Point, Scan, Learn: Comparing AR vs Traditional Vocabulary Learning

Daria-Teodora Godinel and Lena Pickartz

12.01.2026

1 Introduction

Vocabulary acquisition is fundamental to second language learning, directly impacting learners' ability to comprehend and communicate in the target language. However, despite vocabulary's central importance, learners frequently encounter difficulties in acquiring and retaining new vocabulary.

Recent technological advances enable multimodal presentation formats integrating visual, auditory, and interactive elements. While research shows that combined modalities can enhance vocabulary learning through multiple encoding pathways (Perez, 2022; Reynolds et al., 2022; Singh et al., 2021), existing studies on audio output predominantly examine passive contexts like captioned videos (Santos et al., 2016; Parmaxi & Demetriou, 2020; Chen & Chan, 2019) rather than interactive camera-based learning . This experiment investigates how translation delivery modes (audio plus visual text) combined with handheld AR devices affect vocal vocabulary recall in camera-based object recognition for language learning.

2 Related Work

Traditional learning theory has long emphasized the importance of sensory modality in information acquisition and retention. According to models of learning retention, learners retain approximately 20% of what they hear, 30% of what they see, and 50% of what they see and hear simultaneously (Unal, 2022).

As Unal (2022) notes, active engagement tasks combining "doing and saying" are theorized to yield the highest retention rates in the learning hierarchy, suggesting that AR-based vocabulary learning with vocal recall may be more effective than without it. By linking vocabulary directly to physical objects in learners' immediate environment, camera-based approaches could create embodied learning experiences that may enhance memory encoding through engagement and contextual association (Santos et al., 2016).

Research on audiovisual input for second language learning has expanded in recent decades, demonstrating benefits in vocabulary, listening comprehension, grammar, and pronunciation (Perez, 2022). Contemporary language pedagogy increasingly embraces technology-mediated instruction through platforms like YouTube, virtual classrooms, and mobile applications (Singh et al., 2021). Pattermore's synthesis shows that audiovisual input leads to incidental vocabulary acquisition of both single-word and multi-word units, though learning rates tend to be modest, especially at beginner levels (Pattemore & Gesa, 2025). Critically, this review reveals that on-screen text significantly enhances vocabulary learning outcomes, with Reynolds et al. (2022) reporting a large effect size in their meta-analysis, though they point out that the effect appears to vary according to the level of proficiency of the learner, the format of the test in the study (written vs. oral), and the specific linguistic features assessed.

Recent research in Human–Computer Interaction (HCI) highlights the growing role of Augmented Reality (AR) and multimodal feedback in language learning. These approaches create immersive, context-based environments that promote engagement and retention beyond traditional classroom methods.

Weerasinghe et al. (2022) demonstrated that AR-based vocabulary learning, combining visual and auditory cues, improved immediate and delayed recall while reducing mental effort. Similarly, a review by Huang et al. (2021) found that AR and VR tools enhance motivation, reduce learning anxiety, and improve learning outcomes through immersive, interactive experiences.

Studies with secondary and elementary students (Anonymous, 2023; Aldossari & Alsuhaibani, 2023) confirmed that AR increases engagement and supports learner autonomy, even when performance gains are modest. Finally, Jenkins et al. (2020) showed how combining text, speech, and image data supports grounded language learning, emphasizing the value of multimodal input.

Overall, prior research suggests that multimodal AR interaction enhances both motivation and memory, while contextual, object-based experiences improve vocabulary recall. Additionally, interactive feedback that combines visual, auditory, and spoken elements appears to be central to effective digital language learning.

