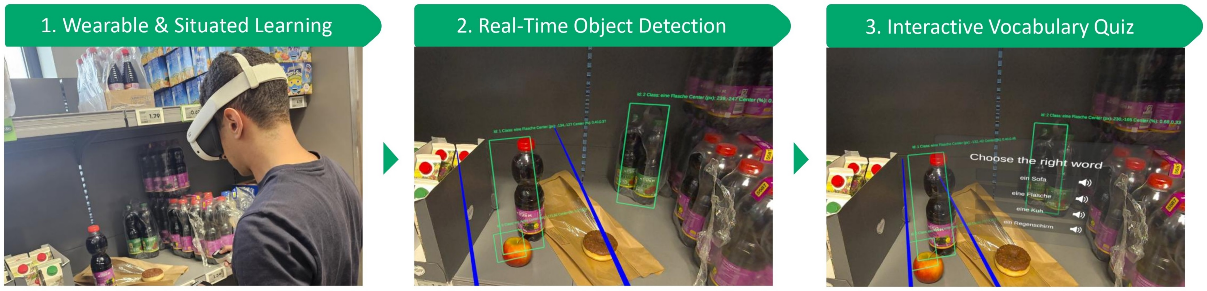
MRLingo: A Mixed Reality Approach for Situated Vocabulary Learning

Ziad Elshereif[*] Busra Balaban[†] Shehabeldin Solyman[‡] Michael Sedlmair[§] Carlos Quijano-Chavez[¶]

University of Stuttgart


1. Wearable & Situated Learning
2. Real-Time Object Detection
3. Interactive Vocabulary Quiz
Choose the right word
ein Sofa
eine Flasche
eine Kuh
ein Regenschirm

User explores real life contexts with MRLingo.

MRLingo detects and highlights objects in-situ.

Users can choose the correct text translation aided by audio pronunciation.

Figure 1: MRLingo enables vocabulary learning in real-life environments using mixed reality, combining object detection and interactive vocabulary practice. People are able to learn new words by text and audio while exploring arbitrary contexts.

Abstract

We present MRLingo, a mixed reality vocabulary learning approach that uses real-time object recognition to offer vocabulary acquisition. MRLingo runs dynamically in real-life contexts, offering more flexibility than traditional devices like smartphones and desktops. Likewise, MRLingo offers visual word translations, audio pronunciation, and quizzes, reinforcing vocabulary memorization. A user study was conducted to evaluate the effectiveness of MRLingo and compare it with a desktop-based prototype. Results show that MRLingo significantly outperforms engagement and usability, though challenges remain in cluttered or occluded contexts.

Index Terms: Mixed Reality, Vocabulary Learning, Object Recognition, Text-to-Speech, Vocabulary Acquisition, User Study.

1 Introduction

Vocabulary learning is the process of acquiring foreign language words and, together with technologies, enhances the user experience and retention [7]. Vocabulary learning tools are commonly used on handheld devices for convenience in spontaneous situations (e.g., Duolingo, Babbel, Anki, etc.). Likewise, Mixed Reality (MR) can leverage active situated contexts, augmenting the real world with dynamic content. Although handhelds are standardized nowadays, MR technologies could provide a better experience due to the hands-free and immersive interactions [6].

The benefits of augmenting the real world for vocabulary learning have been broadly studied. ARLang [2] recognizes outdoor objects for learning through smartphones. Besides, VisionARy [4] captures the surrounding objects to provide personalized conversation using chatbots in MR glasses. Similarly, ConversAR [1] uses language models to offer group conversation using MR devices. However, conversation tools demand a certain level of vocabulary knowledge. In addition, VocabulARy [5] presents audio, text, and animated visuals for vocabulary learning. Nevertheless, it uses specific markers, limiting its usage in arbitrary environments.

