Sign Spotter: Design and Initial Evaluation of an Automatic Video-Based American Sign Language Dictionary System

Matyáš Boháček
maty@stanford.edu
Stanford University
Stanford, California, USA

Saad Hassan
saadhassan@tulane.edu
Tulane University
New Orleans, Louisiana, USA

ABSTRACT

Searching unfamiliar American Sign Language (ASL) words in a dictionary is challenging for learners, as it involves recalling signs from memory and providing specific linguistic details. Fortunately, the emergence of sign-recognition technology will soon enable users to search by submitting a video of themselves performing the word. Although previous research has independently addressed algorithmic enhancements and design aspects of ASL dictionaries, there has been limited effort to integrate both. This paper presents the design of an end-to-end sign language dictionary system, incorporating design recommendations from recent human–computer interaction (HCI) research. Additionally, we share preliminary findings from an interview-based user study with four ASL learners.

CCS CONCEPTS

• **Human-centered computing** → *Accessibility systems and tools*;
• **Information systems** → *Search interfaces.*

KEYWORDS

Sign Languages, American Sign Language (ASL), Dictionary, Search Interfaces, Video Search, User Satisfaction, Search Evaluation, Search System Design

ACM Reference Format:
Matyáš Boháček and Saad Hassan. 2023. Sign Spotter: Design and Initial Evaluation of an Automatic Video-Based American Sign Language Dictionary System. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23), October 22–25, 2023, New York, NY, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3597638.3614497

1 INTRODUCTION AND RELATED WORK

Over 70 million Deaf and hard of hearing (DHH) individuals worldwide use sign languages, with American Sign Language (ASL) being used by approximately 500,000 people in the United States alone [12, 24, 25]. The growing interest in learning sign languages extends beyond the DHH community, as many hearing individuals, such as parents and teachers of DHH children, as well as students in ASL courses, are motivated to learn ASL to facilitate communication and inclusion [14–16, 30, 33].

ASSETS '23, October 22–25, 2023, New York, NY, USA

ACM ISBN 979-8-4007-0220-4/23/10.
https://doi.org/10.1145/3597638.3614497

Traditional methods of searching for signs in sign language dictionaries pose challenges due to the absence of a standardized writing system and the need to recall and specify various linguistic properties of signs [2, 6, 8, 21]. Recent advancements in search-by-video dictionary systems show promise by automatically analyzing sign videos for matches in dictionary collections [9, 10, 13, 22, 27, 28, 34]. However, challenges remain, including difficulties in recognizing complex 3D signs from 2D video, issues related to lighting, camera motion, cluttered backgrounds, and the user's signing accuracy [29, 34]. Consequently, search-by-video systems may not always provide the desired sign as the top result, requiring users to navigate through potentially lengthy result lists [34].

Recent HCI research has explored ASL learners' interaction with Wizard-of-Oz (WOZ) prototype systems for ASL dictionary search, investigating factors that influence user satisfaction, such as search result composition and presentation [1, 17–19]. While these studies have provided valuable insights into design space, interface optimization, and algorithm-independent design solutions, there is limited integration of these findings with functioning recognition systems. Furthermore, there is a lack of user studies on video-based ASL dictionary systems utilizing state-of-the-art sign recognition technology. Understanding user behaviors related to system failures, adaptation to the recorder for better video input quality, communicating system latency, and conveying result accuracy requires the use of end-to-end functioning systems.

In our ongoing research, we are developing a video-based ASL dictionary system using state-of-the-art sign recognition technology [4]. In this demo paper, we showcase the system's design, drawing inspiration from recent HCI research on sign language dictionaries and other look-up systems [11, 19, 31]. Additionally, we present findings from an initial interview-based study involving four ASL learners.

2 SYSTEM DESIGN

2.1 Recognition Model

Although not the main focus of this paper, we briefly discuss the model we use as well as its training and testing to provide some context for our later discussions on design choices and findings from the user study. We chose to use the Transformer-based SPOTER [4] architecture and trained it on a custom subset of WLASL2000 [23]. Compared to earlier approaches for the task of sign language recognition, which usually analyzed raw video data, SPOTER estimates and analyzes the poses of the signer. This step reduces the input data dimensionality, allowing for faster inference and better generalization across diverse signers. SPOTER achieved the state-of-the-art top-1 accuracy on multiple sign language datasets [3]: 30.97% on ASLLVD Skeleton [26], 78.29% on WLASL100 [23], and 18.68% on

