This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3203116

VocabulARy: Learning Vocabulary in AR Supported by Keyword Visualisations

Maheshya Weerasinghe, *Student Member, IEEE*, Verena Biener, Jens Grubert, Aaron Quigley, Alice Toniolo, Klen Čopič Pucihar and Matjaž Kljun

Fig. 1. VocabulARy prototype. (a) Participant interacting with VocabulARy during the study; (b) VocabulARy prototype though HoloLens2 in KEYWORD + VISUALISATION instruction mode (the Japanese word "hagaki" sounds as the English phrase "hug a key" (keyword) and visualised with an animated hand grabbing a key (visualisation)); (c) Participant interacting with non-AR version of VocabulARy in KEYWORD instruction mode (Note there is no visualisation of the keyword). AR and non-AR condition were tested with both instruction modes.

Abstract—Learning vocabulary in a primary or secondary language is enhanced when we encounter words in context. This context can be afforded by the place or activity we are engaged with. Existing learning environments include formal learning, mnemonics, flashcards, use of a dictionary or thesaurus, all leading to practice with new words in context. In this work, we propose an enhancement to the language learning process by providing the user with words and learning tools in context, with VocabulARy. VocabulARy visually annotates objects in AR, in the user's surroundings, with the corresponding English (first language) and Japanese (second language) words to enhance the language learning process. In addition to the written and audio description of each word, we also present the user with a keyword and its visualisation to enhance memory retention. We evaluate our prototype by comparing it to an alternate AR system that does not show an additional visualisation of the keyword, and, also, we compare it to two non-AR systems on a tablet, one with and one without visualising the keyword. Our results indicate that AR outperforms the tablet system regarding immediate recall, mental effort and task-completion time. Additionally, the visualisation approach scored significantly higher than showing only the written keyword with respect to immediate and delayed recall and learning efficiency, mental effort and task-completion time.

Index Terms—Augmented Reality, vocabulary learning, keyword method, contextual learning

1 Introduction

Learning a language is a complex task that requires dedication, perseverance and hard work. The basic learning process consists of comprehension of input (i.e. hearing or reading), comprehensible output (i.e. speaking or writing) and feedback (i.e. identifying errors and making changes in response) [6, 45]. Through these processes we learn vocabulary and grammar enhancing our language comprehension and expression abilities.

Expanding one's vocabulary is an essential element of language learning and in vocabulary learning, methods for improving learners' memory play a vital role. Mnemonics is one such effective method, in which the learner attempts to link new learning with prior knowledge through the use of visual and/or acoustic cues. Keywords are one such practical technique in which the learner attempts to create a symbolic link between new and prior knowledge using associations triggered by keywords, a method shown to be particularly effective in prior research [4].

Furthermore, previous research shows that learning vocabulary can be enhanced through an encounter with words in context [50]. Existing learning environments include formal learning, flashcards, use of a dictionary or thesaurus, all leading to practice with new words in context. For example, in formal learning the context is provided by the instructor or the provided instructional materials, in flashcards it is formed through images depicted on physical cards, in thesaurus it is provided through the provision of synonyms.

Consumer devices such as smartphones, tablets and head mounted displays can be used to enhance existing learning environments or to provide new ones. These systems enable technology driven paradigm shifts such as e-learning [40, 41, 44], and more recently m-learning (mobile learning) [19, 32, 71]. All are capable of enhancement through better provision of learning context and methods for improving learners' memory. Furthermore, these systems are also capable of running Augmented Reality (AR) applications which have the potential to make language learning more intuitive and immersive because of their intrinsic ability to visualise digital information within a real world context.

- *Maheshya Weerasinghe is with the University of Primorska, Slovenia and the University of St Andrews, United Kingdom. E-mail: amw31@st-andrews.ac.uk.*
- *Verena Biener and Jens Grubert are with the Coburg University of Applied Sciences, Germany. E-mail: {jens.grubert | verena.biener}@hs-coburg.de.*
- *Aaron Quigley is with the University of New South Wales, Australia. E-mail: a.quigley@unsw.edu.au.*
- *Alice Toniolo is with the University of St Andrews, United Kingdom. E-mail: a.toniolo@st-andrews.ac.uk.*
- *Klen Čopič Pucihar and Matjaž Kljun are with the University of Primorska, Slovenia. E-mail: {klen.copic | matjaz.kljun}@famnit.upr.si.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

This is particularly important for vocabulary learning because it allows word encounters in real-world context, an important catalyst for vocabulary learning [28, 62, 76].

Despite the fact that prior work looked at AR for vocabulary learning discovering several benefits, such as better improved retention, higher enjoyment, motivation and engagement, none provide a direct comparison of AR applications that run in head-mounted displays to the same technique within a non-AR interface. Furthermore, to the best of our knowledge, no existing evaluation of vocabulary learning that combines keywords with visualisations exist.

This paper contributes to addressing this gap with VocabulARy, an AR application for vocabulary learning that visually annotates objects in AR, in the user's surroundings, with the corresponding English (first language) and Japanese (second language) words. In addition to the written and audio description of each word, VocabulARy also presents the user with a keyword and its visualisation to enhance memory retention. We evaluate the VocabulARy prototype by comparing it to an alternate AR system that does not show an additional visualisation of the keyword and also, we compare it to two non-AR systems on a tablet, one with and one without visualising the keyword. The results show that AR outperforms the NON-AR (tablet) system regarding short-term retention, mental effort and task-completion time. Additionally, the visualisation approach scored significantly higher than only showing the written keyword with respect to immediate and delayed recall and learning efficiency, mental effort and task-completion time.

2 Related Work

Vocabulary learning can be enhanced through methods for improving learners' memory [39, 53] or through an encounter with words in context [50]. AR is an emerging technology for learning in real-world context and to scaffold this we structure our related work into: Learning context, Vocabulary learning in AR and Memory enhancement techniques. To better position our work in the context of language learning in AR environments we also classify prior work based on AR devices, learning content, presentation and learning method (Table 1).

Table 1. Selected prior work related to language learning in AR environments.

Study	AR Device	Content	Presentation	Learning Method
Draxler et al. (2020)	Hand-held	Grammar	Visual, Audio & Text	Context-based
Dalim et al. (2020)	Desktop	Vocabualry	Visual, Audio & Text	Experiential
Arvanitis et al. (2020)	Hand-held	Vocabualry	Visual, Audio & Text	Self-directed
Ibrahim et al. (2018)	HMD	Vocabualry	Visual, Audio & Text	Context-based
Yang & Mei (2018)	Hand-held	Vocabualry	Visual, Audio & Text	Context-based
Hautasaari et al. (2019)	Hand-held	Vocabualry	Audio	Context-based
Vazquez et al. (2017)	HMD	Vocabualry	Visual, Audio & Text	Context-based
Santos et al. (2016)	Hand-held	Vocabualry	Visual, Audio & Text	Context-based
Dita (2016)	Hand-held	Vocabualry	Visual & Text	Context-based
Li et al. (2014)	Hand-held	Vocabualry	Visual & Text	Not Speicified
Liu & Tsai (2013)	Hand-held	Vocabualry	Visual & Text	Context-based
VocabulARy	HMD	Vocabualry	Visual, Audio & Text	Context-based & **Keyword**

2.1 Learning Context

It has been shown that people are more motivated to learn, if they can see the importance of the content with respect to the situation or, if they find the content interesting [49]. For example, being in a bar in a foreign country is likely to increase the interest in learning words and sentences required for ordering a coffee. Additionally, the context makes it possible to form associations that help later retrieval in similar circumstances [28, 62, 76, 80]. In other words, new words relevant to the learning context are more likely to be recalled than unrelated words [15].

AR has the ability to provide context-specific information in an interactive manner. In addition, AR can take any situation, location, environment, or experience to a whole new level by combining digital information with real-world contents. Thus, it has the potential to create more engaging and immersive learning environments. There exists a considerable body of previous work on AR systems that support learning in real-world contexts. For instance, there are systems that provide labels of new words corresponding to real-world objects [13, 62], while others create imaginary settings to describe and enhance the physical properties of everyday objects [27, 73].

2.2 Vocabulary Learning in AR

Previous studies have shown that AR offers many advantages for language and vocabulary learning. For instance, some studies reported that AR improved learning achievements and boosted motivation, engagement and collaboration among learners [12, 13, 25, 28, 62]. Despite, some technical limitations of using AR for learning should be taken into account such as such as educators' limited proficiency with the relatively new technology [29] or the trade-off between connecting the experience to the context of the current location and providing a flexible and portable experience [79].

Fujimoto et al. [18] have shown that users can memorise AR information better if it is shown within the location of a target object in the real world (e.g. AR information about a country shown over a map within this country). However, the information to be memorised in study did not take the context of the real environment and users' surroundings into account.

Several studies presented applications for learning vocabulary using hand-held AR devices. Hautasaari et al. [25] developed the VocaBura smartphone application for learning vocabulary during dead time. The application tracks a users' GPS locations and presents vocabulary related to the current location via audio. A study comparing this to an audio-only method showed that 7 days after the study, participants could recall significantly more words. Santos et al. [62] presented a handheld AR system that displays text, images, animation and sound next to corresponding real-world objects to learn Filipino or German. They compared this system to a non-AR tablet application using a flash card method. Their results indicate that for tests directly after the experiment non-AR users performed better, yet this difference was not detected for long-term retention.

