Effect of Sign-recognition Performance on the Usability of Sign-language Dictionary Search

SAAD HASSAN, OLIVER ALONZO, ABRAHAM GLASSER, and
MATT HUENERFAUTH, Rochester Institute of Technology

Advances in sign-language recognition technology have enabled researchers to investigate various methods that can assist users in searching for an unfamiliar sign in ASL using sign-recognition technology. Users can generate a query by submitting a video of themselves performing the sign they believe they encountered somewhere and obtain a list of possible matches. However, there is disagreement among developers of such technology on how to report the performance of their systems, and prior research has not examined the relationship between the performance of search technology and users' subjective judgements for this task. We conducted three studies using a Wizard-of-Oz prototype of a webcam-based ASL dictionary search system to investigate the relationship between the performance of such a system and user judgements. We found that, in addition to the position of the desired word in a list of results, the placement of the desired word above or below the fold and the similarity of the other words in the results list affected users' judgements of the system. We also found that metrics that incorporate the precision of the overall list correlated better with users' judgements than did metrics currently reported in prior ASL dictionary research.

CCS Concepts: • **Human-centered computing** → **Accessibility design and evaluation methods**; • **Information systems** → *Search interfaces;*

Additional Key Words and Phrases: American sign language (ASL), dictionary, search, information retrieval

ACM Reference format:
Saad Hassan, Oliver Alonzo, Abraham Glasser, and Matt Huenerfauth. 2021. Effect of Sign-recognition Performance on the Usability of Sign-language Dictionary Search. *ACM Trans. Access. Comput.* 14, 4, Article 18 (October 2021), 33 pages.
https://doi.org/10.1145/3470650

1 INTRODUCTION

Nearly one in six adults in the U.S. are **Deaf or Hard of Hearing (DHH)** [4], and about 500,000 people use **American Sign Language (ASL)** as a primary form of communication [26]. Increasing knowledge of ASL may facilitate better communication and inclusion of people who are DHH. In addition, there is growing interest among hearing individuals in learning ASL: ASL is currently

All authors contributed equally to this research.
This material is based upon work supported by the National Science Foundation under Award No. 1763569.
Authors' addresses: S. Hassan, O. Alonzo, and A. Glasser, Golisano College of Computing and Information Sciences, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY, 14623; emails: {sh2513, oa7652, atg2036}@rit.edu; M. Huenerfauth, School of Information, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY, 14623; email: matt.huenerfauth@rit.edu.

1936-7228/2021/10-ART18 $15.00
https://doi.org/10.1145/3470650

the third most studied language at U.S. universities, with the enrollment in ASL courses estimated to be between 100,000 and 200,000 students during academic year 2016 [12]. In addition to the benefits of there being more members of society who can communicate in ASL, parents of DHH children have a particular motivation to learn ASL, as prior research has found positive educational outcomes for DHH children whose parents learn some ASL [25, 35]. Prior research has also found strong interest among these users in technology supporting their ASL acquisition [39].

For written languages, it is relatively straightforward to lookup an unfamiliar word in a dictionary, and for spoken languages that use a standard writing system that has some correspondence to the spoken form, it is also generally possible for someone to look-up a unfamiliar spoken word in a dictionary. In contrast, ASL lacks a commonly used writing system, and if an ASL learner encounters an unfamiliar sign, they cannot type a string to search for it. Moreover, there is no one-to-one correspondence between ASL and English words or any standard convention for English-based gloss labelling of ASL signs. While most ASL dictionaries list signs in an alphabetical order based on their closest English translations, a user who does not know the sign or its closest English translation would not be able to search for it. Some web dictionaries or linguistic tools employ other methods to enable the users to look up an ASL word, including sorting the signs based on different physical features of a sign. For example, some dictionaries sort signs based on the handshape (configuration of the fingers), with some handshape listing defining a sort-order provided at the beginning [36]. Other dictionaries also enable the user to submit a query using other properties of the sign such as the number of hands used, orientation, movement, and so on [7, 22, 28]. Despite these more advanced methods for searching an ASL sign, users still often need to browse a list of possible "matches" to find the word they seek, or a system may erroneously fail to provide a match to the desired word.

Research on developing technology for sign language recognition from video is ongoing [37]. While the ability of existing systems to understand entire ASL sentences still remains limited, based on recent improvements in computer-vision and linguistic technologies, researchers have begun to build systems that could automatically analyze a video of an individual sign and seek a match for this sign in a dictionary collection [10, 38]. Once fully implemented, this technology could allow the users to sign in front of a camera (from memory) or submit a clip of a video of someone else performing the unfamiliar sign, to search for a word in an ASL dictionary. Afterward, the user can browse a list of search-results to look for their desired word. Previous research, e.g., References [3, 13], has investigated prototypes or provided proof-of-concepts of such systems. Other researchers, e.g., References [5], have also gathered and investigated users' requirements for an ASL dictionary. However, to date, prior published research has not included usability evaluations of these camera-based sign-language dictionary-search systems. No prior work has established requirements *for the designers of sign-recognition technologies*. Without such empirical studies, researchers investigating this underlying technology cannot gauge the level of accuracy or precision that the sign-recognition technology must have to support a dictionary-search application.

There are two main contributions of this study. First, we have identified a set of properties of the output of an ASL dictionary-match algorithm that affect users' perception of the system's quality, including: **placement of the desired result in the search-results list**, **whether the desired result lied above the fold or not** (on the first page before users have to scroll down or on later pages), and **the overall precision of the search results**. It is important to note that our study is not intended to guide developers of sign-match algorithms in regard to implementation details of their technology, but rather it sheds light on how to decide when this technology is ready to deploy, how to report the results of this technology, and how search-results interfaces that use this technology may present results. Specifically, these findings will also be useful for researchers reporting the performance of their dictionary systems so that they can assess the quality of results and know when the performance of underlying sign-recognition technology in their ASL

dictionary is of suitable accuracy for deployment with users. We evaluated whether these properties had an effect on users' judgement of an ASL dictionary system, by using a Wizard-of-Oz prototype in three different studies with a non-overlapping set of participants recruited for each:

- We investigated whether variation in where the desired word appears within the search results (i.e., top-5 or top-10), affects users' judgements about the system. We found that as the position rank of the desired result increases, users' satisfaction with the system's results-ranking and their perception of the overall relevance of the results decreases. This finding informs the creators of sign-recognition and matching algorithms on how to report the accuracy of their systems.
- We investigated whether users' satisfaction with the system dramatically drops if the desired result appears below the fold (after the first screen and thus requiring the users to scroll down). For creators of sign-recognition systems and matching algorithms, this finding informs what the key threshold should be for how high in the search results the desired item should appear. This also informs them about how to report their results. For **User Interface (UI)** designers of interfaces for such dictionaries, this finding suggests that the number of results displayed per page is an important design consideration, which may be based upon the performance of the sign-recognition technology used for this matching.
- We investigated whether the precision of the overall search results list affects users' judgement about the system. We found that varying the precision of the items in the list had a significant impact on users' judgement about the system, as measured by their satisfaction with the way the results were ranked and their perceived overall relevance of the search results. This result again has important implications for researchers on how to report the quality of sign-recognition systems in the context of dictionary search.

Finally, based on users' responses in the empirical studies above, we conducted an analysis to identify an information retrieval metric that correlates better with the users' perception of the system, as compared to a metric previously used among ASL dictionary-search researchers.

1.1 A Continuing Line of Research

This article is an extended version of a paper originally presented at the 2019 ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'19) [2]. The original conference paper presented two of the studies in this journal article (referred to as Study 1: Placement Study and Study 3: Precision Study in this article). The original conference paper had investigated the effect of placement of the desired word and overall precision of search-results list on the usability of video-based search interfaces for sign language dictionaries. At ASSETS'19, we received a thought-provoking question from an audience member as to whether another factor could also explain some of the variation in our users' judgment of the system: **Does a user's judgment about the quality of the system's output depend upon whether the sought-after sign appears on the computer screen before the user needs to scroll down to view lower results?** This issue is often described in information retrieval literature as the "above-the-fold" problem (using a paper newspaper metaphor). This question motivated us to conduct a follow-up study (referred to as "Study 2" in this article), in which we investigate if the users' satisfaction with the search system dramatically drops when the desired word does not appear above the fold. This study further consists of two sub-studies conducted with 16 participants each. The results from this new study provide additional guidelines for the creators of sign-recognition technology and designers of search interfaces as well.

In addition, this article describes our experimental studies in greater detail than was possible in the original conference paper. The Appendix provides examples of the user-interface of the Wizard-of-Oz prototype that we used. We have provided high-resolution images of each page of

the prototype at the end of this document. We are sharing videos of ASL signs that were used as prompts for participants by one of the co-authors on this article who is a fluent signer in online resources. In addition, this article includes additional materials in the Appendix and in electronic Supplemental Material, which had not been included in the original ASSETS'19 paper. Specifically, we share samples of the list of signs included in search results during our studies, detailed images of our prototype, and additional raw numerical data used in our correlation analysis. This additional information may be useful for future researchers who seek to replicate this work.

1.2 Structure of this Article

The article is structured as follows: Section 2 describes relevant linguistic information about American Sign Language, existing sign language dictionary systems and their evaluations, and some relevant prior work on information retrieval systems. Section 3 enlists our set of studies and research questions. Section 4 describes Study 1, originally presented at ASSETS'19, on the placement of the desired word in a results list. Section 5 describes the motivation, experimental protocol, and search results of the "above-the-fold" Study 2 mentioned above. Section 6 describes Study 3, originally presented at ASSETS'19 paper, examining the precision of the results list. Section 7 discusses our analysis of several information retrieval metrics that correlate output of matching algorithms with user judgement of ASL dictionary search systems. Section 8 discusses the overall results of this study, and finally, Section 9 outlines limitations and future directions of this work.

2 RELATED WORK

2.1 American Sign Language

ASL is a natural language that is primarily used among the community of people who are DHH in the U.S., Canada, and some other regions of the world. There are several other sign languages used throughout the world, e.g., **British Sign Language (BSL)** and **German Sign Language (DGS)**, and such languages are very rarely mutually intelligible. ASL linguists agree that individual ASL sign consists of a basic set of parameters: handshape (one of a set of approximately 90 different finger configurations of hand), orientation of the palms, location of the hand with respect to the body or in the signing space with a focus on the starting and ending of a sign, movement properties, and non-manual expressions such as facial expressions or bodily movements [8].

Estimating the total number of commonly used ASL signs is not a trivial task due to a number of reasons. ASL words are produced in different ways depending on the context. There are specialized jargon associated with certain fields. Signs have regional dialectical variations in how they are performed, as well as different productive methods for the formation of new words in ASL. The performance of each individual word may also vary based on how it is used in a sentence. The adjacent words in the sentence as well as overall spatial and grammatical aspects of the sentence effect the performance of individual ASL signs [23]. These linguistic aspects of ASL pose challenges for both the learners of the language and the designers of any technology that attempts to recognize ASL words from videos. Distinct internal structures in ASL signs are subject to various linguistic constraints that require recognition strategies that are different from other human activity recognition [40].

2.2 Inconsistency in Evaluating ASL Dictionary Search

Current ASL dictionaries use various input modalities to support the task of sign-language look-up, which can be broadly classified into two types search by feature selection and search by example. Search by feature selection dictionaries, e.g., Handspeak [22] or SLinto [34], allow users to create a query by selecting a set of linguistic features of the desired sign. Users can manually select parameters of a sign such as the handshape, the location, and the movement of the sign they are seeking

to formulate a search query. Previous research has explored how interfaces that support search by feature selection systems are often cumbersome to use, provide poor feature to word matching, and overly constrain how users select features. These findings have motivated researchers to develop machine learning-based systems and also improve the submission interface, by allowing users greater freedom in selecting features [5]. Some of the systems have been able to achieve an accuracy of desired word lying in top-10 results 84.93% of the times in experimental evaluations.

More recently, computer vision-based *search by example* dictionary systems have been proposed that enable users to search for a sign by demonstrating its motion in front of a color or depth camera, or by wearing motion-capture sensors gloves [3, 10, 11, 13]. Some initial research has shown that these systems have been able to correctly identify the desired ASL word in top-5 search-results in 97.6% of searches [24]. However, there is still inconsistency in how the accuracy results are reported. The results depend on at least two factors: the diversity of the human appearance and movement that the system is evaluated against and the size of the vocabulary, which can vary across different systems. Some of these accuracy results might be influenced by the size of data-sets used, which can be relatively small and the type of testing. User-dependent testing in which data from same participants is used in both testing and training can result in higher levels of accuracy.

