc...)

Reading

- Did you primarily look at the ASL or English part? [Follow up to estimate percentage (0% ASL 100% English vs 100% ASL 0% English)]
- How did viewing different signers affect your experience? (If applicable)
- On a scale from 1-5 (1- very difficult), how understandable was the ASL content you viewed? Can you explain why you chose this number?

- On a scale from 1-5 (1-very difficult), how understandable was the English content you viewed? Can you explain why you chose this number?
- Was it helpful to view the content in both English and ASL? Why or why not?
- Did you use the upvote/downvote feature? Why or why not?
- How easy was the interface to use? (1-5: Very difficult – Very easy) If difficult, did information overload contribute to difficulties?
- What did you like or dislike about the interface?

Recording

- Did you find any content challenging to record? If so, what made it challenging?
- Did you use any strategies while recording content? If so, what were they?
- On a scale from 1-5, how easy was the interface to use? (1-5: Very difficult– Very easy)
- What did you like or dislike about the interface?

Desirability

- Do you wish more content online provided both English and ASL? (1-5: Strongly disagree– Strongly agree) If so, can you give some examples of when you wanted content provided in both languages?
- Would you be interested in generating content available in both English and ASL? (1-5: Strongly disagree– Strongly agree)
- What benefits do you feel this site offer to you as a user, if any?
- What concerns do you have in using a website like this, if any?
- How enjoyable was using the website, overall? What did you like/dislike?

- Would you want to use a website like this to read content in the future? Why or why not? Is there different content you would want to read (e.g. movie scripts, podcast, etc.)?
- What type of *Wikipedia* content would you want translated (i.e., picking from the list of topics on the Wikipedia landing page – food, math, Deaf culture, etc.)?
- Would you want to use a website like this to contribute recordings in the future? Why or why not?
- How likely are you to recommend this website to others? (1-5: Very likely – Very unlikely) If so, who would you recommend this to, and for what purpose (e.g. ASL students for learning, people with certain English/ASL fluency, etc.)?

Appendix B

Supplemental Materials for DHH Interest

In this appendix, we provide supplemental materials for *DHH Users' Interest in Sign-Language Interaction with Personal-Assistant Devices* (chapter 3).

B.1 Interview Study Demographics Questionnaire

Figure B.1 shows the questionnaire used for collecting participant demographics in the "interview study".

Participant Code: _____

NAME: _____ DATE: _____

EMAIL: _____

INFORMATION ABOUT YOU

Gender: Male Female Other: _____

How old are you? _____

When did you first take an ASL class? _____

Where? Which class? What semester? _____

How long have you studied it, if applicable? _____

Are you a current university student or alumni? What year and major? _____

Which describes you best? hearing hard-of-hearing deaf/Deaf other: _____

When did you first learn ASL? (How old?) _____

Did/do your parents use ASL at home? _____

Are your parents deaf? _____

Other deaf family? _____

What languages do you use and how much? (For example: 100% English vs 75% English and 25% ASL vs 100% ASL, etc.)

At home: _____

At work/school: _____

Other connections to deaf community? (husband or wife or partner / friends / community / clubs):

_____

Figure B.1: Interview Study Demographics Questionnaire

B.2 Interview Study Demographics Data

Table B.1 shows the participant responses to the demographics questionnaire (fig. B.1).

ID	Gender	Age	When did you first take an ASL class?	Where? Which class? What semester?	How long have you studied it, if applicable?	Are you a current university student or alumni? What year and major?	d/D/HH?	When did you first learn ASL? (How old)	Did/do your parents use ASL at home?	Are your parents deaf?	Other deaf family?	What languages do you use and how much? [At home]	What languages do you use and how much? [At work/school]	Other connections to deaf community?
1	M	18	N/A	N/A	N/A	Yes, a current university student. 1st year marketing	d/D	One year and half	Yes	No	2 deaf brothers and 2 deaf cousins	ASL and English (50% both)	ASL 75% and 25% English	Partner and Friends
2	F	24	Never took it	N/A	N/A	Alumni	d/D	i think 4 years old	No	No	No	80% English Vs 20% ASL	Work: 100% English, School: 100% ASL	N/A
3	NB	24	Summer 2013	New Signer Program	7 days	Student, MS Professional Studies (Project Management and Enterprise)	HH	18	No	No	No	0% ASL	60%ASL	Clubs / work / friends / alumni
4	M	24	4-5 age	Iowa School for the deaf- ASL class	One year	Current university student (6th year CIT majors)	D	∼4	Yes	No	One deaf aunt	75% ASL VS 25% English	50% ASL VS 50% English	Friends
5	M	24	5 years old	Idaho school for the Deaf	half of my whole life	alumni - 2019 (Business Administration)	HH	5 years old	Yes	dad-deaf, mom-hearing	only my family included my sister	ASL 50% English 50	ASL 80% and English 20%	Deaf club, deaf institution, deaf friends, conferences and move
6	F	21	5 years old	The learning center for the deaf	until i entered high school	4th year biomedical sciences	D	5 years old	sometimes	Yes	No	100% Pakistani sign language	50% English 50% ASL	School + university
7	F	22	2nd Grade	Rochester school for deaf	N/A	May 2019, business admin support tech	D	Since I were born	Yes	No	Yes, sister, aunt, uncle, grandma, grandpa	ASL 100% Mom-Spanish 40%	ASL 100%	Sports community and friends
8	M	22	5 years old	elementary school	17 years	Yes, fourth year mechanical engineering technology	d/D	5 years old	yes	No	My cousins and 2 brothers	30% ASL	70% ASL	Clubs
9	F	21	N/A	N/A	N/A	4th year Biomedical sciences	d/D	Birth	Yes, partially	Mom is deaf	Mother's siblings and her parents and my parents and some cousins	50% ASL 50% English	75% ASL 25% English	Friends / Community / clubs
10	M	23	9th grade	Leigh high school, ASL	a year	3D digital design (3DDD)	HH	14	No	No	N	50% English 50% ASL	100% ASL	Partner, friends, community, clubs, teachers, work
11	F	23	Pre-school	Deaf program	maybe 2 years	university student graduate 2020 / graphic design	d/D	7	yes	no	my twin sister	90% ASL 10%English	100% ASL	Friends, club, partner
12	F	25	16 years old	highschool junior + senior year	1 and half years	alumni, BS-MS Biomedical sciences 2017, Professional Studies 2019	d	16 years old	some (for 1 year when i was a baby)	No	No	100% english	50% ASL and 50% English	Friends, club, boyfriend
13	F	36	Since Birth	N/A	N/A	Alumni 2018 - MIS	HH	since birth	No	No	Deaf brother and Hard of Hearing sister	ASL 100% Mom-Spanish 40%	100% ASL	Deaf brother and Hard of Hearing sister
14	F	23	a few months	day care	23	6th year mechanical engineering	d/D	a few months yes	yes	no	deaf brother	50/50 English - ASL	25/75 English ASL	Friends, community
15	F	24	N/A	N/A	N/A	5th and Communications	D	2 years old	Yes	No	Yes, my brother	100% ASL	100% ASL	Friends / Community
16	F	25	N/A	N/A	N/A	Alumni Biology 2017	d/D	1 year old	my mom	no	-	50% English 50%ASL	50% English 50%ASL	90ASL / 10 English deaf events
17	M	28	Since Birth	N/A	N/A	No	HH	since birth	Yes	No	N/A	85% ASL 15% English	75% ASL 25% English	Friends and deaf community and club
18	M	24	2014	LATTC (Los Angeles) ASL1 Fall Semester	One semester	Three year and half and accounting major	d/D	11 years old	yes only my mother and 2 brothers	No	No	90% ASL 10% English	95%ASL and 5% English	
19	F	29	N/A	N/A	N/A	2019 Student Project mgmt and modeling and communication	d/D	19 and Norwegian Sign Language	Norwegian sign language	No	no	ASL 10%English 50%Norwegian	ASL	Friends
20	F	23	middle school teens	Atmy Hometown Brooklyn	I was in deaf institute middle school	University student - 3rd year and Web and mobile Computing	HH	1-2 years old	a few times / sometimes	No	No	90% English and 10% ASL	50% English and 50% ASL	50% English and 50% ASL
21	M	23	8 months old	N/A	N/A	6 sois in civil engineer, sustainable designer	d/D	8 months old	yes home ASL	No	Aunt	50% English 50%ASL	75%ASL 25% English	NTID / Sigma nv, committee under NSC

Table B.1: Demographic data from the interview study

B.3 Interview Study Questions

Figures B.2 to B.5 show the "interview study" interview questions.

