Building User-Centered ASL Communication Technologies for Parent-Child Interactions

Ashley Bao[§]
Amherst College
Amherst, United States
abao26@amherst.edu

Kaleb Newman[§]
Brown University
Providence, United States
kaleb_newman@brown.edu

Madeleine Mann
University of Rochester
Rochester, United States
mmann8@u.rochester.edu

Ekram Hossain
Computer Science
University of Rochester
Rochester, United States
ehossai2@ur.rochester.edu

Chigusa Kurumada
Brain and Cognitive Sciences
University of Rochester
Rochester, United States
ckuruma2@ur.rochester.edu

Zhen Bai
Computer Science
University of Rochester
Rochester, United States
zhen.bai@rochester.edu

2023 IEEE MIT Undergraduate Research Technology Conference (URTC) | 979-8-3503-0965-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/URTC60662.2023.10534918

***Abstract*—Since most deaf or hard-of-hearing (DHH) children are born into hearing families, they are at risk of lifelong language deficits from a lack of natural language exposure in early childhood. The Tabletop Interactive Play System (TIPS) is a real-time American Sign Language (ASL) communication aid, for face-to-face joint play between hearing parents and their DHH children. We investigate user preferences for different ASL recommendation methods, providing insights into optimizing ASL communication technologies for hearing parent-DHH child interactions. We introduce Tap-to-Sign, an AR projection-based ASL sign retrieval system that grants parents complete autonomy in selecting the word to sign to their DHH child. Additionally, we propose three distinct strategies to automatically recommend a word for the parent to sign. To gain insights into our diverse methods, preliminary feedback was gathered from key stakeholders.**

***Index Terms*—American sign language, Augmented reality, Communication aid, Human-computer interaction, Parent-child interaction**

I. Introduction

Deaf or hard-of-hearing (DHH) children who are born into hearing families with limited to no knowledge of sign language encounter a heightened risk of language deprivation during their crucial early years of language development. This is especially concerning since over 90% of DHH children, in the United States, are born to hearing families [1]. This deprivation can lead to a myriad of long-term effects, including constrained first language development [2], cognitive delays [3], and mental health challenges [4].

Existing research primarily aims to assist hearing parents and DHH children in learning sign language in a directly didactic manner. Researchers have investigated games [5], mobile apps [6], bilingual storybooks [7], and avatar-based systems [8]–[11] as potential avenues of teaching American Sign Language (ASL). However, improving real-time, face-to-face communication between hearing parents and DHH children has had little focus despite the fact that high levels of language competence in DHH children is related to high degrees of parental involvement in language acquisition [12]. To address the issue of language deprivation while also providing a communication aid so hearing parents can interact with their DHH child in natural play scenarios, we previously introduced the design of the Tabletop Interactive Play System (TIPS) [13]. When using TIPS, a parent can speak naturally, and a word is chosen from their utterance, followed by a video of the corresponding ASL sign being played. The parent can then mimic that sign while interacting with their child, providing linguistic input to their child.

There remains, however, a critical research gap when designing a real-time communication aid for hearing parent-DHH child interactions: ***what word should be shown to the parent and, subsequently, the child?*** Imagine a scenario in which the parent says, "It's a sunny day and the duck is in the water!" The current configuration of TIPS must select one word from many candidates, such as *sunny*, *day*, *duck*, or *water.* Human-computer interaction research stresses the importance of both minimizing cognitive load [14] and supporting user autonomy when designing user-centered systems [15]. But as a communication aid specifically meant to address the issue of language deprivation in DHH children, we must also consider how to best facilitate satisfactory communication from the parent to the child and how to best support first language acquisition in the child. Addressing this gap is paramount to ensuring that the system offers optimal ease of use and empowers parents to actively participate in the communication process with their DHH children.

To explore this question, we introduce **Tap-to-Sign**, a novel AR projection-based ASL sign retrieval system that grants parents complete control over the recommended word to sign to their DHH child. In addition to Tap-to-Sign, we propose three distinct automatic recommendation strategies,

This research was funded by NSF award #1950460, the University of Rochester Pump Primer II, and the Google Inclusive Research Program.