Despite these advancements in technology, it is still pertinent to note that regardless of the input method for searching, sign-recognition remains a challenging task. A student who needs to search within an ASL dictionary may only vaguely remember a sign that they encountered somewhere. The student may struggle to accurately perform the sign, to search by example. Moreover, as discussed above there are a number of reasons why the appearance of an ASL sign might vary as it is used in the context in a sentence. Poor lighting while recording a sign, camera motion, diverse camera viewpoints, occlusions, poor video quality, and cluttered background can impact the quality of query submitted [31]. Owing to these reasons, it is unrealistic to expect that a dictionary-searching system will be able to identify an ASL sign from a student input and provide just one result. ASL dictionary systems thus provide users with a "page of search-results," which include a set of possible matches for the sign that they are seeking. The correct match might still not always be present at the top of search results.

The performance of dictionary-searching systems that are under development (e.g., ASL dictionaries) is typically measured using metrics of the percentage of trials in which the system satisfies a binary condition: whether the desired word is within the top-k results in the search-results list [10, 11, 24, 38]. The rank value of the desired word k in the search results list is used to make an evaluation of these systems. Among previously published papers in the field, there has been variation in what values of k should be reported, i.e., with some researchers citing value as low as top-4, while others reporting value up to top-375 results [3].

Few evaluations have been performed to determine how systems perform with potential users [10, 13]. However, to the best of our knowledge, no study has focused on how the performance of a dictionary-search system affect user satisfaction. This gap in past literature can explain lack of consensus about the reporting metrics mentioned above. Sign-language recognition research lacks a set of guidelines and requirements that can help us determine what level of performance the technology needs to achieve to support ASL dictionary search applications. There is a need to conduct empirical studies that can help us understand how the accuracy or precision of search results, and their presentation, may affect users' judgements about the performance and usability of a dictionary-search system.

2.3 Information Retrieval and Usability

Our goal is to understand how users' satisfaction may be affected by the performance of the automatic recognition component that underlies a dictionary search system. Given the similarities

of this setting to other forms of search, it is valuable to consider research in the field of information retrieval. Various metrics have been explored to characterize the performance of information retrieval systems and users' judgements of system quality, e.g., Reference [17]. One prior study investigated the relationship between users' satisfaction with search results and several metrics commonly used in information retrieval [1]. Examples of such metrics include accuracy, precision, or **Discounted Cumulative Gain (DCG)** [16]. Their findings suggest that the ranking of the results strongly correlate with the precision of results list, but no other metric significantly correlated with other user responses. In another study, a strong correlation was found between the relevance of results provided by a search engine and user satisfaction with the results [14]. It was also found that the nature of the query, i.e., whether it was informational, navigational, or transactional [6, 33], also affected how relevance correlated with satisfaction. Despite all these studies, we still do not know how these various metrics of quality of search output correlate with user satisfaction in the context of an ASL dictionary-searching system. We cannot directly generalize findings from other information retrieval studies here as well, since the task of looking up an ASL word is inherently different from the task of finding a website using a search engine: In this ASL dictionary search setting, a user may be attempting to find a single desired word that they (partially) recall, rather than search for a set of pages that satisfy an information query.

3 LIST OF STUDIES AND RESEARCH QUESTIONS

Based on our literature review of prior evaluations of ASL dictionary search systems, we identified a set of properties of the output of dictionary matching algorithms that may affect users' judgement of the system. Our next step was to conduct an empirical investigation into whether these factors influence users' satisfaction of ASL dictionary systems. Specifically, we focus on dictionary systems in which users are seeking a particular desired word in a dictionary, submit a video of themselves performing the word, and they view a list of search-results for that query. The results were in the form of short, looping videos of ASL signs.

Given current limits in accuracy or in vocabulary size of video-based ASL sign look-up technologies, Wizard-of-Oz methods (in which the underlying technology is mimicked for a study) are suitable for quickly testing user-interfaces [30]. We therefore implemented a Wizard-of-Oz prototype to simulate an ASL dictionary in which a user can look up the meaning of a desired word by performing it in front of a webcam. It is important to clarify that the prototype did not include actual automatic sign-recognition technology. The set of results shown to the users were pre-determined, which helped us control for the apparent performance of the recognition technology in our studies. We controlled where each sign in the list of results was placed, e.g., 10th in the list.

The first property we investigated was the placement of the desired result in the search-results list. To this end, we conducted a user study (henceforth referred to as the "Study 1: Placement Study") to investigate how the position of the desired sign in a list of results impacts users' judgements about the overall system. Our first research question was as follows:

> **1. In empirical testing of a prototype ASL dictionary search system based on automatic-recognition technology, how does the position of the desired word in the list of search results influence: (a) users' reported satisfaction with the system and (b) their perception of the overall relevance of the search results?**

However, there was a need to further investigate whether the position rank of the desired word was the sole factor that affected the participants' perceptions of search results in the first study, or whether it also mattered whether the desired result appeared "above the fold" on the first screen

(before the participants had to scroll down to view later results). This insight motivated our second study henceforth referred to as "Study 2 : Above-the-fold Study."

> **2. In empirical testing of a prototype ASL dictionary search system based on automatic-recognition technology do users' satisfaction with the quality of the video-based search system's output depend on whether the sought-after sign appears "above the fold"?**

Open-ended feedback that we received at the end of our first study had also suggested that another factor influenced participants opinion about the system: the degree to which the overall list of results appeared similar to the desired word (a property we henceforth refer to as "precision"). This motivated our another study described in this article as "Study 3: Precision Study."

> **3. In empirical testing of a prototype ASL dictionary search system based on automatic-recognition technology, how does the overall precision of the search results influence: (a) users' satisfaction with the system and (b) their perception of the overall relevance of the search results?**

Having found that both the placement of the desired word in the list and the overall precision of the results influence users' opinion of search quality, we wanted to understand whether DCG metrics previously proposed for use in reporting the performance of ASL dictionary-search matching methods [5] actually relate to users' judgements of quality:

> **4. When comparing specific metrics for reporting the performance of ASL dictionary search technology, including metrics that do and do not consider the precision of the results list (i.e., DCG with or without binary relevance weighting), which metrics correlate with users' (a) reported satisfaction with the system and (b) perception of the overall relevance of the results?**

Using the data from the first three studies, we examined whether metrics from the information retrieval literature correlated to users' judgements of the quality of the results in an ASL dictionary search application.

4 STUDY 1: PLACEMENT STUDY

Given that sign-recognition technology is imperfect, we wanted to understand how one aspect of the performance of that technology may influence users' opinion of the quality of an overall ASL dictionary-search system. We conducted an experimental study with the independent variable being the placement of the desired word in the search-results list.

4.1 Details of the Prototype in Study 1

A Wizard-of-Oz prototype of an ASL-to-English dictionary search system was designed for this study, which enabled the user to perform a sign in front of their webcam and the system returns a list of likely matches that look like the sign that was performed. The prototype used in this study was briefly described in Section 3 above, to provide sufficient context for the reader before the presentation of our research questions in this study. This section provides additional details about the prototype, and the reader is also encouraged to consult the Appendix, which contains additional images of the prototype.

The prototype consisted of a sequence of web pages viewed using Google Chrome on a 15.6-inch Levovo ThinkPad P52 Mobile Workstation with a built-in webcam. As discussed above, the prototype did not employ any automatic sign language recognition technology. Since we knew in advance the sign that user will be searching for (presented to them in the form of a video prompt

Fig. 1. Still image from one of the stimulus videos.

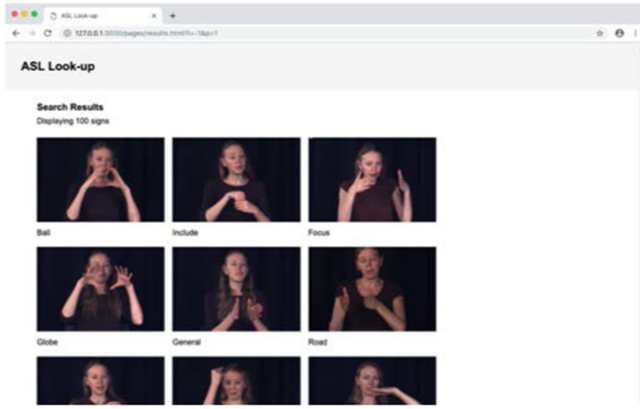

Fig. 2. Sample list of results obtained from our prototype.

before each search), our prototype returned a predetermined list of search-results, regardless of whether the student signed it in front of the webcam or not. This approach enabled us to control the presentation of search-results list so that we can investigate how the accuracy of sign-recognition technology may affect the perception of the users of the quality of the overall system.

The prototype consisted of a series of webpages, as illustrated in Figure 1, Figure 2, and the Appendix. When a user entered the system, they were prompted with a stimuli video of a native signer preforming a sign. They were asked to imagine that they had encountered this sign somewhere but did not know its meaning and needed to search for it. A total of 32 different stimuli videos of individual signs were used in the study, which we had recorded in a studio setting from native ASL signer who grew up using ASL since early childhood years. Figure 1 shows an image of a stimuli video that appeared in our prototype. The words used as stimuli consisted largely of advanced vocabulary that a beginner ASL student would unlikely be familiar with.

The users were asked to press the "next" button after viewing the stimuli video, which took them to a screen where they were asked to press a "record" button, to begin a 3-2-1 countdown timer on-screen. The screen also showed them an option to submit a video clip but the participants were instructed to always record the sign using web-cam instead of recording a video clip and submitting it. Since our prototype was of Wizard-of-Oz nature, our purpose was only to make the users feel like they were interacting with an actual dictionary search system. Once they were done

recording the sign, they were asked to click on the stop button that would take them to the results page, where they were shown 100 results for their search query.

The results were displayed in a scroll-able web-page consisting of approximately three rows onscreen at a time. Figure 2 shows the appearance of the search-results list. The position rank of the items were defined such that the first row contains results 1, 2, and 3, and the second row contains 4, 5, and 6, and so on. The format and layout of presentation of search-results page was designed to mimic image/video search engines, e.g., Google Images or YouTube as much as possible. The video could be played by clicking on each result and text labels were placed below the image. The videos on the results page consisting of a set of 291 videos were extracted from **Boston University's (BU) American Sign Language Lexicon Video Dataset (ASLLVD)** [27, 28]. The text label that appeared underneath each video consisted of the first English word or phrase listed as the translation of the ASL word in ASLLVD.

4.2 Collection of ASL Videos Appearing on Results Page

The ASLLVD contains over 3,300 words, yet we selected a subset of 291 words for use on our results page of the prototype. This subset was selected to carefully include a variety of words that are similar in appearance to the words we are using in stimuli videos. A native signer in our team searched through ASLLVD collection, and added words to our subset. It should also be noted here that although the native signer in our team carefully selected these set of signs using a protocol, in a real system an algorithm would decide what the results would look like. For the purpose of this studies, we are simply trying to simulate an automatic search matching algorithm therefore the choices made by native signer do not hold a lot of significance. We did consider whether to ask a non-native learner of ASL to select the set of similar-looking signs for each item—under the view that there could be a difference in what signs a learner believes look similar, as compared to what a native signer thinks looks similar. Upon consideration, we chose a native signer to do this Wizard-of-Oz aspect of our prototype, since sign-matching algorithms used by computer-vision researchers in sign recognition systems are typically trained on videos labeled by native signers; therefore it is more reasonable to use a native signer as a "wizard" in this study, rather than a non-native signer, to mimic the behavior of an automatic algorithm.

The protocol used to the native signer to search for these signs is described below:

(1) For each of our 32 stimuli words, the native signer on our team selected approximately 3 words considered to be extremely similar in appearance to each stimulus.
(2) Our team member identified the handshape our 32 stimuli videos began with. For each of a total of 8 such handshapes, we identified 15 corresponding signs from the ASLLVD that also used this handshape.
(3) Our stimuli videos also differed in the location of the sign relative to the body. Some of the signs were performed near the head while others were performed near the torso. We selected 30 signs each from ASLLVD with the location near the head and near the torso.
(4) In the end, we added 70 words to our subset of the ASLLVD data set. The rationale behind adding these random signs was to have some words to show in our search-results list that would seem unrelated to the desired word.