Inspired by the importance of vocabulary learning in random contexts and MR opportunities, we present MRLingo (Figure 1), a bilingual MR approach that combines real-world object detection to provide vocabulary learning using text and auditory feedback. To ensure the feasibility of MRLingo for vocabulary learning, we developed two complementary learning modes: an educational mode, which overlays translation labels and plays the sound in situ, and a quiz mode, which challenges learners to identify objects using multiple-choice questions. Therefore, we conducted a user study with 15 participants to evaluate the usability and vocabulary recall of MRLingo. Our results suggest that integrating MR technology, object recognition, and text/audio feedback significantly enhances vocabulary learning in arbitrary contexts. Finally, we discuss how MRLingo could be improved for more general situational cases.

2 Approach Overview

MRLingo follows design choices rooted in creating an MR application that (1) recognizes real-time objects, (2) teaches vocabulary in a foreign language, showing translated texts and audio feedback, and (3) provides two interactive modes of vocabulary learning.

Object detection. MRLingo uses an object detection model to process the real-time camera feed, highlighting objects with bounding boxes and displaying bilingual labels positioned nearby, following common practices in situated contexts [3].

Vocabulary feedback. MRLingo offers both visual and auditory feedback to assist in vocabulary acquisition. The recognized object is initially tagged in the native language and then translated, resulting in dual visual tags: the original word and its translation. Text-to-speech is then applied to audibly read each word aloud, providing multimodal reinforcement through auditory feedback.

Learning modes. Motivated by learning strategies [6], we implemented two modes: educational and quiz. For the educational mode, recognized objects are overlayed with visual labels, and the learner can select the label to hear its pronunciation. In quiz mode, the learner is shown detected objects and prompted to pick the right

[*] e-mail: Elshereifziad@gmail.com
[†] e-mail: st189878@stud.uni-stuttgart.de
[‡] e-mail: st196462@stud.uni-stuttgart.de
[§] e-mail: michael.sedlmair@visus.uni-stuttgart.de
[¶] e-mail: quijancr@visus.uni-stuttgart.de

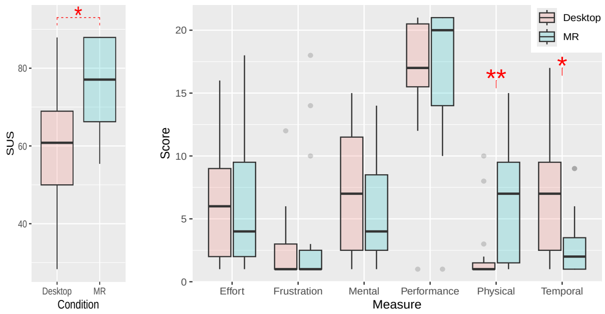


Figure 2: (Left) Mean SUS score evaluated by condition. (Right) Workload scores calculated across six NASA-TLX measures.

translation among a list of multiple-choice options via spatial pointers or hand tracking. This mode reinforces vocabulary retention.

MRLingo is developed for Meta Quest 3, using Unity3D and YOLOv9 for object detection via access to facing cameras. The Google Translate API handles the detected words to provide both text translations and audio output through text-to-speech. MRLingo operates standalone and can be applied in diverse contexts

3 Study

We aimed to investigate whether MRLingo offers usability comparable to that of conventional tools. Thus, we conducted a within-subjects study to compare the effectiveness of MRLingo with a desktop-based condition. The main language is English, and the learning language is Spanish due to the local population. Fifteen participants (8 male, 6 female, one non-binary; ages 18–34) were recruited, with varied educational backgrounds. All were fluent in English and non-Spanish speakers.

The participants were aware of the conditions before performing the sessions. Each participant performed two sessions, where they spent 10 minutes with MRLingo and a desktop-based form (or vice versa), and then answered three distinct vocabulary quiz questions. UMUX-Lite for usability and NASA-TLX for workload were required at the end of each condition.