3 Prototype

3.1 Tech Stack

The prototype application was developed using *Flutter* with the *Dart* programming language for cross-platform deployment on Android and iOS devices. Flutter provides flexibility for designing a responsive user interface while supporting integration with camera and speech. For core functionality, the following libraries and APIs were utilized. For the Frontend (Flutter/Dart) the mobile scanner package was used to access the device's camera and detect QR codes. The translations were outputted by a platform-compatible TTS library for audio delivery and native device microphone access was enabled for vocal recall capture. For the Backend, based on Python/FastAPI, a SQLite database was implemented for logging participant sessions, translations (object, target word, modality, timestamp), and recall attempts, which were additionally saved in audio files. Afterwards a Pandas-based pipeline converted the database records to CSV format for statistical analysis. The source code of the app can be found an accompanying GitHub Repository (Godinel & Pickartz, 2025).

4 Features

The app provides the following functionality:

1. Uses the device camera to display real-world object translations using QR codes.
2. Displays and outputs the translation to the user in both text and vocal.
3. Allows users to record their speech.

4.1 Frontend Design

The user interface is designed for intuitive interaction. A fixed reticle in the center of the screen allows users to select objects for translation. Translated terms are displayed in a box with an option to listen to the pronunciation and a button is provided to record a recall attempt and to start a new translation session.

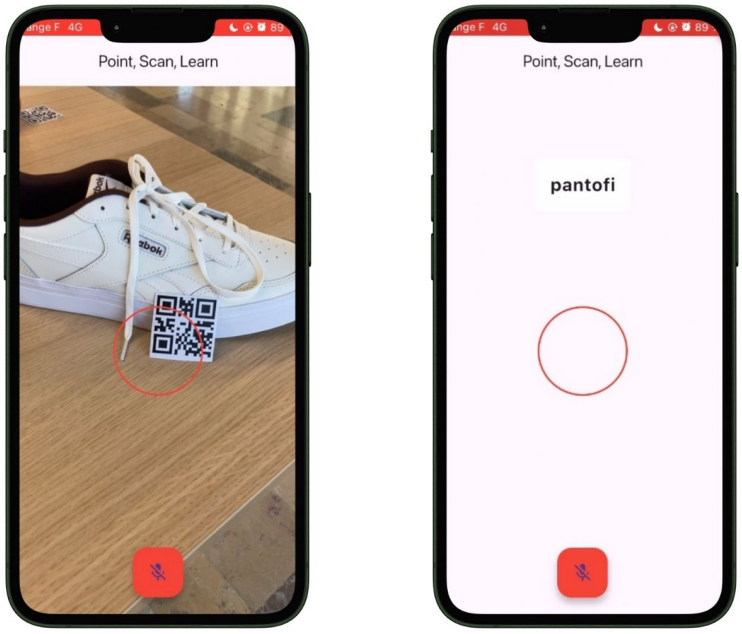

Figure 1: Frontend of the AR application

5 Experimental Plan / Setup

5.1 Research Question

The present study addresses the following research question: *"Does AR statistically significantly improve vocabulary recall in multimodal language learning?"*

This question is motivated by the growing use of augmented reality (AR) in language learning applications (Santos et al., 2016; Parmaxi & Demetriou, 2020; Chen & Chan, 2019). The study seeks to determine, if AR supports and improves vocabulary acquisition and memory retention during interactive camera-based learning tasks.

Participants interact with two experimental conditions: in both cases, they see the translated word displayed as text and hear it. In the first experiment, they will scan a QR code with the actual object near (e.g., *cup, apple, shoes*); in the second, they will scan a QR code with no objects around, but only the word in English (e.g., *spoon, cucumber, jacket*).

5.2 Participants, Language etc.

A total of ten participants is recruited for a within-subjects, between-items experiment. Each participant completes both of the two delivery conditions described above. The group is split in two, so that five people start with traditional delivery and the other five with the AR delivery first. The target words, real life objects, is translated into Romanian. All participants are exposed to the same set of objects to ensure comparability between conditions.

Participants include international students and peers familiar with English with no base knowledge of Romanian, ensuring that the target language is new for them. Demographic information such as age, nationality, gender and previous language learning experience will be recorded.

5.3 Materials

The materials used in this experiment include a mobile application prototype specifically developed to document translation recall across the two modes. As described in section 3 the app integrates QR-code recognition and displays translations as text with audio.