[§]Equal contribution

drawing insights from language acquisition research: **part-of-speech**, **semantic weight**, and **word prevalence**. Each of these strategies leverages specific linguistic cues to suggest the most contextually appropriate ASL signs for the parent to use in their communication with their child. We aim to explore user preference for different methods, providing valuable insights into optimizing ASL communication technologies for hearing parent-DHH child interactions.

In initial feedback, we found that participants generally liked Tap-to-Sign, but expressed concerns about disruptions to the flow of interaction when playing with a child. They had diverse preferences on the three different automatic recommendation strategies, indicating that future work may focus on customizing TIPS to the individual user.

II. System Overview

TIPS is built for parent-child toy play over a tabletop surface to provide isolated ASL vocabulary to the parent and child (see Figure 1). To remain unobtrusive, TIPS utilizes an AR projection system to visualize the ASL videos. By combining natural language processing and object detection pipelines, we are able to also project the video near the object of interest, i.e. the toy the parent is likely talking about. We are thus able to minimize attention diversion and gaze-switching as the parent and child will be jointly attending to the current object of interest [13].

The system takes audio input from a lapel microphone and feeds it into an automatic speech recognition (ASR) engine, implemented with Whisper [16]. It also gives the top-view camera feed to our fine-tuned YOLO-v7 object detection model [17]. The transcript is then passed into the ASL recommendation module that returns a single-word recommendation from the sentence based on a recommendation strategy e.g., *duck*. The recommended word is fed into the ASL video retrieval and object of interest detection modules. The former uses cosine similarity with word embeddings from a pre-trained BERT model [18], [19] to find the most similar ASL sign in our ASL video database. In a similar manner, the latter takes the results of the object detection module and determines the object on the table that is most semantically similar to the recommended word. Finally, the retrieved video and the object of interest are fed into the rendering system which projects the ASL video next to the object of interest. An example of this process is in Figure 1.

Additionally, the current interface includes a live interaction log that projects the real-time transcript and highlights the recommended word. Because Whisper could inaccurately transcribe the conversation or the recommended word could be misaligned with a user's communicative intent, the live interaction log allows the user to validate the accuracy of the system's performance.

III. Tap-to-Sign

Autonomy has long been a focal point in Human-Computer Interaction (HCI) and assistive technology research. Autonomy is a multifaceted concept, subject to various interpretations and meanings [15]. In our context, we define autonomy as: the experience in which outcomes align with personal values, while also perceiving oneself as the causal agent [20] [21]. Previous research examined autonomy as it pertains to user trust, intentions and control in many scenarios [22] [23] [24]. However, understanding how autonomy plays a role in facilitating communication between hearing parents and DHH children is an underexplored topic, although it is instrumental in informing the design of an effective ASL communication tool. The rationale behind our approach is that, by promoting autonomy, we foster a sense of agency over what ASL signs that the parents express, thereby allowing them to actively engage in communication with their child.

A. Tap-to-Sign System Overview

To do this, we developed the **Tap-to-Sign (T2S)** system, which presents a novel approach that enables hearing parents to communicate more autonomously with their DHH children. T2S utilizes the same system architecture as the TIPS system, including a microphone, overhead camera and projector, in a table-top setup. As the parent plays with their child, a live transcript of what they are saying is projected onto the table (Figure 2). With Tap-to-Sign, parents can tap on words within their speech that they want to see the corresponding ASL video for. This design offers flexible communication options, enabling parents to select multiple words within a single sentence. This capability allows them to convey their communicative intent effectively and explore a diverse range of ASL signs during interactions.

Our technical implementation of Tap-to-Sign required rapid performance and low latency. The efficiency of this communication tool directly impacts the fluidity and effectiveness of interactions between hearing parents and their Deaf or hard-of-hearing (DHH) children. Ensuring that the T2S system operates with speed and minimal delay is crucial in delivering a seamless user experience, where parents can effortlessly tap on words to generate ASL videos.