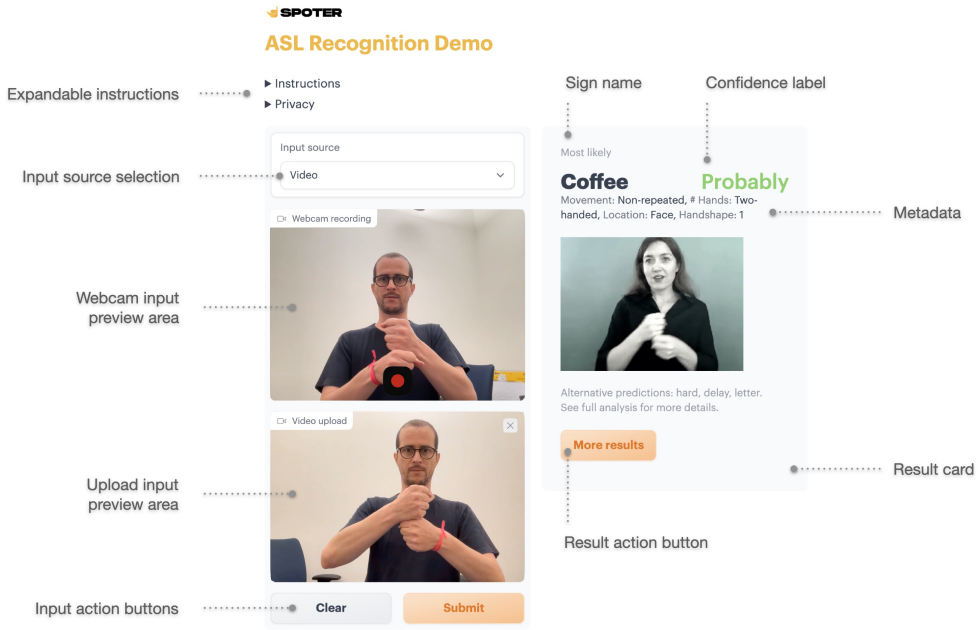

Figure 1: Screenshot of the web application, open on the main dictionary interface page, with a prediction based on an uploaded video recording.

UWB-SL-Wild [5]. Each of these datasets examines different capabilities of the model: ASSLVD requires that the model learns from a small number of examples, WLASL100 requires that the model learns from recordings whose setting and user demographics are largely variable, and UWB-SL-Wild requires that the model learns from in-the-wild dictionary data. Together, these datasets ensure versatility and robustness.

2.2 Interface Design

We designed a web application consisting of two primary web pages: (a) the main dictionary interface and (b) the detailed results page. Besides expandable user instructions at the top, the main dictionary interface is divided into the input area (on the left) and the results area (on the right), as shown in Figure 1.

To start using the application, the user needs to indicate whether they intend to use the webcam or upload a video file as the input source in a drop-down menu. Once the input is uploaded, they hit the 'Submit' button, which starts the recognition process. After a few seconds, the closest-matching result is presented in the result card on the right. Following prior HCI research on individual snippet design, we implemented various interface elements. These include an auto-playing video result that loops, a concise English gloss, and easily interpretable linguistic features [7, 20]. The linguistic meta-data with the result snippet helps users quickly browse the dictionary's prediction and verify that the critical attributes of the results match their desired sign. In a post-query filtering step (on the detailed analysis page), these annotations could be used to inform filtering choices, which has been shown to boost user satisfaction, confidence, and performance when searching for an unfamiliar sign [7, 20].

The model's confidence in this prediction is shown in the top right of the card. While the model yields percentual results for each sign in its vocabulary, the application converts these digits into confidence labels (66-100% to 'Probably', 33-66% to 'Possibly', and 0-33% to 'Unlikely'). We set the confidence label thresholds based on our observations of the model behavior in three common scenarios. The model is usually accurate in predicting one sign with high confidence. If the model gives the same high score to two signs, it usually means that the signs are alike, and either one could be right. If the confidence score is close to zero, the prediction is unlikely to be correct.