Three real-world objects will serve as learning stimuli for each condition. These objects were selected because they are common and easily recognizable items that participants are unlikely to know in Romanian. The app also includes a built-in voice recording, which allows users to pronounce translated words.

5.4 Procedure

Each participant completes the experiment individually in a quiet environment. The session begins with a short introduction explaining the goal of the task and how to use the mobile app. The introduction includes training sessions for using the app, to ensure that every participant is familiar with the application before starting the experiment.

For the experiment, participants are instructed to point the phone's camera at each target object or a word only (using a flashcard). They then read the translated word on the screen and listen to it through the device's speaker. After each word they are asked to repeat the Romanian translation which is recorded by the app. After learning the three words, participants are asked to provide their demographic data and get engaged in a conversation spanning three minutes. After this waiting period, the participants try to recall what was learned. A Romanian speaker judges the recall and documents which specific words were remembered or missed to collect quantitative data for analysis.

5.5 Measurement

Vocabulary recall performance was quantified by counting the number of correctly remembered words for each participant in each learning condition. Recall accuracy was scored binarily (correct = 1, incorrect = 0) based on participants' verbal reproduction of target Romanian words when shown the corresponding objects.

The primary analysis compared recall performance between AR and traditional learning conditions. Because vocabulary recall was measured as a binary outcome and different words were presented in the AR and traditional conditions, recall data were analysed using a mixed-effects regression model. It was fitted with recall accuracy (0 = incorrect, 1 = correct) as the dependent variable and learning condition (Traditional = 0, AR = 1) as a fixed effect. Participant was included as a random effect to control for individual differences in memory and language aptitude. The model was fitted using restricted maximum likelihood estimation (REML) based on 60 observations.

Mathematically, the model can be written as:

$$recall_{ij} = \beta_0 + \beta_1 \cdot condition_{ij} + u_i + \epsilon_{ij} \tag{1}$$

where (β_0) = the predicted recall probability in the Traditional condition (baseline); β_1) = the difference in recall probability between AR and Traditional, averaged across participants; i = participant; j = word; u_i = participant-specific random intercept (variation from average) and ϵ_{ij} = residual error.

The mixed-effects model revealed a statistically significant effect of learning condition on vocabulary recall. The estimated intercept (= 0.300, p < .001) represents the predicted recall probability in the traditional text-audio condition, corresponding to a recall rate of approximately 30%. The AR condition showed a significant positive effect (= 0.467, z = 3.98, p < .001, 95% CI [0.237, 0.697]), indicating that

AR increased recall probability by approximately 46.7 percentage points compared to the traditional condition. This yields a predicted recall rate of approximately 76.7% for AR-based learning.

6 Results

Participants remembered almost twice as many Romanian words when using AR (76.67% recall) compared to traditional text-based learning (30.00% recall). This difference was statistically significant with all ten participants showing better performance with AR.

An exploratory nationality analysis (Figure 2) revealed considerable individual variation, ranging from 33.3% (Egyptian participant) to 83.3% (one Italian participant). However, with only single participants representing six of seven nationalities, these differences cannot be interpreted as meaningful group effects.

Word-level analysis (Figure 3) showed that AR words achieved consistently high recall: *pantofi* (shoes) 90%, *măr* (apple) 80%, and *cupă* (cup) 60%. Traditional condition words showed more variable performance. The high recall for *jachetă* (jacket, 80%) likely reflects cross-linguistic phonological similarity to English "jacket", making it easier regardless of presentation method. Conversely, *castravete* (cucumber, 0%) proved exceptionally difficult for all participants. Despite these word-specific variations, the consistent improvement across all participants (Figure 4) demonstrates that the AR advantage extends beyond individual word characteristics.