B. Implementation Details

To transcribe parental speech, we use the $\text{Whisper}_{\text{tiny}}$ automatic speech recognition model (ASR) [16], for its fast inference speeds. To enable tapping on the table, T2S automatically calculates a Homography Matrix between the camera and table coordinates using the OpenCV library [25]. This is calculated before use, so coordinates can be quickly transformed for each frame. T2S then locates the position of the parent's hand by using a carefully tuned color mask that is only meant to encompass a range of skin colors. By leveraging the top down view of the camera, T2S uses this mask to detect if the parent is tapping a specific word on the table. Using a mask is a low latency option that enables rapid localization of the finger tip tapping on a word. At the onset of a parent selecting a word, T2S uses the BERT [18] language model to find the word in our ASL video database that is most similar to the selected word. Specifically, T2S performs semantic search, using the *all-MiniLM-L6-v2* model from the SentenceTransformer

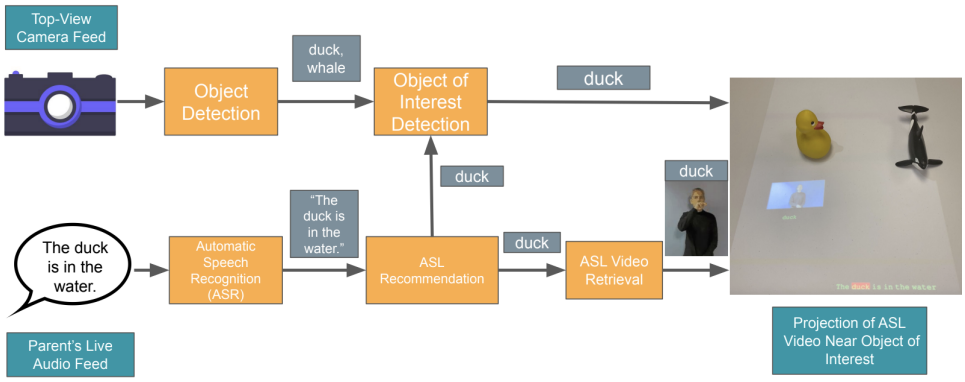

Fig. 1. Overview of TIPS design

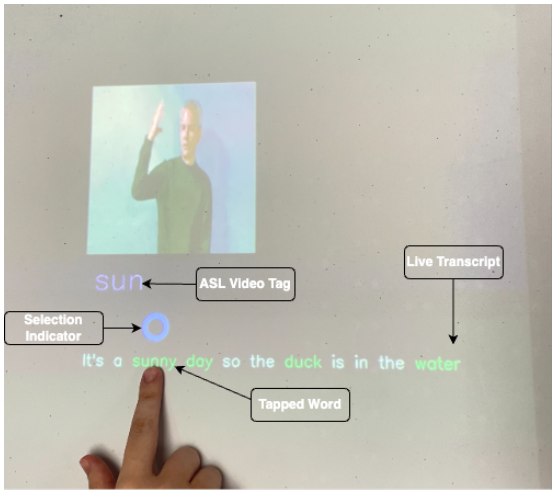

Fig. 2. Interaction with Tap-To-Sign

library [19]. T2S can also utilize multiple methods to render the ASL video, such as projecting all videos at a fixed point or projecting the video above the selected word.

IV. Automatic Real-Time ASL Recommendations

Although autonomy is a crucial consideration in the design of our system, we also have to consider the user's cognitive load. Cognitive load reduction is an integral design principle when creating human-computer interaction systems [14]. Providing users full control of the recommended word may overwhelm them with choice, especially since TIPS' target user is a parent with little to no previous ASL experience. Additionally, having users always choose the recommended word could negatively impact the flow of interaction between the parent and the child, placing focus on the communication system rather than the social interaction at hand.

In consideration of these potential drawbacks, we have developed three distinct automatic recommendation strategies to minimize TIPS' cognitive load and interaction intrusion. All the automatic recommendation strategies focus on single word, vocabulary-based output in order to minimize cognitive load and to align with the natural progression of language learning, where children and adult alike learn individual words before developing more complex linguistic expressions [26].