1. Familiarity
    - Are you familiar with personal assistant technologies, such as Amazon Alexa, Google Home, or Microsoft Cortana?
        - If yes -
            1. Where have you seen it?
            2. Do you own one of these devices?
            3. How often do you use it?
        - If no -
            1. *[Print an image of the devices and show the pictures]* These are a few popular personal assistant devices.
            2. *[Video]* If you want I could you show you a small video - https://www.youtube.com/watch?v=ufs_aDDIgIY
            3. *[Explanation]* I can explain what the device does - Amazon Echo Show is a smart speaker that is part of the Amazon Echo line of products with a touchscreen display that can be used to display visual information to accompany its responses, as well as play video and conduct video calls with other Echo Show users. The echo users usually give voice commands to the device and the device provides information on the screen at the same time answering using voice output.

2. Usage History
    - Have you used the device? *[If yes, ask the follow-up questions]*
        - In what ways have you used it?
        - How did you interact with the device, e.g. with your voice, by typing, etc.?
        - Did you face any problems? Can you tell me about it?
        - Does the device work as per your expectations?
    - *[If no, ask these follow-up questions]*
        - Will the device be useful for you in any way? How?
        - How would you want to interact with the device?

2. Usage Expectations
    - Would you be interested in a personal assistant device that would allow you to communicate with it using American Sign Language (ASL)?
        - Why do/don't you think so?
    - Can you imagine how you would use a device like this, e.g. in your home, at work, or in other settings?
    - For the device to understand the ASL commands, it will use the camera function. Can you tell me where you would place the device in your house and where would you be present when interacting with it in ASL?
    - Can you think of other ways how the personal assistant device will be of any help to you?
        - *(If yes)* Can you say the commands for me in ASL?

Figure B.2: Interview Study Questions – page 1 of 4

- Can you imagine any situation where you might use a personal assistant device for something that hearing people would not do?

3. Commands
   - What commands would you like your device to understand?
   - Can you suggest some commands you might give it? Can you say the commands for me in ASL?
   - How often do you see yourself using these devices (e.g. every day, once a week, once a year, never)?
   - Research has shown that the most common commands that users in the U.S. issue to personal assistant devices include: playing music, asking about the weather, playing a fun quiz game, asking questions about facts, getting directions, setting alarms/reminders, or shopping. Would you be interested in any of these features? Why or why not?
   - Can you categorize the use of the commands based on how frequently would you use them [Always, Often, Occasionally, Rarely, Never]
     - Calendar - Create a reminder, set alarm, check the calendar
     - Communication - Make a phone call, send a text message
     - Location - find or navigate to a certain place
     - Device Control - Launch apps
     - Requesting information or facts - Weather
     - Taking notes, maintain lists
     - Entertainment - Playing a quiz game
     - Connecting with other smart devices
     - Shopping
   - I have a list of the commands which are being popularly used for interacting with these devices. Can you mention which of the commands you are useful for you? *[Show the commands document]*

4. System Acknowledgment
   For the next few questions, I would like to ask you questions about how would you want the personal assistant devices to interact with you.
   We have broken down the process of interacting with the device in 3 stages. I will show you a video after I explain all the 3 stages.
   First is when you call Alexa also known as “waking up the device”, second when you give the command and the device reacts to let you know that it is listening and third when the device responds to the command.
   Please see this video - https://www.youtube.com/watch?v=LYS4QpGOFDs
   If required explain the 3 stages while viewing the video -
   (0.02sec - Alexa wakes up
   0.04sec - Alexa acknowledges the command
   0.06sec - Alexa executes the command)

Figure B.3: Interview Study Questions – page 2 of 4

I will ask you a few questions about how you want the device to be reacting to your commands.

Wake up the system -

- In the case of ASL interactions with the device, how would you wake up the system and how do you want the system to let you know that is it ready for the command?

- We identified a few possible ways in which you can wake up the system
  - Contactless wake up - Similar to how to get a person's attention, you can wave your hand and get the devices' attention.

  - Touch and activate - Using a mobile app or a smartwatch, you can turn ON the device and then interact with it. It will be similar to system where the mic option needs to be pressed for the system to start listening.

- Which of the above-mentioned ways would you choose and why?

Listening State

- Can you imagine how the system would act when you are signing?
  - How do think you will know that the system is actually listening to what you are saying? For example, English translations of your command would be generated on the device screen.

Response from the system

- How do you want the device to show you results and answer your queries? For example - ASL animation, text on the screen, pictures, audio, ...

- Would you be interested in a personal assistant device that was able to show you sign language video or animations on the screen?

5. Concerns -
   - Do you have any concerns using ASL interactions with personal assistant technologies?
     - In terms of placement of the device
     - In terms of waking up the device
     - In terms of reading or understanding the response
     - In terms of connecting the device to other devices
     - In terms of the device understanding the commands

Figure B.4: Interview Study Questions – page 3 of 4

- To interact with personal assistant devices using ASL, the device camera will be on to understand what you are signing. Do you have any concerns regarding that?

6. Do you have any additional comments that you want to share with me?

Some resources used to create these questions and supplementary files:
Resources - https://www.businessinsider.com/s?q=siri-vs-google-assistant-cortana-alexa
A resource explaining the type of questions personal assistants answer - https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/3549/1661
Understanding User Satisfaction with Intelligent Assistants
https://dl.acm.org/citation.cfm?id=2854961
Rating the smarts of the digital personal assistants in 2019
https://www.perficientdigital.com/insights/our-research/digital-personal-assistants-study
System Acknowledgement -
https://dl.acm.org/citation.cfm?id=3291783
Commands -
https://doi.org/10.1177/0961000618759414
https://dl.acm.org/citation.cfm?id=2741669

Figure B.5: Interview Study Questions – page 4 of 4

B.4 Affinity mapping of interview transcripts

Figure B.6 shows the affinity mapping thematic analysis of the interview transcripts from the "interview study".

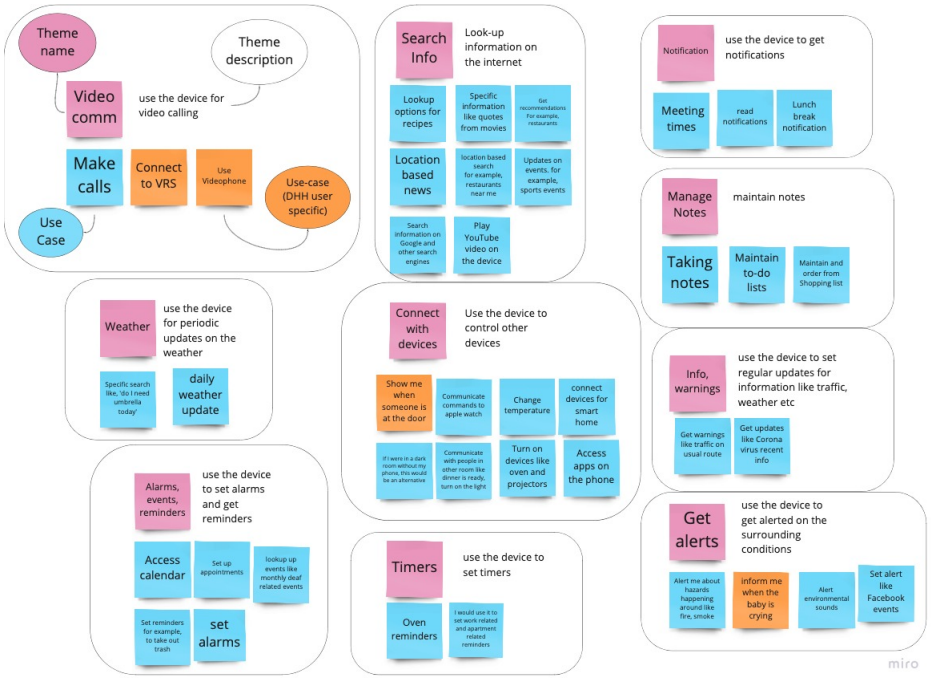


Figure B.6: Affinity mapping of interview transcripts

B.5 Affinity mapping of participant usage suggestions

Figure B.7 shows the affinity mapping thematic analysis of the participant-suggested usage of personal assistants.

