Supporting ASL Communication Between Hearing Parents and Deaf Children

Ekram Hossain
ehossai2@ur.rochester.edu
University of Rochester,
Rochester, NY, USA

Ashley Bao
abao26@amherst.edu
Amherst College,
Amherst, MA, USA

Kaleb Slater Newman
kaleb_newman@brown.edu
Brown University,
Providence, RI, USA

Madeleine Mann
mmann8@u.rochester.edu
University of Rochester,
Rochester, NY, USA

Hecong Wang
hwang99@ur.rochester.edu
University of Rochester,
Rochester, NY, USA

Yifan Li
yli287@ur.rochester.edu
University of Rochester,
Rochester, NY, USA

Chigusa Kurumada
ckuruma2@ur.rochester.edu
University of Rochester,
Rochester, NY, USA

Zhen Bai
zbai@cs.rochester.edu
University of Rochester,
Rochester, NY, USA

ABSTRACT

The vast majority of deaf or hard-of-hearing (DHH) children are born to hearing parents. Due to a lack of immersive exposure to their natural language - sign language - they are at severe risk of language deprivation. In response to this challenge, this paper presents a novel computer-mediated communication platform named Tabletop Interactive Play System (TIPS). It serves as a test-bed to investigate technical and ethical solutions that enable hearing parents to use American Sign Language (ASL) during face-to-face play with their with their DHH children. The TIPS platform offers a variety of user options in three key aspects: (1) ASL recommendation in alignment with hearing parents' real-time speech; (2) ASL display through different form-factors (Augmented Reality (AR) projection, tablet, and smart glasses); and (3) autonomy support to enhance users' sense of agency and trust in the system. In this paper, we will describe the system's design, implementation, and preliminary evaluation results.

CCS CONCEPTS

• **Human-centered Computing → Human-centered Computing; Accessibility**.

KEYWORDS

Computer-Mediated Communication, Deaf or Hard-of-Hearing, American Sign Language, Parent-Child Interaction, Augmented Reality

ACM Reference Format:

Ekram Hossain, Ashley Bao, Kaleb Slater Newman, Madeleine Mann, Hecong Wang, Yifan Li, Chigusa Kurumada, and Zhen Bai. 2023. Supporting ASL Communication Between Hearing Parents and Deaf Children. 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/10.1145/3597638.3614511

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

ACM ISBN 979-8-4007-0220-4/23/10.
https://doi.org/10.1145/3597638.3614511

1 INTRODUCTION

Over 90% of deaf or hard-of-hearing (DHH) children are born to hearing families [22]. Without an immersive sign language environment at home, they face an increased risk of language deprivation and other negative lifelong impacts [7, 10, 11, 19]. Therefore, pediatric research emphasizes the importance for caregivers and family members of DHH children to become involved in the promotion of sign language during the initial phases of child development [14].

Existing work primarily focuses on facilitating individualized learning of American Sign Language (ASL) through various methods, including dictionaries [1, 12, 13], interactive games [4, 23], bilingual storybooks [21], social agents [29], and mixed-reality learning systems [30]. However, there is a lack of research on increasing hearing parents' ASL exposure through real-time, face-to-face interactions with their DHH children. Meanwhile, there are emerging technologies that focus on improving real-time, face-to-face communication for other purposes, such as supporting parents to apply communication strategies during real-time interactions with children with language delays [15], encouraging immigrant parents to speak in a foreign language with their children [18], and raising DHH individuals' awareness of speech and sound in their surroundings [16, 17, 24].

