Effect of Automatic Sign Recognition Performance on the Usability of Video-Based Search Interfaces for Sign Language Dictionaries

Oliver Alonzo*, Abraham Glasser*, Matt Huenerfauth
Golisano College of Computing and Information Sciences
Rochester Institute of Technology (RIT) Rochester, NY, USA
oa7652@rit.edu, atg2036@rit.edu, matt.huenerfauth@rit.edu

ABSTRACT
Researchers have investigated various methods to help users search for the meaning of an unfamiliar word in American Sign Language (ASL). Some are based on sign-recognition technology, e.g. a user performs a word into a webcam and obtains a list of possible matches in the dictionary. However, developers of such technology report the performance of their systems inconsistently, and prior research has not examined the relationship between the performance of search technology and users' subjective judgements for this task. We conducted two 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, which is what is often reported in literature; the similarity of the other words in the results list also 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 Classification
• Human-centered computing~Accessibility design and evaluation methods

Author Keywords
American Sign Language (ASL); Dictionary; Search.

INTRODUCTION
There are 28 million people who are Deaf or Hard of Hearing (DHH) in the U.S., and about 500,000 who use American Sign Language (ASL) as a primary form of communication [20]. Increasing knowledge of ASL may facilitate greater communication and inclusion of people who are DHH, and there is increasing interest in learning ASL among U.S. university students [12]. In addition, 90% of DHH children are born to hearing parents [11, 19], and there are well-documented educational benefits for those children if their parents learn ASL (even if not fluently) [25]. Prior research has found enthusiasm among parents of deaf children about using technology to support learning ASL [30].

If a learner of a spoken/written language encounters an unfamiliar word when reading, it is relatively straightforward to look up this unfamiliar word in a dictionary. In contrast, it is surprisingly difficult to look up an unfamiliar ASL sign in a dictionary. Most ASL dictionaries list signs in alphabetical order based on approximate English translations, but a user who does not understand a sign or know its English translation would not know how to find it. ASL lacks a commonly used written form or an intuitive "alphabetical sorting." There is no one-to-one correlation between ASL signs and English words, and no standard conventions for English labeling of ASL signs. Given this complexity, some dictionary creators invent sometimes-cumbersome methods for searching. Some dictionaries sort signs based on handshape (finger pose), with a handshape-listing defining a sort-order provided at the beginning [27]. Some Web dictionaries or linguistic tools, e.g. [6, 16, 22], enable users to select properties of the sign (handshape, number of hands used, etc.) and submit a search query. But a user must often still look through search results containing a large list of signs to find a match to the unfamiliar sign.

Researchers are developing technology for sign recognition from video, e.g. [1]. While systems' ability to understand entire ASL sentences is still limited, researchers have made more progress on identifying an individual ASL sign performed in isolation, e.g. [7, 29]. This technology could enable users to search for words in ASL dictionaries by performing an ASL sign they do not understand (from memory) into a webcam, or they could submit a video clip of such a word. The system would return a set of results of the most likely matches for the word. The user could then browse this set of results to look for the desired word. While prototypes of such systems have been investigated, e.g. [3, 10], usability evaluations have not been conducted with users. Some prior research has investigated users'

ASSETS '19, October 28–30, 2019, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-6676-2/19/10…$15.00
https://doi.org/10.1145/3308561.3353791

* O. Alonzo and A. Glasser contributed equally to this work.

requirements for an ASL dictionary, e.g. [4], but prior work has not established requirements *for the creators of sign-recognition technology,* to help these researchers know if their technology has sufficient accuracy or precision to support this dictionary-search application.

We conducted two studies in which users interacted with a prototype ASL dictionary search system, in which a user performs a desired ASL word into a webcam, and the system shows a results list (videos that may "match" the desired word). We found that the placement of the desired word in the list (e.g. 5th position) and the precision of the list (overall similarity of items to the desired word) both affected users' opinion of quality of the results. We also found that some information-retrieval metrics of search quality correlate better with users' preferences in this application context.

The contributions of this work are threefold:

- We identify properties of the output of ASL dictionary match algorithms that affect users' opinion of the system's quality; this finding informs designers of match algorithms as to which characteristics to optimize.
- Further, we identify ranges of where the desired item appears in the results (i.e. top-5 or top-10), where there is drop-off in users' satisfaction with the system's ranking of results or perception of result relevance. This finding informs designers of match algorithms what result they should optimize and report in evaluations.
- Finally, we identify an information-retrieval metric for evaluating the output of a match algorithm that correlates with user judgements better than a metric used in prior work on ASL dictionary search systems.

BACKGROUND AND RELATED WORK

As background, ASL is a natural language that is used among the community of people who are DHH in the U.S., Canada, and some other regions of the world. Other sign languages, e.g. British Sign Language (BSL), are used in other regions, and such languages are generally not mutually intelligible. Linguistic researchers generally agree that individual ASL words ("signs") consist of a set of basic parameters: handshape (one of a set of approximately 90 configurations of the fingers of the hand), orientation of the palms, location of the hands relative to the body or in the signing space (especially at the start and end of signs), movement properties, and non-manual expressions (movements of the face and body) [26].

While many ASL dictionaries or databases contain over 3,000 entries [21, 28], it is difficult to estimate the number of ASL signs in common use among signers, due to challenges in differentiating and counting individual words: There are variations in how a word may be produced in context, specialized jargon particular to certain fields, signs with regional dialectical variations in how they are performed, as well as productive methods for new-word-formation in ASL. A challenge for learners of the language (and for any technology that attempts to automatically recognize ASL words from video) is that the performance of an individual word may vary based on its context of use in a sentence, due to influence from the adjacent words or various phenomena by which the movement of a sign may depend upon spatial or grammatical aspects of the sentence, e.g. [17].