Positive effects of AR technology in the context of increased motivation and enjoyment have also been detected. For example, Dalim et al. [12] presented a system combining desktop-AR and speech recognition (TeachAR) and found that it increases children's knowledge gain and enjoyment. Similarly, Li et al. [34] explored an AR application for language learning and found that it increased motivation in the beginning, yet for most participants motivation decreased at the end of the study.

The existing body of literature also includes applications for learning vocabulary on AR head-mounted devices. While most previous systems used some sort of marker to align virtual content with physical objects, Vazquez et al. [76] presented a platform (WordSense) that detects objects in the physical environment and augments them with additional content for language learning including words, sentences, definitions, videos and audio. However, no formal user study was conducted to evaluate the system.

Another example of using AR head-mounted devices for language learning is ARbis Pictus, a system presented by Ibrahim et al. [28] which labels objects in the user environment with the corresponding vocabulary in the target language. They compared this system to a conventional flashcard-based system and found that AR was more effective and enjoyable and participants could remember words better both shortly after the experiment and four days later. However, the significance of these findings is limited, because the flashcard and AR systems were inherently different. For example, with flashcards the word was shown on the opposite side to the image depicting word meaning thus the image and word were never shown together. This was not the case in the AR condition where word annotations were always visible for all objects in the scene. Therefore it is not clear if AR accounted for the improved performance or the different learning approach.

In contrast to the presented studies we compare our AR prototype to a non-AR system that is as similar as possible to enable us to measure only the effect of AR without confounding variables like the learning method. To our knowledge, such an experiment has not yet been

conducted for AR applications that run on head-mounted devices.

2.3 Memory Enhancement Techniques

Research on memory and learning has shown that learning performance and retention depend on different strategies and techniques that can be used to process information in learning [14]. "Mnemonic" is one such technique where the memory capabilities are enhanced by connecting new information to prior knowledge through the use of visuals and/or acoustic cues [39, 53]. Several researchers experimentally showed that "Mnemonic" techniques improve memory and recall, especially in the area of language learning [1, 10, 47, 54].

In the field of language learning, mnemonics have mostly been used for vocabulary learning [2]. One such mnemonic method is the "keyword method" in which learners connect the sound of a word they want to learn to one they already know in either their first language or the target language. Through this process learners create a mental image that helps them remember the association [50]. For example, the Japanese word for bread is "sokupan" which in English sounds like "Sock + Pan". As a result, the learner can imagine a sentence that links a mnemonic keyword with the foreign word. For example, "sokupan" can be imagined as frying a sock and putting it on a slice of bread.

A wide range of existing studies in the broader literature have explored the effectiveness of the keyword method [2, 4, 54, 77]. In this context, comparing the keyword method against other methods in vocabulary learning is one of the most common research designs. There, the keyword method has been compared with learning words in context or learning words with no given strategy. For example, Pressley et al. [51] found the keyword method to be significantly more effective in learning over the context method. Also, Sagarra and Alba [58] compared rehearsal, semantic mapping displays and the keyword method, and found that the keyword method resulted in the best retention. It has also been shown that the keyword method is superior over systematic teaching [30, 51]. In 1975, Atkinson and Raugh [4] found that participants who were given a keyword along with the translation learned more words and also remembered more words after 6 weeks. In the same sense, Raug et al. [55] evaluated the use of the keyword method over a long period of 8 to 10 weeks to teach Russian vocabulary and found it to be highly effective.

Altogether, a significant number of research studies have shown that the keyword method of vocabulary learning is highly effective, yet others showed mixed results [48, 77]. For example, a study conducted by Zheng Wei [77] found no significant differences between the keyword method, the word-part technique (recognizing part of a word) and the self-strategy. From the perspective of the learning method, VocabulARy builds upon the work of Anonthanasap et al. [2] in which the authors propose an interactive vocabulary learning system to teach Japanese that automatically creates keywords using phonetic algorithms. There, if the learner selects an image in the system, the phonetically similar words with image representations will gather around the selected image. Results showed that the keyword-based vocabulary learning system required significantly lower workload than the other compared methods (e.g. paper dictionary and static visualisation in a form of an image).

In summary, our work was inspired by Santos et al. [62], Vazquez et al. [76] and Ibrahim et al. [28] who already used AR devices to augment real world objects with annotations for vocabulary learning. We combined this approach of providing context with a keyword method which has proven to work well in various experiments thus far [2, 4, 58]. To further advance this learning method we augment keywords with visualisations. AR provides ideal conditions for that, because the keyword and its visualisation can be shown in the same context as the corresponding physical object. However, in contrast to existing visualisations of keyword approaches we make a careful selection of keywords so that they not only sound similar, but can also be visualised with an animation in a meaningful way. For example, a Japanese word for bread is "sokupan" and sounds similar to keyword "Sock+Pan" which can be visualised with an animation of frying a sock in a pan and putting it on a slice of bread. We do this to uncover if one can improve vocabulary learning beyond the influence of the traditional keyword method, by augmenting the keywords with animated visualisations. According to Shapiro and Waters [67] the level of visual imagery of a word enhances vocabulary learning suggesting our approach should work, however no formal evaluation of this exists (Table 1). This makes our work both ground breaking and timely.

3 Augmented Vocabulary Learning / Learning Vocabulary with Visualisations in Real Life Context

This work presents two prototype systems for vocabulary learning developed on an AR head-mounted-display (HMD) (i.e. Microsoft HoloLens 2) and an 10.5 *in* Android tablet device (i.e. Samsung Galaxy Tab S4) (Fig. 1). To the best of our knowledge, AR HMD systems for vocabulary learning have not yet been evaluated against comparable non-AR systems (see Sect. 2.2). Both AR and tablet systems combine the keyword method with physical objects. With the AR HMD, our system allows the user to look around a physical room where certain objects are labelled with a button indicating that their translation is available. Upon clicking these buttons with a hand gesture, the English and Japanese word, as well as a keyword with or without visualisation, appears. In addition to the words in both languages and a keyword, an audio of the pronunciation is played. The details of application design and implementation of both prototypes are described in the following sub sections.

3.1 Application Design

In this section we describe key design decisions regarding annotations, animated visualisations and interactions. Careful consideration was given to the selection of annotation and visualisation size. Previous research showed the size of images affects our ability to remember image content during naturalistic exploration [38]. However in such exploration individuals are first asked to freely explore an image without any instructions and are then asked about the details observed. As such there is no guarantee that the visual attention is equally distributed across the observed image and as the image gets smaller so does the key information, which makes it easier to miss it.

To prevent participants from missing key information we design our application to effectively guide visual attention. The application shows only one word visualisation at a time, which avoids cluttering the scene and overloading participants with too much information. Furthermore, we specifically chose to place AR buttons on the physical surface at close proximity to the object of which word was being memorised. This led users to the appropriate physical location from which AR visualisations are clearly visible as the corresponding annotation and animated visualisation size were appropriated for such viewing. To make the NON-AR and AR condition as comparable as possible we made sure that the relative size of annotations and animation was roughly the same in both conditions.

Furthermore, in one of the instruction modes, the keyword is also accompanied by its animated visualisation. Such a visualisation consists of a 3D model resembling the keyword and a short animation involving the objects in question. Besides, the user can also listen to the pronunciation of the Japanese word again by clicking a virtual button that is displayed next to the keyword. For example, Figure 1b shows how the English word "Postcard" and the corresponding Japanese word "Hagaki" are displayed. "Hagaki" sounds like "Hug + A Key", so it is displayed as a keyword and visualised through "hugging a key".

3.2 Implementation of AR Prototype

The AR prototype was implemented for Microsoft HoloLens 2 [42] using the Unity3D game development engine [75]. For camera pose tracking, the HoloLens inbuilt tracking system was used. To initialise the positions of augmentations in our application, we used Vuforia [52] and our custom-made image markers (see Figure 2). We opted for markers in order to support a reliable and accurate detection of physical objects that correspond to the set of vocabulary. These markers were removed from the scene after initialisation.

It would be technically possible to use object recognition techniques to perform object identification and localisation as in [7, 59, 60]. Such implementation could support arbitrary environments without prior preparation, which would enable wide implementation of the system.

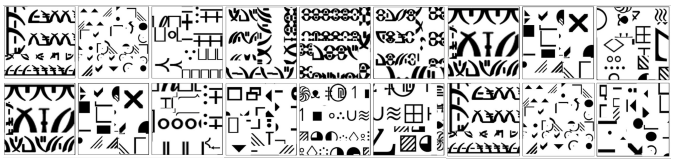


Fig. 2. Custom-made image markers.

However, this was not the scope of this work, as we focus on the effect of learning vocabulary using AR and visualisations.

To interact with the virtual contents, we use HoloLens' built-in hand-tracking and gesture inputs, which allow the user to interact with virtual content by moving the hands or fingers to content's corresponding positions. More specifically we chose to use virtual buttons placed on top of planar physical surfaces such as a table or a wall. In this way, touching a surface acted as a tangible feedback making the button press more realistic.