There remains, however, a critical research gap in enabling hearing parents to communicate fluidly in ASL with their DHH children. There are several challenges involved. (1) ***High cognitive demand:*** The visuospatial aspects of sign language (e.g., handshape, location, movement, and facial expression) may place a high cognitive burden on a parent who only knows spoken language. (2) ***Flow of face-to-face interaction:*** Computer-mediated communication with digital displays often comes with attention diversion [33]. (3) ***Autonomy-supportive intelligent system:*** Computer-mediated communication often involves advanced AI technologies such as natural language processing and computer vision. Users may feel frustrated or mistrustful when a system's recommendations do not align with their needs or expectations. It is therefore critical to explore the design space that respects user autonomy - a sense of willingness and acting in accordance with one's goal, as opposed to being controlled, manipulated, or forced [6].

In this paper, we present a novel computer-mediated communication platform named **TIPS** (Tabletop Interactive Play System). It serves as a test-bed for investigating critical technical and ethical solutions that enable hearing parents to learn and communicate in ASL during face-to-face joint play with their DHH children. Within the platform, a variety of technical solutions are implemented in addressing the aforementioned challenges. (1) ***ASL displays***: Informed by prior work of user preferences of display form-factors for real-time ASL recommendation [3], We have developed three ASL display solutions that minimize communication interference: *AR projection* near the toy of interest during parent-child joint play, *tablet*, and *smart glasses*. (2) ***ASL recommendation***: We have developed three different recommendation strategies to select an ASL video based on *part of speech*, *semantic weight*, or *word prevalence* of the perceived spoken utterance. (3) ***Autonomy-supportive interactions***: We have developed two additional interaction mechanisms to increase respect for user autonomy. (a) The *system behavior monitor* displays the output of key AI components including real-time transcription of Automatic Speech Recognition (ASR), the keyword highlighted for recommendation, and the semantic meaning of the ASL video; (b) the *Tap-to-Sign* feature allows the user to directly request an ASL video by tapping on any single word within the transcribed sentence, providing the user with flexibility and control.

2 SYSTEM DESIGN

The TIPS platform offers a framework that monitors real-time parent-child interactions, extracts one relevant word from each parental utterance, and displays ASL that corresponds to the extracted word (Figure 1(a)). For example, if a parent says, "The duck is in the water," the system may extract and show the sign for "duck". The parent wears a wireless microphone to capture their utterances. A webcam is positioned to capture a top view of the play area to record the play contexts (i.e., types and locations of physical toys).

2.1 ASL Displays

TIPS supports three display form-factors, *AR projection*, *tablet*, and *smart glasses* (Figure1(b-d)), all built on a shared computational framework. Each form-factor has a different impact on attention division during face-to-face interactions.

The **near-object AR projection** utilizes an overhead projector (Optoma ML1050ST+) to display the recommended ASL video (Figure 1(b)). By displaying the video close to the relevant toy (object of interest), attention diversion is likely minimized. Further, the coupling of a physical object with its semantic representation is expected to have a positive effect on language learning [3].

The **tablet** allows users to view the same information on a stand-mounted tablet device (e.g. an iPad) that can be placed anywhere on a table (Figure 1(c)). The mobility of the tablet offers flexible placement options to accommodate users' preferences and diverse interaction settings (e.g., floor play [18]). The tablet solution offers high accessibility and affordability, at the cost of potential obtrusion for face-to-face interaction due to frequent attention switching between the display, the child, and toys on the table.

The **smart glasses** allow users to view the ASL video along with the system behavior through a pair of light-weight and monocular smart glasses (Vuzix Blade 2)(Figure 1(d)). The advantages of smart glasses identified in [3] include high portability and reduced attention diversion as they display the ASL video close to the parent's line of sight. Smart glasses also minimize occlusion of the face compared to other head-mounted AR displays such as Hololens.

The tabletop projection display is limited to a fixed location. The tablet and smart glasses display solutions are more portable, but the performance (e.g., latency) will be impacted by the computational power and network capability

2.2 ASL Recommendation Strategies

The automatic ASL recommendation strategies aim to minimize the decision-making burden imposed on the parents during joint play with their children. TIPS currently only recommends isolated ASL vocabulary, which is considered a reasonable first step based on early feedback from domain experts such as educators in early education programs and members of the Deaf community. In subsequent phases, we plan to expand the range of ASL input we provide to include multi-word phrases and full utterances with grammatical markers. The automatic ASL recommendation takes into account key linguistic features, including parts of speech, semantic weight, and word prevalence in child-directed speech [5, 9, 32].