Inconsistency in Evaluating ASL Dictionary Search

Some online ASL dictionaries allow users to manually select parameters of a sign they are seeking to formulate a search query (e.g. Handspeak or SLinto)[1]. In these systems, users can select, for instance, the handshape, the location and the movement of the sign. Researchers have discussed challenges users face with the interfaces of such systems, which may be cumbersome, overly constrain how users select features, and provide poor matching results [4]. This finding has motivated researchers to develop machine-learning based systems and to improve the submission interface, by allowing users greater freedom in selecting features [4], which they found allowed users to obtain a desired word within the top-10 results 84.93% of the time in an experimental evaluation.

A growing body of research is also investigating the use of computer-vision based ASL dictionary systems in which the user can use a camera to perform the sign they are looking for as a query [9, 7, 10, 29]. Some work has reported being able to identify the correct word among the top-5 in 97.6% of searches [18]. However, there is inconsistency is how results are reported in this field. Results depend upon the size of the vocabulary (which differs across systems) and the diversity of the human appearance and movement that the system is evaluated against. Further, some researchers test their systems against relatively small datasets, and they may perform "user-dependent" testing, in which the same humans are in the training and testing data.

It is important to note that regardless of the input method for searching, recognizing a sign is a challenging task: a student may not precisely remember how a word they had seen was performed, and thus, they may make a slight error when performing it or manually selecting its features. In addition, as discussed above, there are various linguistic reasons why the appearance of an ASL sign may vary, as it is used in context in a sentence. For both of these reasons, it may not be reasonable to expect that a dictionary-searching system identifying an ASL sign from a student's input would be able to identify a single word that exactly matches what the student is seeking. There are many potential sources of error and ambiguity. Thus, the systems mentioned above present users with a "page of results" that shows the user a set of possible matches for the sign they are seeking; the correct match may not always be the first result.

Researchers investigating the development of these kinds of dictionary-searching systems typically measure the performance of their system based on metrics of the

[1] www.handspeak.com/ and www.slinto.com/us/

percentage of trials in which the system satisfies a binary condition: whether the desired word is within the top-k results in the list they provide [7, 9, 18, 29]. In other words, the systems are evaluated by focusing on the value of rank k of the desired word in the list of results provided by the system. However, there is no consensus as for what values of k to report, with some studies reporting up to the top-4 results [10], while others report up to the top-375 results [3].

Some evaluations have been performed to determine how systems perform with potential users [7, 10]. However, to our knowledge no user studies have focused on how the *performance of a system may affect user satisfaction*, which may help explain the lack of consensus about the reporting metrics mentioned above. Thus, researchers in this field of sign-language recognition lack a good set of requirements, as to what level of performance their technology would need to achieve in order to support ASL dictionary search applications. Research is needed on how differences in the accuracy or precision of the search results would affect the judgements of end-users as to the performance and usability of the system.

Information Retrieval and Usability

Since we seek to understand how the performance of the automatic recognition component underlying a dictionary search system that returns a list of results would affect user satisfaction, it was also natural to consider related work in the field of information retrieval (IR). In prior work, researchers have attempted to understand the relationship between different metrics of performance of IR systems, and user judgements of system quality, e.g. [15]. One study asked users about their satisfaction with a system's ranking, accuracy, and coverage of the results; researchers examined if there was a relationship between user responses and metrics of search results commonly used in IR, e.g. accuracy, precision, or Discounted Cumulative Gain (DCG) [2]. They found that user satisfaction with ranking of results correlated strongly with the precision of the results list, but no other metrics significantly correlated with other user responses. In another study, researchers found a strong correlation between the relevance of the results provided by a search engine and user satisfaction with the results [13]. They also found that the nature of the query (i.e. whether it was navigational, informational or transactional [5, 24]) affected how well relevance correlated with satisfaction. However, in the context of an ASL dictionary-searching systems, it is unknown how different metrics of the quality of search output may correlate with user satisfaction. In addition, because the task of finding a match for a word may be different in nature from the task of finding a website using a search engine, it is unknown how the relevance of the results may be related to user satisfaction in this context.

RESEARCH QUESTIONS AND LIST OF STUDIES

We are investigating factors that may influence the usability of imperfect systems for providing users a list of results (videos of ASL signs) for a given query in which the user is seeking a particular desired word in a dictionary. Specifically, our first research question is 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?

We designed a Wizard-of-Oz prototype to simulate an ASL dictionary in which a user can look up the meaning of a word by performing the desired word into a webcam. There was no actual use of automatic recognition technology in this prototype: Instead, we pre-determined the set of results that were displayed, to control for the apparent performance of the recognition technology (we faked it via a Wizard-of-Oz approach). Thus, we controlled where each sign in the list of results was placed, e.g. 10th in the list. We conducted a user study (henceforth referred to as the **"placement study"**) to investigate how the position of the desired sign in a list of results impacts users' judgements about the overall system. While we measured an effect of the position of the desired sign on users' judgements, participants also commented that another factor influenced their 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 suggested a second research question:

2. 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?

To empirically investigate this issue, we conduct a final user study (henceforth referred to as **"precision study"**), using a similar Wizard-of-Oz dictionary-search prototype as the placement study above. As discussed below, in this final study, we did observe an effect of the overall search results precision on users' judgements of the system. 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 [4] actually relate to users' judgements of quality:

3. 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 both of our 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.