Figure B.7: Affinity mapping of participant usage suggestions

B.6 Sample affinity mapping: use-case of connecting with other devices

Figure B.8 contains visualization of an Example of the affinity mapping of the use-case of connecting with other devices.

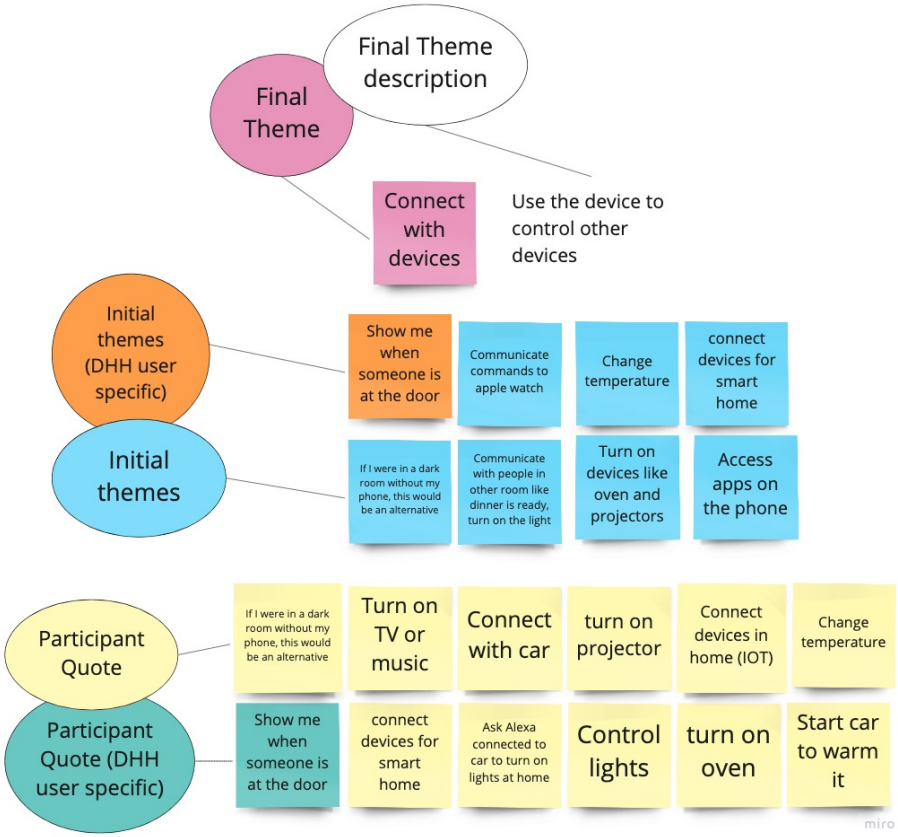


Figure B.8: Sample affinity mapping: use-case of connecting with other devices

B.7 Survey Study Questionnaire

Figures B.9 to B.30 show the online survey questionnaire that was used in the "survey study" portion.

ASL Personal-Assistant Interaction Survey

You are being asked to participate in a research study because you are someone who identifies as Deaf or Hard of Hearing (DHH). The goal of this research is to understand the interests and preferences of people who are DHH in regard to personal assistant technologies, such as Amazon Alexa, Google Home, etc., especially if such technologies may one day be able to understand sign language. Participation in the project consists primarily of answering questions about your interest in using personal assistant devices and how you might use such technology.

If you decide to participate, we will ask you some questions about your professional and educational background, your familiarity and interest in using personal assistant technologies, and how you might imagine using this technology, especially if it could understand questions or commands in sign language.

Finally, we would like to thank you for investing your time in helping us with our research.

The primary investigator of the project is Matt Huenerfauth, Ph.D., Professor, Golisano College of Computing and Information Sciences, Rochester Institute of Technology, matt.huenerfauth@rit.edu

* Required

1. *Mark only one oval.*

- ◯ I agree to participate
- ◯ I do not agree to participate

ASL Translations

ASL Translations have been made for each of the questions in this survey. The videos for the questions appear before each question. You do not have to watch the videos if you don't want to.

Demographics

This section will collect basic, standard information about you. This survey will not collect your name and is anonymous.

http://youtube.com/watch?v=3wb6J4hyzps

Figure B.9: Survey Study Questions – page 1 of 22

2. 1. What is your gender? (Select "Other:" if you prefer to self-describe) *

   Mark only one oval.

   - ◯ Woman
   - ◯ Man
   - ◯ Non-binary
   - ◯ Prefer not to disclose
   - ◯ Other: _____

 http://youtube.com/watch?v=ZlQn3aFXcaQ

3. 2. How old are you? *

   _____

 http://youtube.com/watch?v=hfs6RdADQPM

Figure B.10: Survey Study Questions – page 2 of 22

4. 3. How do you describe yourself? *

   Mark only one oval.

   - ◯ Deaf
   - ◯ deaf
   - ◯ Hard of Hearing
   - ◯ Hearing
   - ◯ Other: _____

http://youtube.com/watch?v=YaCh_2lC2eM

5. 4. At what age did you become D/deaf or hard of hearing? *

   _____

http://youtube.com/watch?v=F5q1XgZmYWY

6. 5. At what age did you begin to learn ASL (e.g. informally from adults/parents or from school)? *

   _____

Figure B.11: Survey Study Questions – page 3 of 22

 http://youtube.com/watch?v=w1x-s_QZrEw

7. 6. Are your parents D/deaf or hard of hearing? *

Mark only one oval.

- ◯ Yes
- ◯ No

 http://youtube.com/watch?v=HUGlTtO8keM

8. 7. Did your parents use ASL at home? *

Mark only one oval.

- ◯ Yes
- ◯ No

 http://youtube.com/watch?v=JblZqEw9FKI

Figure B.12: Survey Study Questions – page 4 of 22

9. 8. Please consider your elementary school, i.e. the school you attended before age 12: In school, did you use ASL? *

Mark only one oval.

- ◯ Yes
- ◯ No

http://youtube.com/watch?v=gx07sxtXMww

10. 9. Please consider your elementary school, i.e. the school you attended before age 12: What type of school did you attend as a child? *

Mark only one oval.

- ◯ Residential school for Deaf students (school with dorms)
- ◯ Daytime school for Deaf students (commute from home to this school)
- ◯ Mainstream school (majority of students are hearing)
- ◯ Other: _____

http://youtube.com/watch?v=X1zUUA-TQrw

Figure B.13: Survey Study Questions – page 5 of 22

11. 10. What describes your current level of education? *

Mark only one oval.

- ◯ Did not graduate high school
- ◯ Graduated high school
- ◯ Graduated college
- ◯ Have bachelor's degree
- ◯ Have graduate degree

 http://youtube.com/watch?v=nIA0wQSY7cQ

12. 11. In what location (city, state, country) have you lived in the longest? *

_____

13. How long did/have you lived there?

_____

 http://youtube.com/watch?v=nKMIDpbqBsg

14. 12. How many people currently live in your household (including yourself)? *

_____

Figure B.14: Survey Study Questions – page 6 of 22

 http://youtube.com/watch?v=kDbFjI_padc

15. 13. How many people in your household use ASL daily (including yourself)? *

 http://youtube.com/watch?v=qx29bkly5XY

14. For the following three situations (home, work/school, friends/family), please answer the question: What languages do you use and how much? (For example: 100% English vs 75% English and 25% ASL vs 100% ASL, etc.)