Part of speech recommendation extracts one of the following from the parent's spoken utterance, in order of priority: first noun, first verb, and first adjective. The design is motivated by the observation that nouns often carry more information contents in conveying the communication of the utterance meaning [32]. For example, if a parent utters, "It's a sunny day, so the duck is in the water," the strategy will return the first noun: *day*.

Semantic weight recommendation extracts the word in the utterance that conveys the core meaning of the utterance. The semantic weight is determined by a pre-trained language model, such as BERT, which has previously been used to extract important words in live captioning for DHH individuals [2]. This identifies the word that is most likely to represent the gist of the parent's communicative intent. For example, again considering the utterance, "It's a sunny day, so the duck is in the water," the strategy will return the word the model determines best captures the sentence's meaning: *duck*.

Word prevalence recommendation extracts the word from the utterance that most frequently appears in the North American English corpora of the Child Language Data Exchange System (CHILDES) [20], a well-established corpus of child and child-directed speech, as an approximation of the word's prevalence during parent-child interactions. For example, for the same example utterance as above, the strategy will return the most commonly found word in the database: *water*.

2.3 Autonomy-Supportive Interactions

Figure 1(b) shows the design for the system behavior monitor feature, which reveals the real-time transcription of users' utterances,

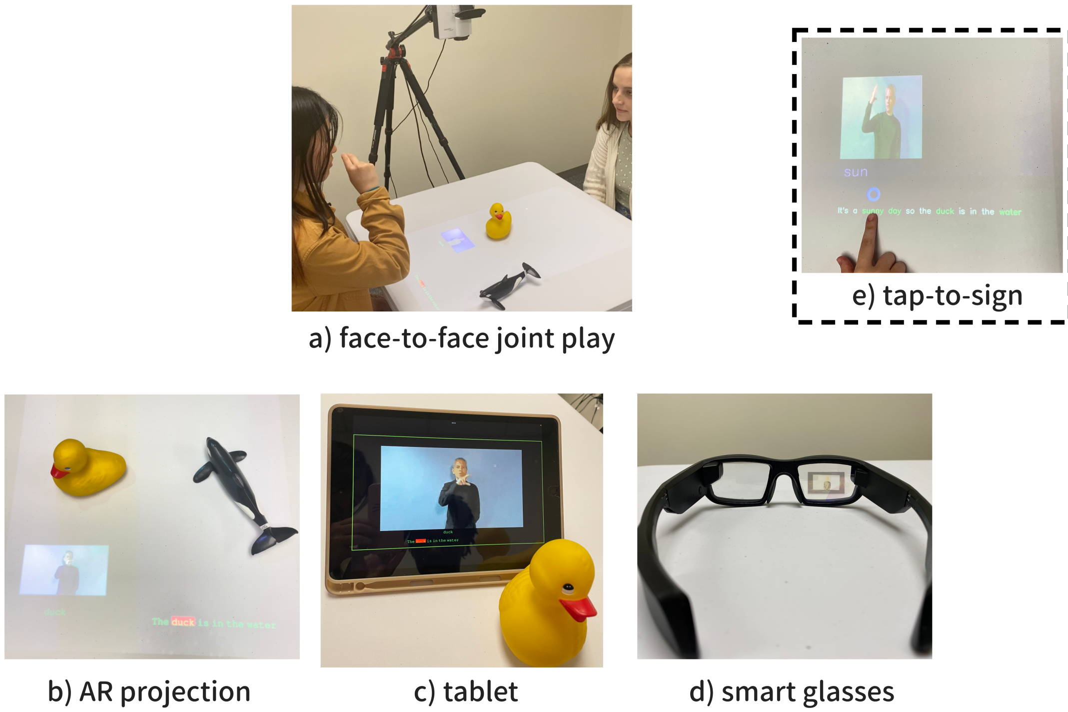

Figure 1: a) An example of user interaction; b) AR projection interface; c) tablet interface; d) smart-glass interface; e) Tap-to-Sign interface