3.3 Implementation of Android Prototype

The Android version of the prototype was also implemented using the Unity 3D development environment, but deployed on a Samsung Galaxy Tab S4 [36] tablet device. Its functionality is similar to the AR prototype. However, instead of seeing the real world environment, an image of an environment is displayed on the screen. In our prototype this was either a kitchen or an office environment. As in the AR prototype certain objects are accompanied with a button. If the user touches the button, the corresponding English and Japanese words, keywords, and animated visualisations appear (Figure 1c). Visualisations only appear in one instruction mode. As the size of all objects was adapted to be clearly visible on the screen and to ensure that the relative size of annotations and animation was roughly the same in both conditions (see Sect. 3.1), we did not provide a feature to zoom into the scene.

Compared to the AR implementation, the tablet application can be used anywhere as it can also show scenes that are not related to the real-world environment around the learner. Because all kinds of virtual scenes can be presented, the tablet allows a more flexible use, such as learning words related to a forest while sitting in the living room.

3.4 Generating Keywords

For generating the keywords, we conducted a small informal survey with 7 participants. They were presented with 28 Japanese-English word pairs and were asked to come up with English words sounding similar to the Japanese words. At the end, we selected 20 words for the study for which the participants could come up with good keywords. As already mentioned, the process of finding keywords could also be automatized [2], but for the scope of this study, this was not needed.

4 Research Method

This section describes the study conditions, study design, participants' sampling, study procedure, data collection instruments, and analysis.

4.1 Study Conditions

We designed four study conditions based on two distinct vocabulary learning scenarios. The first scenario displays ten physical objects related to a kitchen environment, while the second shows an office environment with ten relevant physical objects. In each of these scenarios a different INSTRUCTION MODE is provided. This is either a KEYWORD instruction mode or a KEYWORD + VISUALISATION instruction mode. In the KEYWORD condition, only the written keywords are displayed to support the participants in remembering the word. In the KEYWORD + VISUALISATION condition, an animated 3D visualisation of the keyword is displayed in addition to the written keyword. These variations are presented to the participants on two different INTERFACES, one in AR on a HoloLens2 and one in NON-AR on an Android tablet. These study conditions are illustrated in Figure 3.

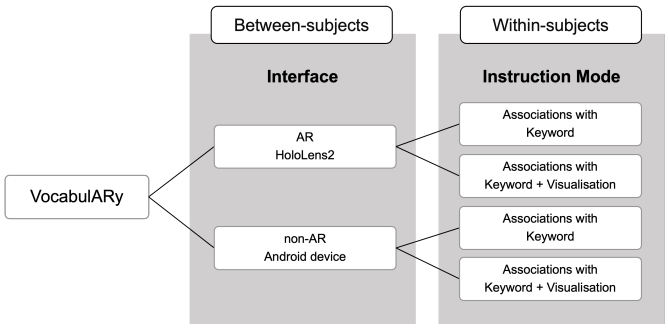


Fig. 3. Study design and conditions.

4.2 Study Design

Our study design has two independent variables: INSTRUCTION MODE which is KEYWORD or KEYWORD + VISUALISATION and INTERFACE which is either AR or NON-AR. We used a 2×2 mixed design (see Figure 3) because a within-subjects design would make the study, which is mentally demanding, too long (i.e. approximately two hours). We believe that such a long duration could intensify the fatigue and hinder the performance of the participants. Furthermore, running all 4 conditions in a within-subject design would require 2 additional learning scenarios making it more difficult to counterbalance for scenario effects. Reducing the study length by running the study in several sessions also introduces other biases and practical issues. Therefore, the INSTRUCTION MODE was evaluated as a within-subjects variable and the INTERFACE as a between-subjects variable. This means each participant was either using the AR-prototype or the NON-AR-prototype, but all participants experienced the KEYWORD and the KEYWORD + VISUALISATION conditions.

To avoid the "order effects", which may have an influence on participants' performance due to the order in which the conditions are presented [66], the order of the INSTRUCTION MODE as well as the order of the learning scenarios (the kitchen and the office environments) was counter balanced. Special care was given to counterbalance the learning scenario across all independent variables.

4.3 Participants

The study was completed by 32 participants, all voluntarily recruited from our university. None of the participants had any prior knowledge of the Japanese language (identified via a short competency test questionnaire). The between subject sample comprised of 16 participants for the AR condition (10 females) and 16 participants for the NON-AR condition (7 females). All the participants were between the age of 19 to 30 years, with the mean of $\bar{x} = 21.6$ and $SD = 2.1$.

All our participants were computer science undergraduate and graduate students. No student had previous experience with AR HMDs. The mean age for the AR group was $\bar{x} = 22.13$ ($SD = 2.68$), and for NON-AR group $\bar{x} = 21.13$ ($SD = 1.26$). The percentage of females in the AR group was $\bar{x} = 62.5\%$, and in the NON-AR group $\bar{x} = 43.75\%$. The groups were comparable.

4.4 Procedure

On arrival participants were first randomly assigned to one of the two groups (AR or NON-AR). Next, we randomly selected which instruction mode would be used first (KEYWORD or KEYWORD + VISUALISATION). Finally, the learning scenario was also randomly chosen (kitchen or office environment). All randomisations were counterbalanced.

After being assigned to a particular condition, participants were given a consent form to sign, together with the Participant Information Sheet (PIS) outlining the entire research process in simple language. After briefly explaining the vocabulary learning task with the two learning scenarios, they were asked to fill in the Questionnaire on Current Motivation (QCM) [57].

Before starting the actual task they completed a five-minute training session on a separate demo application to understand the VocabulARy interface. Participants were then given up to 15 minutes to complete

the first language learning scenario with 10 words. After finishing the first scenario, they filled in the NASA Task Load Index (NASA-TLX) questionnaire [24] and, then answered a post-test questionnaire developed by the researchers to assess their immediate recall performance. After taking a 5 minutes break, they were again given up to 15 minutes to complete the second language learning scenario with 10 words. Subsequent to the second scenario, they filled in the same NASA-TLX and the immediate recall questionnaires.

After finishing the experiment, participants were given another two standard questionnaires – a system usability (SUS) [33] and a user experience questionnaire (UEQ) [64]. At the end, participants filled in a short post-questionnaire with demographic questions, questions about previous experience with AR technology and vision problems they might have. The entire experiment took 45 to 60 minutes.

One week later, participants were again requested to answer the same post-test questionnaire developed by the researchers to assess their delayed recall performance as undertaken in prior work [25].

4.5 Data Collection

In order to measure the task completion time, the time stamp data (start time and end time) were logged by the system. To measure the motivation, the short form of Questionnaire on Current Motivation (QCM) with 12 items/questions [17, 56] was used. Anxiety, challenge, interest, and probability of success were measured on a five-point Likert scale, with the labels "strongly disagree" at 1 and "strongly agree" at 5. Rather than aiming for constructing sub-dimensions (i.e., anxiety, challenge, interest, and probability of success), we used the mean score of the 12 items as an indicator for the overall motivation.

The NASA Task Load Index (NASATLX) [23, 43] was used to measure participants' subjective level of workload/mental effort. Participants rated five of its six dimensions (mental demand, physical demand, temporal demand, effort and frustration) on a 20-point scale ranging from 0 (very low) to 20 (very high). The endpoints of the sixth dimension (own performance) were success and failure. Finally, the overall workload/mental effort was calculated across these six dimensions.

In the retention questionnaires, participants were asked for the Japanese translations of the vocabulary they learned. This was undertaken immediately after interacting with the prototype (Immediate Recall) and after one week (Delayed Recall).

Learning efficiency was determined based on the ratio of performance to the difficulty of the learning task as proposed in [46]. The performance of each study condition was based on the recall scores participants obtained after completing the task. The difficulty of the task was based on the mental effort they invested during the learning phase (see Sect. 5.3). Performance and task difficulty data were then standardised using Formula 1 where z = Z-score, r = Raw data score, M = Population mean, and SD = Standard deviation.

$$z = \frac{r - M}{SD} \tag{1}$$

Next, the learning efficiency (E) was assessed for each of the four study conditions (Sect. 4.1) using Formula 2 [9, 22, 46] where E = Learning efficiency, z_P = Average performance in Z-scores, and z_M = Average task difficulty in Z-scores. This was done for both immediate recall performance (immediately after participants had completed the task) and delayed recall performance (a week after participants had completed the task). Note that square root of 2 in this formula comes from the general formula for the calculation of distance from a point, $p(x,y)$, to a line, $ax+by+c=0$.

$$E = \frac{z_P - z_M}{\sqrt{2}} \tag{2}$$

To measure the usability of the system, we used the System Usability Scale (SUS), a ten question questionnaire originally created by Brooke, 1996 [5], on a five-point Likert scale, ranging from "Strongly" agree at 1 to "Strongly disagree" at 5. For measuring the user experience we used the short version of the User Experience Questionnaire (UEQ-S) [63,64] with eight items/questions, reported on a 7-point Likert scale. The first four represent pragmatic qualities (Perspicuity, Efficiency and Dependability) and the last four hedonic qualities (Stimulation and Novelty) [63].