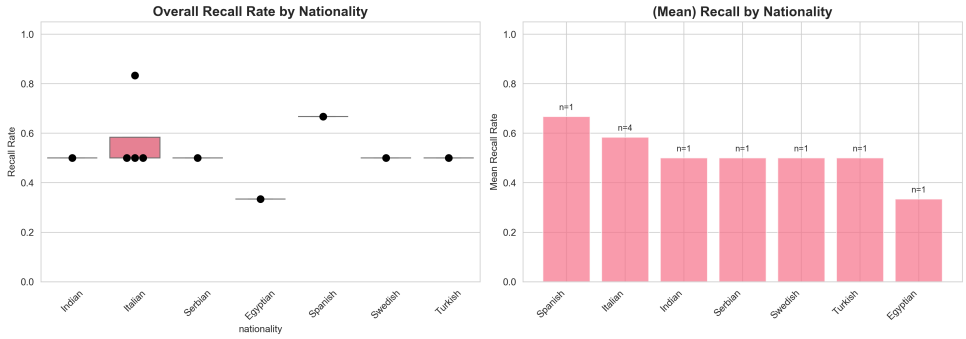

Figure 2: Overall recall performance by participant nationality

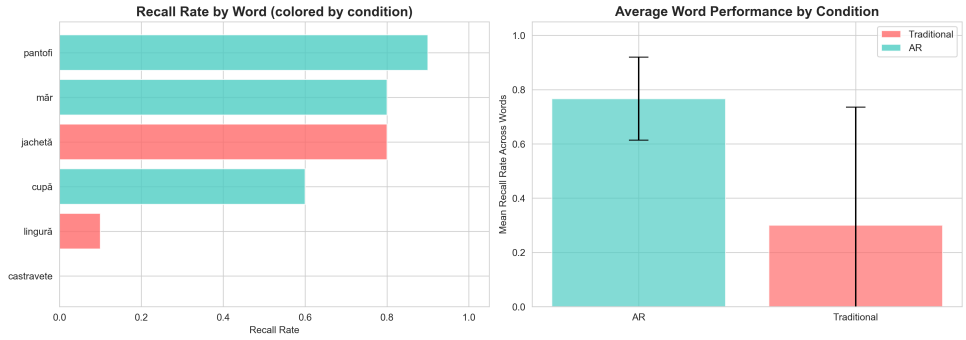

Figure 3: Vocabulary recall rates by word and condition

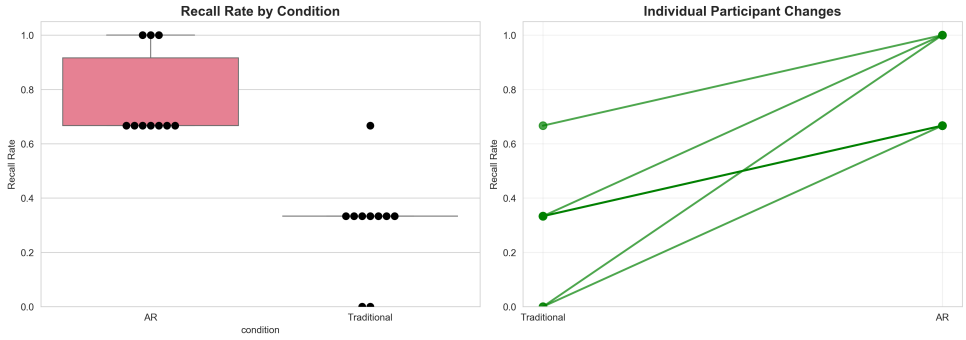

Figure 4: Recall performance by learning condition

6.1 Limitations

While the results indicate a positive effect of AR on short-term vocabulary recall, there are three main limitations. First, the sample size (n = 10) is relatively small, which makes it difficult to apply these findings to a wider group of people with certainty.

Second, we did not account for how hard the specific words were to learn. For example, the Romanian word *jachetă* sounds very similar to the English "jacket," making it easy to remember in either method. In contrast, *castravete* (cucumber) sounds completely different from English. This suggests that the choice of words might have influenced the comparison between the AR and traditional methods.

Finally, the backgrounds of the participants may have created a bias. Four of the ten participants were Italian speakers. Since Italian and Romanian are similar languages, these users likely had an easier time guessing or remembering words compared to those with different backgrounds (e.g., Arabic or Hindi speakers).