Unlike prior video-based ASL dictionary systems and WOZ prototypes, we opted to display only the top prediction in the main interface, benefiting from the significant improvements in video-based search accuracy. This approach enhances the user experience and aligns with the tendency of most online dictionary users to prioritize the primary prediction [11, 31, 32]. A list of runner-up predictions is presented in a detailed analysis. The user may expand the prediction and see alternative signs by hitting the 'More results' button, which takes them to the detailed results page shown in Figure 2. This list is ordered by the model's prediction confidence. The individual cards present the same information as the top prediction on the main results page. The linguistic meta-data are filterable using three drop-down menus and a specialized handshape selection modal, with the available handshapes displayed visually, as shown in Figure 3 [20]. The user could reset filters or return to the main dictionary interface using the buttons on the top right.

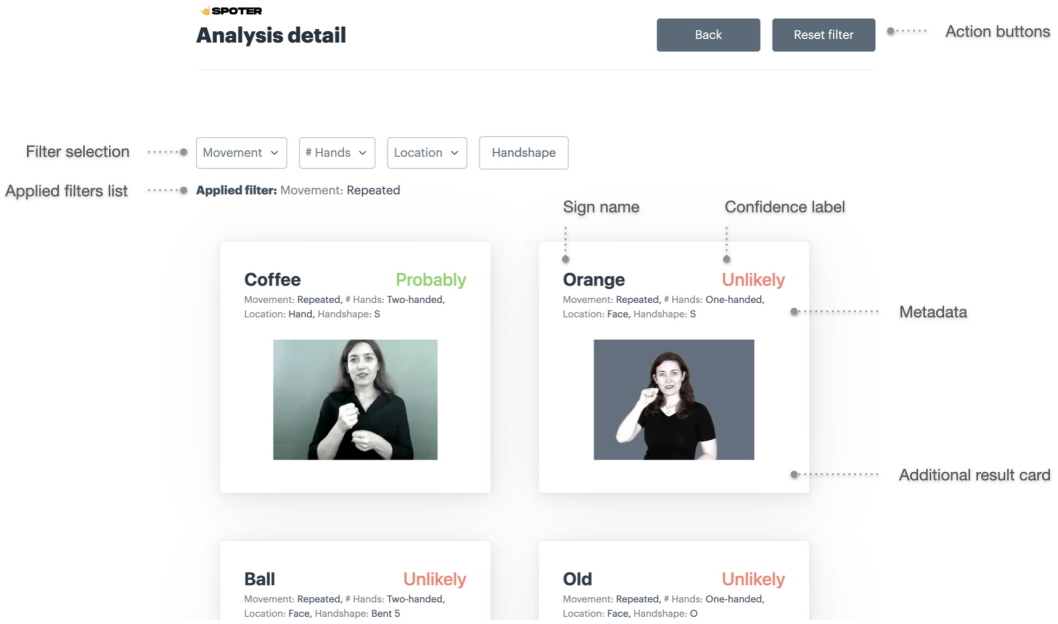

Figure 2: Screenshot of the web application, open on the detailed results page, with some filters selected and the top prediction visible.

3 ASL LEARNERS' FEEDBACK

3.1 Participants and Study Design

Participants for the study were recruited by emailing an advertisement to students enrolled in introductory and intermediate-level ASL courses. The screening questions asked about their current or past enrollment in introductory or intermediate ASL courses. Four participants were recruited: P1 (23-year-old female, started learning ASL in 2019, took 8 classes), P2 (22-year-old male, started learning ASL in 2019, took 5 courses), P3 (19-year-old male, started learning ASL in 2021, took 2 courses), and P4 (38-year-old female, started ASL in 2018, took 1 course). None of the participants had any DHH family members. Participants signed an IRB-approved consent form before the start of the study and were paid $40 for participation.

Participants provided feedback on the system design through a structured interview. Prior to the interview, we shared our ASL dictionary and a list of signs for them to try performing. The list included a subset of non-trivial signs that the model was trained on (see Appendix A). Additionally, we provided videos of these signs in a folder for reference. During the interview, we discussed their experience using ASL dictionaries and how ours differed. We gathered feedback on the recorder, system latency, individual result snippets, detailed analysis, and how result confidence was conveyed. We also inquired about their perception of system accuracy and their behavior when the system failed to provide the correct result.

The interviews had an average duration of 35 minutes and were recorded and transcribed. Instead of conducting a formal, thematic analysis of the interview responses, we present participants' comments on various aspects of the system, which provide insights for future research and design endeavors.