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

RESEARCH QUESTION 1: PLACEMENT STUDY

Since the technology for automatically recognizing an ASL word in a video is imperfect, in this study, we wanted to understand how the performance of that technology may influence users' opinion of the quality of the overall system. Thus, the independent variable in this study was the placement of the desired word in the results list. For use in this study, we created a Wizard-of-Oz prototype of an ASL-to-English dictionary search system, where a user performs an ASL sign into a webcam and the system returns a list of results (of likely "matches" for words in its dictionary that look like what the person had performed). Specifically, the prototype consisted of sequences of web pages viewed using Google Chrome on a 15.6-inch Lenovo ThinkPad P52 Mobile Workstation with a built-in webcam. Notably, this prototype system did not use actual automatic recognition technology: Instead, we knew in advance which desired words the user would be searching for, and our prototype simply returned a predetermined set of results, regardless of what the user actually signed into the webcam. This approach enabled us to have control over the results list so that we could investigate how the accuracy of the automatic recognition technology may affect the judgements of users as to the quality of the overall system.

Upon entering the system, users were prompted with a stimulus video of a person performing an ASL sign. They were asked to imagine that they had seen someone perform this ASL sign, but they did not know what it meant and needed to search for it. There were 32 different stimuli videos of individual ASL words used in the study, which we had recorded from a native ASL signer (who grew up using ASL since early childhood) in a studio setting. An image of a stimulus video appears in Figure 1. The set of words used as stimuli consisted of relatively advanced vocabulary, which a beginning ASL student would be unlikely to know.[2]

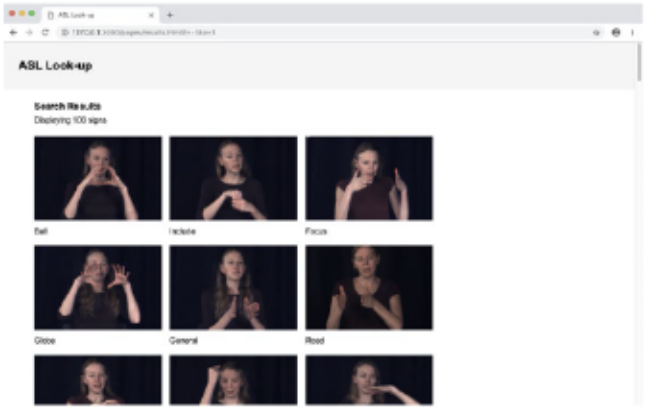

Figure 2. Sample list of results obtained from our prototype.

After viewing a stimulus video and pressing a "next" button, participants were taken to a screen where they were asked to press a "record" button, to begin a 3-2-1 countdown timer on-screen. Participants were asked to perform the desired word (from memory) into the webcam, to search for the word in the dictionary. Given the Wizard-of-Oz nature of our system, the purpose of this video submission was simply to make the users feel as if they were truly querying the system. Once they were done performing the sign, they could click on a stop button that would take them to the results page, where they were shown 100 results for their search query.

As shown in Figure 2, the results were displayed as a scrollable webpage, with approximately three rows of results onscreen at a time. The first row contains items 1, 2, and 3, and the second row contains items 4, 5, and 6, etc. The layout of the results page mimicked image/video search engines, e.g. Google Images or YouTube. The videos could be played by clicking on each, with small text labels below each. The videos on the results page consisted of subsets of a set of 291 videos extracted from Boston University's (BU) American Sign Language Lexicon Video Dataset (ASLLVD)[3] [21, 22]. The short text label that appeared below each video consisted of the first English word or phrase listed as a translation for that word in the ASLLVD database.

Collection of ASL Videos Appearing on Results Page

The ASLLVD contains over 3300 words, yet we selected a subset of 291 words for use on our results page of our prototype. The selection of this subset was done to carefully include a variety of "look-alike" words that might be relatively similar in appearance to some of our stimuli videos (the words participants were asked to search for). To build this subset of 291 words from the ASLLVD, a native signer on our team manually searched through the ASLLVD collection, to add words to this subset, as follows:

2 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.

3 http://secrets.rutgers.edu/dai/queryPages/search/search.php

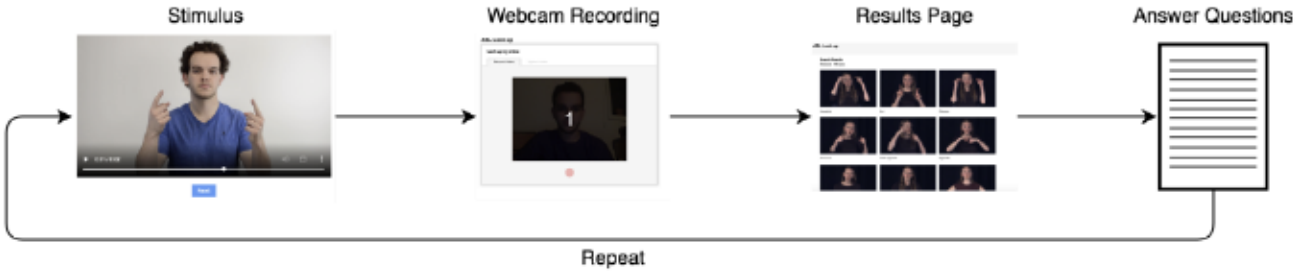

Figure 3. Flowchart of the procedure our participants followed during the experiment.

- For each of the 32 stimuli words, approximately 3 other words were selected from the ASLLVD that were "extremely similar in appearance" to each stimulus.
- Our 32 stimuli words began with a total of 8 different ASL handshapes. Thus, for each of these 8 handshapes, we identified an additional 15 signs from the ASLLVD that also used this handshape.
- Some of our 32 stimuli words were performed near the head and some in front of the torso. Thus, another 30 signs from the ASLLVD were selected with a location near the head, and 30 that were in front of the torso.
- Lastly, we selected 70 words at random from the ASLLVD to add to our subset. This was done so that we would have some words to show in our results page that would seem unrelated to the desired word, if this was necessary for a particular experimental study design.