4.6 Data Analysis

The analysis was completed in R studio. Each data set collected in the study was first checked for mixed ANOVA assumptions. The normality assumption was checked using the Shapiro–Wilk normality test [68]. Most of the data sets were normally distributed with some of them only approximately normally distributed. The homogeneity of variance assumption of the between-subject factor (INTERFACE) was checked using the Levene's test [65] that confirmed homogeneity of variances for each variable ($p > 0.05$). Finally, the homogeneity of covariances of the between-subject factor (INTERFACE) was evaluated using the Box's M-test of equality of covariance matrices. The test showed homogeneity of covariances for each variable ($p > 0.001$). Considering the fact that some of the data sets were only approximately normally distributed, we used robust statistical methods implemented in the "WRS2" R package to conduct the analysis [37], which is a standard practice in such cases.

In all statistical analyses we used a significance level $p-value > 0.05$ and a restrictive confidence interval (CI) of 95%. For immediate recall, delayed recall, mental effort, task completion time and learning efficiency, the statistical significance was examined using a robust two-way mixed ANOVA on the 20% trimmed means–"bwtrim" [37].

For motivation, system usability and user experience, the statistical significance was examined using a Mann–Whitney U test [74]. Asterisk notation is used in tables to visualise statistical significance (ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$).

To assess the reliability of motivation and mental effort questionnaires, we performed a Cronbach's alpha test. Estimated reliability for each questionnaire (motivation Cronbach's $\alpha = 0.77$ and mental effort $\alpha = 0.85$) is acceptable for research purposes [3]. To measure the reliability of retention questionnaires, we conducted a Kuder-Richardson 20 test [31]. The $KR = 0.61 > 0.5$ value indicates that the reliability of the retention questionnaire is also acceptable.

We also conducted a power analysis to check and validate the results and findings of the study. We calculated the effect size (Cohen's d) for each data set collected [11], selected the minimum effect size (Cohen's $d = 0.69$) and estimated the statistical power $(1-\beta)$ of data to check whether the type II error probability (β) is within an acceptable range for a given sample size ($n = 16$ per group) and a significance level ($\alpha = 0.05$). The estimated power value 0.96 shows that with the given sample size, we can have more than a 90% chance that we correctly reject the null hypothesis with a significance level of 0.05.

5 Results

The results and analysis are based on the 32 participants who had completed all the facets of the study, i.e., the motivation questionnaire, the basic training, the vocabulary learning task and the post-questionnaires include mental effort, immediate recall and delayed recall tests. Participants had not undertaken any extra study for the delayed recall test.

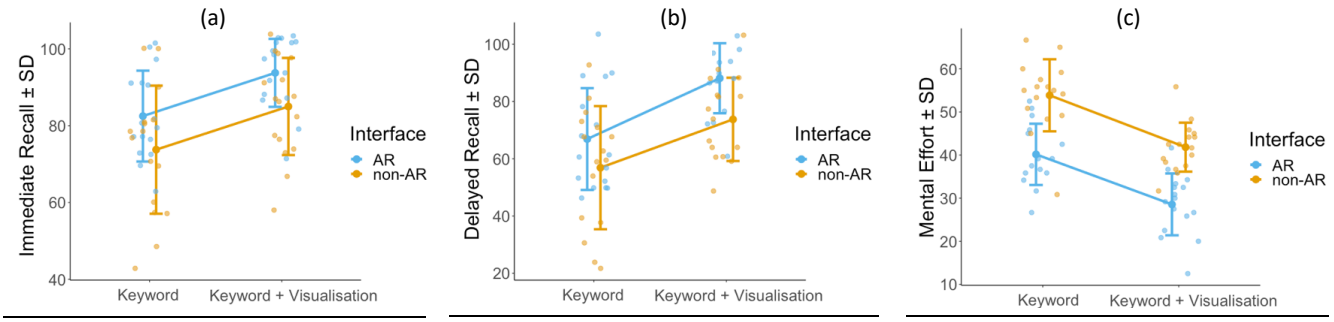
We conducted a statistical analysis including gender as a between-subject factor (2 (GENDER) x 2 (INTERFACE) x 2 (INSTRUCTION MODE) mixed design) in order to exclude possible gender-based differences. The results did not indicate any statistical significant effect of GENDER on any dependant variable: immediate and delayed recall, mental effort, task completion time and, immediate and delayed learning efficiency. The results related to these variables are presented in the following subsections.

5.1 Immediate Recall

The mean values of immediate recall performance and the ANOVA results across study conditions, i.e., the INTERFACE (AR and NON-AR) and, the INSTRUCTION MODE (KEYWORD and KEYWORD+VISUALISATION) are shown in Figure 4a.

A significant main effect of INTERFACE on immediate recall performance could be detected ($F(1,60) = 7.46$, $p < 0.05$, $n^2 p = 0.11$). Here, participants' immediate recall scores were significantly better

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3203116



Immediate Recall

	SS	df	MS	F	p	n^2p
Interface	1225	1	1225	7.46	0.01 **	0.11
Instruction Mode	2025	1	2025	12.34	0.00 ***	0.17
Interface * Instruction Mode	0	1	0	0.00	1.00	0.00

Delayed Recall

	SS	df	MS	F	p	n^2p
Interface	77	1	77	3.86	0.65	0.00
Instruction Mode	5814	1	5814	20.39	0.00 ***	0.25
Interface * Instruction Mode	76	1	76	0.27	0.61	0.00

Mental Effort

	SS	df	MS	F	p	n^2p
Interface	2010	1	2910	57.05	0.00 ***	0.49
Instruction Mode	2236	1	2236	43.83	0.00 ***	0.42
Interface * Instruction Mode	0.72	1	0.72	0.01	0.906	0.00

Fig. 4. Means with standard deviation and ANOVA results for: (a) immediate recall performance in percentage of correctly remembered words; (b) delayed recall performance in percentage of correctly remembered words; (c) Mental effort.

in AR condition ($\bar{x} = 88.13\%$, $SD = 10.34$) compared to the NON-AR condition ($\bar{x} = 79.38\%$, $SD = 14.67$). Also, a significant main effect of INSTRUCTION MODE on immediate recall performance could be detected ($F(1,60) = 12.34$, $p < 0.001$, $n^2p = 0.17$). Results indicated that participants' immediate recall scores in KEYWORD + VISUALISATION ($\bar{x} = 89.38\%$, $SD = 10.75$) were significantly better than in KEYWORD ($\bar{x} = 78.13\%$, $SD = 14.26$). No significant interaction effect could be found between INTERFACE and INSTRUCTION MODE ($F(1,60) < 0.001$, $p > 0.05$, $n^2p < 0.001$).

5.2 Delayed Recall

The mean values of delayed recall performance and the ANOVA results across all study conditions, i.e., the INTERFACE (AR and NON-AR) and, the INSTRUCTION MODE (KEYWORD and KEYWORD + VISUALISATION) are shown in Figure 4b.

No significant main effect was found between INTERFACE on delayed recall performance ($F(1,60) = 3.86$, $p > 0.05$, $n^2p < 0.001$). The significance was only marginally missed. In addition, the mean values for the AR condition ($\bar{x} = 77.50\%$, $SD = 18.50$) were higher than for the NON-AR condition ($\bar{x} = 65.30\%$, $SD = 20.00$).

A significant main effect of INSTRUCTION MODE on delayed recall could be detected ($F(1,60) = 20.39$, $p < 0.001$, $n^2p = 0.25$). Results indicate that participants' delayed recall scores in KEYWORD + VISUALISATION ($\bar{x} = 80.88\%$, $SD = 13.60$) were significantly better than in KEYWORD ($\bar{x} = 61.88\%$, $SD = 19.65$). No significant interaction effect could be found between INTERFACE and INSTRUCTION MODE ($F(1,60) = 0.27$, $p > 0.05$, $n^2p < 0.001$).

5.3 Mental Effort

The mean values of mental effort (measured by NASA-TLX) invested to carry out the learning task and the ANOVA results over the study conditions are illustrated in Figure 4c.

A significant main effect of INTERFACE on mental effort could be detected ($F(1,60) = 57.05$, $p < 0.001$, $n^2p = 0.49$), such that the mental effort was significantly lower for AR condition ($\bar{x} = 34.36$, $SD = 7.14$) compared to the NON-AR condition ($\bar{x} = 47.85$, $SD = 7.02$). Also, a significant main effect of INSTRUCTION MODE on mental effort could be detected ($F(1,60) = 43.83$, $p < 0.001$, $n^2p = 0.42$). Here, participants' mental effort in KEYWORD + VISUALISATION ($\bar{x} = 35.19$, $SD = 6.42$) was significantly lower than in KEYWORD ($\bar{x} = 47.01$, $SD = 7.73$). No significant interaction effects was found between INTERFACE and INSTRUCTION MODE ($F(1,60) = 0.01$, $p > 0.05$, $n^2p < 0.001$).

5.4 Motivation

The mean values of motivation between INTERFACES (AR and NON-AR) is illustrated in Figure 5a. The data summarised in Figure 5a is analysed using a Mann–Whitney U test.

