ARbis Pictus: A Study of Language Learning with Augmented Reality

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

This paper describes "ARbis Pictus" –a novel system for immersive language learning through dynamic labeling of real-world objects in augmented reality. We describe a within-subjects lab-based study (N=52) that explores the effect of our system on participants learning nouns in an unfamiliar foreign language, compared to a traditional flashcard-based approach. Our results show that the immersive experience of learning with virtual labels on real-world objects is both more effective and more enjoyable for the majority of participants, compared to flashcards. Specifically, when participants learned through augmented reality, they scored significantly better by 7% (p=0.011) on productive recall tests performed same-day, and significantly better by 21% (p=0.001) on 4-day delayed productive recall post tests than when they learned using the flashcard method. We believe this result is an indication of the strong potential for language learning in augmented reality, particularly because of the improvement shown in sustained recall compared to the traditional approach.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Language Learning, Education, Augmented Reality, HCI, Experimentation

INTRODUCTION

This paper addresses the problem of facilitating and understanding the process of language learning in immersive, augmented reality (AR) environments. Recent heavy investment in AR technology by industry leaders such as Google, Microsoft, Facebook and Apple is an indication that both device technology and content for this modality will improve rapidly over the coming years. Looking forward, we believe that AR can have significant impact on the way we learn foreign or technical languages, processes and workflows, for example, by creating new personalized learning opportunities in a

physical space that is modeled, processed and labeled by automated machine learning (ML) classifiers, assisted by human users. These augmented learning environments can include annotations on real objects, placement of virtual objects, or interactions between either type to describe complex processes. AR devices will eventually become affordable and portable enough to be commonly used in day-to-day tasks. In this setting, learning can occur passively as people interact with objects and processes in their environments that are annotated to support personalized learning objectives.

To study the impact of AR on language learning, we are in the process of developing *ARbis Pictus* –an emerging interactive learning platform that supports personalized, dynamic labeling of objects in a real world environment using AR modalities. It was named after Orbis Sensualium Pictus (Visible World in Pictures), one of the first widely used children's picture books, written by Comenius and first published in 1658. The concept includes a server-side deep neural network that communicates in real-time with an AR device (such as a tablet-based AR magic lens, or a HoloLens). Image data is streamed to the server, which returns object labels and bounding boxes that are used to annotate items that the end user sees. We have implemented an early prototype version of this system [1] with Microsoft's HoloLens[1] and YOLO [22], a deep-learning based object recognizer, that is capable of labeling objects in a scene with reasonable accuracy for a proof-of-concept technology demonstration in real time. A personalized learning module would later set a learning goal for the participant based on manually provided data, or data from a linked course information system such as Moodle or SIS, and objects in their world view will be labeled according to the educational goal. The system is targeted towards basic learning of noun terms in a foreign language, with the long term goal of facilitating more complex learning tasks such as scientific language or workflows. In the latter case, real or virtual objects can interact to achieve some educational goal, such as learning an experiment with scientific equipment, or preparing a recipe in a kitchen, similar to [25]. Figure 1 shows an example labeled scene from a manually populated system (no ML component) where objects are labeled in the target language. This is a captured view from the device stream. The actual view that the learner sees through the Microsoft HoloLens device we used will have a much more restricted field of view, as indicated by the red box in Figure 3. In this paper, we describe a user study to evaluate the impact of learning simple noun terms in a foreign

---

[1]https://www.microsoft.com/en-us/hololens

language with augmented reality labeling using the system. We do not here discuss the neural network or personalization components of the ARbis Pictus system. The paper focuses on the following three research questions, which we view as important to understand before conducting further studies with adaptive and therefore more complex system components:

- RQ 1: When learning vocabulary [or individual lexical items] in an unknown second language, is there a difference in learner performance in a flashcard-based multimodal environment as compared to an AR environment?

- RQ 2: In the above setting, how does productive and receptive recall vary after some time has passed?

- RQ 3: How do users *perceive* the language learning experience in Augmented Reality compared to traditional flashcards?