16. At home: *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
100% ASL, 0% English	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	0% ASL, 100% English

17. [optional] Explain:

Figure B.15: Survey Study Questions – page 7 of 22

18. At work/school: *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
100% ASL, 0% English	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	0% ASL, 100% English

19. [optional] Explain:

20. With friends/family: *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
100% ASL, 0% English	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	0% ASL, 100% English

21. [optional] Explain:

Familiarity with the device

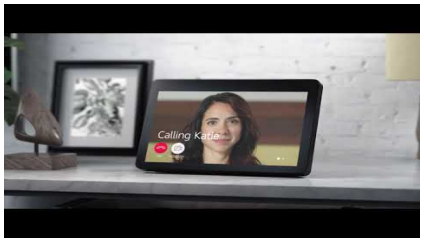
This section will ask you about your familiarity with smart personal assistants.

Figure B.16: Survey Study Questions – page 8 of 22

Images of Google Nest Hub and Amazon Echo Show

Google Nest Hub device Amazon Echo Show

[optional to watch] Here is a video demonstrating what the device is and what it can do.



http://youtube.com/watch?v=ufs_aDDIgIY

http://youtube.com/watch?v=UJj0TS2nDvE

22. 1. Have you ever seen smart personal assistant devices like Amazon Alexa or Google Home, which allow someone to give commands or to ask questions? *

Mark only one oval.

- ◯ Yes
- ◯ No *Skip to question 29*

Familiarity with the device

This section will ask you about your familiarity with smart personal assistants.

Figure B.17: Survey Study Questions – page 9 of 22

 http://youtube.com/watch?v=EF7JwKyUT_I

23. 2. Where have you seen it? *

_____

 http://youtube.com/watch?v=YfutU_oICLE

24. 3. Does your household have one of these devices? *

Mark only one oval.

- ◯ Yes
- ◯ No

 http://youtube.com/watch?v=pgGgVi5XrV4

Figure B.18: Survey Study Questions – page 10 of 22

25. 4. Do you personally own one of these devices? *

Mark only one oval.

- ◯ Yes
- ◯ No

http://youtube.com/watch?v=shDjm5uwAio

26. 5. Have you used it before? How often do you use it? *

Mark only one oval.

- ◯ Daily
- ◯ Weekly
- ◯ Monthly
- ◯ Yearly
- ◯ Less than once a year
- ◯ Never *Skip to question 29*

Usage experience

This section will ask about your experience about using smart personal assistants.

http://youtube.com/watch?v=pO1rZgIrV70

Figure B.19: Survey Study Questions – page 11 of 22

27. 1. In what ways have you used the device? (Select all that apply) *

Check all that apply.

- [ ] Ask weather-related questions (e.g., temperature, need for an umbrella, etc.)
- [ ] Set alarms, events, and reminders
- [ ] Set timers
- [ ] Get alerts (e.g., doorbells, smoke alarms)
- [ ] Search for information (e.g., recipes, movie times)
- [ ] Connect to other smart devices (e.g., lights, TV, cars)
- [ ] Video-based communication (e.g., video calling)
- [ ] Notifications (e.g., read, delete notifications)
- [ ] Information, Warnings (e.g., traffic, weather conditions)
- [ ] Manage notes (e.g., to-do lists, shopping lists)

Other: [ ] _____

http://youtube.com/watch?v=7KlHJya_L6k

28. 2. In what ways have you interacted with the device? (Select all that apply) *

Check all that apply.

- [ ] Speaking to the device with your own voice
- [ ] Using text-to-speech, e.g. typing text into your phone so that it is read aloud in a computer voice, which the device listens to
- [ ] Typing commands or questions on the on-screen device keyboard
- [ ] Typing the commands on your phone app connected remotely to the device, e.g. using the Amazon Alexa application
- [ ] Selecting among the suggestions provided on the screen of the device
- [ ] None of these

Other: [ ] _____

Interest and expected usage

This section will ask you questions about whether you are interested and how you would use the device.

Figure B.20: Survey Study Questions – page 12 of 22

 http://youtube.com/watch?v=eaR8-htWz0Q

29. 1. Please indicate whether you agree with this statement: I would be interested in using sign language interaction with a personal assistant device, such as Alexa or Google Home *

Mark only one oval.

- ◯ Strongly Disagree
- ◯ Disagree
- ◯ Neutral
- ◯ Agree
- ◯ Strongly Agree

 http://youtube.com/watch?v=BzI62QNQs8c

30. 2. These devices are often used by people who are hearing, and they often place the device in different rooms throughout their house, including the kitchen, the living room, or bedroom. If the device could understand American Sign Language (ASL) commands, where would you place the device and why? *

Figure B.21: Survey Study Questions – page 13 of 22

 http://youtube.com/watch?v=lFgVoAdGoio

3. Imagine you have a personal assistant device that understands ASL, can you please suggest some ideas of commands you would like to give it, questions you would like to ask, or things you'd like to do with it?

31. Idea 1

32. Idea 2

33. Idea 3

34. More ideas

Figure B.22: Survey Study Questions – page 14 of 22

http://youtube.com/watch?v=okOrVU6s9VQ

35. 4. Please indicate whether you agree with each of these statements: If the device could understand ASL, I would be interested in using the device in the following ways: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Ask weather-related questions (temperature, need for an umbrella, etc.)	◯	◯	◯	◯	◯
Set alarms, events, and reminders	◯	◯	◯	◯	◯
Set timers	◯	◯	◯	◯	◯
Get alerts (doorbells, smoke alarms)	◯	◯	◯	◯	◯
Search for information (recipes, movie times)	◯	◯	◯	◯	◯
Connect to other smart devices (lights, TV, cars)	◯	◯	◯	◯	◯
Video-based communication (video calling)	◯	◯	◯	◯	◯
Notifications (read, delete notifications)	◯	◯	◯	◯	◯
Information, Warnings (traffic, weather conditions)	◯	◯	◯	◯	◯
Manage notes (to-do lists, shopping lists)	◯	◯	◯	◯	◯

Figure B.23: Survey Study Questions – page 15 of 22

http://youtube.com/watch?v=13hjyOzeWmo

36. 5. The previous question asked you to consider how interested you were in these types of commands. Now, please consider how often would you use the following commands, if your device could understand ASL: *

Mark only one oval per row.

	Daily	Weekly	Monthly	Yearly	Less than once a year	Never
Ask weather-related questions (e.g., temperature, need for an umbrella, etc.)						
Set alarms, events, and reminders						
Set timers						
Get alerts (e.g., doorbells, smoke alarms)						
Search for information (e.g., recipes, movie times)						
Connect to other smart devices (e.g., lights, TV, cars)						
Video-based communication (e.g., video calling)						
Notifications (e.g., read, delete notifications)						
Information, Warnings (e.g., traffic, weather conditions)						
Manage notes (e.g., to-do lists, shopping lists)						

Figure B.24: Survey Study Questions – page 16 of 22

Commands

This section will ask you about commands you would use with the device, if it could understand ASL.

http://youtube.com/watch?v=qgoud58d1bg

1. Considering the device can understand ASL, for each of the following scenarios, can you think of a specific command that you would give to the device? For example in the case of 'search for information', a possible command for the device can be 'Can you show me some chicken pasta recipes'. Please feel free to share as many ideas as you would like for each category!

37. Ask weather-related questions (e.g., temperature, need for an umbrella, etc.)

38. Set alarms, events, and reminders

Figure B.25: Survey Study Questions – page 17 of 22

39. Get alerts (e.g., doorbells, smoke alarms)

40. Search for information (e.g., recipes, movie times)

41. Connect to other smart devices (e.g., lights, TV, cars)

42. Video-based communication (e.g., videophone/VRS)

43. Notifications (e.g., read, delete notifications)

Figure B.26: Survey Study Questions – page 18 of 22

44. Information, Warnings (e.g., traffic, weather conditions)

45. Manage notes (e.g., to-do lists, shopping lists)

46. Other ideas

Response

This section asks about how the device would display it's responses and results.

http://youtube.com/watch?v=5pNiFgowNqA

Figure B.27: Survey Study Questions – page 19 of 22

47. 1. How do you want the device to show you results and answer your queries? (Please select as many options as you want) *

Check all that apply.