A significant effect was found between INTERFACES for participants' motivation ($U(N_{AR} = 16, N_{non-AR} = 16) = 48.50$, $p < 0.001$). There, the motivation was significantly higher for the AR condition ($\bar{x} = 3.69$, $SD = 0.36$) compared to the NON-AR condition ($\bar{x} = 3.29$, $SD = 0.27$).

5.5 System Usability

The answers to System Usability Scale (SUS) questions/items are reported on a 5-point Likert scale. The SUS scores are calculated as follows: for each of the odd numbered questions, subtract one from the user response, while for each of the even numbered questions, subtract their response from five and, add up the converted responses for each user and multiply that total by 2.5. This converts possible values to the range of 0 to 100 instead of 0 to 40. These adjustments are kept throughout the rest of the analysis.

The average SUS scores for AR and NON-AR INTERFACES are illustrated in Figure 5b. A Mann–Whitney U test indicated that there was no significant effect between the INTERFACES for SUS ($U(N_{AR} = 16, N_{non-AR} = 16) = 91.50$, $p > 0.05$).

5.6 User Experience

The UEQ-s questionnaire provides a benchmark to compare user experience between different systems [26]. It measures pragmatic qualities of a system (including efficiency, perspicuity and dependability) and hedonic qualities (including stimulation and novelty). The overall value was calculated from all 5 UEQ scale means. We adapted the standard method as suggested in [63,64] for calculating the scale means for each factor individually (efficiency, perspicuity, dependability, stimulation and novelty) and to obtain values for pragmatic quality, hedonic quality and overall user experience for both AR and NON-AR systems.

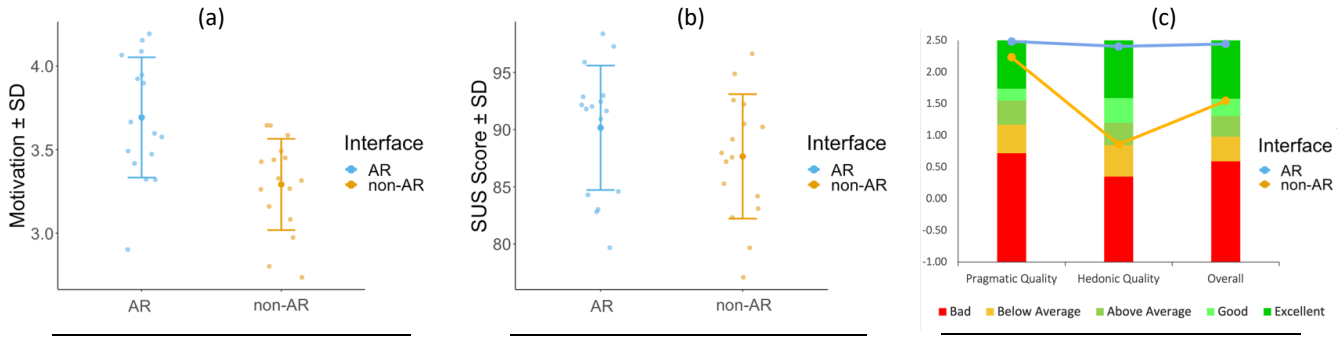
In the AR condition the pragmatic ($\bar{x} = 2.48$) and hedonic ($\bar{x} = 2.41$) qualities, as well as overall user experience ($\bar{x} = 2.45$) were all perceived as excellent (benchmarks for an excellent score: pragmatic > 1.73, hedonic > 1.55, overall > 1.58). In the NON-AR condition the mean value of pragmatic quality was perceived as excellent ($\bar{x} = 2.23$), and the overall experience as good ($\bar{x} = 1.55$) (benchmarks for a good score: pragmatic between 1.55 - 1.73, hedonic between 1.25 - 1.55, overall between 1.4 - 1.58). However, the hedonic factor was perceived as below average ($\bar{x} = 0.88$) (benchmarks for below average score: pragmatic between 0.73 - 1.14, hedonic between 0.88 - 1.24, overall between 0.68 - 1.01).

The overall user experience was analysed using a Mann–Whitney U test. The data is presented in Fig. 5c. A significant effect was found between INTERFACES ($U(N_{AR} = 16, N_{non-AR} = 16) = 16.00$, $p < 0.001$).

5.7 Task Completion Time

The mean values of task completion time across study conditions, i.e., the INTERFACE (AR and NON-AR) and, the INSTRUCTION MODE

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3203116



Motivation	Statistics	p	Sample Size
Interface	48.50	0.00 ***	16

System Usability	Statistics	p	Sample Size
Interface	91.50	0.17	16

User Experience	Statistics	p	Sample Size
Interface	16.00	0.00 ***	16

Fig. 5. Means with standard deviation for: (a) Motivation before starting the experiment and Mann–Whitney U test results; (b) SUS score and Mann–Whitney U test results; (c) UEQ factors (pragmatic and hedonic) and all item/question together (overall) with Mann–Whitney U test results.

(KEYWORD and KEYWORD + VISUALISATION) are shown in Figure 6a. The data summarised in Figure 6a is analysed using a between-within subjects ANOVA on the 20% trimmed means [37].

A significant main effect of INTERFACE on task completion time could be detected ($F(1,60) = 14.06$, $p < 0.001$, $n^2p = 0.19$). Here, the completion time was significantly lower for AR condition($\bar{x} = 618.57s$, $SD = 77.92s$) compared to the NON-AR condition ($\bar{x} = 698.41s$, $SD = 90.59s$). Also, a significant main effect of INSTRUCTION MODE on task completion time could be detected ($F(1,60) = 31.19$, $p < 0.001$, $n^2p = 0.34$), such that KEYWORD + VISUALISATION ($\bar{x} = 599.04s$, $SD = 74.99s$) resulted in a significantly lower completion time than KEYWORD ($\bar{x} = 717.95s$, $SD = 93.52s$). There was no significant interaction effect found between INTERFACE and INSTRUCTION MODE ($F(1,60) = 0.25$, $p > 0.05$, $n^2p < 0.001$).

5.8 Learning Efficiency

The average learning efficiencies for immediate recall and delayed recall across study conditions are shown in Figure 6a and Figure 6b respectively. For the definition of learning efficiency refer to Sect. 4.5. The data summarised in Figure 6a and Figure 6b are analysed using a between-within subjects ANOVA on the 20% trimmed means [37].

Statistical analysis in Figure 6b and Figure 6c showed no significant effect of INTERFACE for participants' learning efficiency for immediate recall ($F(1,60) = 9e-6$, $p > 0.05$, $n^2p < 0.001$) or delayed recall ($F(1,60) = 9e-5$, $p > 0.05$, $n^2p < 0.001$). A significant main effect of INSTRUCTION MODE on participants' learning efficiency for immediate recall could be detected ($F(1,60) = 34.14$, $p < 0.001$, $n^2p = 0.36$). There, the learning efficiency was significantly higher in KEYWORD + VISUALISATION support ($\bar{x} = 0.92$, $SD = 0.23$) compared to the KEYWORD ($\bar{x} = -0.92$, $SD = 0.23$). A significant main effect on participants' learning efficiency for delayed recall also could be detected ($F(1,60) = 41.25$, $p < 0.001$, $n^2p = 0.41$). There, the learning efficiency was significantly higher in KEYWORD + VISUALISATION instruction mode ($\bar{x} = 1.07$, $SD = 1.15$) compared to the KEYWORD ($\bar{x} = -1.07$, $SD = 1.32$) mode. There was no significant interaction effect found between INTERFACE and INSTRUCTION MODE for in immediate recall ($F(1,60) = 0.07$, $p > 0.05$, $n^2p < 0.001$) or delayed recall ($F(1,60) = 0.18$, $p > 0.05$, $n^2p < 0.001$).

6 Discussion and Future Directions

For this study, we developed an AR system called VocabulARy, that supports learning new Japanese words, but can be expanded to support other languages. The system was used to evaluate user experience, system usability, mental effort, motivation and memory recall when shown keywords over the objects vs. keywords together with a visualisation of the objects. These were compared in two interfaces: AR (Microsoft HoloLens 2) and a NON-AR (Android tablet computer). We used the two interfaces to investigate whether showing keywords and visualisations in context of immediate surrounding compared to the context provided on the virtual scene on the screen results in any performance difference.

6.1 Usability and User Experience

The results of the study show that participants evaluated both AR and NON-AR prototypes with good usability scores, clearly higher than average (68) and no significant difference between the two could be found. In addition, during the study we did not observe any readability problems (e.g. none of the users tried to zoom in on the tablet computer in order to make it easier to view presented information and none of the AR HMD users were observed to move very close to augmentations). This provided a good basis for further investigation, as we wanted to make the comparison as fair as possible by trying to not influence learning performance with usability issues as well as by making both conditions as comparable as possible (see Sect. 3.1).

It has been shown for example that the unfamiliarity with AR could result in lower performance as has been reported in prior work [78]. In addition, the user experience in both conditions has been rated very positively with a higher score for AR regarding the hedonic factors represented by Stimulation and Novelty. This makes sense, as AR is still an exciting and less widely used technology compared to tablet computers for many users.