In the process of answering these research questions, we make the following key contributions:

- Design and implementation of a system that supports foreign language learning with augmented reality and with traditional flashcards.

- Design and implementation of a user experiment to evaluate the impact of AR-based learning for second language acquisition.

  - Statistically significant results that show better recall (7%) for AR learning compared to traditional flashcards.

  - Statistically significant results that show an increased advantage (21%) for AR in productive recall four days after the initial test, compared to traditional flashcards.

  - Analysis of interaction data (Clicks, Eye-tracking (flashcards) and head tracking (AR)) that reveal learning patterns in each modality.

  - Qualitative survey and interview data that show participants believe that AR is effective and enjoyable for language learning.

BACKGROUND

Multimedia Learning

Our framework is motivated by Mayer et al.'s cognitive theory of multimedia learning (CTML) [14][13][15], one of the most compelling learning theories in the field of Educational Technology. The theory posits, first, that there are two separate channels (auditory and visual) for processing information, second, that learners have limited cognitive resources, and third, that learning entails active cognitive processes of filtering, selecting, organizing, and integrating information. Certain basic principles comprise the theory, and these principles address the optimal ways in which multimodal information, e.g., text, images, audio, and video, can and should be presented to learners to ensure retention, and, more importantly, to ensure transferability to new learning situations.

The CTML predicts, based on extensive empirical evidence, that people learn better from a combination of words and pictures than from words alone (the Multimedia Principle)[16].
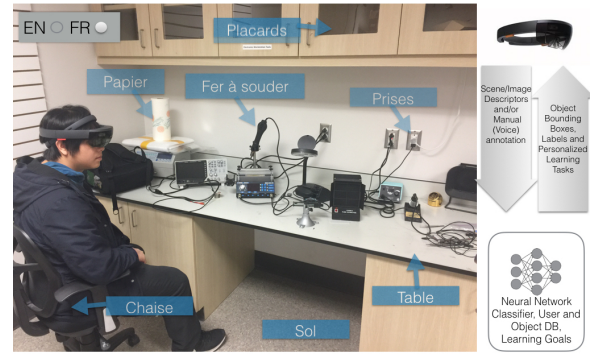

Figure 1. Mixed reality screen shot of a language learner using the ARbis Pictus system. Note that the user will only see annotations in a 30 degrees field of view.

In the field of Second Language Acquisition (SLA), studies using the CTML as their theoretical basis have shown that when unknown vocabulary words are annotated with both text (translations) and pictures (still images or videos), they are learned and retained better in post tests than words annotated with text alone [3][21][30]. A second principle of the CTML is that people learn better when corresponding words and pictures are presented near rather than far from each other on the page or screen (the Spatial Contiguity Principle), as the easy integration of verbal and visual information causes less cognitive load on working memory, thereby facilitating learning [19]. SLA research has found that simultaneous display of multimedia information leads to better performance on vocabulary tests than an interactive display [28]. A recent study by Culbertson et al. in [6] describes an online 3D game to teach people Japanese. Their approach used situated learning theory, and they found excellent feedback on engagement. Specifically, people were learning 8 words in 40 min on average. Experts who already knew some Japanese were the most engaged with the system. Learning results from that study informed the design and complexity of the learning tasks in our experiment. The broader vision for our ARbis Pictus system, including personalized learning and real-time object recognition was influenced by work by Cai et al. in [2], which found that we can leverage the small waiting times in everyday life to teach people a foreign language, e.g. while chatting with a friend electronically. Cai et al. implement an IM messenger that detects some of the words in the conversation and prompts the user about them based on some a-priori knowledge of the users learning goals and objectives.

Virtual and Augmented Reality in Education

The use of Augmented Reality for second language learning is in its infancy [24][9], and there are only a small number of studies that link AR and second language learning. For example, in [12], Liu et al. describe an augmented reality game that allows learners to collaborate in English language learning tasks. They find that the AR approach increases engagement in the learning process. In contrast, our experiment is an evaluation of the effects of immersive AR on lexical learning, using simple noun terms only, analogous more to a traditional flashcard-based learning method. Flashcards are a well-known

tool for language learning, and their benefits and shortcomings are documented in the second language learning literature [20]. In this study, we employ this method as a simple benchmark, purposely chosen to minimize effects of user interactions, and to expose the impact of immersion in AR on a set of performance metrics during a vocabulary learning task.