There was some overlap between the different words that we identified from ASLLVD using the aforementioned protocol; so, this process yielded a total of 291 ASL videos for use in the search-results pages of our prototype. The characteristics of the set of individual signs on each search-results list were carefully engineered based on the query and for each participant. This control over our selection of search-results list and the order of placement of these signs makes our prototype design Wizard-of-Oz in nature. In the placement study (that we are discussing in

detail in this section), the ordinal position of the desired result (the sign participant is looking for and shown in stimulus video) was controlled by us. When the participants made a query to search for the sign shown in stimulus video, the specific position k of the desired word in the list was already predetermined. After placing the word at a specific ordinal position, we then had to select 99 other signs to fill the entire search results list. These "distractor" words (words that did not match our desired word) were again carefully chosen to make the search-results list look as realistic as possible so that the participant would believe that an actual automatic system had been matching their query to signs in a database and returning results.

Our stimuli included these ASL signs: AUSTRALIA, BRIDGE, CHARACTER, CHICAGO, CIGARETTE, COW, CURLY, DIRTY, DYE, FAMOUS, FANCY, FORK, FREE, FUNERAL, GIRAFFE, INTERNET, JESUS, MIX-UP, OLYMPICS, PIG, PUFF-SMOKE, RAINBOW, SALT, SAVE-MONEY, SCOTLAND, SENTENCE, SILLY, STRUGGLE, SUBWAY, TEND, WHEEL, and YAWN. For each stimulus video, we created a sorted list of 100 items that would be presented in the search-results list, drawn from the set of 291 signs that had been extracted from the ASLLVD. The protocol of selection of the entire results list is as follows:

(1) The matching video for the desired word was set aside, since at the end of the process we would place it at a specific position k in the search-results list.
(2) The native signer on our team manually selected a set of signs that were "extremely similar" to the desired word.
(3) We then selected signs with the same handshape as our desired word. The order of these signs was randomized, and they were placed after the items mentioned in 2.
(4) To complete the list of a total of 100 signs, we took signs with the same location (near head or near torso) as our desired word and placed them after the items in 3 in a randomized manner.
(5) In some cases, steps 2–4 did not yield a total of 99 words. We then selected the remaining words from those that had not yet been selected, to appear at the end of the search-results list.
(6) In the end, we inserted the "matching video" we had set aside at the specific position k in the final search-results list.

4.3 Data Collection Procedure

Participants were shown stimuli videos in the form of a video recorded by the native signer on our team. Afterward, they were asked to search for this sign by recording it in front of the webcam. For each word they searched for, they were asked to identify the best match on the search-results list and write down the English label appearing beneath each video on a sheet of paper provided to them. The desired word was always present in the search-results list. However, in case participant was unable to find the sign that they were looking for, they were instructed to write "not found." The participants were then asked to answer two questions. The participants rated their **satisfaction with the way the results were ranked** on a five-point Likert-scale from "Strongly Disagree" to "Strongly Agree". They also rated their **perceived relevance of the search-results list** on a ternary scale: *highly relevant, relevant, and not relevant.* These two questions were adapted from the methodology of Reference [1]. After answering these two questions, the participants repeated this entire process for the rest of stimuli videos. Figure 3 illustrates this iterative process.

After searching for the 32 words and answering the follow-up questions, a semi-structured interview was conducted. The participants were asked how they would describe the list of results and what they wish had been different about this system. We also asked participants about their thought process while they had been searching for a word and viewing the search-results list,

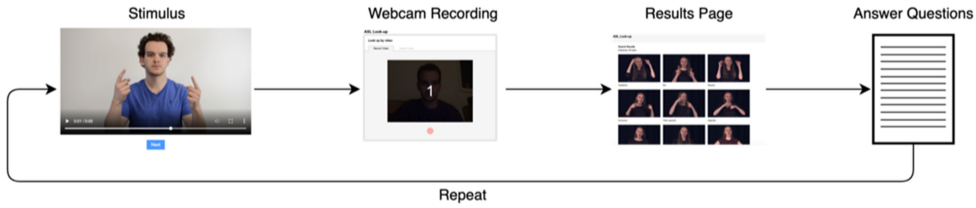

Fig. 3. Flowchart of the procedure our participants followed during the experiment.

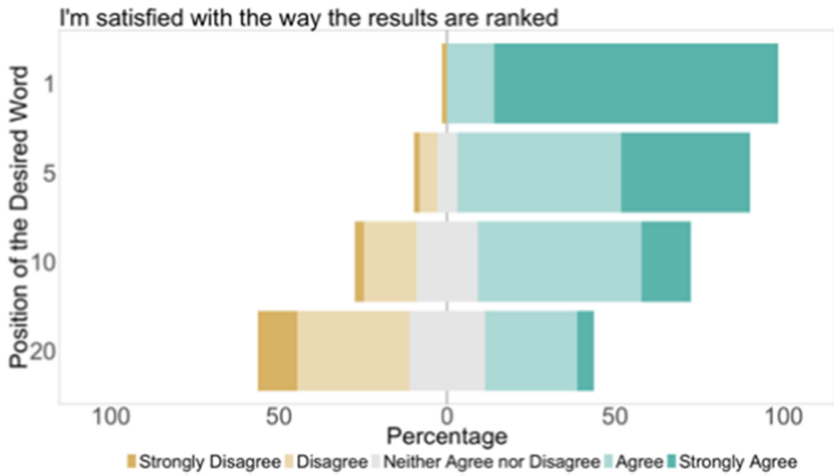

Fig. 4. Placement Study: Users' satisfaction with the way the results were ranked.

especially in the case in which their desired word had not been near the top of the search-results list. These questions were included to better understand what aspects of search results were important to the participants. In the end, the participants were then informed about the Wizard-of-Oz nature of the prototype.

4.4 Participants

Participants were recruited by contacting professors of introductory ASL courses at Rochester Institute of Technology, who shared an advertisement by email with their students, with two screening questions: "(1) Are you currently taking an introductory or intermediate course in American Sign Language?" and "(2) Have you completed an introductory or intermediate ASL course in the past five years?" Participants are recruited if they responded with yes to at least one question.

Participants received $40 cash compensation for participation in this 70-min study. A total of 16 individuals participated, including 15 females and 1 male. The mean age of the participants was 22 years. The number of years of experience in using ASL among our participants varied widely: from 0.5 to 15 years. All the participants identified as hearing. We collected data from a total of 512 searches, since each of the 16 participants performed 32 searches. A total of 59 responses were set aside for separate analysis, because users either did not find the sign at all or wrote down a sign that was similar to the sign shown in stimuli video, but not the one we intended.

4.5 Findings

4.5.1 Effect of Placement on User Satisfaction. Figure 4 displays users' satisfaction with how the results were ranked, for search results lists in which the desired word appeared at various rank

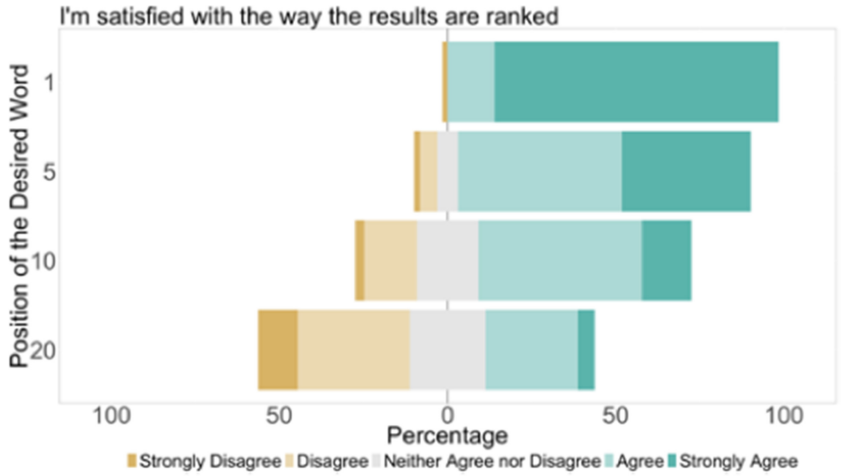

Fig. 5. Placement Study: Users' judgements of the relevance of the results.

positions in the results list: 1, 5, 10, and 20. This Likert response data is displayed using a diverging stacked bar graph, as recommended in Reference [32], in which the percentage of people who agreed with the statement is shown to the right of the zero line and the percentage who disagreed is shown to the left, with the percentage of people who neither agreed nor disagreed centered around the middle.

Users' satisfaction with the way the results were ranked was higher when the desired sign appeared near the top of the search-results list. A Friedman test (χ^2 = 182.682, DF = 3, p < 0.01) revealed that there was a significant difference in reported satisfaction for results that lied at different levels k = 1,5,10,20. In post-hoc testing, pairwise Wilcoxin Signed Ranks tests, with Bonferroni corrections, revealed significant differences between all four levels (p < 0.01).

4.5.2 Effect of Placement on Perceived Relevance. Participants' responses to the question about the relevance of the results are shown in Figure 5. A Friedman test (χ^2 = 80.678, DF = 3, p < 0.01) indicated that there were differences in perceived relevance across different levels of k = 1,5,10,20. A post-hoc Wilcoxin Signed Ranks test with Bonferroni corrections indicated that there were significant difference between all levels (p < 0.01). These results indicate that the position of the desired sign in the list has an impact on users' perceived relevance of the overall list.

4.5.3 Results from the Discarded Searches. In this subsection, we present results for the subset of searches during Study 1 that had been set aside for separate analysis. On 33 occasions, participants were unable to find the word they were seeking in the results list (even though the correct result was always included in the list of results). These errors occurred only 4 times when the desired sign was at position $k = 1$; 8 times when at $k = 5$; 9 times when at $k = 10$, and 12 times when $k = 20$. For these 33 cases, the median and mode response to the satisfaction question was *"disagree,"* and the median and mode response to the perceived relevance question was *"not relevant."*

On 26 other occasions, the participants identified a sign on the list of search results that was similar in appearance to the desired word but not the correct match (to the sign that had been displayed to the participant at the beginning). No such instances occurred when the desired result was placed at position $k = 1$; 5 when at $k = 5$; 9 when at $k = 10$; and 12 when $k = 20$. For these 26 occasions, the median and mode response for the satisfaction was *"agree,"* and the median and mode response for relevance was *"relevant."*

4.5.4 Open-ended Feedback Comments from Participants. When we asked the participants about how they would describe the search-results, they often mentioned the position of the desired word in the results list or how far they had to scroll down to find the desired word. We had been expecting some comments of this nature, since the independent variable in this study had been the position of the desired word on the search results list.

> "There was some that [the sign I was looking for] was the first one so that was good; I think that should be the goal. But that's ambitious. So, at least in like the first 10, that would make it more efficient."—P6

Given that the design of Study 1 had focused on the position of the desired word in the results list, it was notable that participants also commented about another factor: Specifically, some participants mentioned how their impression of the search-results was influenced by how similar the other signs on the list were to their desired word:

> "They're pretty much spot on I'd say. All that they're getting for me is what it looks like, but for most of them, it was coming up with signs that are similar. But that's understandable [...] So even if I had to like scroll down to find the right word, it was still pretty accurate."—P10

> "Most of them had the same handshape, I'd say the first 6 had almost the same sign. But the one I was looking for wasn't always in the top, but it was somewhere in the results."—P7

While several participants indicated that seeing this similarity to the desired word among many items on the search results was desirable, one participant did note that if the results included words that looked very similar to the desired word, a beginning ASL student searching for a particular word might be confused.

> "[...] with a new signer, they may see what they think it is and not keep scrolling, so when it isn't as precise, I think this app could like mislead people."—P14

These participants' comments suggest that the position of the desired word on the results page may not be the only relevant factor influencing user perception of a dictionary search system. This suggested that we should also examine whether users' satisfaction would also be affected by the degree to which the other surrounding words in the results list appear similar to the desired word. For the sake of consistency, we will refer to this property of the search results as **precision** of the results. Study 3, which will be presented later in this article, will specifically investigate this precision factor.