- [ ] Text output on the screen
- [ ] ASL animation shown on the screen
- [ ] Videos
- [ ] Photos
- [ ] Drawings
- [ ] Computer-generated speech from the device
- [ ] Sound effects or audio

Other: [ ] _____

http://youtube.com/watch?v=2BCK2Yk5S4s

48. 2. Please indicate whether you agree with this statement: I would be interested in a personal assistant device that was able to show you sign language video or animations on the screen *

Mark only one oval.

- ◯ Strong Disagree
- ◯ Disagree
- ◯ Neither agree nor disagree
- ◯ Agree
- ◯ Strongly agree

Concerns

This section records your concerns about using smart personal assistant devices that understands ASL input.

Figure B.28: Survey Study Questions – page 20 of 22

Instructions for this section:

http://youtube.com/watch?v=odMzcxp4oFE

http://youtube.com/watch?v=AHcpm_zpuRU

49. 1. Please indicate whether you agree with these statements: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
From a privacy perspective, I would be concerned about having a device with a camera.					
The device might pick up on some signs that were not meant for it.					
It is important to have an option of turning off the microphone sometimes for privacy.					
It is important to have the option of turning off the camera sometimes for privacy.					
It is important to have a physical cover to block the camera sometimes for privacy.					

Figure B.29: Survey Study Questions – page 21 of 22

Raffle

Participants who complete the survey will be entered into a raffle for a $100 gift card. The chances of winning the raffle are approximately 1 in 40.

The raffle is optional. Your name and email address will only be used to contact you for the raffle. The survey itself is anonymous.

Please provide your name and email address so we can contact you if you win the raffle!

50. Name

51. Email address

Feedback

52. [not required] Do you have anything else you want to say? Do you have feedback about this survey? Let us know here!

53. [not required] Would you like to be contacted for future research studies? Provide your contact information here:

Figure B.30: Survey Study Questions – page 22 of 22

B.8 Survey Study Demographics Data

In this section are aggregated participant responses to *selected* questions from the "survey study" questionnaire (figs. B.9 to B.30).