Since there were some overlaps among the words identified through this procedure above, this yielded 291 ASL videos, which we used on the results page of our prototype. These characteristics of each sign in our set of 291 videos were used to carefully engineer the set of results that were shown to individual participants. Thus, our control over the selection and order of the signs within a results list is what makes our prototype a *Wizard-of-Oz*. In this *placement study*, we controlled the ordinal position in the results list where the desired word (the sign the person was looking for) appeared. When a participant performed a search and saw a list of results, the desired word was placed at a specific position k in the list. In addition to selecting where to place the desired word, we also had to select how to fill the rest of the list with "distractors" (words that did not match the desired word). Our goal was to make the results seem realistic, as if they had been sorted based on how well an automatic system believed each word matched the query. For each stimulus, we created a sorted list of 100 items for the results page using our 291 ASLLVD videos, as follows:

1. We set aside the video that was a match for the desired word, since at the end of this process we would place this result at a particular position k in the results list.
2. We began our list with the words that had been hand-selected by a native signer as being "extremely similar" to the desired word.
3. From the remaining words, we took those with the same handshape as the desired word. The order of these was randomized, and they were placed after the items in 2.
4. From the remaining words, we took those with the same location (near head or torso) as the desired sign. These were placed after the items in 3, in random order.
5. If steps 2-4 above had not yet yielded 99 words, then we selected the remaining words at random from those not yet selected, to appear at the end of the list.
6. Finally, we inserted the video that was a match for the desired word at a particular position k in the final list.

Conditions and Sequence of Presentation

We had to select the specific values of k to use across the conditions in this experimental study. During some pilot testing of our prototype system with four participants (whose response data is not included in the findings below), we had included a larger set of conditions $k = \{1, 5, 10, 20, 40, 90\}$, yet we noticed that users' responses to questions about satisfaction appeared to drop substantially as k rose above 10. To avoid our study from being underpowered, we needed to reduce the number of conditions, and we therefore selected $k = \{1, 5, 10, 20\}$. Across the participant trials in our study, the assignment of conditions to each stimulus (the word being searched for) and the order of these conditions was counterbalanced using a Latin Squares schedule.

Data Collection Procedure

For each word that they searched for, participants were asked to identify the best match on the results list and to write down on a paper answer sheet the English label appearing below that video. Even though the desired word always appeared somewhere on the results list, participants were instructed to write down 'not found' if they believed the desired sign was not in the list of results. Next, participants rated their **satisfaction with the way the results were ranked** using a 5-point Likert-scale and also rated their perceived **relevance of the results** using a ternary scale: *highly relevant*, *relevant*, or *not relevant*. These two questions were used following the methodology of [2]. After providing their ratings, participants repeated the entire process for another stimulus video; this iterative procedure is illustrated in Figure 3.

Once all the searches were completed, a semi-structured interview was conducted, which included questions about how they would describe the list of results obtained and what they wish were different about the system. These questions were included to understand what aspects of the results participants focused upon. Participants were then informed about the Wizard-of-Oz nature of the prototype.

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

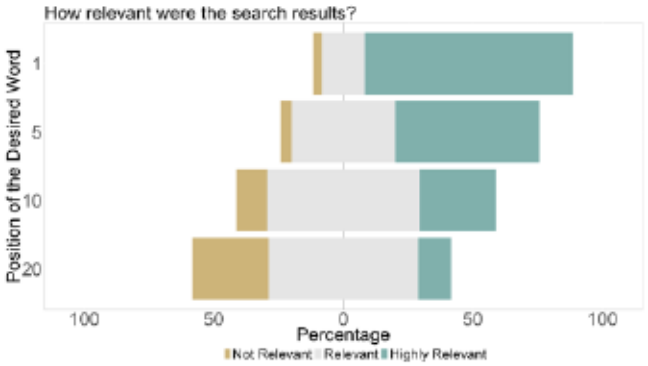

Figure 5. Placement Study: Users' judgements of the relevance of the results.

Recruitment and Participants

Participants were recruited by email advertisement shared through professors of ASL at Rochester Institute of Technology. The advertisement included two key criteria: having studied ASL in the past 5 years and started learning ASL after the age of 5. Participants received $40 cash compensation for this in-person, 70-minute study. Sixteen people participated in the study: 15 females and 1 male, mean age of 22, and experience learning ASL varying from 0.5 to 15 years. All identified as hearing. As each participant performed 32 searches, we collected data for 512 total searches. Five responses were left blank, while 59 responses were excluded for separate analysis because users indicated that they did not find the sign or wrote down the meaning of a sign that was similar to, but not the one we intended.[4]

Findings: Effect of Placement on User Satisfaction

When the desired word was closer to the top of the results list, users' satisfaction with the way the results were ranked was higher. A Friedman test ($\chi2$=182.682, DF=3, $p<0.01$) indicated significance for the differences in satisfaction across the levels of k = {1, 5, 10, 20}. To compare levels pair-wise, we used a Wilcoxon Signed Ranks test (with Bonferroni corrections); this post-hoc test indicated significant differences across all levels ($p< 0.01$).

Of course, finding that user satisfaction with a search system drops as the desired item appears lower in the list of results was not a surprising result. However, it is useful to consider these findings in light of prior work: As discussed earlier, there is disagreement among researchers studying ASL dictionary search technologies about how to report their results: Researchers often report the percentage of time that a desired word is in the top-k of their results, with different papers presenting results for various values of k [3, 10]. For instance, some systems reported how often the desired word appeared in the top-20 of their ranked search results [3, 8].