AR has been used in classrooms in a variety of situations, including support of language learning. For example, AR textbooks have been studied by Grasset [10] and Scrivner et al. [24]. The latter describes an ongoing project for testing AR textbooks in the classroom for undergraduate Spanish learners. Their approach differs from our experiment in that we use minimal virtual objects (labels only), but incorporate physical objects in the real world as a pedagogical aid, including their spatial positioning in the augmented scene. Godwin [9] provides a review of AR in education, focusing on popular games such as Pokemon Go! and on general AR devices and techniques marker-based tracking. However, there is no discussion of formal evaluation of AR for second language learning, although the LearnAR website linked in the study does have module listings for English, French and Spanish. Going beyond simple learning of lexical terms, the European Digital Kitchen project [25] incorporates process-based learning with AR to support language learning. They apply a marker-based tracking solution to place item labels in the environment to help users prepare recipes, including actions such as stirring, chopping or dicing, for example. Dunleavy et al. [8] discuss AR and situated learning theory. They claim that immersion helps in the learning process, but also warn about the dangers of increased cognitive overload that comes with AR use. In our experimental design, we consider this advice and allow ample time for familiarization with the AR device to reduce both cognitive overload resulting from the unfamiliar modality, and other novelty effects.

Interactive Applications for Learning with AR

There have been several interactive games involving AR for learning in a variety of situations. Costabile [5] discuss an AR application for teaching history and archaeology. Like [8], they hypothesized that engagement would be increased with AR compared to more traditional displays. However, the results found that a traditional paper method was both faster and more accurate than AR for the learning task. Another benefit of AR is that it brings an element of gamification to the learning task, making it particularly suitable for children to learn with. A notable example of this is Yannier et al.'s study [29] on the pedagogy of basic physical concepts such as balance, using blocks. In their study, AR outperformed benchmarks by about a 5-fold increase, and was reported as far more enjoyable. A similar, but much earlier approach that applied AR to collaboration and learning was Kaufman's work [11] on teaching geometry to high-school level kids. An updated version of this system was applied to mobile AR devices by Schmalstieg et al. in [23]. Now that we have described relevant related work that has informed our experimental design and setup, we can proceed with details of our designs. This will be followed with a discussion of results.

EXPERIMENTAL DESIGN

52 participants (33 females, 19 males, mean age of 21, SD of 3.8) took part in a 2 by 2 counterbalanced within-subject study. 30 Basque words were divided into two groups of 15, called A and B, further divided into fixed subgroups of 5 referred to as A1, A2, A3 and B1, B2 and B3. Each subject saw one of the two word groups on one of the devices, and the other group on the other device. In total, 13 people saw the word group A in AR first, 13 word group B in AR first, 13 word group A with the flashcards first, and 13 word group B with the flashcards first, as described in Table 1.

After answering a few background questions, the participants were told what the objects were in English and started using one of the devices after the instructors informed them about the learning tasks and the specifics of the tests. On the AR device, the participants first undertook a training task where they could take as much time as they wanted to setup the device, get used to the controls and reduce the novelty aspect of it while interacting with virtual objects. Before using the flashcards, an eye-tracker was calibrated for each participant. Then, the participants moved on to the learning task, which consisted in 3 learning phases and 4 tests (3 receptive, 1 productive) per device. In the first learning phase, the participants had 90 seconds to learn the 5 words of the first subgroup of one of the word groups on a given device. After a distraction task, they took a receptive test. Afterwards, they undertook have a second learning phase on the same device, and had 90 seconds to learn the 5 new words from the second subgroup of the same word group, along with the 5 previous words. Following a distraction task, they took a receptive test on the 5 new words. They then had a third learning phase on the same device during which they saw for 90 seconds the 5 words of the last subgroup of the selected word group, alongside the 10 previous ones. After a distraction and a receptive test on the 5 words from the last subgroup, they took a productive test on all 15 words from the word group chosen. They then had another, similar set of 3 learning phases and 4 tests on the other device using the other word group, as illustrated in Table 1. The AR learning task, flashcard learning task and tests took place in 3 different rooms to avoid potential biases.