with the recommended keyword highlighted as well as the English tag for the ASL video. Users are thus able to determine if the AI components give erroneous outputs, avoiding any potential self-blame if they receive incorrect or undesired ASL recommendations [18]. Figure 1(e) shows the **Tap-to-Sign** feature, which is designed to offer users full control over the selection of the words they wish to learn. Users can view a specific ASL video by tapping on any word within the real-time utterance transcription.

3 SYSTEM IMPLEMENTATION

The TIPS computational framework is based on the AR projection display form-factor and composed of five key components in addition to the Tap-to-Sign framework:

Speech transcription: The Whisper-small model [26] is deployed to transcribe parent utterances in real-time.

Object recognition: We fine-tuned the YOLOv7-tiny [31] on a small set of pre-defined toys to identify toys on the table.

ASL recommendation: We utilize the recommendation strategies described in 2.2 to extract a word to recommend based on the transcription. We first filter out function words (e.g., "the" or "and"), because they have little semantic meaning and do not have significant equivalents in ASL. For part of speech recommendation, we used Stanza [25], a natural language processing toolkit, to parse through utterances and to determine parts of speech. For semantic weight-based recommendation, we used the all-MiniLM-L6-v2 BERT model [8, 27] to determine semantic weight. For word prevalence recommendation, we used the childespy Python package to extract the most frequently appeared word in the CHILDES database [20, 28].

Object of interest detection: We take the recommended ASL word and determine the object on the table that is most likely the topic of the utterance. This is implemented by also using the all-MiniLM-L6-v2 BERT model to create text embeddings for all toys detected on the table, and then using cosine similarity to match the most relevant toy to the recommended keyword chosen word from the parent's speech.

Interface rendering: We render the ASL video and system behavior information on the different display devices described in 2.1. We retrieve the ASL video by comparing the recommended word's text embedding with the text embeddings of each of the signs in the University of Rochester ASL corpus, as determined by the all-MiniLM-L6-v2 BERT model [8]. As a rapid prototyping approach, the current tablet and smart glasses display the ASL video via Zoom using the computational architecture described above based on the AR projection display.

Tap-to-Sign: For prototyping purposes, Tap-to-Sign runs on a slightly different framework. Tap-To-Sign only needs to utilize the speech transcription and interface rendering components. Additionally, this method uses the OpenCV library to implement the capability for parents to tap words on the table. Tap-to-Sign tracks users' word selections by utilizing a carefully tuned color mask that captures skin color. We define the mask in the HSV color space using upper and lower bounds. The bounds were selected through visualizing the mask (in-range colors) during internal trials on a range of skin colors. Utilizing a color mask establishes a low latency baseline for the tapping functionality. Then, when the mask covers the table coordinates of a projected word, the system promptly displays the corresponding ASL video for that word.