The experiment lasted approximately 30 min. A total of 10 quiz words were chosen, and the participant randomly answered three per condition (counterbalancing). We use real products for MRLingo and pictures for the desktop form. Finally, participants were compensated with 14 EUR

4 Results

Results revealed that MRLingo is more usable than the desktop condition with strong evidence (Figure 2-left). For usability, we calculated the SUS score from Umux-lite. Friedman test was performed to compare usability (normal distribution invalidated by Shapiro-Wilk), resulting in significant differences. Post-hoc analysis by Wilcoxon suggests that MRLingo is more usable than desktop (p = .0174). Similarly, the workload is calculted using NASA-TLX (Figure 2-right). Friedman test resulted in significant evidence, and post-hoc analysis by Wilcoxon suggests that MRLingo is more physically demanding than desktop (p = .0057), but less time pressure than desktop (p = .0123). No differences were found in effort, frustration, and mental demand measures.

Regarding the accuracy, MRLingo was more error-prone. Participants answered incorrectly more in MRLingo (eight participants) than in the desktop condition (three participants). However, we suspect that the participants were more distracted by exploring more words using MRLingo, which differed from the desktop condition.

Additionally, participants' comments revealed a substantial preference for the multimodal feedback provided by MRLingo, such as real-time object recognition and audio pronunciation, valued for supporting context-aware vocabulary exploration on demand. Furthermore, they expressed interest in extending the system to cover more complex vocabulary or sentence-level practice.

5 Discussion and Conclusion

We have presented MRLingo, an MR vocabulary learning approach for situated contexts. The results show greater user engagement with MRLingo compared to the desktop condition, supported by the user feedback. Nevertheless, the visuals introduce obstruction challenges. Free object detection and continuous outlines created visual clutter at the expense of accuracy. Future research should introduce filtering mechanisms to deal with cluttering and occlusion in complex contexts. In addition, results show that MRLingo does not increase the workload, presenting similar effort, frustration, and mental demand measures to the desktop condition. Likewise, we infer that the physical demand of MRLingo is higher due to the device's lightness, while the temporal demand of MRLingo is lower due to the real-time occurrence of words. Future work could incorporate adaptive learning features, such as adjustable difficulty levels and pedagogical performance metrics. Enhancing interaction modalities—through gaze selection, hand tracking, or other natural inputs—may further streamline the user experience. Moreover, our study did not quantitatively assess performance; measures such as completion time and accuracy warrant more detailed analysis. Despite these limitations, MRLingo demonstrates strong potential for vocabulary acquisition in real-life contexts.

Acknowledgments

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting MS and CQ.

References

[1] J. Bendarkawi, A. Ponce, S. C. Mata, A. Aliu, Y. Liu, L. Zhang, A. Liaqat, V. N. Rao, and A. Monroy-Hernández. Conversar: Exploring embodied llm-powered group conversations in augmented reality for second language learners. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706599.3720162 1

[2] A. Caetano, A. Lawson, Y. Liu, and M. Sra. Arlang: An outdoor augmented reality application for portuguese vocabulary learning. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, DIS '23, p. 1224–1235. ACM, July 2023. doi: 10.1145/3563657.3596090 1

[3] B. Lee, M. Sedlmair, and D. Schmalstieg. Design patterns for situated visualization in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1324–1335, Jan. 2024. doi: 10.1109/TVCG.2023.3327398 1

[4] H. Lee, C.-C. Hsia, A. Tsoy, S. Choi, H. Hou, and S. Ni. Visionary: Exploratory research on contextual language learning using ar glasses with chatgpt. pp. 1–6, 09 2023. doi: 10.1145/3605390.3605400 1

[5] M. Weerasinghe, V. Biener, J. Grubert, A. Quigley, A. Toniolo, K. Pucihar, and M. Kljun. Vocabulary: Learning vocabulary in ar supported by keyword visualisations. 07 2022. doi: 10.48550/arXiv.2207.00896 1

[6] M. M. Zhang, H. Hashim, and M. M. Yunus. Analyzing and comparing augmented reality and virtual reality assisted vocabulary learning: a systematic review. *Frontiers in Virtual Reality*, Volume 6 - 2025, 2025. doi: 10.3389/frvir.2025.1522380 1

[7] F. Zulfiqar, R. Raza, M. O. Khan, M. Arif, A. Alvi, and T. Alam. Augmented reality and its applications in education: A systematic survey. *IEEE Access*, 11:143250–143271, 2023. doi: 10.1109/ACCESS.2023.3331218 1