As shown in Figure 4, somewhere between k=10 and k=20, user satisfaction drops below the midpoint of the scale (*neither agree nor disagree*)[5]. This result suggests a range of values of k that may be of particular importance for researchers to report the accuracy of their systems.

Findings: Effect of Placement on Perceived Relevance

Figure 5 displays participants' responses to the question about the relevance of the results. A Friedman test ($\chi2$=80.678, DF=3, $p<0.01$) indicated significance for the differences in perceived relevance across the different levels of k = {1, 5, 10, 20}. Post-hoc Wilcoxon Signed Ranks tests with Bonferroni corrections indicated significant differences across all levels ($p<0.01$). These results indicate that the position of the queried sign in the list has an impact on users' perceived relevance of the overall list.

Open-Ended Feedback Comments from Participants

When asked about how to describe the results, participants often talked about the position of the queried sign in the list or how far down the list they would have to scroll:

> "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

We had expected comments like those above, since our study had been focused on the position of the desired word on the results list. However, participants also commented about how their impression of the 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

[4] On 33 occasions, participants were unable to find a match (4 occurred when the match for the desired sign was at position k = 1; 8 when at k = 5; 9 when at k = 10; and 12 when at k = 20). For these 33, the median and mode response to the satisfaction question was 'disagree,' and the median and the mode response to the perceived relevance question was 'not relevant.' On 26 occasions users identified matches that were similar in appearance to the desired word but not the match we had intended (none occurred when the correct match was at position k = 1; 5 when at k = 5; 9 when at k = 10; and 12 when at k = 20). For these 26, the median and mode response for satisfaction was '*agree*,' and the median and mode response for relevance was '*relevant*.'

[5] Figure 4 displays Likert response data using a diverging stacked bar graph, as recommended in [23], which centers the neutral response.

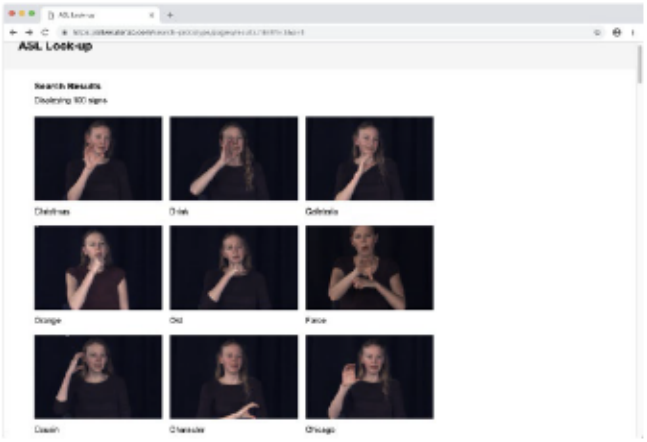

Figure 6. Sample list of results with high precision for the sign for Chicago.

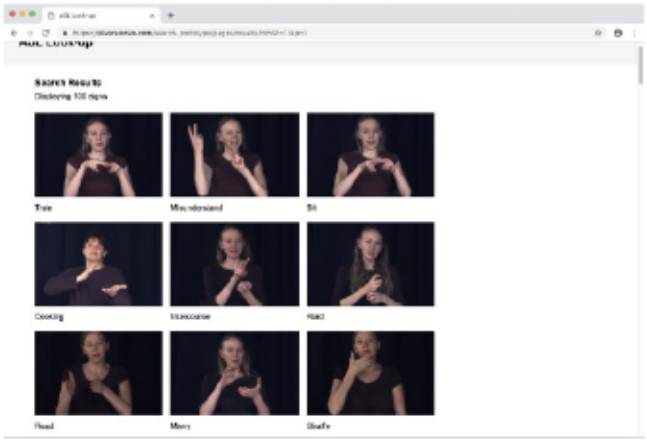

Figure 7. Sample list of results with low precision for the sign for giraffe.

> "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

One participant was concerned that if results included words that looked similar to the desired word, a novice searching for a particular word could 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

Based on these participants' comments, we realized that the ranking on the results page of where the desired sign appears may not be the only relevant factor influencing user satisfaction with a dictionary search. We may also need to examine whether user satisfaction and perception of the relevance of the results is affected by the degree to which the surrounding words on the results list appear similar to the desired word. For concision, we will henceforth refer to this property of the search results as the **precision** of the results.

RESEARCH QUESTION 2: PRECISION STUDY

Based on these comments, we wanted to examine how the precision of the results may influence user judgements about a system's quality. Our motivation was that 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 was the top-k ranked results of the system. However, we did not find studies in which researchers had conducted an analysis of the surrounding words on their results list to determine how close they appear to the desired word. If users' judgements of the quality of a dictionary search system are affected by the precision of the results, then it may be important for researchers to report such data.

We conducted a follow-up study, nearly identical in design to our earlier placement study. The appearance of the prototype, the task that users were asked to perform, and the questions they were asked were identical to the prior study.

The difference was that in this new *precision study*, we kept the position of the desired word on the results list (nearly) fixed and controlled the precision of the overall results list.

Conditions and Sequence of Presentation

While we could have planned a two-factor study that examined both the variables of placement and precision (and thereby been able to investigate if there were interaction effects), we were skeptical that we could recruit a sufficiently large sample of ASL students (who had not participated in our prior placement study) to ensure that a two-factor study would be sufficiently powered. For this reason, we decided to hold the variable of placement constant and explore the variable of precision in this study. We wanted to select a value for placement from our prior study with relatively middle values for satisfaction and which reflected likely improvements to the state-of-the-art of automatic recognition technology (i.e. considering the top-20 basis of reporting in [3, 8]). For this reason, we selected a placement value of k=10. However, we were concerned that participants in the study might notice that the desired word always appeared in the same position on the results list; thus, we allowed the placement of the desired word on the results page to vary randomly among values of $k = 10 \pm 2$.