At the conclusion of the learning task, the participants answered a questionnaire on how efficient and engaging they perceived each device. A short interview allowed us to gather more feedback on their preferences. Four days after the learning phases, the participants were asked to take again the same 8 tests they took the day of the study to assess long-term recall. 32 users accepted to take the tests.

Every participant was compensated $10, and the study lasted a total of 40 to 65 minutes for every user (with most of the variance due to the AR training phase's flexible length).

More details about the tasks and tests are given in the following sections.

EXPERIMENTAL SETUP

Flashcards

The flashcard modality was designed as a web application emulating traditional physical flashcards, running on a desktop

Order	Device used | word subgroup(s) seen during each learning phase					
I	AR - A1	AR - A1, A2	AR - A1, A2, A3	FC - B1	FC - B1, B2	FC - B1, B2, B3
II	AR - B1	AR - B1, B2	AR - B1, B2, B3	FC - A1	FC - A1, A2	FC - A1, A2, A3
III	FC - A1	FC - A1, A2	FC - A1, A2, A3	AR - B1	AR - B1, B2	AR - B1, B2, B3
IV	FC - B1	FC - B1, B2	FC - B1, B2, B3	AR - A1	AR - A1, A2	AR - A1, A2, A3

Table 1. Table of conditions and balancing across the six learning phases. AR shows the augmented reality conditions and FC represents flashcards. A and B are distinct term groups for the within-subject design, and the group number indicates one of the subgroups of 5 words.
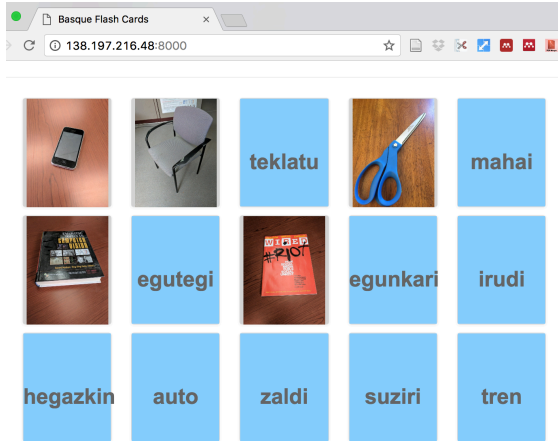

Figure 2. Screen shot of the web-based flashcard application that was used in the study.

Figure 3. Example of the Basque labels shown in the AR condition of the experiment, with the AR field of view in red.

computer that the user interacted with using a mouse. After entering a user ID and one of the combinations of word subgroups seen in Table 1, the instructor let the participants interact with 1, 2 or 3 rows of 5 flashcards, all visible on a single page, with each flashcard consisting in a word in the foreign language on the back and an image of the corresponding object on the front. The images used were pictures of the real objects used in the AR condition. A recording of the word being pronounced was automatically played through speakers every time the user clicked on the back of a flashcard. The same recording of the Basque word being spoken by a human (male) was used in both modalities. Clicks were logged during every phase to track possible learning strategies. Additionally, an eye-tracker was calibrated before the learning task with the flashcards to track the participants' gaze during the learning phases.