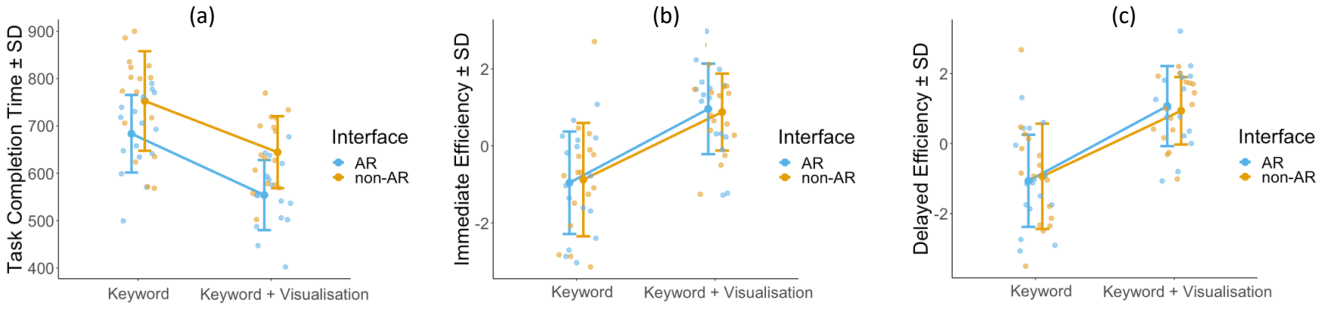
A recent study also revealed that the size of images affects our ability to remember image content during naturalistic exploration [38]. Although our study did not involve naturalistic exploration, it clearly steered participants' attention, and we made extra care that both AR and NON-AR showed comparable imagery, it would still be interesting to explore if the size of imagery has an effect on the ability to memorise vocabulary words.

The prolonged use of HMDs in the current form factor can also influence the usability and thus performance of users. Some studies have already investigated the effects of the HMDs weight, their pressure on the face, latency, image quality and the authenticity of the representation of digital objects [20, 21, 69, 72]. However, these issues will likely be addressed with future development of HMDs.

6.2 AR vs Non-AR

Our results show that immediate recall (a recall of words right after the study) in the AR system is significantly higher compared to the NON-AR system. However, no statistical significance was detected for delayed recall (a recall of the same words a week after the study). Nonetheless, it is important to note that significance was only marginally missed for delayed recall. The results can thus not confirm the outcomes of a previous study conducted by Ibrahim et al. [28] who report a significantly better performance of the AR system compared to FLASHCARDS for both immediate and delayed recall. One of the reasons, and also

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3203116



Task Completion Time	SS	df	MS	F	p	n^2p
Interface	1e+5	1	1e+5	14.06	0.00 ***	0.19
Instruction Mode	2e+5	1	2e+5	31.19	0.00 ***	0.34
Interface * Instruction Mode	1811	1	0	0.25	0.62	0.00

Immediate Efficiency	SS	df	MS	F	p	n^2p
Interface	1e-5	1	1e-5	9e-6	0.99	0.00
Instruction Mode	53.93	1	53.93	34.14	0.00 ***	0.36
Interface * Instruction Mode	0.11	1	0.11	0.07	0.79	0.00

Delayed Efficiency	SS	df	MS	F	p	n^2p
Interface	1e-4	1	1e-4	9e-5	0.99	0.00
Instruction Mode	64.37	1	64.37	41.25	0.00 ***	0.41
Interface * Instruction Mode	0.28	1	0.28	0.18	0.68	0.00

Fig. 6. Means with standard deviation and ANOVA results for: (a) Task-completion-time in seconds; (b-c) immediate recall and delayed recall learning efficiency.

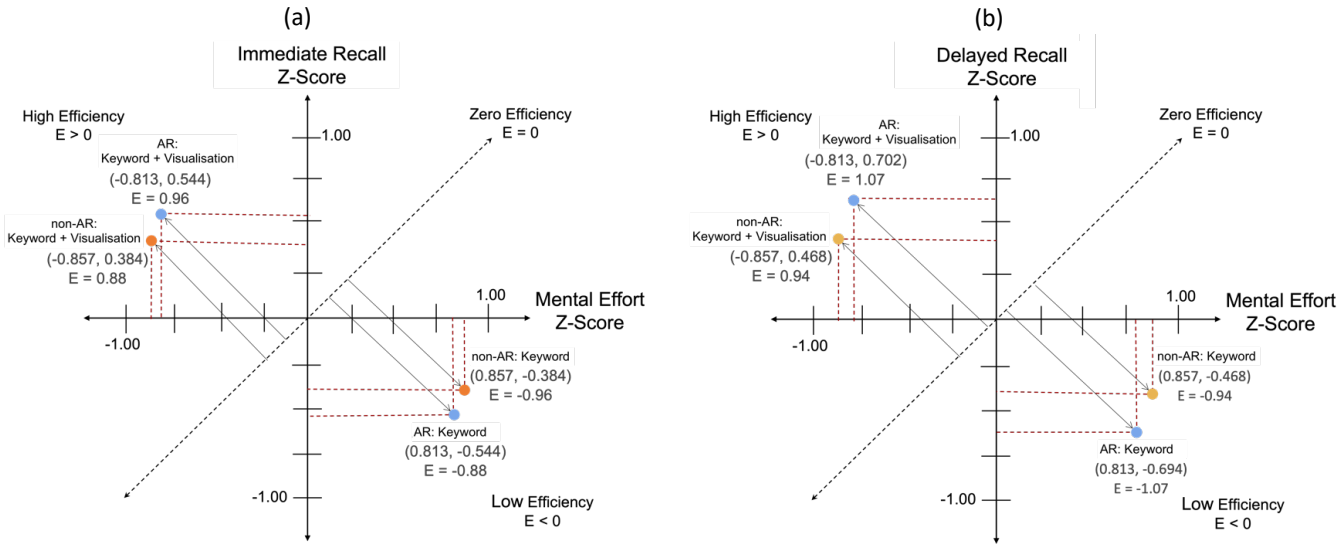


Fig. 7. Learning efficiency: (a) immediate recall learning efficiency; (b) delayed recall learning efficiency.

a major difference between this and our study, might be that in the aforementioned work, the learning methods were not identical in both conditions, which could have placed the AR system at an advantage. For example, with flashcards the word was shown on the opposite side to the image depicting word meaning; thus, the image and word were never shown together. This was not the case in our AR condition where the word annotations were always visible for a selected object in the scene. In comparison, we carefully designed our experiment to minimise any such confounding variable that might influence the results.

Furthermore, participants expressed a significantly higher level of motivation in the AR condition. This is in line with previous work [12, 34] and should be considered when interpreting our results since motivation can be an important factor in learning [8] and technology can play a significant role in this [35]. What causes higher motivation falls out of the scope of this study; however, one could hypothesise that the novelty of the AR plays an important role. The observations show that participants were excited about testing the AR HMD compared to using a tablet. This introduces a need for a longitudinal study of using AR in vocabulary learning, as the influence of motivation might decrease with increasing familiarity with the system.

Interestingly, the AR condition also outperformed the NON-AR condition in task completion time (about 11% faster). This results show that participants were able to learn all words faster in AR condition compared to the NON-AR condition whilst also achieving higher immediate recall scores. This result is also somewhat surprising as AR is at a disadvantage to NON-AR for activating objects. That is, target selection in our setup was typically faster on a tablet computer compared to in mid-air tapping on a HMD. Furthermore, a tablet computer also offered instant access to all buttons at the same distance, whereas users need to physically move to activate some of the buttons in AR. One future direction could involve making users spend the same amount of time in both conditions and explore if this would further improve the performance of AR condition.

6.3 Keyword vs. Keyword + Visualisation

Regarding our second independent variable INSTRUCTION MODE, our results clearly show that vocabulary learning can be improved beyond the traditional keyword method, by augmenting the keywords with animated 3D visualisations. We found overwhelming evidence for this in all metrics, such as immediate and delayed recall, learning efficiency, mental effort, and even task completion time.

This is in line with observations by Shapiro and Waters [67], who reported that the level of visual imagery of a word enhances vocabulary learning. In this work, we go beyond simple imagery and show that the 3D animated content is potentially an even more effective approach. This opens up another direction for future work involving the necessity for detailed comparison of the effect of different visualisation techniques.

As mentioned, our results showed that providing visualisations for

keywords reduced the mental effort for vocabulary learning. However, reducing mental effort in learning scenarios can also result in reduced learning outcomes. For example, Salmon showed that the amount of invested mental effort positively correlates with learning efficiency [61]. Knowing this, we could expect a decline in performance of immediate and delayed recall. One reason why this did not happen in our case might be the fact that enough effort was needed in order to complete the task (moving, tapping, remembering). One way to increase the mental effort would be to require users to come up with their own associations for keywords instead of providing predefined keywords as in our study. Providing predefined keywords might not be in line with user's mental model, thus making it difficult for the user to (mentally) visualise them. This could have made visualisations in our study more important. However, previous research suggests that users might have difficulties coming up with their own keywords and that predefined keywords lead to better learning outcomes [4]. Despite, future studies should explore if the difference persists also when personalised keywords are used in learning scenario presented in this study.