5 STUDY 2: ABOVE-THE-FOLD STUDY

In Study 1, we found that as the position-rank of the desired result increases, both of the following decrease: users' satisfaction with the presentation of search results and their perception of the relevance of overall results. However, we still did not know whether users' judgements were solely a function of the position rank of the desired results, or whether users' satisfaction also depended upon whether the desired result appears above the fold, i.e., on the first screen before users had to scroll down to view later results.

While no prior ASL dictionary search research had examined the issue of how much content to fit in a single page of an ASL dictionary, there has been related research in the context of web search. Information-retrieval researchers have investigated how the design of a search-results page may influence user's behavior, e.g., References [18–21]. This work has examined the design of search-engine result pages, including: the number of results presented per page, screen sizes, and how the

positioning of results affect users' judgement of search systems. Some researchers have found that regardless of how many items are displayed per screen of results, users seem predisposed towards examining a certain number of queries for each search [19].

Other work has found that items that are presented on the first screen of results do tend to receive more attention from users. In a prior study on position bias and click-through rates, researchers found that users examined 20%–70% of results that were "above the fold" (in the top 6 of the search-results items displayed), and only examined 5%–10% of items that were displayed "below the fold" (in position 7 to 10 on the list of results) [18]. There may also be an interaction between the number of results that the person desires to view and the effort required to paginate or scroll down to view more results. The additional cost of scrolling through results or paginating may lead to a more in-depth inspection of the results (particularly the top-ranked results) [21].

Search-result pages also serve as feedback to the users about the quality of their query. If you increase the number of results per page, then users can make a quicker assessment of the quality of their query and spend more time formulating queries rather than viewing results [20]. In addition to the users having quicker visual access to more results when more on each page, there is also less memory burden across pages when seeking a desired result item, which leads to users viewing more results.

However, squeezing more results onto a single page is not necessarily advantageous: Research in cognitive psychology on the "paradox of choice" has found that an overload of choices can lead to poor decisions or lower satisfaction with good choices that they make [29]. While the memory burden on users might be lesser for those viewing more results per page, they may still experience more significant overall cognitive workload due to more results per page, thereby making them prone to rating their searching skills lower.

To summarize, prior literature has revealed that the number of results presented per page can have an effect on users' perception of search systems. However, there are differences between navigational web search and the task of someone searching for a sign in an ASL dictionary, e.g., different input and output modalities, which do not permit directly generalizing findings from studies in the domain of web-search to ASL dictionary search. Therefore, in this study, we are investigating whether users' satisfaction with the system drops dramatically when their desired result does not appear on the first screen (and they have to scroll down further). If such an effect is observed, then this would suggest that designers of the user-interface of ASL dictionary search systems should consider the accuracy of the underlying sign-match algorithm (and how likely it is for the desired word to appear in the top-k search results) when selecting how many search results items to display per page.

5.1 Conditions and Sequence of Presentation

Our experimental study used two web-based prototypes, similar to the one in the placement study, which consisted of a series of web pages that simulated users submitting a video of themselves performing a sought-after sign and viewing search results. In this between-subjects study, each group of participants used one of the two prototypes; the prototypes were identical except that: (a) in one prototype six results appeared on the first page before the users had to scroll down to next page, while (b) in the other prototype, there were eight results on the first page before users had to scroll down to the next page. The prototypes were deployed online, since the experiment had to be conducted remotely, due to social distancing during the COVID-19 pandemic. A researcher sent an informed consent form to the participant through email, which the participant read and reviewed, prior to a video conference meeting between the researcher and the participant. A link to the correct version of the web-prototype was provided to each participant, and a calibration screen appeared at the beginning of each web-based prototype, to ensure that the size and aspect ratio

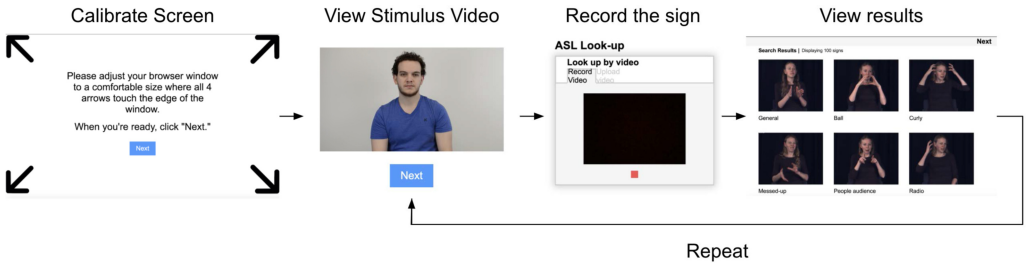

Fig. 6. Modified flowchart of the procedure our participants followed during above-the-fold experiment.

of the web browser window was consistent across participants. Figure 6 illustrates this iterative process.

To investigate our research question, we needed the desired result to appear at 16 different position-ranks on the results list, with the remaining "distractor" items on the list remaining in their original order (same as in Study 1). Each participant had to complete 32 queries with the desired word placed at each of the 16 different position-rank values twice. For instance, the rank-position order (where the designed sign appeared on the results) randomly determined for the first participant was 17, 1, 8, 6, 25, 4, 7, 21, 3, 5, 13, 11, 2, 19, 15, 9. This order was counterbalanced for subsequent participants.

After each sign query, participants were asked to look through the results page and identify the item that seemed like the best match for the sought-after sign (and to note this on a separate response form). Although the matching sign was always present in the results list, if a participant believed that the desired result was not on the list, they could write "not found." After each query, participants rated their **satisfaction with the way results are ranked** on a Likert scale from *strongly disagree* to *strongly agree* and **their perceived relevance of the results** on a ternary scale: *highly relevant, relevant, not relevant.* These questions were adapted from Reference [1].

5.2 Recruitment and Participants

Participants were recruited by contacting professors of introductory ASL courses at Rochester Institute of Technology, who shared an advertisement by email with their students, with the same two screening questions as Study 1: "(1) Are you currently taking an introductory or intermediate course in American Sign Language?" and "(2) Have you completed an introductory or intermediate ASL course in the past five years?" Participants are recruited if they respond with yes to at least one question. A total of 32 participants were recruited (16 per group), with each participant performing 32 searches, yielding a total of 512 searches. These participants were different from participants recruited for Study 1 or Study 3.

As in Study 1, our analysis examines user responses in those cases in which participants eventually found the desired sign, with 61 searches removed from the responses from the group that interacted with prototype with six results on first page, in cases in which participants either did not find the word that was shown in the stimulus video (44 cases) or did not report finding the word in the entire list (19 cases). This left us with 449 searchers, for which we report the results in the findings section. Similarly, 77 searches were removed from the responses from the group that interacted with prototype with eight results on the first page as the participants either did not find the word that was shown in the stimulus video (56 cases) or did not report finding the word in the entire list (21 cases). This left us with 435 searchers, for which we report the results in the findings section.

Table 1. Regression Discontinuity Design (RDD) Results for Both Questions with Cutoff at 6

Question	Bandwidth	Observations used	Estimate	Std. Error	Lower CL	Upper CL	z value	$p(>\|z\|)$
Satisfaction with ranking	6.119	289	−0.4217	0.1667	−0.7485	−0.09494	−2.529	0.01143 **
Perceived overall relevance	11.174	373	−0.04295	0.1291	−0.2960	0.2101	−0.3327	0.7394

Table 2. Regression Discontinuity Design (RDD) Results for Both Questions with Cutoff at 8

Questions	Bandwidth	Observations used	Estimate	Std. Error	Lower CL	Upper CL	z value	$p(>\|z\|)$
Satisfaction with ranking	13.748	409	−0.3381	0.1546	−0.6412	−0.03503	−2.1865	0.02878 *
Perceived overall relevance	8.853	359	−0.3458	0.2051	−0.7478	0.05630	−1.6855	0.09189

Each participant was interviewed after they completed their searches, and they received $40 cash compensation for this remotely conducted, 70-min study.

5.3 Findings

We analyzed our data to address our research question "Does users' satisfaction with the quality of the video-based search system's output depend on whether the sought-after sign appears on the computer screen before the user needs to scroll down to view lower results?" As the two factors, (1) the rank-position (k) of the desired result on the search-result list, and (2) whether the item appears on the first screen of the results or the participants need to scroll, are intrinsically linked, we used a **regression discontinuity design (RDD)** [9]. In our case, the input to the regression model was the position-rank of the desired result, and the output was user satisfaction with the results. The intuition behind RDD was that it may reveal whether there exist different slopes and intercepts that fit data on either side of some sharp "cutoff" (here, whether or not the result is visible in the top-six or top-eight results, before the user needs to scroll down on the page). RDD can reveal whether user-satisfaction with the presentation of the results "dramatically" falls, if the sought-after sign appears on the results list after the cutoff value, i.e., 6 or 8, the number of results that can fit on one page.

We used the optimum non-parametric regression to fit our data for both questions. Table 1 and 2 summarize the results. The cutoff was set to 6 and 8 and the type was set to "sharp." We omitted the "bw" argument, which is a numeric vector that specifies the bandwidths at which to estimate the regression discontinuity. The bandwidth is a tuning parameter or the discontinuity sample (small neighborhood to the left and right of the cutoff point used in RDD analysis). As the bandwidth is increased the bias in data increases and variance decreases as we have more data points farther from the cutoff. By default, if the "bw" argument is omitted, the bandwidth is calculated using Imbens-Kalyanaraman 2012 method [15]. The values of bandwidth that this method returned were 6.119 and 11.174 for "satisfaction with ranking" and "perceived overall relevance" questions when cutoff was set at 6, and 13.748 and 8.853 for both questions, respectively, when cutoff was set at 8.

Overall our results showed that as the position-rank of the desired word increases, both users' satisfaction with the results ranking and their perceived overall relevance of the search results decrease as shown in Figure 7 for the $k = 6$ group, and in Figure 8 for $k = 8$ group. For the $k = 6$ group, the ($p(>|z|)$ value was: 0.011426 ($p < 0.05$) in case of the "satisfaction with ranking of results" question ($z = -2.529$), and 0.7394 ($p > 0.05$) in case of "perceived relevance of overall results" question ($z = -0.3327$). We observed a significant result for the first question, but we did not observe a significant difference for the second question. For the $k = 8$ group, the ($p(>|z|)$ value was: 0.02878 ($p < 0.05$) in case of the "satisfaction with ranking of results" question ($z = 2.187$), and 0.0920 ($p > 0.05$) in case of "perceived relevance of overall results" question ($z = -1.685$). Like the $k = 6$ group, we observed a significant result for the first question but did not observe a significant difference for the second question.

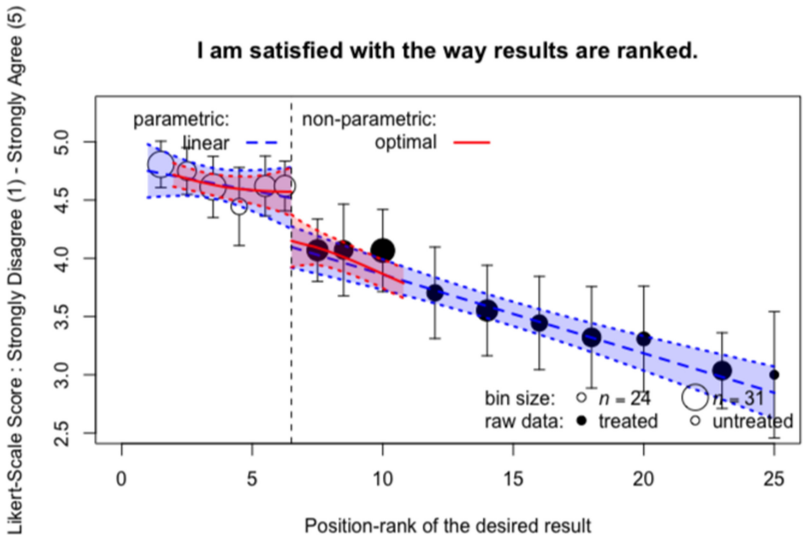

Fig. 7. Results of RDD for "satisfaction with ranking" question with $k = 6$, indicating that there was a significant discontinuity at position 6 but not at position 8.

