Embodied AR Language Learning Through Everyday Object Interactions: A Demonstration of EARLL

Jaewook Lee[1,*] Sieun Kim[2,*] Minji Park[3] Catherine L Rasgaitis[1] Jon E. Froehlich[1]
[1]University of Washington, Seattle, USA [2]Seoul National University, Seoul, Korea
[3]Sungkyunkwan University, Seoul, Korea

Figure 1: Example interactions with EARLL. EARLL recognizes everyday object interactions such as reading a book and provides foreign vocabulary corresponding to that object. Optionally, the system also vocalizes the translation to support pronunciation.

Abstract

Learning a new language is an exciting and important yet often challenging goal. To support foreign language acquisition, we introduce *EARLL*, an embodied and context-aware language learning application for AR glasses. EARLL leverages real-time computer vision and depth sensing to continuously segment and localize objects in users' surroundings, check for hand-object manipulations, and then subtly trigger foreign vocabulary prompts relevant to that object. In this demo paper, we present our initial EARLL prototype and highlight current challenges and future opportunities with always-available, wearable, embodied AR language learning.

CCS Concepts

• Human-centered computing → Mixed / augmented reality; **Gestural input**; **• Applied computing → Education**.

Keywords

augmented reality, embodied language learning, computer vision

ACM ISBN 979-8-4007-0718-6/24/10
https://doi.org/10.1145/3672539.3686746

ACM Reference Format:
Jaewook Lee[1,*] Sieun Kim[2,*] Minji Park[3] Catherine L Rasgaitis[1] Jon E. Froehlich[1]. 2024. Embodied AR Language Learning Through Everyday Object Interactions: A Demonstration of EARLL. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3672539.3686746

1 Introduction

Learning a new language is an important yet often challenging life endeavor [22]. Today, many people rely on computer- and mobile-assisted language learning applications like *Duolingo* and *Rosetta Stone* [4, 25]. While these platforms enable learning from anywhere, there has been a growing interest in embodied [2, 17], context-aware [1, 7, 8, 14, 15], and augmented reality (AR)-based [3, 6, 9, 21, 24] approaches to bring language learning into real-world contexts. For example, *MicroMandarin* by Edge *et al.* is a mobile app that provides city-specific content to facilitate language microlearning [7]. Most closely related to our work, previous studies have used wearable AR and computer vision (CV) to display foreign vocabulary near objects in users' surroundings [3, 6, 9, 21, 24]. While promising, these studies have not examined embodied techniques like how a users' own physical interactions in the world—such as grabbing a cup, holding a book, or eating food—can assist learning.

Research in learning sciences highlights the importance of tangible manipulatives and physical interactions in learning [12, 19],

[*]These authors contributed equally to this study.

with some studies specifically noting its positive impact on learning foreign vocabulary [2, 17]. Informed by prior work, we designed and built EARLL to use interactions with everyday objects, such as grabbing, as cues for teaching foreign vocabulary (Figure 1). EARLL is an embodied and context-aware language learning application for wearable AR that leverages recent advances in AR, CV, and depth sensing that continuously segments and localizes objects in a user's vicinity, checks for grabbing gestures, and prompts foreign vocabulary when appropriate. Our vision is to support language learning subtly through everyday object interactions.

In this demonstration paper, we showcase an initial EARLL prototype then highlight its current challenges and future opportunities. The accompanying video demo highlights EARLL working in both a kitchen and office scenario. Beyond just suggesting foreign vocabulary, future iterations of EARLL may suggest context-dependent sentences (*e.g.,* instead of simply "*cup*", EARLL observes the user action and says "*a person is drinking water from a cup*" overlaid in AR in the foreign language). Leveraging user's physical behavior alongside object contexts can be valuable for learning [2, 12, 17, 19], and we encourage researchers to explore this space further. As a UIST Demo submission, we will invite attendees to learn a few new words as they interact with everyday objects.

2 System Implementation

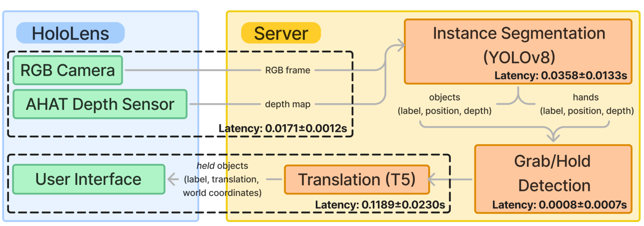

Figure 2: System overview of EARLL showing how data flows from the HoloLens to a local server for CV, then sent back for rendering foreign words.

We prototyped EARLL on a *Microsoft HoloLens 2* headset with the Mixed Reality Toolkit (MRTK3)[1] and Unity 2022.3.25f1[2]. We describe key components below (See Figure 2).