6.4 Implications and Design Recommendations

The benefits of the keyword method over other learning techniques are well known [30, 51]. We have shown that the keyword method can provide even better results by adding animated visualisations that depict the keyword itself. This is an important implication for designing such applications for vocabulary learning. However, this also opens up several questions. For example, do animations of visualisations of keywords contribute to the learning outcome or would visualisation without an animation result in comparable efficiency?

One of the most important things to consider in designing such a system is the keywords or words from the language a learner speaks that sound similar to the word being learnt. In our prototype, we only used a limited set of vocabulary for which we were able to find appropriate keywords and accompanying visualisations. Finding these keywords and visualisations takes time, which needs to be considered when thinking about applying this method in practice. And it is no necessary that every word would have an appropriate keyword. Crowd-sourcing could be one approach to tackle this problem. Additionally, approaches for automating the process of finding keywords already exist [2]. Also, our future direction will involve investigating the effect of asking users to choose their own keywords and visualisations. A system that would use a combination of these approaches could probably satisfy a variety of learning types.

Another thing to keep in mind, and we are not aware of any study investigating it, is the fact that the vast number of keywords might be overwhelming for users. One of the unanswered question is thus how many keywords is recommended to provide at one time (in our study only one was shown at the time to direct users).

For the purpose of this study we used marker based tracking to initialise the settings in AR. As such, our system was linked to a particular physical space. To enable wider adoption, another space-independent object recognition technique should be used as discussed in Sect. 3.2. Such a system would also need to have a database of objects with the corresponding keywords available upfront so when users look at a physical scene AR visualisations would be fetched on the fly.

Only two participants used all the available time to learn and all 10 words were correctly remembered in 23% of immediate and 4% of delayed recall tests. For immediate recall we might have reached the ceiling effect, and making the task more difficult would highlight even greater changes between the test conditions. This was even more obvious for delayed recall, where only very few users finished the test with no errors. This could be taken in consideration when building such a system – during the testing phase the system should try to increase the level of engagement and encourage users to take more time, while and after testing the system should encourage users to rethink about wrong answers.

6.5 Limitations

As explained in Sect. 5 gender did not have a significant effect on the results of the study. However, future work should look into a possible gender bias in more detail with a higher number of participants, as our result on this is not conclusive.

Another thing to consider in our study is age bias. However, the age group studied is highly mobile, spending extended periods of time in foreign countries (for example, the EU Erasmus+ programme alone funds more than half a million exchanges yearly [16]). As such, this group could benefit from an improved vocabulary learning system. Nevertheless, the results cannot be generalised over the whole population, and expanding the study to other age groups and exploring the effect of age on the proposed learning system is an important future direction.

Further, we only used nouns in our prototype. More specifically, all nouns were associated with objects. In fact, a number of studies have shown that concrete terms (e.g., nouns such as bread) are better remembered than abstract terms (e.g., abstract nouns and verbs) [70]. The benefits of in-situ learning with AR will therefore be reduced when abstract terms are considered as it becomes difficult to make them relevant to context of users' immediate environment. Nevertheless, future studies could focus on exploring the potential of 3D AR animation to make abstract terms visually more accessible.

As mentioned in the paper, our prototype was only tested for a short time on a limited vocabulary. To further validate our findings, the vocabulary should be expanded and tested over a longer period of time. Especially the effect of higher motivation in AR could wear off as users become more familiar with the system. Additionally, we only measured the immediate recall (immediately after participants had completed the task) and a delayed recall (a week after participants had completed the task) of the vocabulary learned. Future work should also consider recall after longer periods of several weeks. This could also be combined with repeating the learning phase in certain intervals, as it is normally done when learning vocabulary.

7 Conclusion

Learning vocabulary can be enhanced when encountering words in context. This context can be afforded by the place or activity people are engaged with. For this purpose we developed VocabulARy, a HMD AR system that visually annotates objects in the user's surroundings, with the corresponding English (first language) and Japanese (second language) words to enhance the language learning process. In addition to the written and audio description of each word, we also present the user with a keyword and its animated 3D visualisation to enhance memory retention.

We evaluated our prototype by comparing it to an alternate AR system that does not show any additional visualisation of the keyword, and also, we compare it to two non-AR systems on a tablet, one with and one without visualising the keyword. Our results indicate that AR outperforms the NON-AR system regarding short-term retention, mental effort and task-completion time. Additionally, the visualisation approach scored significantly higher than only showing the written keyword with respect to immediate and delayed recall, learning efficiency, mental effort and task-completion time. Visualisation of keywords thus proved more efficient compared with the traditional keyword method only and opens new avenues for future improvements in AR enabled vocabulary learning systems.

Acknowledgments

The authors wish to thank Cuauthli Campos for helping with preparing the video and all the volunteers who participated in the user study.

This research was supported by European Commission through the InnoRenew CoE project (Grant Agreement 739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (investment funding of the Republic of Slovenia and the European Union of the European Regional Development Fund). We also acknowledge support from the Slovenian research agency ARRS (program no. BI-DE/20-21-002, P1-0383, J1-9186, J1-1715, J5-1796, and J1-1692).

REFERENCES

[1] M. Amiryousefi and S. Ketabi. Mnemonic instruction: A way to boost vocabulary learning and recall. *Journal of Language Teaching & Research*, 2(1), 2011.

[2] O. Anonthanasap, C. He, K. Takashima, T. Leelanupab, and Y. Kitamura. Mnemonic-based interactive interface for second-language vocabulary learning. *Proceedings of the Human Interface Society, HIS*, 14, 2014.

[3] D. Ary, L. C. Jacobs, C. K. S. Irvine, and D. Walker. *Introduction to research in education*. Cengage Learning, 2018.

[4] R. C. Atkinson. Mnemotechnics in second-language learning. *American psychologist*, 30(8):821, 1975.

[5] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[6] H. D. Brown and S. T. Gonzo. *Readings on second language acquisition*. Allyn & Bacon, 1995.

[7] R. O. Castle and D. W. Murray. Object recognition and localization while tracking and mapping. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pp. 179–180, 2009. doi: 10.1109/ISMAR.2009.5336477

[8] Y.-C. Chen and P.-C. Chen. The effect of english popular songs on learning motivation and learning performance. *WHAMPOA-An Interdisciplinary Journal*, 56:13–28, 2009.

[9] R. C. Clark, F. Nguyen, and J. Sweller. *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. John Wiley & Sons, 2011.

[10] A. D. Cohen and E. Aphek. Retention of second-language vocabulary over-time: Investigating the role of mnemonic associations. *System*, 8(3):221–235, 1980.

[11] J. Cohen. Statistical power analysis for the behavioral sciences. *England: Routledge*, 1988.

[12] C. S. C. Dalim, M. S. Sunar, A. Dey, and M. Billinghurst. Using augmented reality with speech input for non-native children's language learning. *International Journal of Human-Computer Studies*, 134:44–64, 2020.

[13] F. Draxler, A. Labrie, A. Schmidt, and L. L. Chuang. 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*, pp. 1–12, 2020.

[14] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58, 2013.

[15] D. Edge, E. Searle, K. Chiu, J. Zhao, and J. A. Landay. Micromandarin: mobile language learning in context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 3169–3178, 2011.

[16] Erasmus+. Factsheets and statistics on Erasmus+. `https://erasmus-plus.ec.europa.eu/node/2585`. 2022-05-24.

[17] P. A. Freund, J.-T. Kuhn, and H. Holling. Measuring current achievement motivation with the qcm: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, 51(5):629–634, 2011.

[18] Y. Fujimoto, G. Yamamoto, H. Kato, and J. Miyazaki. Relation between location of information displayed by augmented reality and user's memorization. In *Proceedings of the 3rd Augmented Human International Conference*, pp. 1–8, 2012.

[19] T. Georgiev, E. Georgieva, and A. Smrikarov. M-learning-a new stage of e-learning. In *International conference on computer systems and technologies-CompSysTech*, vol. 4, pp. 1–4, 200.

[20] J. Guo, D. Weng, Z. Zhang, H. Jiang, Y. Liu, Y. Wang, and H. B.-L. Duh. Mixed reality office system based on maslow's hierarchy of needs: Towards the long-term immersion in virtual environments. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 224–235. IEEE, 2019.

[21] J. Guo, D. Weng, Z. Zhang, Y. Liu, and Y. Wang. Evaluation of maslows hierarchy of needs on long-term use of hmds–a case study of office environment. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 948–949. IEEE, 2019.

[22] A. K. Halabi. Applying an instructional learning efficiency model to determine the most efficient feedback for teaching introductory accounting. *Global Perspectives on Accounting Education*, 3(1):6, 2006.

[23] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[24] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[25] A. Hautasaari, T. Hamada, K. Ishiyama, and S. Fukushima. Vocabura: A method for supporting second language vocabulary learning while walking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–23, 2019.

[26] A. Hinderks, M. Schrepp, and J. Thomaschewski. A benchmark for the short version of the user experience questionnaire. In *WEBIST*, pp. 373–377, 2018.