The independent variable in this study was the precision of the results, i.e. how similar the distractors on the result list (the other 99 words) are to the desired word, as follows:

- **High precision:** The list of 99 distractors began with the words that had been manually selected by a native ASL signer as being "extremely similar" in appearance to the desired word. The next 15 words on the list consisted of words that had an identical handshape to the desired word. This was followed by approximately 30 signs that had a similar location to the desired sign, and the end of the list contained randomly selected words. Figure 6 shows an example of list of results with high precision.

Figure 8. Precision Study: Users' satisfaction with the way the results were ranked.

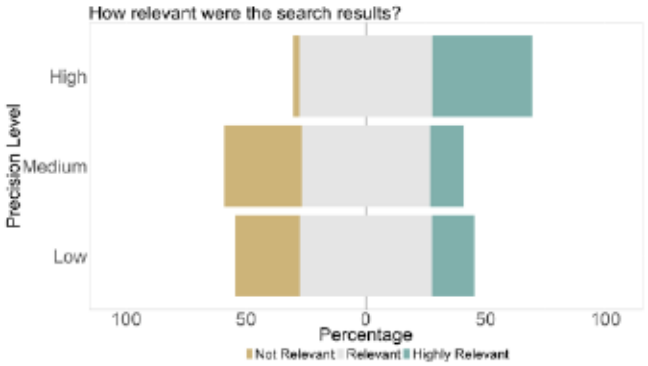

Figure 9. Precision Study: Users' judgements of the relevance of the results.

- **Medium precision**: In this case, no words that had been determined to be "extremely similar" to the desired word were included in the list. The list consisted of a random sequence of signs that contained an even mix of signs that appeared at a near-the-face and near-the-torso location, followed by randomly selected words.
- **Low precision:** The distractor list was intentionally filled with signs with a different handshape and a different location (than the desired sign). Figure 7 shows an example of list of results with low precision.

Each participant engaged in a total of 30 searches, with 10 at each precision level. The sequence of each stimulus was randomized, and the assignment of condition was counterbalanced across participants.

Recruitment and Participants

As before, participants were recruited by email advertisement at our university, and there were two inclusion criteria: having studied ASL in the past 5 years and started learning ASL after the age of 5. Participants received $40 cash compensation for this in-person, 70-minute study. There was no overlap between the set of participants in the placement study and those in the precision study.

A total of 10 ASL students participated. Participants' included 8 females and 2 males, mean age of 23.3 and experience learning ASL ranging from .5 to 15 years. Nine participants identified as hearing, and one identified as deaf. With each participant performing 30 searches, we gathered data for 300 total searches. As in the previous study, three participants left their responses blank, and a total of 46 responses were excluded for separate analysis because users either did not find the desired sign or identified a sign that was different than what we had intended as the match.[6]

Findings

Precision had a significant impact on users' **satisfaction** with the way the results are ranked. A Friedman test ($\chi2$=16.526, DF=2, $p<0.01$) indicated a significant effect of precision on user satisfaction. Pairwise post-hoc comparison using Wilcoxon Signed Ranks with Bonferroni corrections indicated significant differences between the *high* level and both the *medium* ($p = 0.0002$) and the *low* ($p = 0.0146$). However, no significant difference was observed between the *middle* and the *low* levels ($p = 0.058$). Figure 8 shows the percentages of responses across the different levels.

We also observed a significant difference in users' rating of the **relevance of the results**: A Friedman test ($\chi2$=35.438, DF=2, $p<0.01$) indicated a significant effect of precision on perceived relevance. Pairwise post-hoc comparison using Wilcoxon Signed Ranks with Bonferroni corrections indicated significant differences between the *high* level and both the *medium* ($p = 5.5767E-7$) and *low* ($p = 0.000027$). However, no significant difference was observed between the *middle* and the *low* levels ($p = 0.353160$). Figure 9 shows the percentages of responses across the different levels.

Our findings indicate that users not only want a list of results that includes exact match for the query within the top k results, but they also prefer a coherent list containing signs that are similar to the query towards the top. This result has important implications on the way in which researchers studying ASL dictionary search systems should report the performance of their search-matching algorithm.

RESEARCH QUESTION 3: METRICS FOR ASL SEARCH

Researchers who design algorithms for searching for matches to a query often report the performance of their system using a metric that provides a single composite score that indicates the overall quality of a set of results that are returned. As discussed in [4], there are several metrics used within the field of information retrieval, which may be

[6] In 34 cases, participants did not find the sign (7 occurred at the high precision level, 14 at medium, and 13 at low). Of the 34, the median and mode response for satisfaction was *'disagree,'* and the median and mode response for perceived relevance was *'relevant.'* On 12 occasions, users identified matches that were similar to the stimulus but not what we had intended (11 occurred at the high precision level, none at medium, and 1 at low). Of the 12, the median mode response for satisfaction was *'strongly agree' and the mode, 'agree,'* while the median and mode response for perceived relevance was *'highly relevant.'*

User's Judgement	Placement Study bDCG	Placement Study nDCG	Precision Study bDCG	Precision Study nDCG
Satisfaction	0.665 **	0.646 **	0.208 **	0.196 **
Relevance	0.511 **	0.530 **	0.099	0.295 **

Table 1. Spearman correlation between different metrics and users' satisfaction with the ranking of the results, and their perceived relevance of the results (indicates $p < 0.01$).**

suitable for use in reporting the performance of ASL dictionary systems. Based on the findings for RQ2 above, we advocate for the use of the Discounted Cumulative Gain (DCG) [14], which considers both the placement and the precision of the overall list of results, both of which our studies suggest are important for users. DCG considers both the positions and the relevance of each item in a list of search results, so that a composite score can be calculated of the overall "quality" of the result. The metric depends upon the length of the list of results shown p, as shown in Equation 1. The relevance of each result is given as rel_i.