A. Part-of-Speech

Research shows that the part of speech of a word is an indicator of how much semantic information it carries. Content words (i.e., nouns, verbs, adjectives, and adverbs) carry more semantic information compared to function words (e.g., articles, prepositions, or pronouns) [27]. Nouns in particular usually carry more information in a sentence [28]. Content words are also thought to promote children's engagement in interactions [26] and form a large proportion of young children's expressive vocabulary [29].

To reflect the above considerations, the part-of-speech recommendation strategy extracts words according to their part of speech. In order of priority, it will recommend the first noun i.e., the one that is closest to the start of the utterance in terms of linear surface order. If there is no noun in a given utterance, it will recommend the first verb, and if there are no nouns or verbs, the first adjective. Since ASL and English encode adverbs in different manners, TIPS only considers nouns, verbs, and adjectives. For example, if a parent utters, "It's a sunny day, so the duck is in the water," this strategy will return the first noun in the sentence, "day."

B. Semantic Weight

Semantic weight is a composite term that integrates multiple dimensions of individual and inter-word semantics. Like the part-of-speech strategy discussed above, the semantic weight strategy generally prioritizes content words such as such as nouns and verbs, especially those that play critical roles such as the grammatical subject of the main clause. Importantly, the semantic weighting strategy can address the a language learner's core need, which is to communicate the intended meaning with a limited set of vocabulary [30]. By approximating semantic weight, this strategy could also return the word that captures the gist of the utterance which could be most representative of the parents' communicative intent.

The semantic weight recommendation strategy extracts the word that contains the largest amount of semantic information in the sentence as determined by a pre-trained language

model. Following previous research that uses a pre-trained BERT model to determine important words in live captioning [31], TIPS uses the all-MiniLM-L6-v2 BERT model [19] to generate sentence and word embeddings. We then use cosine similarity to find the word that is most similar to the sentence as a whole. For example, if a parent utters, "It's a sunny day, so the duck is in the water," this strategy will return "duck" because it is most semantically similar to the full sentence.

C. Word Prevalence

Using the North American English corpora of the Child Language Data Exchange System (CHILDES) [32], a 5 million-word database of child and child-directed speech, this strategy approximates how prevalent a word is in child and child-directed speech. This allows us to recommend more age-appropriate words, as more frequently appearing words tend to be acquired earlier [33], and thus more closely align to the needs of the parent-child dyads TIPS targets since parents may want to communicate the simpler, more commonly-occurring words to their children.

The word prevalence recommendation strategy extracts the word from the utterance that occurs most frequently in the North American English corpora of CHILDES. We determine the frequency that each word occurs in the corpus and return the most frequently occurring word. For example, if a parent utters, "It's a sunny day, so the duck is in the water," this strategy will return "water" because it is more frequently used in child-directed speech than other words in the same utterance.

V. Feedback Study

In order to understand users' preferences, we plan to invite key stakeholders such as language development experts, ASL teachers, and hearing parents with DHH children. After having them experience T2S and the three automatic recommendation strategies, we will conduct an interview. The goal is to determine how T2S compares with automatic recommendation and how the three automatic recommendation strategies compare with each other in terms of their effectiveness in capturing communicative intent and the extent they facilitate users' ASL learning and use.

A. Preliminary Feedback

We conducted an internal pilot study with a language development expert and an undergraduate student in our research team to gather feedback on the system. Neither had any prior experience with ASL.

1) Comparison of T2S to Automatic Recommendation: When we asked the language development expert to compare automatic recommendations to T2S, she said that she liked the ability to choose a word and that it made her learning process feel "more active." However, she also commented that T2S may be distracting to parents when they interact with their children since she focuses on her utterance whereas automatic recommendation allowed her to focus on the interaction at hand. The undergraduate student strongly appreciated the degree of autonomy T2S gave her, and that it was good at capturing her communicative intent. She commented that automatic recommendations felt "less personal" in comparison. However, she did think that T2S interrupted the play experience because she had to focus on choosing and tapping the word to sign. She liked that automatic recommendations relieved some of that cognitive burden.