[27] M. B. Ibáñez, Á. Di Serio, D. Villarán, and C. D. Kloos. Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education*, 71:1–13, 2014.

[28] A. Ibrahim, B. Huynh, J. Downey, T. Höllerer, D. Chun, and J. O'donovan. Arbis pictus: A study of vocabulary learning with augmented reality. *IEEE transactions on visualization and computer graphics*, 24(11):2867–2874, 2018.

[29] B. Khoshnevisan and S. Park. Affordances and pedagogical implications of augmented reality (ar)-integrated language learning. In *Designing, Deploying, and Evaluating Virtual and Augmented Reality in Education*, pp. 242–261. IGI Global, 2021.

[30] M. E. King-Sears, C. D. Mercer, and P. T. Sindelar. Toward independence with keyword mnemonics: A strategy for science vocabulary instruction. *Remedial and Special Education*, 13(5):22–33, 1992.

[31] G. F. Kuder and M. W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937.

[32] S. Kumar Basak, M. Wotto, and P. Belanger. E-learning, m-learning and d-learning: Conceptual definition and comparative analysis. *E-learning and Digital Media*, 15(4):191–216, 2018.

[33] J. R. Lewis and J. Sauro. The factor structure of the system usability scale. In *International conference on human centered design*, pp. 94–103. Springer, 2009.

[34] S. Li, Y. Chen, D. M. Whittinghill, and M. Vorvoreanu. A pilot study exploring augmented reality to increase motivation of chinese college students learning english. In *2014 ASEE Annual Conference & Exposition*, pp. 24–85, 2014.

[35] M.-H. Lin, H.-g. Chen, et al. A study of the effects of digital learning on learning motivation and learning outcome. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(7):3553–3564, 2017.

[36] S. E. C. LTD. Samsung galaxy - the official samsung galaxy site. `https://developer.vuforia.com/`. Accessed: 2022-07-27.

[37] P. Mair and R. Wilcox. Robust statistical methods in r using the wrs2 package. *Behavior research methods*, 52(2):464–488, 2020.

[38] S. Masarwa, O. Kreichman, and S. Gilaie-Dotan. Larger images are better remembered during naturalistic encoding. *Proceedings of the National Academy of Sciences*, 119(4):e2119614119, 2022. doi: 10.1073/pnas.2119614119

[39] M. A. Mastropieri and T. E. Scruggs. *Teaching students ways to remember: Strategies for learning mnemonically.* Brookline Books, 1991.

[40] T. Mayes and S. De Freitas. Learning and e-learning. *Rethinking pedagogy for a digital age*, pp. 13–25, 2007.

[41] C. McLoughlin and M. J. Lee. Personalised and self regulated learning in the web 2.0 era: International exemplars of innovative pedagogy using social software. *Australasian Journal of Educational Technology*, 26(1), 2010.

[42] Microsoft. Microsoft hololens — mixed reality technology for business. `https://www.microsoft.com/en-us/hololens`. Accessed: 2022-07-27.

[43] NASA. Nasa tlx: Task load index, 2006.

[44] P. Nicholson. *A History of E-Learning*, pp. 1–11. Springer Netherlands, Dordrecht, 2007. doi: 10.1007/978-1-4020-4914-9_1

[45] D. Nunan. *Second Language Teaching & Learning.* ERIC, 1999.

[46] F. G. Paas and J. J. Van Merriënboer. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human factors*, 35(4):737–743, 1993.

[47] A. Paivio and A. Desrochers. Mnemonic techniques in second-language learning. *Journal of Educational Psychology*, 73(6):780, 1981.

[48] I. Pearlman. Effectiveness of keyword versus direct instruction on vocabulary acquisition by primary-grade handicapped learners. *Bulletin of the Psychonomic Society*, 28(1):14–16, 1990.

[49] P. R. Pintrich. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of educational Psychology*, 95(4):667, 2003.

[50] M. Pressley, J. R. Levin, and H. D. Delaney. The mnemonic keyword

method. *Review of Educational Research*, 52(1):61–91, 1982.

[51] M. Pressley, J. R. Levin, N. A. Kuiper, S. L. Bryant, and S. Michener. Mnemonic versus nonmnemonic vocabulary-learning strategies: Additional comparisons. *Journal of Educational Psychology*, 74(5):693, 1982.

[52] PTC. Vuforia developer portal. `https://developer.vuforia.com/`. Accessed: 2022-07-27.

[53] A. L. Putnam. Mnemonics in education: Current research and applications. *Translational Issues in Psychological Science*, 1(2):130, 2015.

[54] M. R. Raugh and R. C. Atkinson. A mnemonic method for learning a second-language vocabulary. *Journal of Educational Psychology*, 67(1):1, 1975.

[55] M. R. Raugh, R. D. Schupbach, and R. C. Atkinson. Teaching a large russian language vocabulary by the mnemonic keyword method. *Instructional science*, 6(3):199–221, 1977.

[56] F. Rheinberg, R. Vollmeyer, and B. D. Burns. Fam: Ein fragebogen zur erfassung aktueller motivation in lern-und leistungssituationen (langversion, 2001). *Diagnostica*, 2:57–66, 2001.

[57] F. Rheinberg, R. Vollmeyer, and B. D. Burns. Qcm: A questionnaire to assess current motivation in learning situations. *Diagnostica*, 47(2):57–66, 2001.

[58] N. Sagarra and M. Alba. The key is in the keyword: L2 vocabulary learning methods with beginning learners of spanish. *The modern language journal*, 90(2):228–243, 2006.

[59] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1352–1359, 2013. doi: 10.1109/CVPR.2013.178

[60] M. Salman and M. J. Pearson. Whisker-ratslam applied to 6d object identification and spatial localisation. In V. Vouloutsi, J. Halloy, A. Mura, M. Mangan, N. Lepora, T. J. Prescott, and P. F. Verschure, eds., *Biomimetic and Biohybrid Systems*, pp. 403–414. Springer International Publishing, Cham, 2018.

[61] G. Salomon. Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of educational psychology*, 76(4):647, 1984.

[62] M. E. C. Santos, T. Taketomi, G. Yamamoto, M. M. T. Rodrigo, C. Sandor, H. Kato, et al. Augmented reality as multimedia: the case for situated vocabulary learning. *Research and Practice in Technology Enhanced Learning*, 11(1):1–23, 2016.

[63] M. Schrepp. Ueq-user experience questionnaire, 2019.

[64] M. Schrepp, A. Hinderks, and J. Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *IJIMAI*, 4(6):103–108, 2017.

[65] B. B. Schultz. Levene's test for relative variation. *Systematic Zoology*, 34(4):449–456, 1985.

[66] H. Schuman, S. Presser, and J. Ludwig. Context effects on survey responses to questions about abortion. *Public Opinion Quarterly*, 45(2):216–223, 1981.

[67] A. M. Shapiro and D. L. Waters. An investigation of the cognitive processes underlying the keyword method of foreign vocabulary learning. *Language teaching research*, 9(2):129–146, 2005.

[68] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[69] R. Shen, D. Weng, S. Chen, J. Guo, and H. Fang. Mental fatigue of long-term office tasks in virtual environment. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 124–127. IEEE, 2019.

[70] R. N. Shepard. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, 6(1):156–163, 1967.

[71] M. Spitzer. M-Learning? When it comes to learning, smartphones are a liability, not an asset. *Trends in Neuroscience and Education*, 4(4):87–91, 2015. doi: 10.1016/j.tine.2015.11.004

[72] F. Steinicke and G. Bruder. A self-experimentation report about long-term use of fully-immersive technology. In *Proceedings of the 2nd ACM symposium on Spatial user interaction*, pp. 66–69, 2014.

[73] M. P. Strzys, S. Kapp, M. Thees, P. Klein, P. Lukowicz, P. Knierim, A. Schmidt, and J. Kuhn. Physics holo. lab learning experience: using smartglasses for augmented reality labwork to foster the concepts of heat conduction. *European Journal of Physics*, 39(3):035703, 2018.

[74] G. M. Sullivan and A. R. Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013.

[75] U. Technologies. Unity real-time development platform. `https://unity.com/`. Accessed: 2022-07-27.

[76] C. D. Vazquez, A. A. Nyati, A. Luh, M. Fu, T. Aikawa, and P. Maes. Serendipitous language learning in mixed reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2172–2179, 2017.

[77] Z. Wei. Does teaching mnemonics for vocabulary learning make a difference? putting the keyword method and the word part technique to the test. *Language Teaching Research*, 19(1):43–69, 2015.

[78] M. Wille, L. Adolph, B. Grauel, S. Wischniewski, S. Theis, and T. Alexander. Prolonged work with head mounted displays. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers: Adjunct Program*, pp. 221–224, 2014.

[79] H.-K. Wu, S. W.-Y. Lee, H.-Y. Chang, and J.-C. Liang. Current status, opportunities and challenges of augmented reality in education. *Computers & education*, 62:41–49, 2013.

[80] S. Yang and B. Mei. Understanding learners' use of augmented reality in language learning: insights from a case study. *Journal of Education for Teaching*, 44(4):511–513, 2018.