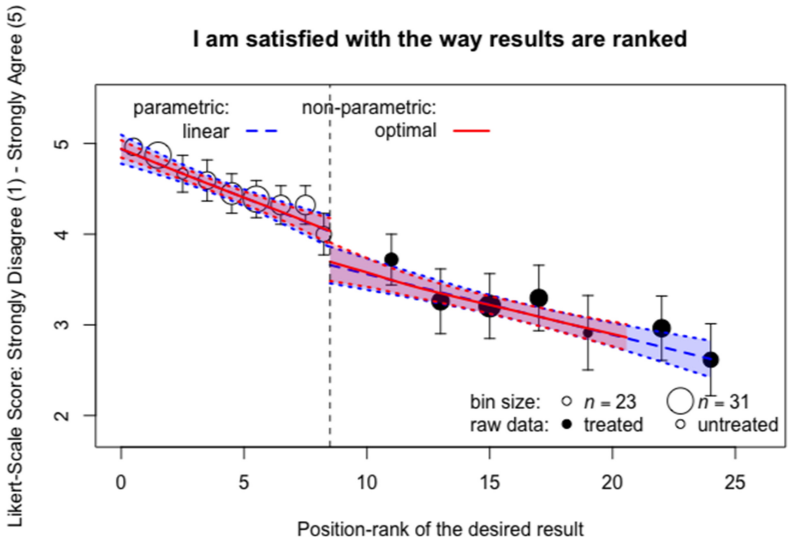

Fig. 8. Results of RDD for "satisfaction with ranking" question with $k = 8$, indicating that there was a significant discontinuity at position 8 but not at position 6.

These findings reveal that there was a discontinuity in users' subjective impressions of search ranking after position 6 when users interacted with the prototype that showed only 6 items on the first page before scrolling, and there was a discontinuity after position 8 when they interacted with the prototype showing eight items on the first page. The magnitude of discontinuity at $k = 6$ position is -0.4217 when we displayed six results on one screen and -0.3381 at $k = 8$ position when we displayed eight results. However, there was no discontinuity after position 8 for the six-per-page prototype, nor any discontinuity after position 6 for the eight-per-page prototype.

These findings suggest that there is an above-the-fold effect for ASL dictionary search-results pages.

6 STUDY 3: PRECISION STUDY

Based on the open-ended comments of our participants in Study 1, as discussed in Section 4.5.4, we wanted to examine how the precision of the results may influence user judgements about a system's quality. We had noticed that prior research on automatic recognition technologies for identifying ASL words from video generally report their results in terms of accuracy, i.e., whether the desired word appears within the top-k ranked results of the system. To the best of our knowledge, no prior study looked at the effect of other surrounding words in the search-results list to determine whether the they affect users' perception of the search-results as well. To this end, we conducted a follow-up study, very similar to our earlier Study 1 (placement study). The same prototype was used and the participants were asked the same questions about their satisfaction with the way the results were ranked and their perceived relevance of the overall system. The only difference was that in this study we did not vary the placement of the desired word a lot and instead kept it in a very narrow range.

6.1 Conditions and Sequence of Presentation

In this study, we kept the variable of placement constant and explored the variable of precision of the results surrounding our desired word in the search-results list. We could have potentially conducted this study in a two-factor manner where we examined both the placement of the desired word and the precision of the results-list, but we were skeptical if we will be able to recruit a sufficiently large group of ASL students who had not participated in our other studies mentioned in this article to ensure that the two factor study is sufficiently powered. We selected the placement of the desired result value to be 10. There were two reasons behind this decision. We wanted to select a value of k for which we had received relatively middle values of satisfaction (in Study 1) such that the placement factor would not pull our participants' responses to to the end of our response scale too much. Our choice of rank 10 also reflects the likely improvements to the state-of-the art of automatic recognition technology, considering the top-20 basis of reporting in References [3, 8]. In considering keeping the placement constant, we had a concern that participants might notice that the desired result always appeared at the exact same position in the search-results list, which could have made them become suspicious about our prototype. Therefore, we varied the position of the desired word slightly by placing it in values $k = 10 \pm 2$.

6.2 Varying the Independent Variable Precision

The independent variable in this study was the precision of the results surrounding our desired result on the search-results list (the other 99 words). These "distractor" surrounding words were classified into three categories as follows:

- **High Precision:** The list of 99 distractor words begins with the words that were manually selected by the native signer on our team as "extremely similar" in appearance to the desired word. The next 15 items on this list consisted of words that had the same handshape as the desired word. This was followed by another 30 signs that had the same location (i.e., near head or near torso) as the desired word. Last, the list contained randomly selected words to make up a total of 99. An example of a search results list with high precision is shown in Figure 9.
- **Medium Precision:** In this condition, no words that were extremely similar to the desired word were included in the set of 99 surrounding search results. We did include signs that

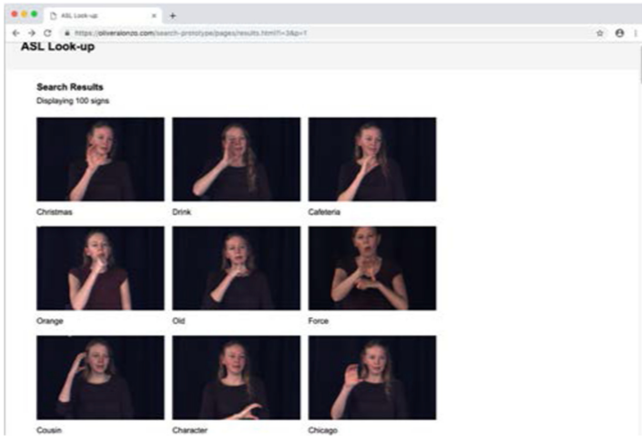

Fig. 9. Sample list of results with high precision for the sign for Chicago.

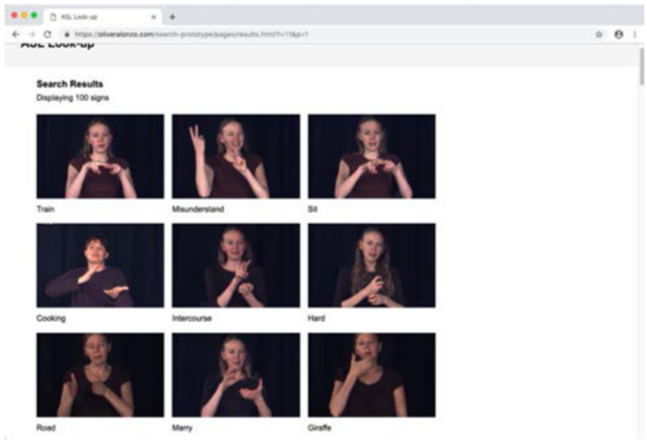

Fig. 10. Sample list of results with low precision for the sign for giraffe.

had the same location (near head or near torso). Finally, at the end of the list, we added signs randomly until the list had a total of 99 signs.

- **Low Precision:** This list was intentionally filled with signs that had a different handshape and a different location than the desired sign. An example of a search results list with low precision is shown in Figure 10.

The sequence of presentation of each stimulus video was randomized for participants. We also randomized the assignment of condition across participants. Each of our participants engaged in a total of 30 searches, with 10 at each precision level.

6.3 Recruitment and Participants

We followed the same recruitment strategy as Studies 1 and 2. Participants were recruited by email at our university, and there were two inclusion criteria: having taken an ASL course in the past 5 years and having started learning ASL after the age of 5. These participants were different from

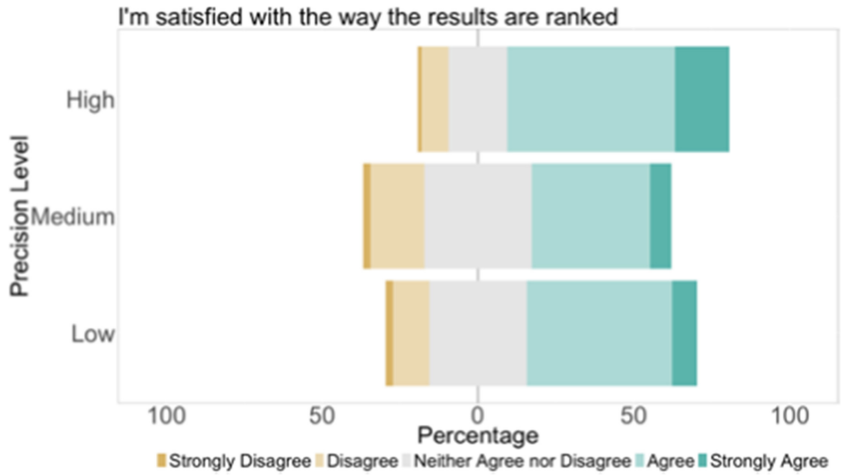

Fig. 11. Precision Study: Users' satisfaction with the way the results were ranked.

participants recruited for Studies 1 and 2. Participants received a $40 compensation for participation in a 70-min study.

Ten beginning ASL students participated in this study, including 8 women and 2 men. The mean age of the participants was 23.3, and participants' years of experience learning ASL ranged from 0.5 to 15 years. Nine participants identified as hearing, and one participant identified as deaf. Since each participant conducted a total of 30 searches, we gathered data for a total of 300 searches. Similar to Studies 1 and 2, there were occasions when participants could not find the desired word or occasions in which they identified a similar but wrong word. A total of 46 such responses were set aside for separate analysis.

6.4 Findings

The precision of the search-results list had a significant impact on users' satisfaction with the way the results were ranked, as revealed by a Friedman test (χ^2 = 16.526, DF = 2, $p < 0.01$). The test indicated that there was significant effect of precision on user satisfaction. Pairwise post-hoc comparison with Wilcoxon Signed Rank tests, with Bonferroni corrections, indicated significant differences: (a) between *high level* and *medium level* with a p-value of 0.0002 and (b) between *high level* and the *low level* (p = 0.0146). We did not observe a significant difference between the *middle level* and *low level*. The percentages of responses across the different levels are illustrated in Figure 11.

We also observed a significant difference in users' rating of the relevance of the results. A Friedman test (χ^2 = 35.438, DF = 2, $p < 0.01$) revealed that there was a significant effect of precision on the perceived relevance of the entire search-results list. We again performed pairwise post-hoc comparison using Wilcoxon Signed Rank tests, with Bonferroni corrections, which indicated that there were significant differences: (a) between the *high level* and the *medium level* (p = 5.5767E-7) and (b) between the *high level* and *low level* (p = 0.000027). However, we did not find a significant difference between the *middle level* and the *low level* (p = 0.353160). Figure 12 illustrates the percentages of responses across the different levels.

6.4.1 Results from the Discarded Searches. In this subsection, we present the results for the searches that were set aside for separate analysis. On 34 occasions, participants reported that their

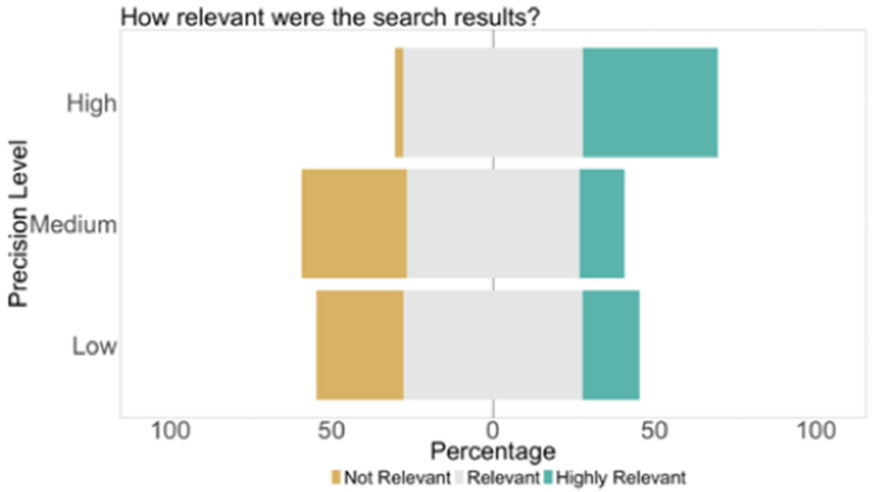

Fig. 12. Precision Study: Users' judgements of the relevance of the results.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Fig. 13. Traditional formula for DCG.

desired word did not appear on the results list (although it was always included on the list). Seven such instances occurred at high precision level, 14 at medium precision level, and 13 at low precision. For these 34 cases, the median and mode response to the satisfaction question was *"disagree,"* and the median and mode response to the perceived relevance question was *"relevant."*

On 26 other occasions, participants identified matches that were similar in appearance to the desired word but not the match that we intended. Eleven such instances occurred at the high precision level, none at medium precision level, and 1 at the low precision level precision. For these 26 occasions, the median response for the satisfaction was *"strongly agree*," and the mode was *"agree."* The median and mode response for relevance was *"highly relevant."*

7 METRIC FOR ASL SEARCH

In information retrieval research, metrics are often used to provide a single composite score to indicate the overall quality of search results returned by a system. Researchers have designed algorithms to gauge the performance of a variety of search systems [5]. Several of these metrics can be used for reporting the performance of ASL dictionary systems as well. In our Studies 1, 2, and 3, we established that both the position of the desired result and the precision of the surrounding results in the search-results list affect the user perception of these systems. Based on these findings, we suggest using **Discounted Cumulative Gain (DCG)** as a performance evaluation metric for ASL dictionary systems, since DCG considers both the position of the desired result as well as the precision of the overall search-results list. A composite score calculated using this metric might be a better estimation of the overall "quality" of the result, as compared to a metric that considered only one of these factors. This metric is a function of the length of the list of results shown p, as shown in Figure 13.

$$IDCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(\pi(i) + 1)}$$
$$nDCG_p = \frac{DCG_p}{IDCG_p} \ ;$$

Fig. 14. Formula for nDCG, where $\pi(i)$ is the position of the ith result in relevance order.

$$bDCG_p = DCG_p \ , \ rel_i \in \{0, 1\}$$

Fig. 15. Formula for bDCG.