4 PRELIMINARY SYSTEM EVALUATION

We are in the process of evaluating the performance of the overall system as well as its individual components. To achieve this, we analyze the log that is generated when users are interacting with the system. So far, we have conducted system testing with one pair of participants from our research group using AR projection and part of speech-based recommendations, without the Tap-to-Sign feature. The total duration of the system interaction was one hour. We evaluated the following aspects of the system: **(1) Accuracy of ASR transcription:** The Word Error Rate (WER) of the ASR transcription is 9.58%. It is calculated as the sum of the number of substitutions, insertions, and deletions divided by the total number of spoken words. **(2) Accuracy of object recognition:** The performance of the object recognition model had an F-1 score of 82%, a precision score of 82.5%, and a recall score of 81.5%. This is accessed by comparing the recognized objects with the ground truth classes. **(3) Correctness of ASL video retrieval:** The correctness of the retrieved ASL video scored 2.49 on a scale of 0 to 3. This is assessed based on the semantic match between the tags of retrieved ASL videos and recommended keywords. A score of '3' denotes a perfect match, '2' indicates a strong correlation, '1' suggests a partial relation, and '0' signifies no relation or incorrect tags. **(4) Accuracy of object of interest detection:** The system correctly identified the toy of interest 77% of the time. In this context, the object of interest detection accuracy refers to a binary rating system where '1' indicates the correct identification of the object of interest and '0' denotes an incorrect identification. We determined the object of interest based on the topic of the parents' speech as determined by human annotation. We are in the process of conducting a more comprehensive performance evaluation.

5 CONCLUSION AND FUTURE WORK

The TIPS project aims to address the communication barriers faced by DHH children and their hearing parents. We proposed a context-responsive ASL communication system to increase the amount and quality of natural language exposure given to DHH children in hearing families. To address the challenges of minimizing cognitive load, reducing obtrusiveness, and supporting user autonomy, this system incorporates various components, including multiple ASL displays, various ASL recommendation strategies, and interaction features designed to promote user autonomy and support a seamless user experience. In our upcoming user studies, we plan to obtain feedback from stakeholders (e.g., hearing parents of DHH children, novice ASL learners, and early childhood educators for DHH children) regarding their design preferences and identify any areas where TIPS may not fully meet their needs. The findings of the user studies will inform system iteration and design guidelines for future systems. In future follow-up user studies, we aim to measure the effectiveness of the system in supporting hearing parents when delivering ASL to their DHH children.

ACKNOWLEDGMENTS

We express our gratitude to the Google Inclusive Research Award, the National Science Foundation REU program (award no. 1950460), and Pump Primer II funds from the University of Rochester for their support in funding this research.

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 *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 56–67.

[2] Akhter Al Amin, Saad Hassan, Cecilia O Alm, and Matt Huenerfauth. 2022. Using BERT Embeddings to Model Word Importance in Conversational Transcripts for Deaf and Hard of Hearing Users. *arXiv preprint arXiv:2206.12368* (2022).

[3] Zhen Bai, Elizabeth Codick, Ashely Tenesaca, Wanyin Hu, Xiurong Yu, Peirong Hao, Chigusa Kurumada, and Wyatte Hall. 2022. Signing-on-the-Fly: Technology Preferences to Reduce Communication Gap between Hearing Parents and Deaf Children. In *Interaction Design and Children*. 26–36.

[4] Dhruva Bansal, Prerna Ravi, Matthew So, Pranay Agrawal, Ishan Chadha, Ganesh Murugappan, and Colby Duke. 2021. Copycat: Using sign language recognition to help deaf children acquire language skills. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.

[5] ARIELLE BOROVSKY and JEFF ELMAN. 2006. Language input and semantic categories: a relation between cognition and early word learning. *Journal of Child Language* 33, 4 (2006), 759–790. https://doi.org/10.1017/S0305000906007574

[6] Rafael A Calvo, Dorian Peters, Karina Vold, and Richard M Ryan. 2020. Supporting human autonomy in AI systems: A framework for ethical enquiry. *Ethics of digital well-being: A multidisciplinary approach* (2020), 31–54.

[7] Naomi K Caselli, Wyatte C Hall, and Jonathan Henner. 2020. American Sign Language interpreters in public schools: An illusion of inclusion that perpetuates language deprivation. *Maternal and Child Health Journal* 24, 11 (2020), 1323–1329.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Nick C. Ellis and Alan Beaton. 1993. Psycholinguistic Determinants of Foreign Language Vocabulary Learning. *Language Learning* 43, 4 (1993), 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-1770.1993.tb00627.x

[10] Matthew L Hall, Wyatte C Hall, and Naomi K Caselli. 2019. Deaf children need language, not (just) speech. *First Language* 39, 4 (2019), 367–395.