Capturing Context. To segment and localize objects near users, we streamed synchronized *RGB PhotoVideo* camera (*i.e.,* PV camera) data and *Articulated HAnd Tracking* (*i.e.,* AHAT sensor) depth data to a local server. To achieve this, we set the HoloLens to *Research Mode* [23] for accessing raw sensor data and used the *hl2ss* library [5] for real-time streaming (640x480 @ 30 FPS). On the server, we first read and integrate the RGB data into the depth map. We then perform YOLOv8 [10] on each frame to recognize objects in 3D space. The depth of each object is later retrieved from the object's instance segmentation mask. Objects identified as hands are used for grab detection.

Grab and Hold Detection. Because we are interested in surfacing language vocabulary only when a user has directly interacted with an object, we needed to create a robust and real-time user "grab-and-hold" detection algorithm using RGB + depth data alone. As this remains an open problem, we opted to employ a heuristic-based approach. Using known hand and object locations, we apply a two-step elimination approach to check if the object is grabbed: (1) bounding box overlap; and (2) intersecting depths. We verify the latter by: (1) checking if the depth ranges (10th to 90th percentile to remove noise and outliers) of hand and object pixels overlap; and (2) ensuring their mean depths are within 5 cm of each other. Finally, we categorize an object as held if, within a sliding window of multiple frames, it is inferred as grabbed more than twice and for over one second, accounting for occasional missed grab events.

Presenting Vocabulary. When objects are detected as held, EARLL displays their names in both L1 (native language) and L2 (second language) for three seconds in world coordinates. These coordinates are derived by projecting image coordinates from instance segmentation results. Object normals are calculated to ensure the words face the user. For translation, we leveraged the *T5* model [20], more specifically its *T5-small* variant. EARLL also speaks the object name in L2 to aid user pronunciation, which can be configured. To achieve multi-language text-to-speech, we used a TTS solution native to Windows and HoloLens 2[3].

Scenarios. We tested EARLL in various scenarios, including holding a wine glass, eating a banana, and washing a cup in the kitchen, as well as moving a mouse, using a cell phone, and reading a book in an office. See Figure 1 and our video figure for more.

3 Discussion and Conclusion

Below, we discuss current challenges and future potential in leveraging object interactions as cues for language learning.

Improved CV and System Latency. Although EARLL runs nearly in real-time, there is some perceived latency primarily due to slow grab-and-hold detection. Action recognition algorithms, including our heuristic approach, need to analyze multiple frames. In EARLL, we waited at least a second to mitigate and prevent false detections. Resolving latency would require more robust object detection and action recognition models.

Context-Dependent Sentence Suggestions. Studies show that learning vocabulary in sentence-level contexts is more effective for knowledge transfer, listening comprehension, and long-term recall compared to word-for-word learning [11]. To facilitate sentence-level learning, EARLL could use image-to-text models like BLIP-2 [16] to produce descriptions of user interactions (*e.g.,* instead of just "*pencil*", EARLL could say "*You are writing a note with a pencil*").

Beyond Object Grab and Hold. EARLL should recognize additional bodily gestures, including touch and pointing. Gestures referring to objects not directly related to the user, such as pointing, could allow users to receive even more vocabulary suggestions (*e.g.,* "cat" far away) [13]. We can even provide sentence-level suggestions (*e.g.,* pointing at a "*cat*" gives "*a cat sleeping on a couch*").

Gamification. By tracking object interactions, EARLL can provide personalized games that extend current gamification features [18].

Why Wearable AR? Unlike smartphones, an AR glasses allows users to use both hands freely as it continuously scans the environment and provide subtle prompts for foreign vocabulary.

We hope the space of embodied AR language learning and its future possibilities excite and inspire the research community.

[1]https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk3-overview/
[2]https://unity.com
[3]https://learn.microsoft.com/en-us/hololens/hololens2-language-support

Acknowledgments

This work was supported by an NSF Graduate Research Fellowship and NSF Award #1763199.

References

[1] Riku Arakawa, Hiromu Yakura, and Sosuke Kobayashi. 2022. VocabEncounter: NMT-powered Vocabulary Learning by Presenting Computer-Generated Usages of Foreign Words into Users' Daily Lives. 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 6, 21 pages. https://doi.org/10.1145/3491102.3501839

[2] Florence Bara and Gwenael Kaminski. 2019. Holding a real object during encoding helps the learning of foreign vocabulary. *Acta Psychologica* 196 (2019), 26–32. https://doi.org/10.1016/j.actpsy.2019.03.008

[3] Arthur Caetano, Alyssa Lawson, Yimeng Liu, and Misha Sra. 2023. ARLang: An Outdoor Augmented Reality Application for Portuguese Vocabulary Learning. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1224–1235. https://doi.org/10.1145/3563657.3596090

[4] Zhenzhen Chen, Weichao Chen, Jiyou Jia, and Huili An. 2020. The effects of using mobile devices on language learning: A meta-analysis. *Educational Technology Research and Development* 68, 4 (2020), 1769–1789. https://doi.org/10.1007/s11423-020-09801-5