In some information retrieval applications, queries may return different numbers of results; to account for this variation and to enable comparisons of performance despite different-length search-results lists, a **Normalized DCG (nDCG)** is used. In principle, while calculating nDCG, **Ideal DCG (IDCG)** is calculated. The IDCG is the maximum possible DCG of a list of results of length p, which is calculated by first sorting the list of results based on the relevance of each item and then computing the DCG score using the position $\pi(i)$ of each result in that sorted list. The relevance rel_i of item is 1 if it is the desired word, and 0, if it is not. Figure 14 describes the calculation of nDCG using IDCG.

While the nDCG metric considers both placement and precision of the results list, this metric is suitable for use in ASL dictionary search systems. However, nDCG has not been utilized in prior research on evaluating the performance of ASL dictionary search systems. The most sophisticated use of a metric in ASL dictionary search system is in the search-by-selecting-features ASL dictionary search system of Bragg et al. [9]. However, even then, a simpler metric was used, which only considered whether each item in the list of results was a perfect match to the desired word. More specifically, that prior study used a simpler version of DCG, which we will refer to as **Binary DCG (bDCG)** here, as shown in Figure 15.

Our data from our earlier studies enables us to investigate various metrics, as to their suitability for ASL dictionary search. In studies 1, 2, and 3, we obtained judgements from users about "their satisfaction with the way the results were ranked" and their opinion of the "relevance of search results." With this data, we can compare these alternative metrics empirically, to determine which *correlates better* with users' preferences.

To conduct this analysis, we calculated both bDCG and nDCG metrics for each list of results that had been displayed to the participants in our earlier studies. For the purpose of this study, we calculate the relevance of individual items in a Wizard-of-Oz manner. For the bDCG metric, to calculate the relevance of individual items in the list, we used a simple binary decision of assigning 1 if it is the desired word, and 0, if it is not. For nDCG, to calculate the relevance of individual items in our search results, we devised a heuristic: 1 if an item is the desired word, 0.5 if it is a sign that a native signer on our team identified as "extremely similar" to the desired word, 0.25 if it is a sign with the same handshape or the same location as the desired word, or 0 otherwise.

While we had implemented this metric manually for the purposes of this study, we note that a metric like this could be automated. Many dictionaries of ASL signs include various linguistic

Table 3. Spearman Correlation between Different Metrics and Users' Satisfaction with the Ranking of the Results, and Their Perceived Relevance of the Results (** Indicates $p < 0.01$)

User's Judgement	Placement Study bDCG	Placement Study nDCG	Precision Study bDCG	Precision Study nDCG	Above-the-fold Study bDCG	Above-the-fold Study nDCG
Satisfaction	0.665 **	0.646 **	0.208 **	0.196 **	0.660 **	0.636 **
Relevance	0.511 **	0.530 **	0.099	0.295 **	0.538 **	0.528 **

attributes for individual ASL signs, e.g., the handshape of each hand at the beginning and the end [5, 22, 34]. Sometimes these dictionaries also include meta-data about the location of hands [5, 22, 34], movement [5, 22], orientation [5], or other details. In future work, it is reasonable that someone could implement an automatic algorithm that would use such meta-data attributes for each entry in the dictionary to calculate the relevance of a search-result and similarity scores between different signs.

Table 3 displays the results of analysis of how both nDCG and bDCG may correlate with users' satisfaction with the way the results were ranked and their perceived relevance of the search-results. This analysis revealed that both metrics correlated to a similar degree with the users' "satisfaction" score in each study, and in our Study 1 (*placement study*), both metrics correlate to a similar degree with the users' opinion of the "relevance" of the results. However, when we examined the response data from Study 3 (*precision study*), we saw a different result. In Study 3, the search-results lists had varied widely in how similar the entire set of signs was to the desired word. In this case, nDCG correlated with users' judgements about the relevance of the results, the bDCG had no significant correlation.

These findings are relevant for researchers who are investigating methods for searching for a match in an ASL dictionary, whether by in search-by-feature-selection (in which users select linguistic elements of words on a form) [5] or through automatic recognition of video input [38]. In either setting, researchers should consider using the nDCG metric with non-binary weighting, which considers not only the placement of the desired word in the results but also the similarity of other items in the list.

8 DISCUSSION

We anticipate the results of the studies presented in this article to be of interest to two audiences. First, our findings provide guidance for **designers of user interfaces of ASL dictionaries**. We reveal that these designers should consider the current performance of the underlying match technology when selecting how many results-per-page to display. Second, our findings provide guidance for **researchers studying underlying technologies for identifying matches**, e.g., sign recognition from video, for sign-language dictionary search systems. While our work does not inform these researchers how to build their algorithms, our work does suggest how these researchers should measure the performance of their systems to decide when they are ready for use in this dictionary application—our findings identify some important metrics and thresholds.

In regard to designers of user interfaces for ASL dictionary search systems, our findings illustrate the need for designers to understand the current accuracy of the underlying matching technology—specifically, how likely it is for the desired word to be within some top-k positions in the ranked list of results. In our follow-up Study 2 presented in this article, we found that the placement of the desired result above the fold (on the first screen before the users have to scroll down) or below the fold (on later screens) impacts users' judgements about the dictionary search system. Specifically, for prototypes with six results per screen, we observed a significant discontinuity in users' satisfaction scores for items above/below position 6, and for prototypes with eight results per screen, we observed a significant discontinuity for items before/after position 8 instead. As we did

not observe any significant discontinuity at position 6 for the eight-results-per-page prototype—nor vice versa at position 8 for the six-results-per-page prototype—our findings indicate that user satisfaction significantly decreased when the desired result was not on the first page.

While our studies investigated prototypes with six and eight results per page, our findings should not be understood as advocating for a specific number of results per page. Instead, our findings suggest that there is a relationship between the ranking quality and the presentation of results. Obviously, if a designer knew in advance that a desired result would always be at position 7 of search results, then they could simply set the fold location to after that position—or simply show users the seventh item on the results list! However, designers cannot predict where a user's desired word will appear on the results list for a particular query. Instead, our findings reveal that designers should plan how many items to appear on each page of search results "above the fold" based on a consideration of the accuracy, in general, of the underlying matching technology being used. For instance, if during evaluation of the matching technology it is known that the system returns the desired word within the top-k of the search results in a high percentage of cases, then designers should consider displaying k results per page, if possible. In essence, designers of ASL dictionary systems should consider how many result items to display on a single screen, as a function of the accuracy of the underlying sign match technology being used, to potentially increase the likelihood of the desired result appearing above the fold. The contribution of our study lies in highlighting this relationship.

While prior research in the field of information retrieval, in the context of web-search, had established the existence of this "above-the-fold" effect [18–21], no prior study had looked at this factor in the context of an ASL dictionary search interface. Given the differences in users' goals in these two task settings, an empirical study was necessary in the ASL dictionary search context. Thus, our findings build upon this prior work in the domain of web-search, to extend it to this domain of ASL dictionary search user-interfaces.

As discussed above, the findings of our study also inform the work of **researchers who are investigating the design of underlying matching technology** for ASL dictionary systems. While our findings do not suggest to these researchers how they should specifically build their matching algorithms, our findings did reveal some metrics these researchers should consider when measuring the performance of their technology, to determine when it is ready to deploy for ASL dictionary-search applications. Specifically, we investigated whether users' judgements of the quality of an ASL dictionary search system vary depending on the placement of the desired word in the list of search results—as well as the precision of the results list (the similarity of the other words on the list to the desired word).

We observed a significant effect of placement on responses to two question items commonly used in the information retrieval literature [1]: (a) users' satisfaction with the way the results are ranked and (b) users' perception of the relevance of the results. We observed that as the position rank of the desired result increases, users' satisfaction with the system, as measured by the two question items, decreases. Thus, ASL dictionary search researchers (or researchers studying underlying technologies, e.g., automatic recognition of ASL signs from video) should focus on optimizing and reporting the performance of their systems regarding placement of the desired word within the top few results.

While finding that user satisfaction with the presentation of the search-results drops as the desired item appears lower in the list is intuitive, our work has a more specific contribution: As our Related Work Section 2 discussed, there is disagreement among researchers studying ASL dictionary search technologies about how to report their results. The percentage of time that a desired word is in the top-k of their results is often reported as a measure of accuracy of these systems [3, 13]. However, researchers report this statistic for different values of k, with researchers

reporting how often the desired word appeared in the top-k of their results for values of k ranging from 20 up to several hundred [3, 10]. Our findings in Study 1 reveal that somewhere between $k = 10$ and $k = 20$, our users' satisfaction drops below the midpoint of the Likert-scale used to assess their satisfaction with how the results were ranked. This result helps to constrain the range of values of k, for which it is of greatest interest for researchers to report the success of their matching algorithms, for this ASL dictionary application.

While it may seem intuitive that users would prefer the output of an ASL dictionary search system if the desired word appears closer to the top of the results list, our Study 3 revealed another factor that influences users' perception of the system: Specifically, we found that even when the placement of the desired word in the list of results is held constant, users' perception of the quality of a search tool is affected by the precision of the *other words* in the results. Sign-recognition researchers generally do not report the performance of their systems in regard to the *precision of the surrounding words*, but our findings suggest that this is another metric that these researchers should begin to measure and report in evaluations of their technology. The quantitative findings in Study 3 aligned with some open-ended feedback from our participants in Study 1, which had suggested that users of ASL dictionaries not only want the exact match to their desired word in the top-k results, but they also prefer having a more coherent list containing similar signs near the top of the list.

Finally, our analysis of the response data from all three studies revealed that metrics previously used in the ASL dictionary search literature (based on a binary decision of whether the match for the query is within the top-k results) do not correlate with user judgements of system quality as well as metrics that also incorporate the relevance of each result in the list. Specifically, we found that the nDCG metric correlated with both users' reported satisfaction with how the search results were ranked and their impression of the overall relevance of the results. This finding provides guidance for researchers designing ASL dictionary search systems, in that it suggests a useful metric to use for concisely assessing the overall quality of a results list. It also provides empirical evidence that this metric correlates to users' judgements of the system's quality.

9 LIMITATIONS AND FUTURE WORK

There were several limitations in our study that may suggest avenues for new research in future. In this article, we report three separate studies each based on a different property of the output of ASL dictionary-match algorithms. Therefore, it did not allow us to investigate any possible interactions between the three properties. In future, multi-factor studies with more participants could enable us to explore this issue.

In all of our studies, to simulate the experience of a student who had encountered an unfamiliar ASL sign, our participants were shown videos of isolated signs, and they attempted to produce these signs in front of their webcam. In future research, it would be useful to provide participants with a more realistic search context: For instance, the participant could be shown a stimulus sentence containing an unfamiliar word, rather than being shown an isolated sign. Such a study would enable us to understand how users may incorporate contextual clues about a word's possible meaning into their searches.