Augmented Reality
The augmented reality modality made use of a Microsoft HoloLens, an augmented reality head-mounted display. The application was set up in a room containing all of the objects from the two word groups, and allowed the participants to see labels annotating the objects from the chosen subgroups with the relevant words in the foreign language. The device's real-time mapping of the room let the users walk around the room while keeping the labels in place, and save the location of the labels throughout the study, between users and after restarting the device. As a precaution, before every learning phase on the HoloLens, the administrators of the study verified that the labels were in place, and after handing over the device to the participants, that they were able to see every label. The
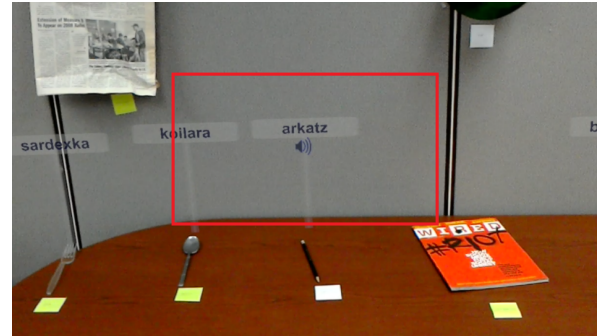
app had two modes: "admin mode", allowing the instructor to place labels with voice commands or gestures, select which words subgroups to display or enter a user ID; and a "user" mode which restricted these functionalities but allowed the participants to interact with labels during the learning task. On the HoloLens, the cursor's position is natively determined by the user's head orientation; in the app, moving the blue circle used as a cursor close to a label would turn the cursor into a speaker, signalling to the user the possibility to click to hear a recording of the word being pronounced through the device's embedded speakers. Each label had an extended, invisible hitbox to allow the users to click the labels more comfortably. Moreover, the labels' and hitboxes' sizes, along with the real objects' locations, were adjusted based on the room's dimensions and the device's field of view to ensure that the participants could not see more than two labels at the same time, and that looking at a label would most likely lead to the cursor being in that label's hitbox. This was used to log the attention given to each word during the learning task, in "user mode". In-between the learning phases, "admin mode" was switched on to display a new subgroup, check on the labels and prevent the app from logging attention data. Due to the HoloLens's novelty, the participants were allowed to interact with animated holograms for as long as they wished before the AR learning task to get used to the controls, adjust the device and overcome some of the novelty factor of the modality.

Learning Task
The Basque language was chosen after ruling out numerous languages that shared too many cognates (words that can be recognised due to sharing roots with other languages) with English, Spanish and other languages that are commonly spoken in the region where the study was administered. Basque presented interesting properties: latin alphabet to facilitate the learning, but generally regarded as a language isolate from

the other commonly spoken languages [27], allowing us to control the number of cognates more easily, with one of the authors being fluent in the language. The 30 words were carefully chosen and split into two groups A and B based on difficulty and length, and further split into 3 subgroups per word group where each subgroup corresponded to a topic: A1 was composed of office related words (pen, pencil, paper, clock, notebook), A2 of kitchen related words (fork, spoon, cup, coffee, water), A3 of clothing related words (hat, socks, shirt, belt, glove), B1 of some other office related words (table, chair, scissors, cellphone, keyboard), B2 of printed items (newspaper, book, magazine, picture, calendar), and B3 of means of locomotion (car, airplane, train, rocket, horse). The study's counterbalancing helped address possible issues arising from A and B potentially not being balanced enough. The learning task on a device was constituted of 3 learning phases, each of which lasted 90 seconds, for a total of 2 learning tasks (one per device) or 6 learning phases across both devices. The limit of 90 seconds was adjusted down from 180 seconds after a pilot study had shown a large ceiling effect with the users reporting having too much time. Once A or B was chosen as a group of words, the users successively saw subgroup 1 (5 words) during the first learning phase, then subgroups 1 and 2 (10 words) during the second learning phase, and then 1, 2 and 3 (15 words) in the last learning phase. The decision to allow the users to review the previous subgroups came as a solution to avoid the flooring effects in the productive test observed in the pilot study.

Distraction task

In order to prevent the users from going straight from learning to testing, a distraction was used to reduce the risk of measuring only very short-term recall. The task needed to have enough cognitive load to distract the participants from the words they had just learnt. The participants' performance at the task should also be correlated to their general performance regarding the study, in order to avoid introducing new effects – for example, a mathematical computation may bias the results as a participant with above average computational skills but below average memory skills may pass it fast enough that they would perform as well as another participant with below average computational skills but above average memory skills. Therefore, the distraction was chosen to be a memorisation task, in which the participants were asked to learn a different alphanumeric string of length 8 before every receptive test. The six codes used were the same for everyone, and were presented in the same order for every participant for the 2 by 2 balancing to take care of possible concerns over ordering.