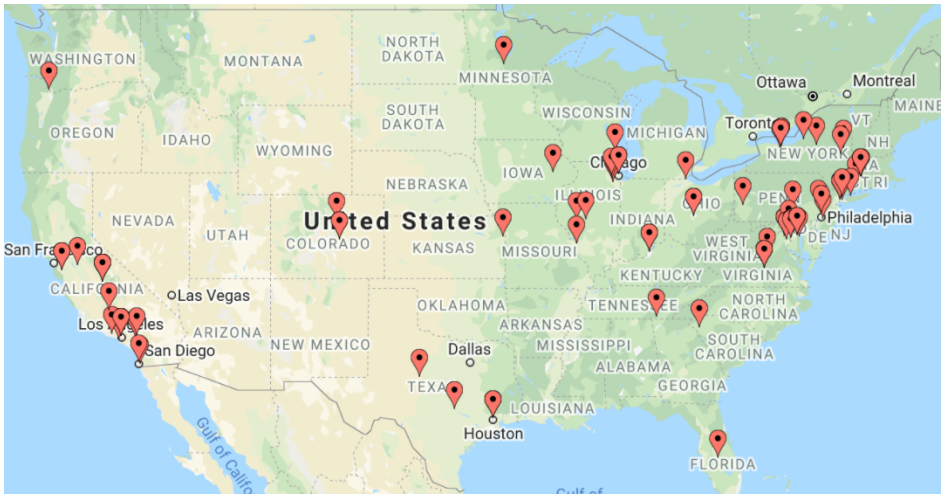


Figure B.31: Map showing locations of survey respondents

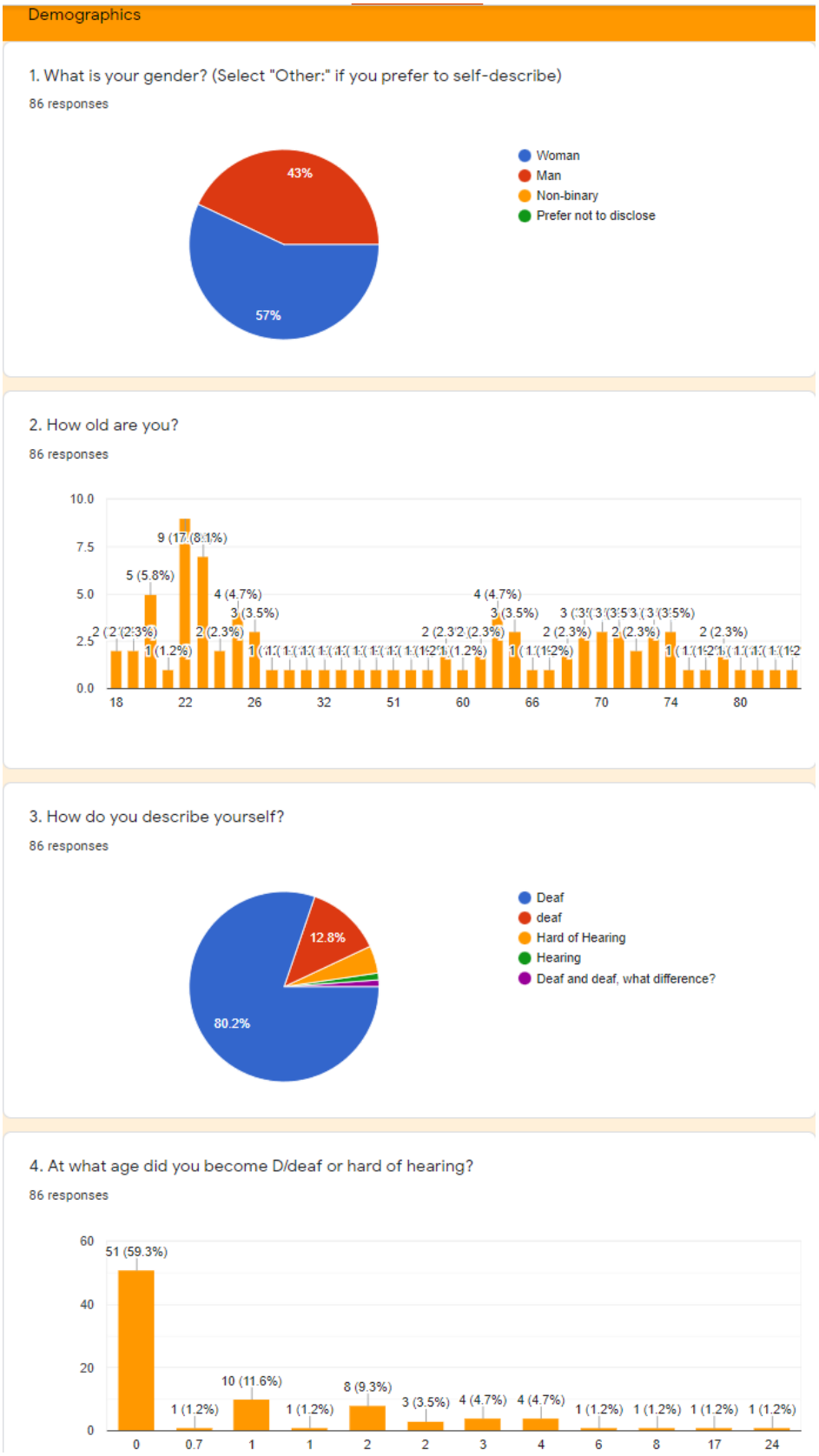


Figure B.32: Survey responses for gender, age, d/D/HH?, age became d/D/HH

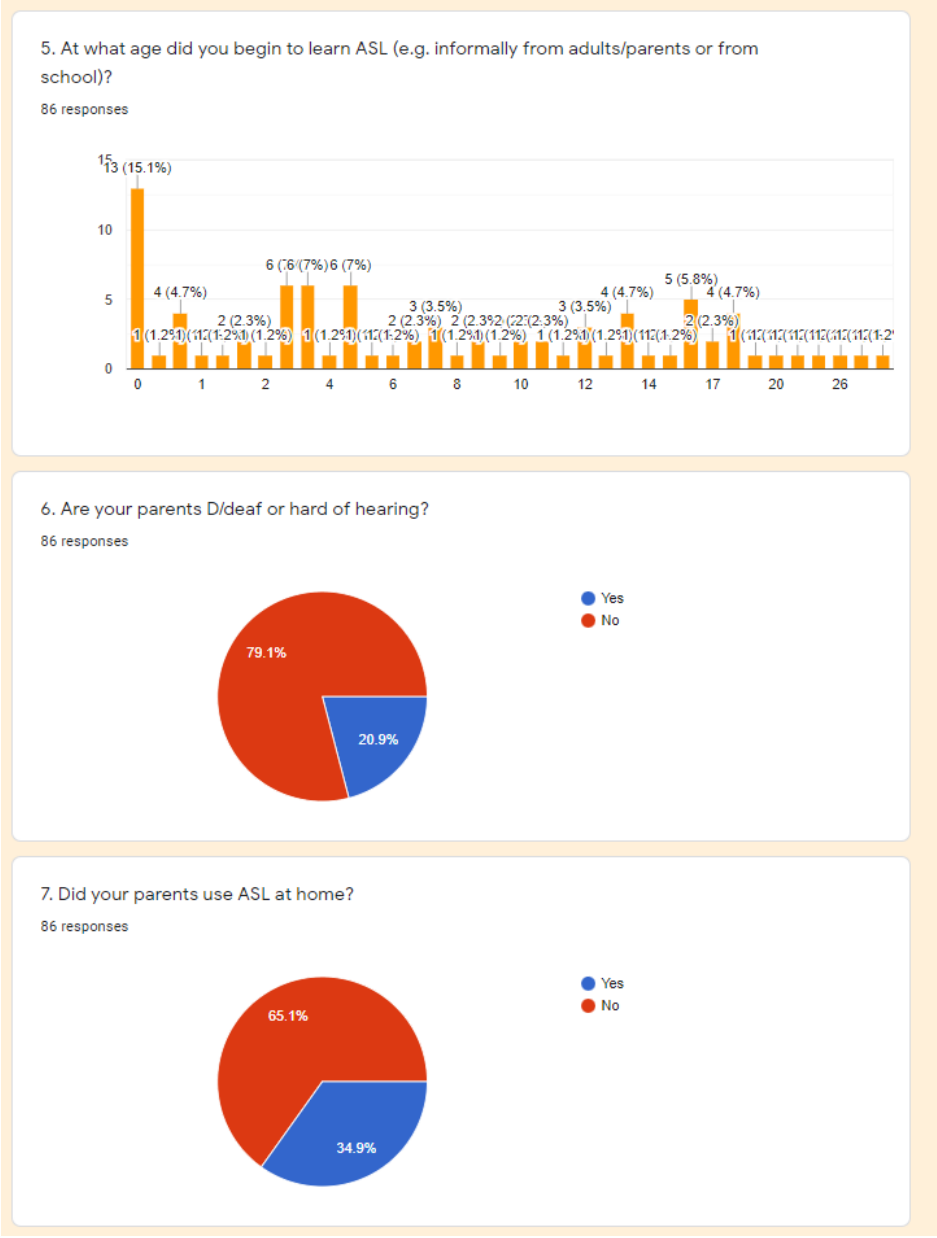


Figure B.33: Survey responses for age learned ASL, DHH parents, ASL-using parents

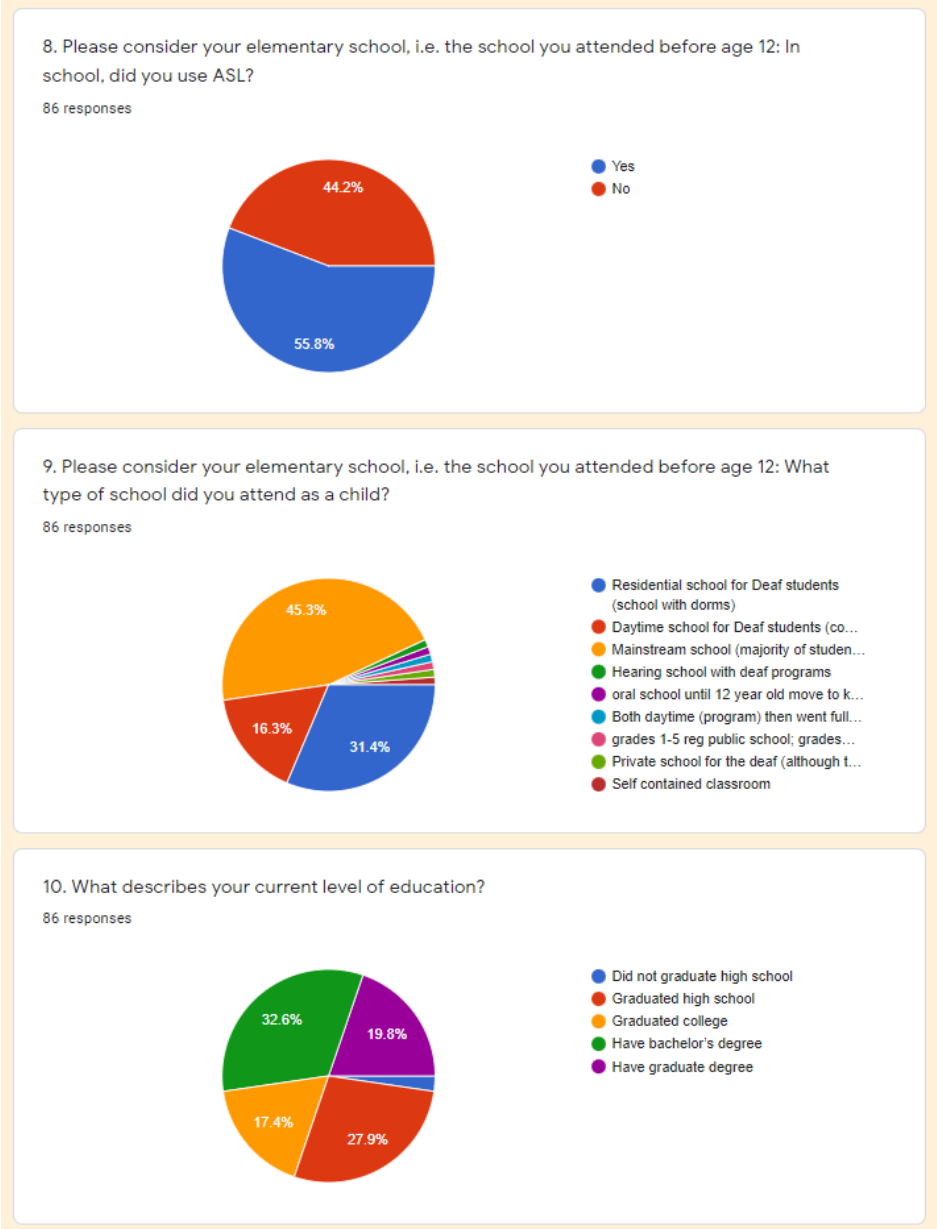


Figure B.34: Survey responses for ASL in elementary school, type of school, education level