3.2 Initial Findings

All four participants expressed interest in using our video-based dictionary system. P1 stated, "*I believe... I would make regular use of it, as many individuals in my class already engage in recording themselves and sharing it with multiple people to determine the understanding of certain signs.*" Participants mentioned employing such a system "*after having a signed conversation with someone*" (P2), "*while watching signed videos*" (P3), and "*during homework for courses*" (P1, P4). Two participants indicated their willingness to use the system even with its current level of accuracy, while the other two expressed a desire for improved accuracy. P2, for instance, stated: "*To be honest, I might actually end up using something like this in its current state, because I feel like in the detailed analysis page, you can really get what you'd like.*" P1, who desired recognition improvements, said that "*A lot of times [the system] would only pick up on one aspect, like only the hand shape, or only the area of space... I think it also may have had trouble with, like, particular hand shapes, I know, like AFRICA.*" Participants also reported accuracy percentages ranging from 25% (P1) to 75% (P2) for the top result.

All participants liked the recorder and input recorded video functionalities. P2 recommended using 'Play' and 'Pause' buttons instead of icons for the recorder. Interestingly, participants mentioned diverse behaviors when they could not find the sign. For example, "*repeating it*"(P1, P3, P4), "*performing it face on*" (P1), "*performing it at three quarter angle*" (P1), "*shifting from one side to*

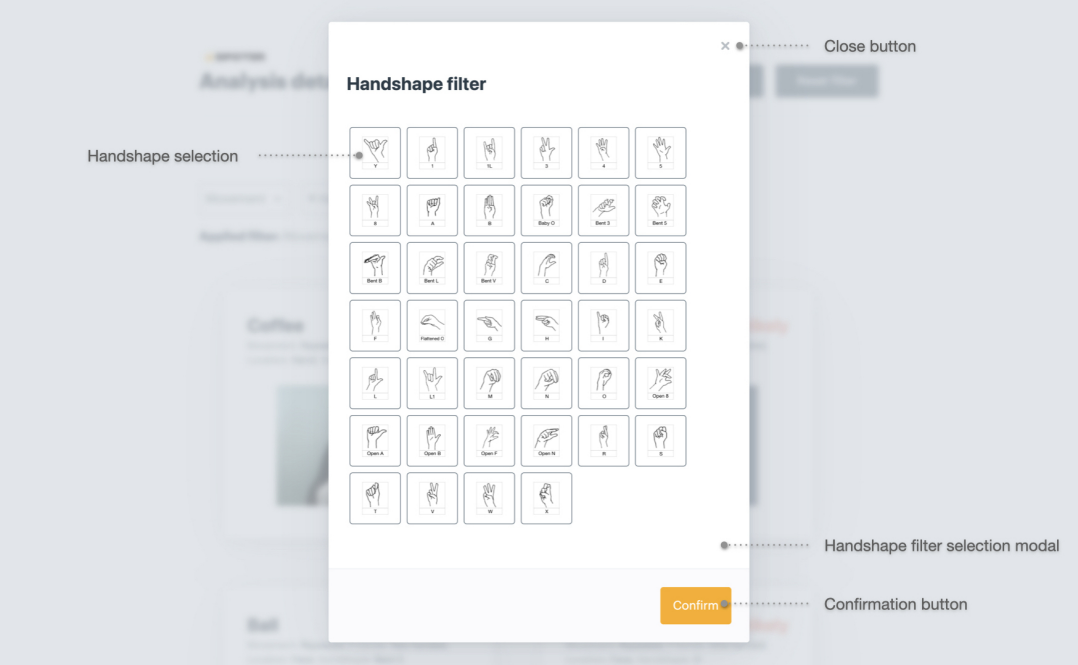

Figure 3: Screenshot of the handshape filter selection modal, expandable from the web application's detailed results page.

another" (P1, P3), "*performing it in front of a dark background of chair*" (P2), and "*performing it on their shirt*" (P2).

Participants had varying suggestions regarding the presentation of results. P2 and P3 preferred a single overall result on the first page, with a detailed analysis available upon request. P2 specifically appreciated the simplicity of having one result and a separate interface for detailed analysis, stating, "*I would prefer to have the one result and separate detailed analysis because then when you would, it would sort of just be one thing... and then click onto another screen to access all those other really useful things, like your filters... [Filters] do a really great job at narrowing down certain stuff... you can take all of the different classifiers and tags that are on each image and just immediately get rid of a whole bunch of other results.*" P1 mentioned that given the current accuracy, they would like to see four results on the first page and have a detailed analysis for the remaining results.