METRICS

Receptive test

The receptive tests were administered on the desktop computer used for the questionnaire, in a different room from the two used for the learning tasks. Figure 4 shows the format of the test. The questions consisted in 5 images, each accompanied by a choice of 4 words from which the participants had to pick the appropriate one. Each image corresponded to one of the 5 new words seen in the preceding learning phase: A1 or B1 after the first learning phase, A2 or B2 after the second learning

Figure 4. Format of Receptive Recall Test.

Figure 5. Format of Productive Recall Test.

phase, and A3 or B3 after the third learning phase, depending on which one of A or B was chosen as the word group for that learning task, for a total of 6 receptive tests across the 2 learning tasks. All 5 images were available on the same page, allowing the participants to proceed by elimination. There was no time constraint, to avoid frustrating the participants, who were encouraged to use the tests as a way to prepare for the productive tests due to the strong flooring effects observed in the pilot study. The performance was measured as either 1 for a correct answer, or 0 for an incorrect answer. Every question was accompanied by a confidence prompt on a scale of 5 ranging from "Lowest Confidence" to "Highest Confidence".

Productive test

The productive tests took place on the same computer used for the receptive tests, immediately after the third receptive test at the end of each learning task. Figure 5 shows the format of the test, which also required a confidence evaluation for each answer. The productive test had 15 images corresponding to the 15 words from the selected word group, and participants were asked to type the corresponding word in Basque below each image. The error on a participant's answer was measured

Dependent Variable (Accuracy)	Z	p	effect size
Same-day Productive	-2.5397	**0.01109**	0.352
Delayed Productive	-3.1959	**0.001394**	0.574
Same-day Receptive	-0.7926	0.42799	0.110
Delayed Receptive	-0.1239	0.901389	0.022
Same-day Productive (FC pref group)	1.1589	0.246488	0.237
Delayed Productive (FC pref group)	-0.0580	0.95367	0.016

Table 2. Key results from statistical analysis. Results highlighted in bold face are statistically significant effects.

using the Levenshtein distance, which counts the minimum number of insertions, deletions and substitutions needed to transform a word into another, between their answers and the correct spellings. Participants were therefore encouraged to try their best guess to get partial credit if they did not know the answer, and had to provide an answer to every question to end the test. The Levenshtein distance was also upper bounded in our analysis by the length of the (correctly spelled) word considered, to prevent answers such as "I don't remember" from biasing a participant's average error, and divided by the length of the correct answer to get a normalised error:

$$\text{AdjLev}(w, \hat{w}) = \frac{\min(\text{Lev}(w, \hat{w}), \text{Length}(\hat{w}))}{\text{Length}(\hat{w})} \quad (1)$$

where w is the participant's answer on a given question, and $\hat{w}$ the correct answer. The score was then computed as

$$\text{Score}(w, \hat{w}) = 1 - \text{AdjLev}(w, \hat{w}) \quad (2)$$

where 1 indicates a perfect spelling, and 0 a maximally incorrect answer. As in the receptive tests, every question was accompanied by a confidence prompt on a scale of 5 ranging from "Do not know" to "Very confident".