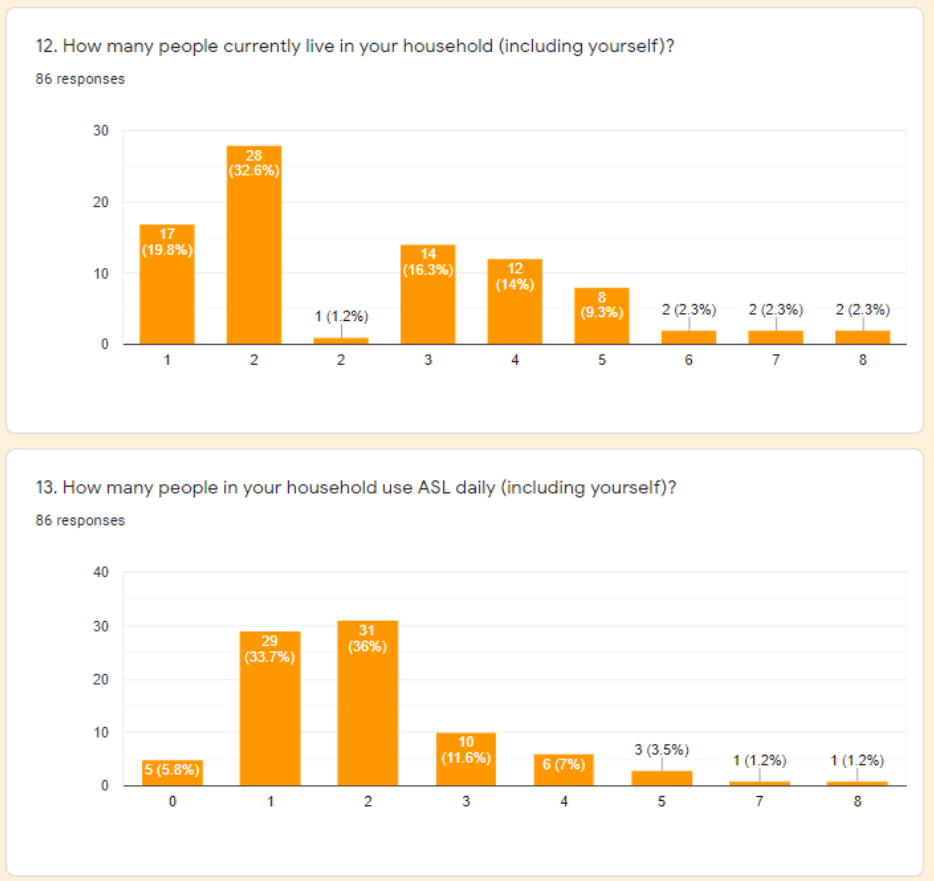


Figure B.35: Survey responses for number of household members and ASL usage

Figure B.36: Survey responses to using ASL vs English in home/work/school/friends/family