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

Equation 1. Traditional formula for DCG.

Of course, in many applications, queries may have different number of results; to compare DCG scores across lists of results of different lengths, a normalized DCG (nDCG) is available, in which the DCG is divided by the Ideal DCG (IDCG). This IDCG is the maximum possible DCG of a list of results of length p, which may be obtained by sorting the list of results according to the relevance of each item and then computing the DCG score using the position $\pi(i)$ of each result in that sorted list, as shown in Equation 2 below.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \ ,$$

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

Equation 2. Formula for nDCG, where $\pi(i)$ is the position of the i^{th} result in relevance order.

While this nDCG metric considers both placement and precision of the results list, it has not previously been utilized to present the results of an ASL dictionary search system. Unfortunately, most prior work on ASL dictionary search systems has not used these metrics from the information retrieval literature: The most sophisticated consideration of this issue to date is in the search-by-selecting-features ASL dictionary search system of Bragg et al. [4]. However, in that work, researchers used a simpler version of this metric that merely considered whether each item in the list of results was (a) a perfect match to the desired word or (b) not a match. Specifically, they used an alternative version of this metric we refer to as Binary DCG (bDCG), as shown in Equation 3, in which the relevance rel_i of items is 1 if it is the desired word, 0 if it is not.

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

Equation 3. Formula for bDCG.

While prior work on ASL dictionary search systems had simply argued that particular metrics may be suitable proxies for the overall quality of search results, we can go further. In our two studies, since we had obtained judgements from users as to their "satisfaction with the way the results were ranked" and their opinion of the "relevance of the search results," we can actually compare these metrics empirically, to determine *which correlates better* with users' preferences. For each list of results that had been displayed to each participant in our studies, we calculated both the bDCG and nDCG metrics. For bDCG, the rel_i of individual items in the list of results is the simple binary decision explained above. For nDCG, to calculate the relevance of individual items in our search results, we used the following heuristic: 1 if an item is the desired word, 0.5 if it is a sign that we had manually identified as being "extremely similar" to the desired word, 0.25 if it is a sign that has the same handshape or the same location as the desired word, or 0 otherwise.

We then compare the correlation between nDCG and bDCG with user satisfaction and perceived relevance of the results. As shown in Table 1, both metrics correlate to a similar degree with the users' "satisfaction" score in each study, and in our *placement* study, both metrics correlate to a similar degree with users' opinion of the "relevance" of the results. However, when we examined the responses from our *precision* study, in which the lists of results shown varied widely in regard to how similar the entire set of results was to the desired word, **there was a difference in how well our metrics correlated with users' judgements:** While nDCG was correlated with users' judgements about the relevance of the results, the bDCG had no significant correlation.

This result suggests that researchers who are investigating methods for searching for a match in an ASL dictionary, whether by enabling users to select linguistic elements of words on a form [4] or through automatic recognition of

video input [29], should report the results of their system using metrics such as nDCG 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.

CONCLUSION, LIMITATIONS, AND FUTURE WORK

Overall, our findings provide guidance for researchers studying sign-language dictionary search systems or for researchers who are developing underlying technologies for identifying matches, e.g. sign recognition from video. 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 and 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 [2]: (a) users' satisfaction with the way the results are ranked and (b) users' perception of the relevance of the results. We found that user satisfaction for a search system dipped below the midpoint of the satisfaction scale somewhere between position 10 and 20. 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-10 or higher.

In a follow-up study, we found that *even when the placement of the desired word in the list of result is held constant,* users' perception of the quality of a search tool is affected by the precision of the other words in the results. However, ASL dictionary search researchers generally do not report the performance of their systems for this metric. Finally, we found 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 incorporate the relevance of each result in the list. Specifically, we found that nDCG correlated with both our users' reported satisfaction with how the search results and their impression of the overall relevance of the results.

There were several limitations in our study:

- A two-factor study (placement and precision) with more participants could allow us to understand any interactions between these two variables.
- We may also investigate a more realistic search context in which the participant sees a stimuli *sentence* containing an unfamiliar word; such a study would enable us to understand how users may incorporate contextual clues about a word's possible meaning into their searches.
- A future study could also examine alternative design choices for how to present other metadata, e.g. the definition of each word, on the results list, since some signs either do not translate directly to a brief English word or phrase or may have multiple translations.
- In addition, it would be useful to consider a wider variety of users in a future study: (a) This study focused on primarily hearing ASL students, but future work is needed to investigate the potentially unique needs of DHH users of an ASL dictionary. (b) Our study included students at the beginner-to-moderate range of ASL skill, but a larger study that examined the skill of students as a variable may enable new insights.
- In this study, while we had engineered our set of stimuli to avoid words that students in a first-semester ASL course may be familiar with, some of our participants indicated that they were familiar with some of our ASL stimuli words. While students do look up known words in ASL dictionaries at times, e.g. to view videos, 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 completely unfamiliar word.
- In our precision study, in which the placement of the desired item in the results list was kept relatively constant, we did vary the placement of the desired word randomly among values of $k = 10 \pm 2$, to prevent participants from noticing that the desired word always appeared in the same placement. Given this variation (and the variation in the composition of the surrounding signs on the results page), it is unlikely that participants noticed that the results were generally near placement 10 ± 2, but in future work, participants should be asked in debriefing interviews whether they had noticed this regularity.
- Lastly, 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. While we had calculated this heuristically in our study, 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 would build upon the contributions of this current paper, which has identified how the performance of dictionary-search technologies affect users' satisfaction with a system, and which has also provided methodological guidance to dictionary-search and recognition researchers on how they should report their results.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under award No. 1763569. We are grateful for the contribution of Aakash Maddi in the collection of data from participants for this work.