Participants provided suggestions regarding the display of system latency. P1 appreciated the current design, which included an estimated processing time. In contrast, P2 suggested using a more prominent UI element with indications of the uploading or processing status, stating, "*a loading wheel or, like some sort of larger UI element to, like, tell you that, you know, it's either uploading or processing*". P4 also mentioned that it could be helpful to inform users if the processing time is longer than usual.

Participants had mixed opinions regarding the presentation of confidence levels. While two participants expressed occasional disagreement with the confidence levels, they still liked the overall presentation. P1 shared an example stating, "*For one video, it said unlikely, and that was the correct answer.*" Participant 2 said, "*And the breakdown of what it thought was... confident versus possible versus unlikely. I thought that was pretty helpful.*"

4 CONCLUSION

We implemented a novel end-to-end video-based ASL dictionary that integrates the latest recognition approach and incorporates design recommendations from HCI research. Through a small user study with four ASL learners, we gained insights into usage contexts and system performance perceptions and received valuable design suggestions. We plan to conduct a demo of our system to further understand user behaviors and gather feedback on design elements.

ACKNOWLEDGMENTS

This material is partially based upon work supported by the National Science Foundation under Award no. 2125362, 2212303, 2235405, and Duolingo. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Duolingo.

REFERENCES

[1] 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 *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 56–67. https://doi.org/10.1145/3308561.3353791

[2] 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 *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, Vol. 2. European Language Resources Association (ELRA), Valletta, Malta, 11–14.

[3] Matyáš Boháček, Zhuo Cao, and M. Hrúz. 2022. Combining Efficient and Precise Sign Language Recognition: Good pose estimation library is all you need. *International Conference on Computer Vision and Pattern Recognition Accessibility, Vision, and Autonomy Meet (AVA) 2022 Workshop* (2022).

[4] Matyáš Boháček and M. Hrúz. 2022. Sign Pose-based Transformer for Word-level Sign Language Recognition. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (2022), 182–191.

[5] Matyáš Boháček and M. Hrúz. 2023. Learning from What is Already Out There: Few-shot Sign Language Recognition with Online Dictionaries. *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)* (2023), 1–6.

[6] 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 and Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1837–1848. https://doi.org/10.1145/2675133.2675226

[7] Danielle Bragg, Kyle Rector, and Richard E. Ladner. 2015. A User-Powered American Sign Language Dictionary *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1837–1848. https://doi.org/10.1145/2675133.2675226

[8] Fabian Bross. 2015. Chereme. In: Hall, T. A. Pompino-Marschall, B. (ed.): Dictionaries of Linguistics and Communication Science (Wörterbücher zur Sprach- und Kommunikationswissenschaft, WSK). Volume: Phonetics and Phonology. Berlin, New York: Mouton de Gruyter. *Proceedings of the XVIII EURALEX International Congress* 1, 1 (01 2015), 9.

[9] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, New York, New York, US, 7784–7793. https://doi.org/10.1109/CVPR.2018.00812

[10] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, New York, New York, US, 10020–10030. https://doi.org/10.1109/CVPR42600.2020.01004

[11] Thanh-Dung Dang, Gwo-Dong Chen, Giao Dang, Liang-Yi Li, and Nurkhamid. 2013. RoLo: A dictionary interface that minimizes extraneous cognitive load of lookup and supports incidental and incremental learning of vocabulary. *Comput. Educ.* 61 (2013), 251–260.

[12] Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2021. Sign language. https://www.ethnologue.com/subgroups/sign-language

[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 Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*. SLTAT, Dundee, Scotland, 8 pages. http://personal.ee.surrey.ac.uk/Personal/H.Cooper/research/papers/SBE_SLTAT.pdf

[14] National Center for Education Statistics (NCES). 2018. Digest of Education Statistics Number and percentage distribution of course enrollments in languages other than English at degree-granting postsecondary institutions, by language and enrollment level: Selected years, 2002 through 2016. https://nces.ed.gov/programs/digest/d18/tables/dt18_311.80.asp

[15] David Goldberg, Dennis Looney, and Natalia Lusin. 2015. Enrollments in Languages Other than English in United States Institutions of Higher Education, Fall 2013.

[16] Wyatte C Hall, Leonard L Levin, and Melissa L Anderson. 2017. Language deprivation syndrome: A possible neurodevelopmental disorder with sociocultural origins. *Social psychiatry and psychiatric epidemiology* 52, 6 (2017), 761–776.