Future studies should control for the participants' native language, and focus on a balanced objective difficulty of translations to ensure that improvements are due to the AR tool, not language similarity.

7 Conclusion

This study compared AR against traditional text-audio methods for vocabulary learning. The results show a significant advantage for AR, with a 76.67% recall rate compared to 30.00% for the traditional approach. This suggests that physical interaction with objects helps users remember words better.

However, these results are preliminary due to the small sample size and variations in word difficulty. Future work requires a larger, more diverse group of participants and a fairer selection of vocabulary to rule out language bias. Despite these limits, the "Point, Scan, Learn" prototype demonstrates that handheld AR is a promising and effective tool for immersive language learning.

References and Related Research

Perez, M. M. (2022). Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching*, *55*(2), 163–192.

Reynolds, B. L., Cui, Y., Kao, C.-W., & Thomas, N. (2022). Vocabulary acquisition through viewing captioned and subtitled video: A scoping review and meta-analysis. *Systems*, *10*(5), 133.

Singh, C. K. S., Singh, H. K. S., Singh, T. S. M., Tek, O. E., Yunus, M. M., Rahmayanti, H., & Ichsan, I. Z. (2021). Review of research on the use of audio-visual aids among learners' english language. *Turkish Journal of Computer and Mathematics Education*, *12*(3), 895–904.

Santos, M. E. C., Lübke, A. i. W., Taketomi, T., Yamamoto, G., Rodrigo, M. M. T., Sandor, C., & Kato, H. (2016). Augmented reality as multimedia: The case for situated vocabulary learning. *Research and Practice in Technology Enhanced Learning*, *11*(1), 4.

Parmaxi, A., & Demetriou, A. A. (2020). Augmented reality in language learning: A state-of-the-art review of 2014–2019. *Journal of Computer Assisted Learning*, *36*(6), 861–875.

Chen, R. W., & Chan, K. K. (2019). Using augmented reality flashcards to learn vocabulary in early childhood education. *Journal of Educational Computing Research*, *57*(7), 1812–1831.

Unal, K. (2022). A study on the effect of visual and auditory tools in foreign language teaching. *Journal of research in Social Sciences and Language*, *2*(2), 108–117.

Pattemore, A., & Gesa, F. (Eds.). (2025, June). *Foreign language learning from audiovisual input: The role of original version television*. Springer. https://doi.org/10.1007/978-3-031-91001-2

Weerasinghe, M., Biener, V., Grubert, J., Quigley, A. J., Toniolo, A., Čopič Pucihar, K., & Kljun, M. (2022). Vocabulary: Learning vocabulary in ar supported by keyword visualisations. *IEEE Transactions on Visualization and Computer Graphics*, *28*(11), 3748–3758.

Huang, X., Zou, D., Cheng, G., & Xie, H. (2021). A systematic review of ar and vr enhanced language learning. *Sustainability*, *13*(9), 4639.

Anonymous. (2023). The impact of augmented reality (ar) on vocabulary acquisition and student motivation. *Electronics*, *12*(3), 749.

Aldossari, S., & Alsuhaibani, Z. (2023). Using augmented reality in language classrooms: The case of efl elementary students. *Advances in Language and Literary Studies*, *14*(1), 24–33.

Jenkins, P., Sachdeva, R., Youssouf Kebe, G., Higgins, P., Darvish, K., Raff, E., Engel, D., Winder, J., & Matuszek, C. (2020). Presentation and analysis of a multimodal dataset for grounded language learning. *arXiv preprint arXiv:2012.03484*.

Godinel, D. T., & Pickartz, L. (2025). *Ar-vocabulary-app* [GitHub repository]. https://github.com/len-rtz/AR-vocabulary-app

Zhan, A. ((n.d.). Speech_recognition/examples/microphone_recognition.py at github.com.

Patel, R. (2019). Giving Lens New Reading Capabilities in Google Go — research.google.