[11] Wyatte C Hall. 2017. What you don't know can hurt you: The risk of language deprivation by impairing sign language development in deaf children. *Maternal and child health journal* 21, 5 (2017), 961–965.

[12] 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*. 1–14.

[13] 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*. 1–13.

[14] Tom Humphries, Poorna Kushalnagar, Gaurav Mathur, Donna Jo Napoli, Christian Rathmann, and Scott Smith. 2019. Support for parents of deaf children: Common questions and informed, evidence-based answers. *International journal of pediatric otorhinolaryngology* 118 (2019), 134–142.

[15] Inseok Hwang, Chungkuk Yoo, Chanyou Hwang, Dongsun Yim, Youngki Lee, Chulhong Min, John Kim, and Junehwa Song. 2014. TalkBetter: family-driven mobile intervention care for children with language delay. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1283–1296.

[16] Dhruv Jain, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E Froehlich. 2015. Head-mounted display visualizations to support sound awareness for the deaf and hard of hearing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 241–250.

[17] Dhruv Jain, Angela Lin, Rose Guttman, Marcus Amalachandran, Aileen Zeng, Leah Findlater, and Jon Froehlich. 2019. Exploring sound awareness in the home for people who are deaf or hard of hearing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[18] Taeahn Kwon, Minkyung Jeong, Eon-Suk Ko, and Youngki Lee. 2022. Captivate! Contextual Language Guidance for Parent–Child Interaction. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[19] Diane Lillo-Martin and Jonathan Henner. 2021. Acquisition of sign languages. *Annual Review of Linguistics* 7 (2021), 395.

[20] Brian MacWhinney. 2000. *The CHILDES project: The database*. Vol. 2. Psychology Press.

[21] Melissa Malzkuhn and Melissa Herzig. 2013. Bilingual storybook app designed for deaf children based on research principles. In *Proceedings of the 12th International Conference on Interaction Design and Children*. 499–502.

[22] Ross E Mitchell and Michaela 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.

[23] Jintao Nie, Yijun Zhao, Bing Yao, Zheng Xu, Jiayan Chen, Changchao Yu, Preben Hansen, Jianhui Liu, Jiadi Wang, Ge Yan, et al. 2022. SignFind: A Synchronized Sign Language and Chinese Character Teaching Game for Chinese Deaf Children Using Gesture Recognition. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[24] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable subtitles: Augmenting spoken communication with lightweight eyewear for all-day captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1108–1120.

[25] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess.AS]

[27] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[28] Alessandro Sanchez, Stephan C. Meylan, Mika Braginsky, Kyle E. MacDonald, Daniel Yurovsky, and Michael C. Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods* 51, 4 (2019), 1928–1941. https://doi.org/10.3758/s13428-018-1176-7

[29] Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, et al. 2018. Teaching language to deaf infants with a robot and a virtual human. In *Proceedings of the 2018 CHI Conference on human Factors in computing systems*. 1–13.

[30] Qijia Shao, Amy Sniffen, Julien Blanchet, Megan E Hillis, Xinyu Shi, Themistoklis K Haris, Jason Liu, Jason Lamberton, Melissa Malzkuhn, Lorna C Quandt, et al. 2020. Teaching american sign language in mixed reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–27.

[31] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7464–7475.

[32] Sandra Waxman, Xiaolan Fu, Sudha Arunachalam, Erin Leddon, Kathleen Geraghty, and Hyun-joo Song. 2013. Are nouns learned before verbs? Infants provide insight into a long-standing debate. *Child development perspectives* 7, 3 (2013), 155–159.

[33] Kristin Williams, Karyn Moffatt, Jonggi Hong, Yasmeen Faroqi-Shah, and Leah Findlater. 2016. The cost of turning heads: A comparison of a head-worn display to a smartphone for supporting persons with aphasia in conversation. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. 111–120.