[17] Saad Hassan, Oliver Alonzo, Abraham Glasser, and Matt Huenerfauth. 2020. Effect of ranking and precision of results on users' satisfaction with search-by-video sign-language dictionaries. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, Vol. 4. Computer Vision – ECCV 2020 Workshops, Virtual, 6 pages.

[18] 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 (oct 2021), 33 pages. https://doi.org/10.1145/3470650

[19] Saad Hassan, Akhter Al Amin, Caluã de Lacerda Pataca, Diego Navarro, Alexis Gordon, Sooyeon Lee, and Matt Huenerfauth. 2022. Support in the Moment: Benefits and Use of Video-Span Selection and Search for Sign-Language Video Comprehension among ASL Learners. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 29, 14 pages. https://doi.org/10.1145/3517428.3544883

[20] Saad Hassan, Akhter Al Amin, Alexis Gordon, Sooyeon Lee, and Matt Huenerfauth. 2022. Design and Evaluation of Hybrid Search for American Sign Language to English Dictionaries: Making the Most of Imperfect Sign Recognition. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 195, 13 pages. https://doi.org/10.1145/3491102.3501986

[21] Robert J Hoffmeister. 2000. *A piece of the puzzle: ASL and reading comprehension in deaf children*. Mahwah, N.J. : Lawrence Erlbaum Associates, New Jersey, USA. 143–163 pages.

[22] Kabil Jaballah and Mohamed Jemni. 2010. Toward Automatic Sign Language Recognition from Web3D Based Scenes. In *Proceedings of the 12th International Conference on Computers Helping People with Special Needs* (Vienna, Austria) *(ICCHP'10)*. Springer-Verlag, Berlin, Heidelberg, 205–212.

[23] Dongxu Li, Cristian Rodriguez-Opazo, Xin Yu, and Hongdong Li. 2019. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), 1448–1458.

[24] Ross Mitchell, Travas Young, Bellamie Bachleda, and Michael Karchmer. 2006. How Many People Use ASL in the United States? Why Estimates Need Updating. *Sign Language Studies* 6 (03 2006). https://doi.org/10.1353/sls.2006.0019

[25] J Murray. 2020. World Federation of the deaf. http://wfdeaf.org/our-work/

[26] C. Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. Challenges in development of the American Sign Language Lexicon Video Dataset (ASLLVD) corpus.

[27] Junfu Pu, Wengang Zhou, and Houqiang Li. 2018. Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, 885–891. https://doi.org/10.24963/ijcai.2018/123

[28] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign Language Recognition: A Deep Survey. *Expert Systems with Applications* 164 (2021), 113794. https://doi.org/10.1016/j.eswa.2020.113794

[29] Kishore K Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Machine vision and applications* 24, 5 (2013), 971–981.

[30] Jerry Schnepp, Rosalee Wolfe, Gilbert Brionez, Souad Baowidan, Ronan Johnson, and John McDonald. 2020. Human-Centered Design for a Sign Language Learning Application. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '20)*. Association for Computing Machinery, New York, NY, USA, Article 60, 5 pages. https://doi.org/10.1145/3389189.3398007

[31] Sayush Shrestha and Pietro Murano. 2022. The Design and Evaluation of an Online Dictionary User Interface. *International Journal of Computing and Digital Systems* (2022).

[32] Yukio Tono. 2000. On the Effects of Different Types of Electronic Dictionary Interfaces on L2 Learners' Reference Behaviour in Productive/Receptive Tasks. *2000 Proceedings of European Association for Lexicography*.

[33] 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* (Dundee, Scotland, UK) *(ASSETS '11)*. Association for Computing Machinery, New York, NY, USA, 91–98. https://doi.org/10.1145/2049536.2049554

[34] Polina Yanovich, Carol Neidle, and Dimitris Metaxas. 2016. Detection of Major ASL Sign Types in Continuous Signing For ASL Recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 3067–3073. https://www.aclweb.org/anthology/L16-1490

A LIST OF SIGNS

The list of signs participants were asked to try first was: BOWLING, AFRICA, CHEAT, DECIDE, LETTER, LAUGH, BLANKET, CUTE, LEAVE, LOSE, PROBLEM, SHARE, APPROVE, CONVINCE, COUNTRY, CRASH, GOVERNMENT, HOPE, ORDER, PRESIDENT, RUSSIA, SINCE, THEORY, WAR, CHAMPION, DELAY, DELICIOUS, DISAPPEAR, FAULT, HUMBLE, KILL, LAW.