The studies presented in this article did not investigate variations in the user-interface or many other design choices in the dictionary search system, which could be explored in future research. For instance, research is needed on the the presentation of the overall search-results page and each individual result snippet (focusing on both the textual and visual parts of each). A future study could also examine how to best present present other metadata, e.g., the definition of each word, on the results list, since we know that: there is no one-to-one correspondence between ASL

and English, some signs either do not translate directly to a brief English word or phrase, and some signs may have multiple translations.

In the studies described in this article, we primarily focused on beginner ASL students who all identified as hearing. Future studies could consider users who have a wider range of ASL skill level, and such studies could also consider DHH users who are looking up words in ASL dictionaries. The findings of our current studies may not generalize to these other groups of users. In addition, in future studies, eye-tracking may be a useful methodology for how users' attention moves across the interface.

In the subjective feedback that we acquired from our participants in all three studies, participants occasionally indicated that they were familiar with the sign shown to them in the original stimuli (the sign they were asked to search for). We had engineered our set of stimuli to avoid words that students in a first-semester ASL course may be familiar with, and we also asked our participants to imagine that the word they are looking for is an unfamiliar word that they had randomly encountered. However, it would be useful for a future study to ensure that all words shown as stimuli were unfamiliar to students, to enable us to determine if there are unique preferences among users who are looking up a truly unfamiliar word.

Based on our finding in Study 2 that the placement of the desired result above the fold impacted users' judgements about ASL dictionary search system, a future study could investigate the performance characteristics of specific underlying sign-matching technology, e.g., a specific sign-recognition system. In order for a designer of an ASL dictionary system to determine the optimal number of results-per-page to display, an analysis of the output of sign-matching algorithms would be needed, to determine how likely it is for the desired word to appear in various rank locations in the search output. As we explain in our Discussion, this property of the performance of sign-recognition technology can guide designers of ASL dictionary-search applications in selecting how many results per page to display.

Last, while we have found that the nDCG metric correlated with users' judgements of the quality of the output of an ASL dictionary search system, this metric requires a method of determining the relevance (the similarity) of each individual item in the list to the desired word. For the purpose of our analysis, we calculated this similarity heuristically. However, research is needed on how to best calculate the relevance of an individual sign based on its similarity to the desired word, as input to this metric.

This future work outlined above would build upon the contributions of this current article. In summary, our work has identified how the performance of sign-recognition technologies affects users' satisfaction with an ASL dictionary-search system, and our findings have provided methodological guidance to dictionary-search and recognition researchers on how to best report results of their research.

APPENDICES

A ONLINE RESOURCES

This Appendix includes information about the contents of the electronic Supplemental Material that will be uploaded to the ACM digital library to accompany this article. These files are available at http://latlab.ist.rit.edu/taccess2020/. These files include:

- **STIMULI VIDEOS**
  These are a set of videos recorded by Abraham Glasser who is a co-author of this article. Abraham identifies as Deaf and is a fluent ASL signer. These videos were shown to participants before each search. Participants were asked to imagine if they had encountered the sign shown in video stimuli somewhere but did not know its meaning.

- **SAMPLE SEARCH RESULTS LISTS**
  These are three CSV files, with one corresponding to each study. Each CSV file contains a sample set of 99 "distractors" (the other signs shown in the search results list that were not the desired word), with a set of distractors provided in this CSV file for each sign that participants searched for. As discussed in the article, the list of search results randomly varied for each participant, so we are sharing sample lists that were generated using the protocols for each study.
- **CORRELATION DATA**
  These are CSV files containing the averaged nDCG and bDCG scores calculated for each search, across all three studies, and the participants' responses to the two questions used to gauge their judgements about the system.
  — Participants' responses to the first question (satisfied with the way results were ranked) are recorded on a 1–5 scale, where 1 corresponds to "Strongly Disagree" and 5 corresponds to "Strongly Agree." These responses appear under the "Satisfaction" Column in each of the three CSVs.
  — Participants' responses to the second question (perceived relevance of the overall results) are recorded on a 1–3 scale where 1 corresponds to "Highly Relevant," 2 corresponds to "Relevant," and 1 corresponds to "Not Relevant." These responses appear under the "Relevance" Column in each of the three CSVs.
  — The nDCG and bDCG averages for each search appear under nDCG and bDCG columns, respectively.
  — For the *Correlation Data for Placement Study*, column labelled K refers to the adjusted position rank of the desired result in the results list while Position column refers to the actual position rank.
  — For the *Correlation Data for Precision Study*, column labelled Position again refers to the actual position rank.
  — For the *Correlation Data for Above-the-fold Study*, the additional columns participant(p) and search(i) refer to participant number and search number, respectively.

B IMAGES OF USER INTERFACE

This Appendix contains additional images of the prototype used within our studies.

Figure 16 displays the screen that participants saw at the beginning of their study, where they entered a specific code number ID.

Figure 17 is a screen that appeared next (only in the above-the-fold study, which occurred remotely), so that the user could configure the ratio of their web browser window, to ensure that the user was viewing only a specific number of sign results per page.

Figure 18 is an example of the "View Video" page, in which the participant was shown an ASL sign, which they were later asked to search for in the dictionary.

Figure 19 displays the user-interface that participants saw when they used their webcam to record themselves producing an ASL sign to search for.

Figure 20 displays an example of a page displaying search results, in this case the results for query in which the participant had searched for the ASL sign ROAD. The desired sign did not appear above the fold on the first screen in this example; so, the participant needed to scroll down to find it.

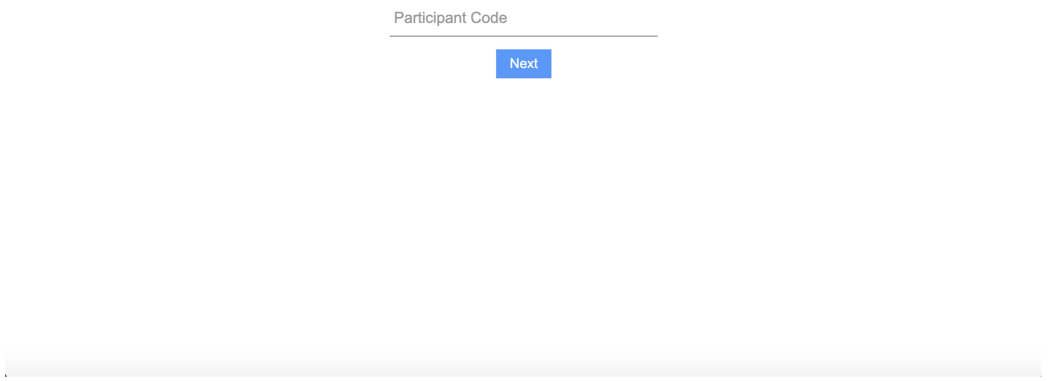

Fig. 16. Enter Participant Code page.

C NOT SIGNIFICANT RESULTS

Figure 21 shows the results of RDD for "overall relevance of the results" question with $k = 6$, indicating that there was no significant discontinuity.

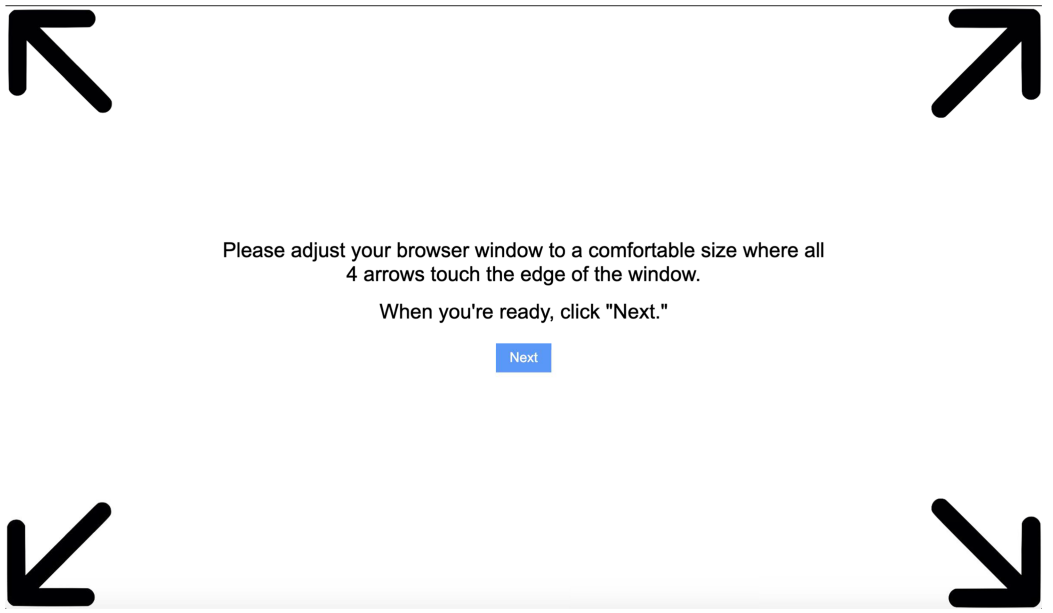

Fig. 17. Calibration page (used for Above-the-fold study).

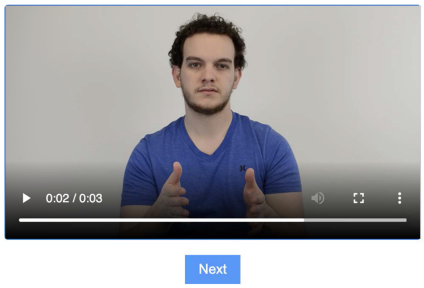

Fig. 18. View Video Stimuli page.

Figure 22 shows the results of RDD for "overall relevance of the results" question with $k = 8$, indicating that there was no significant discontinuity.

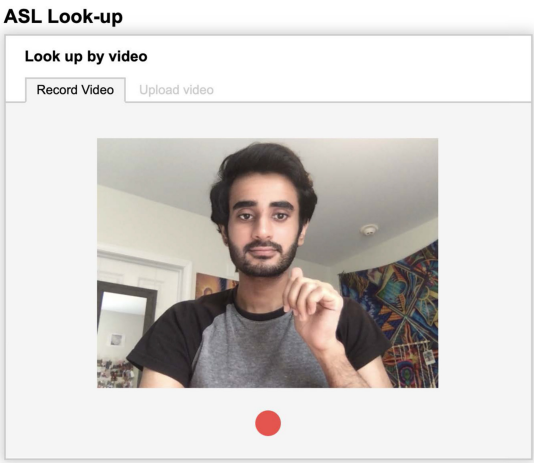

Fig. 19. Record a Sign.

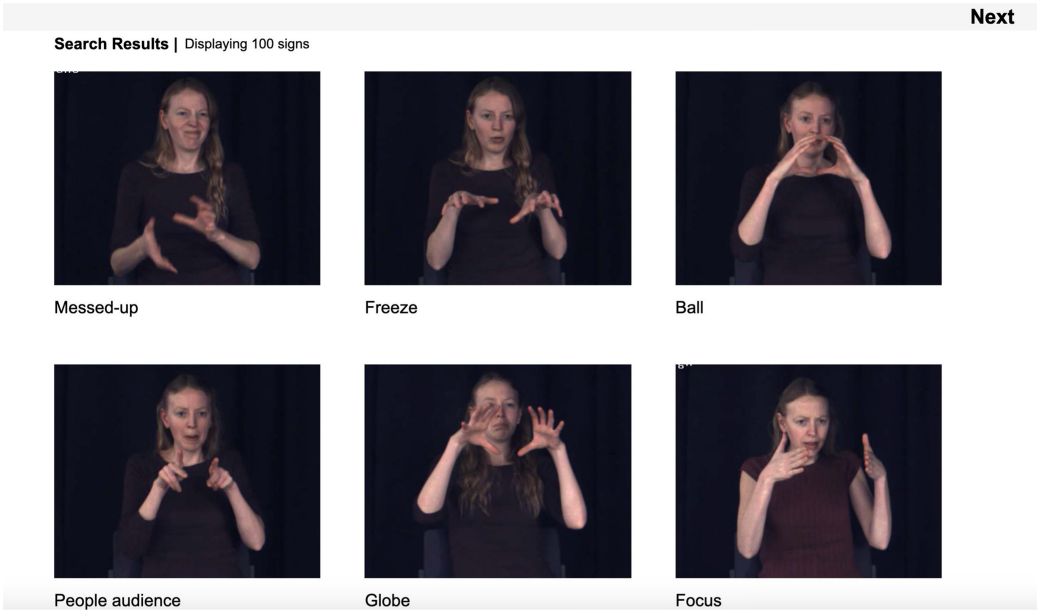

Fig. 20. Search Results page for Sign ROAD.