REFERENCES

[1] Ulrich Von Agris, Christoph Blomer, and Karl-Friedrich Kraiss. 2008. Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, MLLR, and

MAP. *2008 19th International Conference on Pattern Recognition* (2008). DOI: http://dx.doi.org/10.1109/icpr.2008.4761363

[2] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 07*(2007). DOI: http://dx.doi.org/10.1145/1277741.1277902

[3] Vassilis Athitsos et al.2010. Large Lexicon Project: American Sign Language Video Corpus and Sign Language Indexing/Retrieval Algorithms. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*.

[4] Danielle Bragg, Kyle Rector, and Richard E. Ladner. 2015. A User-Powered American Sign Language Dictionary. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW 15*(2015), 1837–1848. DOI: http://dx.doi.org/10.1145/2675133.2675226

[5] Andrei Broder. 2002. A taxonomy of web search. In *SIGIR Forum*. 3–10.

[6] Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2016. ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*49, 2 (2016), 784–801. DOI: http://dx.doi.org/10.3758/s13428-016-0742-0

[7] Christopher Conly, Zhong Zhang, and Vassilis Athitsos. 2015. An integrated RGB-D system for looking up the meaning of signs. *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments - PETRA 15*(2015). DOI: http://dx.doi.org/10.1145/2769493.2769534

[8] Christopher Conly. 2016. *Improving Accuracy in Large Vocabulary Sign Search Systems*. Doctoral Dissertation, The University of Texas at Arlington.

[9] Helen Cooper, Nicolas Pugeault, and Richard Bowden. 2011. Reading the signs: A video based sign dictionary. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*(2011), 914–919. DOI: http://dx.doi.org/10.1109/iccvw.2011.6130349

[10] Ralph Elliot, Helen Cooper, Eng-Jon Ong, John Glauert, Richard Bowden, and François Lefebvre-Albaret. 2011. Search-By-Example in Multilingual Sign Language Databases. In *Procs. of Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*.

[11] Gallaudet Research Institute. 2011. *Regional & national summary report of data from the 2009-10 annual survey of deaf and hard of hearing children and youth*, Washington, DC.

[12] David Goldberg, Dennis Looney, and Natalia Lusin. 2015. Enrollments in Languages other than English in United States Institutions of Higher Education, Fall 2013. *Modern Language Association*(2015).

[13] Scott B. Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction? *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 07*(2007). DOI: http://dx.doi.org/10.1145/1277741.1277839

[14] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*20, 4 (2002), 422–446. DOI: http://dx.doi.org/10.1145/582415.582418

[15] Jiepu Jiang and James Allan. 2016. Correlation Between System and User Metrics in a Session. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR 16*(2016). DOI: http://dx.doi.org/10.1145/2854946.2855005

[16] Jolanta Lapiak. Sign Language • ASL Dictionary. https://www.handspeak.com/

[17] Scott K. Liddell. 2003. Grammar, gesture, and meaning. (2003), 355–362. DOI: http://dx.doi.org/10.1017/cbo9780511615054.012

[18] Dimitri Metaxas, Mark Dilsizian, and Carol Neidle. 2018. Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[19] Ross E. Mitchell and Michael A. Karchmer. 2004. Chasing the Mythical Ten Percent: Parental Hearing Status of Deaf and Hard of Hearing Students in the United States. *Sign Language Studies*4, 2 (2004), 138–163. DOI: http://dx.doi.org/10.1353/sls.2004.0005

[20] Ross E. Mitchell, Travas A. Young, Bellamie Bachleda, and Michael A. Karchmer. 2006. How Many People Use ASL in the United States? Why Estimates Need Updating. *Sign Language Studies*6, 3 (2006), 306–335. DOI: http://dx.doi.org/10.1353/sls.2006.0019

[21] Carol Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*.

[22] Carol Neidle and Christian Vogler. 2012. A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*.

[23] Naomi Robbins and Richard Heiberger. 2011. Plotting Likert and Other Rating Scales. *Proceedings of the 2011 Joint Statistical Meeting*.

[24] Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. *Proceedings of the 13th conference on World Wide Web - WWW 04*(2004), 13–19. DOI: http://dx.doi.org/10.1145/988672.988675

[25] Jenny L. Singleton and Elissa L. Newport. 2004. When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*49, 4 (2004), 370–407. DOI: http://dx.doi.org/10.1016/j.cogpsych.2004.05.001

[26] William C. Stokoe, Dorothy C. Casterline, and Carl G. Croneberg. 1965. *A dictionary of American sign language on linguistic principles*, Washington, D.C.: Gallaudet College Press.

[27] Richard A. Tennant and Marianne Gluszak. Brown. 2010. *The American Sign Language handshape dictionary*, Washington, D.C.: Clerc Books, Gallaudet University Press.

[28] Clayton Valli. 2005. *The Gallaudet Dictionary of American Sign Language*, Gallaudet University Press.

[29] Haijing Wang, Alexandra Stefan, Sajjad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad Kamangar. 2012. A System for Large Vocabulary Sign Search. *Trends and Topics in Computer Vision Lecture Notes in Computer Science*(2012), 342–353. DOI: http://dx.doi.org/10.1007/978-3-642-35749-7_27

[30] Kimberly A. Weaver and Thad Starner. 2011. We need to communicate! *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility - ASSETS 11*(2011). DOI: http://dx.doi.org/10.1145/2049536.204