2) Comparison of Automatic Recommendation Strategies: We also asked the participants to compare the three automatic recommendation strategies in terms of how well they represented their communicative intent, how well they supported language development, and their overall preferences. Because the student had no expertise in linguistics, we only asked questions regarding language development to the language development expert.

The language development expert said that part-of-speech and word prevalence may be better aligned with the trajectory of language acquisition. Interestingly, despite that the strategies gave different recommendations, she thought all of them equally captured her intended communication. She preferred the word prevalence strategy because it was the most intuitive and tended to recommend "easy" words, i.e. basic words such as concrete nouns or colors. On the other hand, the undergraduate student preferred the semantic weight strategy because she thought it best represented her communicative intent. Her least favorite was word prevalence because it felt repetitive to her, recommending words that added no new information.

The results of our preliminary interviews suggest that there is likely no one-size-fits-all design; instead, users will likely need some degree of customization of the system to achieve satisfaction with the results.

VI. Conclusion and Future Work

Our team previously presented TIPS, a communication aid that addresses a critical need to improve the quality of communication between DHH children and their hearing families. Now, we introduce **Tap-to-Sign** and three automatic ASL recommendation strategies: **part-of-speech**, **semantic weight**, and **word prevalence**. These novel systems allow us to tackle the question of what words the system should expose to parents and children. We gathered preliminary feedback on our system which indicated that different users will have different preferences regarding Tap-to-Sign and the different automatic recommendation strategies. Moving forward, we plan to conduct a full feedback study with stakeholders and then conduct a user study with hearing parents and their DHH children.

Acknowledgments

We thank Dr. Wyatte Hall for his valuable feedback on the system and Hecong Wang, Yifan Li and other members of Bai lab for their help and support.

REFERENCES

[1] R. E. Mitchell and M. Karchmer, "Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the united states," *Sign Language Studies*, vol. 4, no. 2, pp. 138–163, 2004. [Online]. Available: http://www.jstor.org/stable/26190985

[2] R. I. Mayberry and R. Kluender, "Rethinking the critical period for language: New insights into an old question from american sign language," *Bilingualism: Language and Cognition*, vol. 21, no. 5, pp. 886–905, 2018.

[3] A. Kumar, M. E. Behen, P. Singsoonsud, A. L. Veenstra, C. Wolfe-Christensen, E. Helder, and H. T. Chugani, "Microstructural abnormalities in language and limbic pathways in orphanage-reared children: a diffusion tensor imaging study," *Journal of child neurology*, vol. 29, no. 3, pp. 318–325, 2014.

[4] W. C. Hall, L. L. Levin, and M. L. Anderson, "Language deprivation syndrome: A possible neurodevelopmental disorder with sociocultural origins," *Social psychiatry and psychiatric epidemiology*, vol. 52, no. 6, pp. 761–776, 2017.

[5] D. Bansal, P. Ravi, M. So, P. Agrawal, I. Chadha, G. Murugappan, and C. Duke, "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*, 2021, pp. 1–10.

[6] K. A. Weaver and T. Starner, "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*, 2011, pp. 91–98.

[7] M. Malzkuhn and M. Herzig, "Bilingual storybook app designed for deaf children based on research principles," in *Proceedings of the 12th International Conference on Interaction Design and Children*, 2013, pp. 499–502.

[8] B. Scassellati, J. Brawer, K. Tsui, S. Nasihati Gilani, M. Malzkuhn, B. Manini, A. Stone, G. Kartheiser, A. Merla, A. Shapiro *et al.*, "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*, 2018, pp. 1–13.

[9] Q. Shao, A. Sniffen, J. Blanchet, M. E. Hillis, X. Shi, T. K. Haris, J. Liu, J. Lamberton, M. Malzkuhn, L. C. Quandt *et al.*, "Teaching american sign language in mixed reality," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–27, 2020.

[10] L. Quandt, "Teaching asl signs using signing avatars and immersive learning in virtual reality," in *The 22nd international ACM SIGACCESS conference on computers and accessibility*, 2020, pp. 1–4.