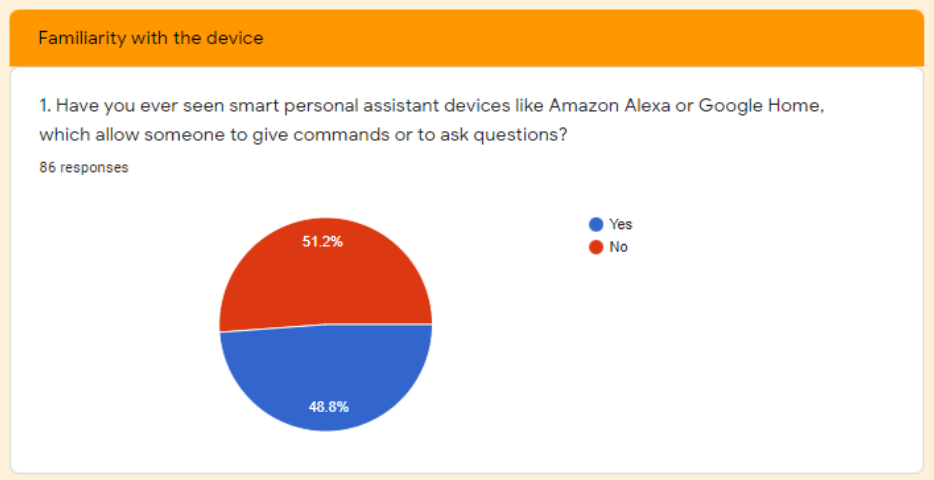


Figure B.37: Survey responses to seeing personal assistants before

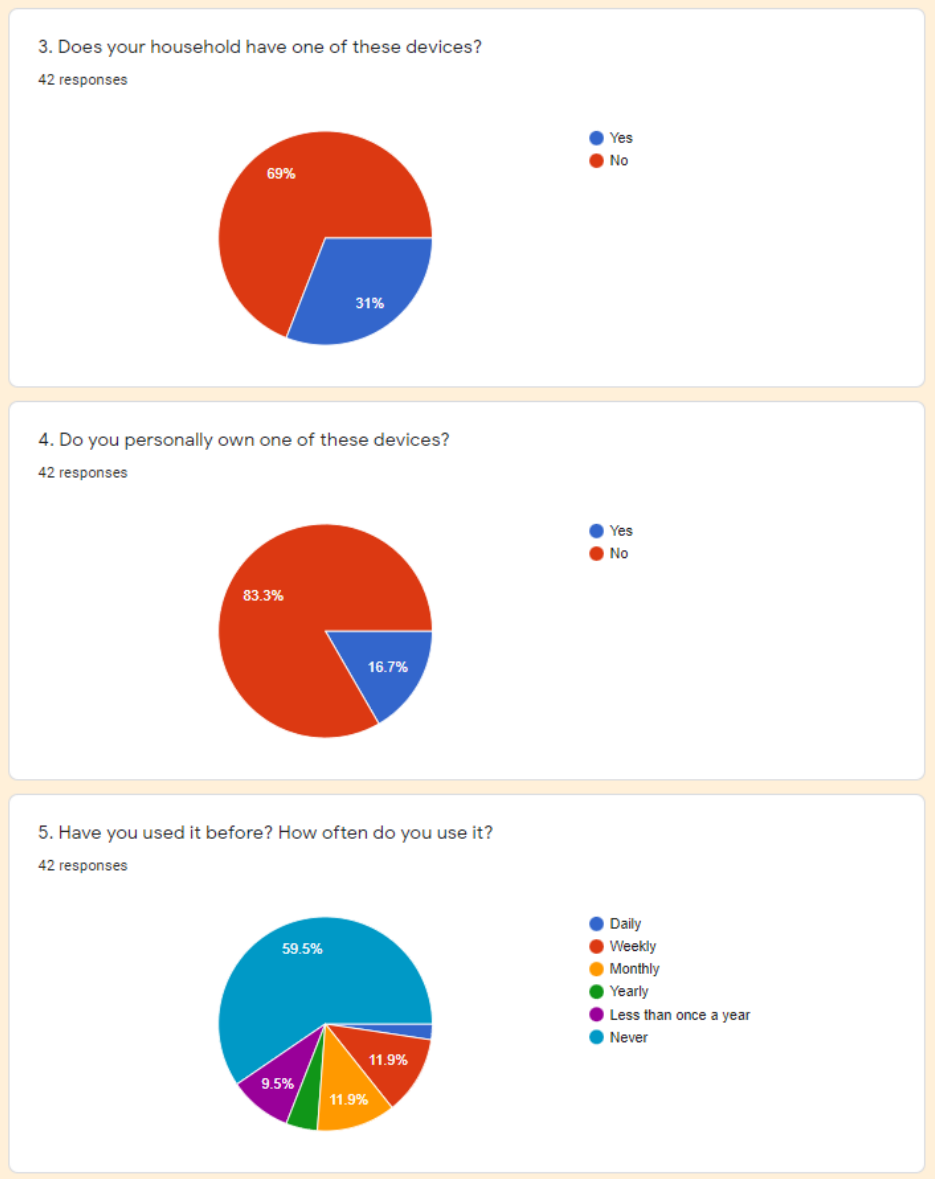


Figure B.38: Survey responses to household or personally-owned device and usage

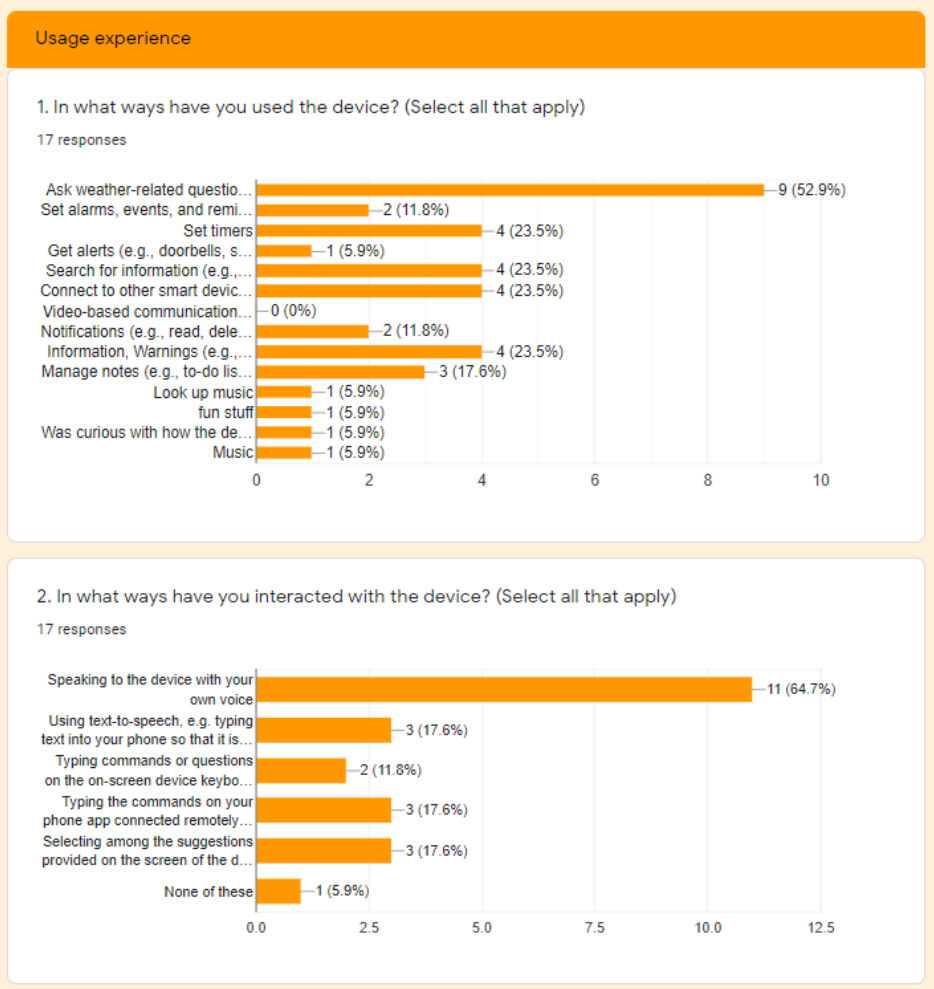


Figure B.39: Survey responses to using the device

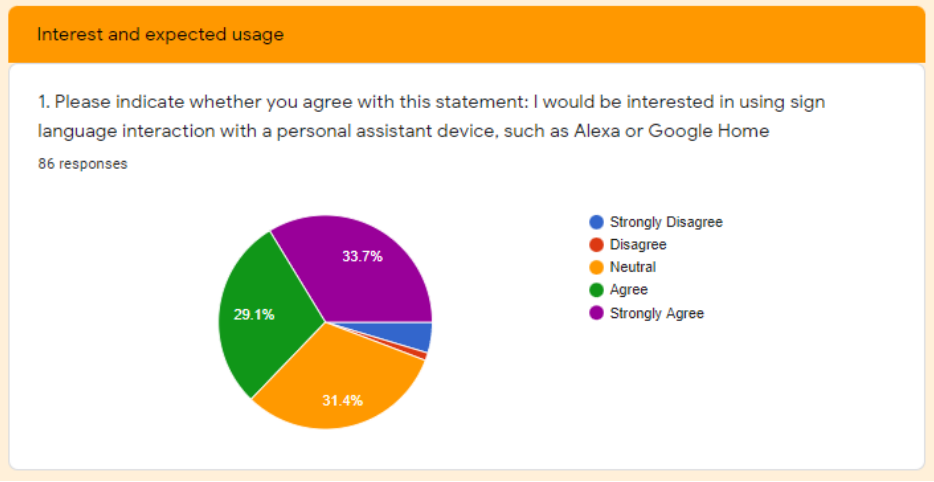


Figure B.40: Survey responses to interest in using a device that can understand ASL

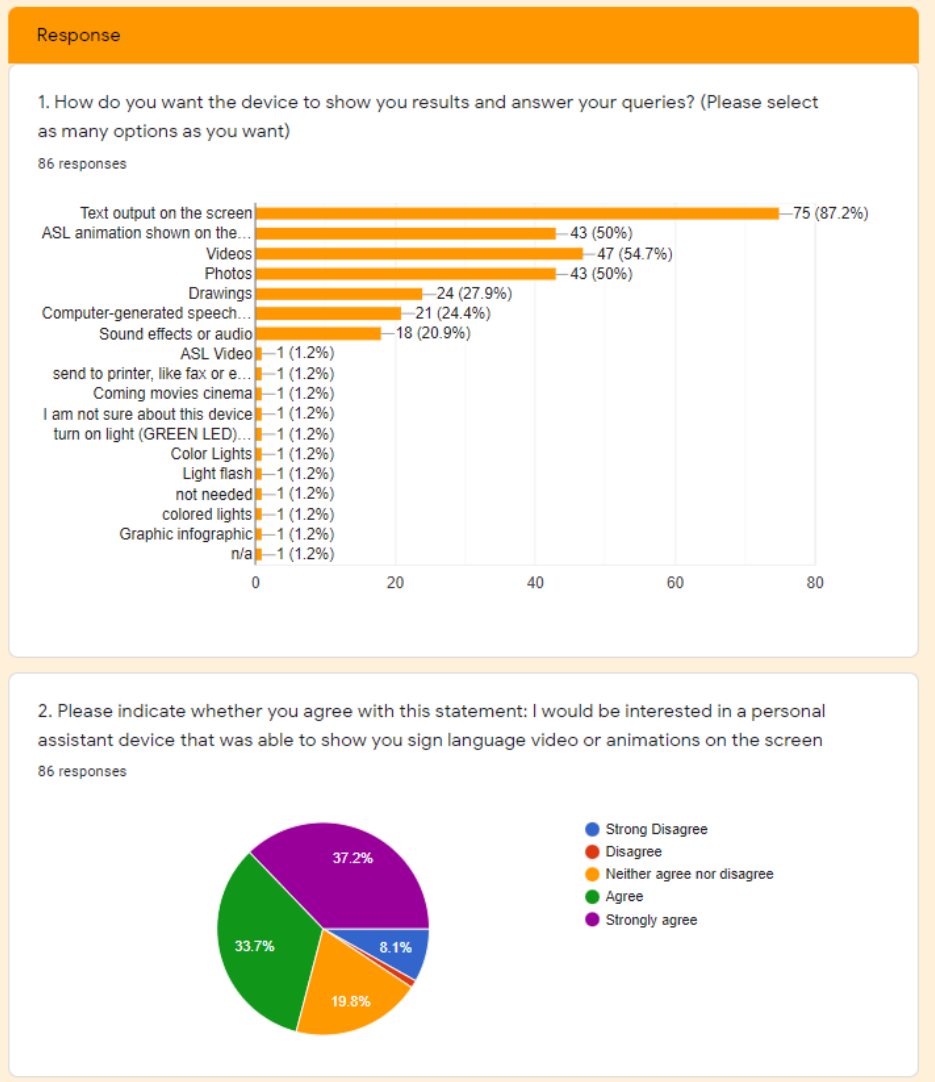


Figure B.41: Survey responses to how the device should show results to the user

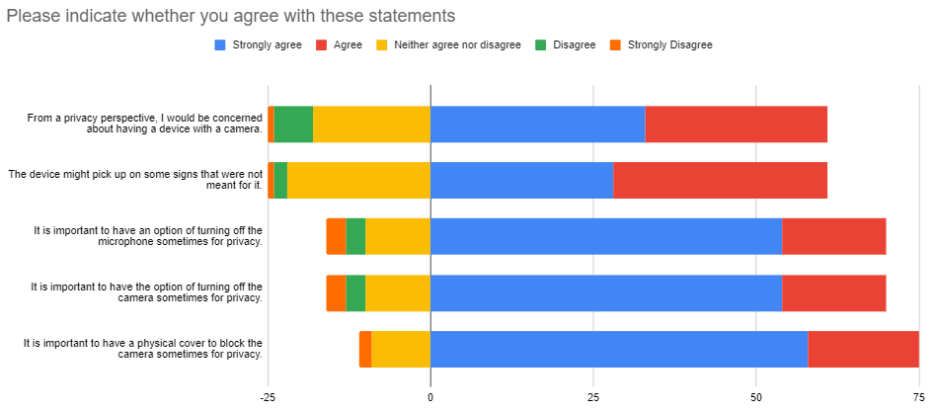


Figure B.42: Survey responses to concerns about having a personal assistant device