[5] Juan C. Dibene and Enrique Dunn. 2022. HoloLens 2 Sensor Streaming. arXiv:2211.02648 [cs.MM] https://arxiv.org/abs/2211.02648

[6] Fiona Draxler, Audrey Labrie, Albrecht Schmidt, and Lewis L. Chuang. 2020. Augmented Reality to Enable Users in Learning Case Grammar from Their Real-World Interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376537

[7] Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A. Landay. 2011. MicroMandarin: mobile language learning in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3169–3178. https://doi.org/10.1145/1978942.1979413

[8] Ari Hautasaari, Takeo Hamada, Kuntaro Ishiyama, and Shogo Fukushima. 2020. VocaBura: A Method for Supporting Second Language Vocabulary Learning While Walking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 135 (sep 2020), 23 pages. https://doi.org/10.1145/3369824

[9] Adam Ibrahim, Brandon Huynh, Jonathan Downey, Tobias Höllerer, Dorothy Chun, and John O'donovan. 2018. ARbis Pictus: A Study of Vocabulary Learning with Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 11 (2018), 2867–2874. https://doi.org/10.1109/TVCG.2018.2868568

[10] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics

[11] Sook-Hi Kang. 1995. The effects of a context-embedded approach to second-language vocabulary learning. *System* 23, 1 (1995), 43–55. https://doi.org/10.1016/0346-251X(94)00051-7

[12] Ziyi Kuang, Wanling Zhu, Meixia Cheng, Fuxing Wang, and Xiangen Hu. 2023. The effectiveness of learning by enacting and its mechanisms. *Advances in Psychological Science* 31, 10 (2023), 1924. https://doi.org/10.3724/SP.J.1042.2023.01924

[13] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. https://doi.org/10.1145/3613904.3642230

[14] Sangmin-Michelle Lee. 2022. A systematic review of context-aware technology use in foreign language learning. *Computer Assisted Language Learning* 35, 3 (2022), 294–318. https://doi.org/10.1080/09588221.2019.1688836 arXiv:https://doi.org/10.1080/09588221.2019.1688836

[15] Jennifer Legault, Jiayan Zhao, Ying-An Chi, Weitao Chen, Alexander Klippel, and Ping Li. 2019. Immersive Virtual Reality as an Effective Tool for Second Language Vocabulary Learning. *Languages* 4, 1 (2019). https://doi.org/10.3390/languages4010013

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV] https://arxiv.org/abs/2301.12597

[17] Manuela Macedonia, AE Lehner, and Claudia Repetto. 2020. Positive effects of grasping virtual objects on memory for novel words in a second language. *Scientific Reports* 10, 1 (2020), 10760. https://doi.org/10.1038/s41598-020-67539-9

[18] Irina Kuznetcova Bethany Martens Mitchell Shortt, Shantanu Tilak and Babatunde Akinkuolie. 2023. Gamification in mobile-assisted language learning: a systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning* 36, 3 (2023), 517–554. https://doi.org/10.1080/09588221.2021.1933540 arXiv:https://doi.org/10.1080/09588221.2021.1933540

[19] Magdalena Novak and Stephan Schwan. 2021. Does touching real objects affect learning? *Educational Psychology Review* 33, 2 (2021), 637–665. https://doi.org/10.1007/s10648-020-09551-z

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG] https://arxiv.org/abs/1910.10683

[21] Marc Ericson C Santos, Arno in Wolde Lübke, Takafumi Taketomi, Goshiro Yamamoto, Ma Mercedes T Rodrigo, Christian Sandor, and Hirokazu Kato. 2016. Augmented reality as multimedia: the case for situated vocabulary learning. *Research and Practice in Technology Enhanced Learning* 11 (2016), 1–23. https://doi.org/10.1186/s41039-016-0028-2

[22] Tom Schuller, John Preston, Cathie Hammond, Angela Brassett-Grundy, and John Bynner. 2004. *The benefits of learning: The impact of education on health, family life and social capital*. Routledge.

[23] Dorin Ungureanu, Federica Bogo, Silvano Galliani, Pooja Sama, Xin Duan, Casey Meekhof, Jan Stühmer, Thomas J. Cashman, Bugra Tekin, Johannes L. Schönberger, Pawel Olszta, and Marc Pollefeys. 2020. HoloLens 2 Research Mode as a Tool for Computer Vision Research. arXiv:2008.11239 [cs.CV] https://arxiv.org/abs/2008.11239

[24] Christian David Vazquez, Afika Ayanda Nyati, Alexander Luh, Megan Fu, Takako Aikawa, and Pattie Maes. 2017. Serendipitous Language Learning in Mixed Reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2172–2179. https://doi.org/10.1145/3027063.3053098

[25] Xiaojing Weng and Thomas K.F. Chiu. 2023. Instructional design and learning outcomes of intelligent computer assisted language learning: Systematic review in the field. *Computers and Education: Artificial Intelligence* 4 (2023), 100117. https://doi.org/10.1016/j.caeai.2022.100117