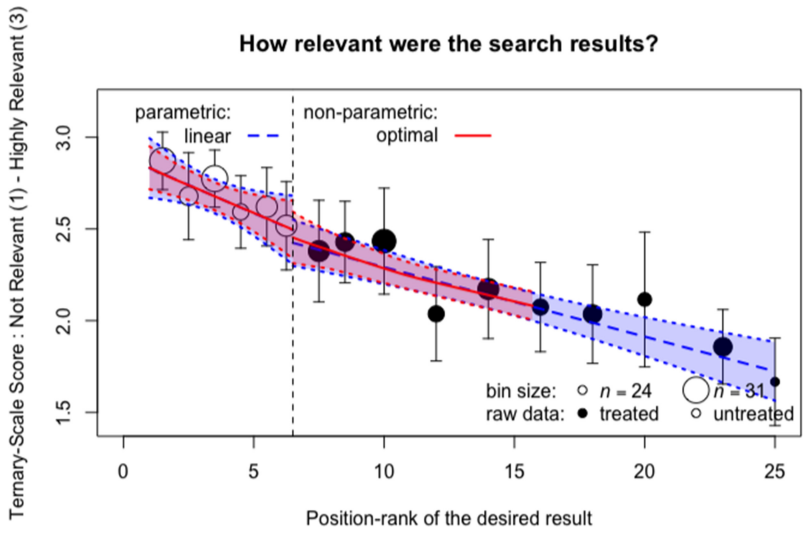

Fig. 21. Results of RDD for "percieved relevance of overall results" question with $k = 6$.

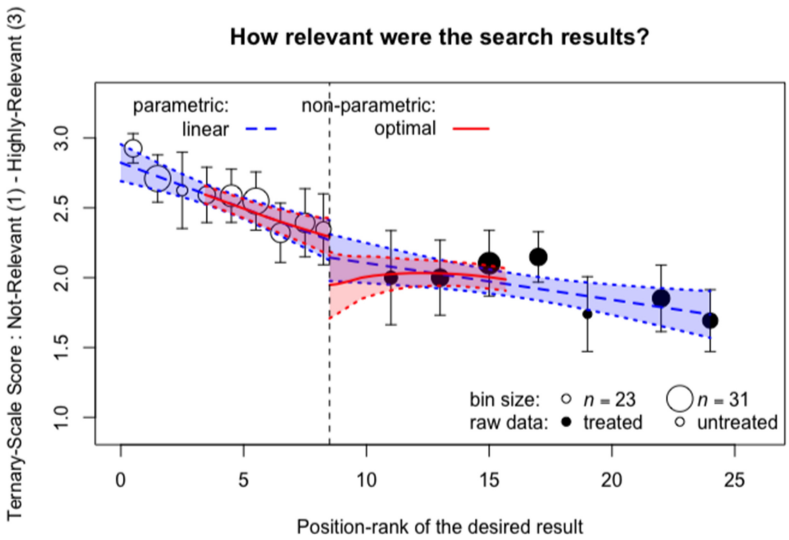

Fig. 22. Results of RDD for "percieved relevance of overall results" question with $k = 8$.

ACKNOWLEDGMENTS

We are grateful for the contribution of Aakash Maddi, Alexis Gordon, and Sarah Andrew in the collection of data from participants.

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. Association for Computing Machinery, New York, NY, 773–774. https://doi.org/10.1145/1277741.1277902

[2] Oliver Alonzo, Abraham Glasser, and Matt Huenerfauth. 2019. Effect of automatic sign recognition performance on the usability of video-based search interfaces for sign language dictionaries. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'19)*. Association for Computing Machinery, New York, NY, 56–67. https://doi.org/10.1145/3308561.3353791

[3] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Ashwin Thangali, Haijing Wang, and Quan Yuan. 2010. Large lexicon project: American sign language video corpus and sign language indexing/retrieval algorithms. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT'10)*, Vol. 2. European Language Resources Association (ELRA), Valletta, Malta, 11–14.

[4] Debra L. Blackwell, Jacqueline W. Lucas, and Tainya C. Clarke. 2014. Summary health statistics for U.S. adults: National health interview survey, 2012. *Vital and Health Statistics. Series 10, Data from the National Health Survey* 1, 260 (Feb. 2014), 171.

[5] 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 (CSCW'15)*. Association for Computing Machinery, New York, NY, 1837–1848. https://doi.org/10.1145/2675133.2675226

[6] Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10. https://doi.org/10.1145/792550.792552

[7] Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2017. ASL-LEX: A lexical database of american sign language. *Behav. Res. Methods* 49, 2 (2017), 784–801.

[8] Dorothy Casterline, Carl Croneberg, et al. 1965. A dictionary of American Sign Language on linguistic principles. Gallaudet College.

[9] Matias D. Cattaneo, Nicolas Idrobo, and Rocio Titiunik. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Retrieved from https://ideas.repec.org/p/arx/papers/1911.09511.html.

[10] Christopher Conly, Zhong Zhang, and Vassilis Athitsos. 2015. An integrated RGB-D system for looking up the meaning of signs. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'15)*. Association for Computing Machinery, New York, NY, Article 24, 8 pages. https://doi.org/10.1145/2769493.2769534

[11] H. Cooper, N. Pugeault, and R. Bowden. 2011. Reading the signs: A video based sign dictionary. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV'11)*. IEEE Computer Society, Los Alamitos, CA, 914–919. https://doi.org/10.1109/ICCVW.2011.6130349

[12] Natalia Lusin and Dennis Looney. 2018. Enrollments in languages other than english in united states institutions of higher education, summer 2016 and fall 2016: Preliminary report. In *Proceedings of the Modern Language Association of America*. 92. Retrieved from https://www.mla.org/content/download/110154/2406932/2016-Enrollments-Final-Report.pdf.

[13] Ralph Elliott, Helen Cooper, John Glauert, Richard Bowden, and François Lefebvre-Albaret. 2011. Search-by-example in multilingual sign language databases. In *Proceedings of the 2nd International Workshop on Sign Language Translation and Avatar Technology (SLTAT'11)*. SLTAT, Dundee, Scotland, 8. Retrieved from http://personal.ee.surrey.ac.uk/Personal/H.Cooper/research/papers/SBE_SLTAT.pdf.

[14] Scott B. Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. Association for Computing Machinery, New York, NY, 567–574. https://doi.org/10.1145/1277741.1277839

[15] Guido Imbens and Karthik Kalyanaraman. 2012. Optimal bandwidth choice for the regression discontinuity estimator. *Rev. Econ. Studies* 79, 3 (2012), 933–959.

[16] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Info. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[17] Jiepu Jiang and James Allan. 2016. Correlation between system and user metrics in a session. In *Proceedings of the ACM on Conference on Human Information Interaction and Retrieval (CHIIR'16)*. Association for Computing Machinery, New York, NY, 285–288. https://doi.org/10.1145/2854946.2855005

[18] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*. Association for Computing Machinery, New York, NY, 154–161. https://doi.org/10.1145/1076034.1076063

[19] Matt Jones, Gary Marsden, Norliza Mohd-Nasir, Kevin Boone, and George Buchanan. 1999. Improving web interaction on small displays. *Comput. Netw.* 31, 11–16 (1999), 1129–1137.

[20] Diane Kelly and Leif Azzopardi. 2015. How many results per page? A study of SERP size, search behavior and user experience. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. Association for Computing Machinery, New York, NY, 183–192. https://doi.org/10.1145/2766462.2767732

[21] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2015. Eye-tracking analysis of user behavior and performance in web search on large and small screens. *J. Assoc. Info. Sci. Technol.* 66, 3 (2015), 526–544.

[22] J. Lapiak. 2021. Handspeak. Retrieved from https://www.handspeak.com/.

[23] Scott Liddell. 2003. Grammar, gesture, and meaning in american sign language. *Grammar, Gesture, and Meaning in American Sign Language* 1 (Mar. 2003), 404. https://doi.org/10.1017/CBO9780511615054

[24] Dimitris Metaxas, Mark Dilsizian, and Carol Neidle. 2018. Linguistically-driven framework for computationally efficient and scalable sign recognition. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA), Miyazaki, Japan, 8. Retrieved from https://www.aclweb.org/anthology/L18-1271.

[25] Ross Mitchell and Michael Karchmer. 2004. Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the united states. *Sign Lang. Studies* 4 (12 2004), 138–163. https://doi.org/10.1353/sls.2004.0005

[26] Ross Mitchell, Travas Young, Bellamie Bachleda, and Michael Karchmer. 2006. How many people use ASL in the united states? Why estimates need updating. *Sign Lang. Studies* 6 (03 2006). https://doi.org/10.1353/sls.2006.0019

[27] Carol Neidle and Joan Cottle Poole Nash. 2015. American sign language. Jullie Bakken Jepsen, Goedele De Clerck, Sam Lutalo Kiingi, and William B. McGregor (eds.), Sign Languages of the World 1, 1 (2015), 31–70.

[28] Carol Neidle and Christian Vogler. 2012. A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAI). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC'12)*. Citeseer, OpenBU, Istanbul, Turkey, 8. https://open.bu.edu/handle/2144/31886.

[29] Antti Oulasvirta, Janne P. Hukkinen, and Barry Schwartz. 2009. When more is less: The paradox of choice in search engine use. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information*

Retrieval (SIGIR'09). Association for Computing Machinery, New York, NY, 516–523. https://doi.org/10.1145/1571941.1572030

[30] John Sören Pettersson and Malin Wik. 2015. The longevity of general purpose wizard-of-oz tools. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction (OzCHI'15)*. Association for Computing Machinery, New York, NY, 422–426. https://doi.org/10.1145/2838739.2838825

[31] Kishore K. Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Mach. Vision Appl.* 24, 5 (2013), 971–981.

[32] Naomi Robbins and Richard Heiberger. 2011. Plotting likert and other rating scales. *Proceedings of the Joint Statistical Meeting*. 9.

[33] Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04)*. Association for Computing Machinery, New York, NY, 13–19. https://doi.org/10.1145/988672.988675

[34] ShuR. 2021. SLintoDictionary. Retrieved from http://slinto.com/us.

[35] Jenny L. Singleton and Elissa L. Newport. 2004. When learners surpass their models: The acquisition of american sign language from inconsistent input. *Cogn. Psychol.* 49, 4 (2004), 370–407. https://doi.org/10.1016/j.cogpsych.2004.05.001 Retrievedfromhttps://www.ncbi.nlm.nih.gov/pubmed/24819891.

[36] Richard A. Tennant, Marianne Gluszak, and Marianne Gluszak Brown. 1998. *The American Sign Language Handshape Dictionary*. Gallaudet University Press, Washington, D.C.

[37] U. von Agris, C. Blomer, and K. Kraiss. 2008. Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, MLLR, and MAP. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08)*. IEEE Computer Society, Los Alamitos, CA. https://doi.org/10.1109/ICPR.2008.4761363

[38] Haijing Wang, Alexandra Stefan, Sajjad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad Kamangar. 2012. A system for large vocabulary sign search. In *Trends and Topics in Computer Vision*, Kiriakos N. Kutulakos (Ed.). Springer, Berlin, 342–353.

[39] Kimberly A. Weaver and Thad Starner. 2011. We need to communicate! helping hearing parents of deaf children learn american sign language. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'11)*. Association for Computing Machinery, New York, NY, 91–98. https://doi.org/10.1145/2049536.2049554

[40] Polina Yanovich, Carol Neidle, and Dimitris Metaxas. 2016. Detection of major ASL sign types in continuous signing for ASL recognition. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 3067–3073. Retrieved from https://www.aclweb.org/anthology/L16-1490.

Received July 2020; revised March 2021; accepted June 2021