[11] S. Al-Khazraji, L. Berke, S. Kafle, P. Yeung, and M. Huenerfauth, "Modeling the speed and timing of american sign language to generate realistic animations," in *Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility*, 2018, pp. 259–270.

[12] J. Kobosko and M. Zalewska, "Maternal identity of hearing mothers of deaf adolescents. empirical studies—an interpersonal approach," *The Volta review*, vol. 111, pp. 39–59, 04 2011.

[13] E. Hossain, M. L. Cahoon, Y. Liu, C. Kurumada, and Z. Bai, "Context-responsive asl recommendation for parent-child interaction," in *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3517428.3550366

[14] A. Duenser, G. Raphael, H. Seichter, and M. Billinghurst, "Applying hci principles to ar systems design," 01 2007.

[15] D. Bennett, O. Metatla, A. Roudaut, and E. D. Mekler, "How does hci understand human agency and autonomy?" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3580651

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[17] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "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*, 2023, pp. 7464–7475.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] N. Reimers and I. Gurevych, "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, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[20] P. Wisniewski, H. Xu, M. B. Rosson, D. F. Perkins, and J. M. Carroll, "Dear diary: Teens reflect on their weekly online risk experiences," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 3919–3930. [Online]. Available: https://doi.org/10.1145/2858036.2858317

[21] S. Deterding, "Contextual autonomy support in video game play: A grounded theory," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 3931–3943. [Online]. Available: https://doi.org/10.1145/2858036.2858395

[22] S. Kafle, A. Glasser, S. Al-khazraji, L. Berke, M. Seita, and M. Huenerfauth, "Artificial intelligence fairness in the context of accessibility research on intelligent systems for people who are deaf or hard of hearing," *SIGACCESS Access. Comput.*, no. 125, mar 2020. [Online]. Available: https://doi.org/10.1145/3386296.3386300

[23] O. Alonzo, M. Seita, A. Glasser, and M. Huenerfauth, "Automatic text simplification tools for deaf and hard of hearing adults: Benefits of lexical simplification and providing users with autonomy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376563

[24] F. Güldenpfennig, P. Mayer, P. Panek, and G. Fitzpatrick, "An autonomy-perspective on the design of assistive technology experiences of people with multiple sclerosis," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–14. [Online]. Available: https://doi.org/10.1145/3290605.3300357

[25] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[26] E. Laubscher and J. Light, "Core vocabulary lists for young children and considerations for early language development: A narrative review," *Augmentative and Alternative Communication*, vol. 36, no. 1, pp. 43–53, 2020.

[27] R. Carnap, *The Logical Syntax of Language*, A. Smeaton, Ed. London,: K. Paul, Trench, Trubner & co., 1937.

[28] S. Waxman, X. Fu, S. Arunachalam, E. Leddon, K. Geraghty, and H.-j. Song, "Are nouns learned before verbs? infants provide insight into a long-standing debate," *Child development perspectives*, vol. 7, no. 3, pp. 155–159, 2013.

[29] J. C. Goodman, P. S. Dale, and P. Li, "Does frequency count? parental input and the acquisition of vocabulary," *Journal of child language*, vol. 35, no. 3, pp. 515–531, 2008.

[30] M. M. Vihman, "Variable paths to early word production," *Journal of Phonetics*, vol. 21, no. 1, pp. 61–82, 1993, phonetic Development. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S009544701931321X

[31] A. A. Amin, S. Hassan, C. Alm, and M. Huenerfauth, "Using BERT embeddings to model word importance in conversational transcripts for deaf and hard of hearing users," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 35–40. [Online]. Available: https://aclanthology.org/2022.ltedi-1.5

[32] B. MacWhinney, *The CHILDES project: The database*. Psychology Press, 2000, vol. 2.

[33] F. Smolík and M. Filip, "Corpus-based age of word acquisition: Does it support the validity of adult age-of-acquisition ratings?" *PLoS ONE*, vol. 17, no. 5, p. e0268504, May 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9132288/