Delayed test
The delayed tests consisted in the same tests used for the same-day testing, in a slightly different order: the productive test of each word group was administered before the 3 receptive tests to prevent participants from reviewing with the receptive tests due to the absence of a time constraint. The tests were sent in a personalized email to the participants 4 days after the study. Only tests completed in the 24 hours after received the email were kept in the analysis. Further, the test did not allow the participants to press the back button, and only tests completed in a similar amount of time as the same-day tests were kept. Participants were informed that the study being comparative, the absolute number of words they remembered did not matter, and that the goal of the study was to measure how many people performed better with either device with no expectation of a modality being better than another. This was done in order to reduce the impact of potential demand effects, and only recall (no feedback) was evaluated in the delayed test to further diminish such biases. In total, 31 participants' delayed test answers satisfied the criteria mentioned above. Note that the 2 by 2 counterbalancing was conserved (8 participants had followed order I, II and IV and 7 order III as defined in Table 1).
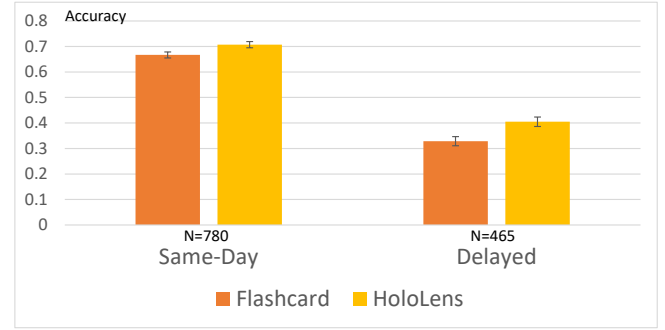

Figure 6. User performance on same-day productive recall tests. The left group shows the flashcard and AR accuracy score for the same-day test and the right side shows the comparison for the 4 day delayed test. Error bars show standard error here.

RESULTS

Productive Recall
Figure 6 shows the accuracy results of the same-day productive recall test compared to the delayed test for both modalities. The AR condition is shown in the lighter color. The delayed test was administered 4 days after the main study, and there was some attrition, with 780 question responses in the main study and 465 for the delayed. Accuracy was measured using the score function previously defined in Eq.2 as 1 minus the normalized Levenshtein distance between the attempted spelling and the correct spelling. In the same-day test, the AR condition outperformed the flashcards condition by 7%, and more interestingly, in the delayed test, this improvement was more pronounced, at 21% better than the flashcard condition. The test results were analyzed in a non-parametric way after Shapiro-Wilk tests confirmed the non-normality of the data. This is due in part to the many occurrences of words perfectly spelled. Both differences are significant with Wilcoxon Signed-rank tests: p=0.011 and p=0.001 for the same-day and the delayed productive results respectively, as seen in Table 2. The table also reports productive recall scores for those users who reported that Flashcards were more effective than AR (FC pref Group). Interestingly, no significant difference was found between the modalities for this sub-group, in contrast to the results for the general population.

Based on interviews with the participants, we believe that the significant improvement in delayed recall is linked to the spatial aspect in the HoloLens condition. Several participants reported qualitative feedback to this effect, such as in the following example: *One reason the AR headset helped me recognize the words better is because of the position of the object. Sometimes, I'm not memorizing the word, I'm just recognizing the position of the object and which word it correlates to.*

Productive Recall by Term
Figure 7 and Figure 8 show the same-day and delayed productive recall scores, respectively, broken down by term. The graphs show box plots with mean accuracy on the y-axis and the 30 Basque words on the x-axis, ranked by the accuracy score, and color-coded based on the exposure group. For instance, orange bars represent terms that appeared only once,
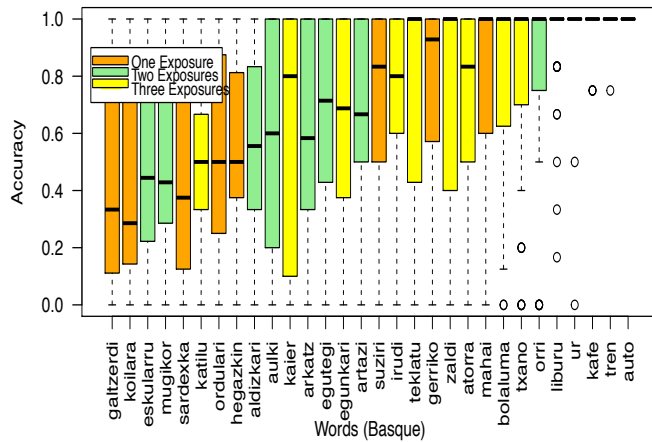
Figure 7. User performance on delayed productive recall tests, ranked by term. Colors show exposure groups. The accuracy score on the y-axis is computed from the mean of the normalized Levenshtein distance between the participant's spelling and the correct spelling.
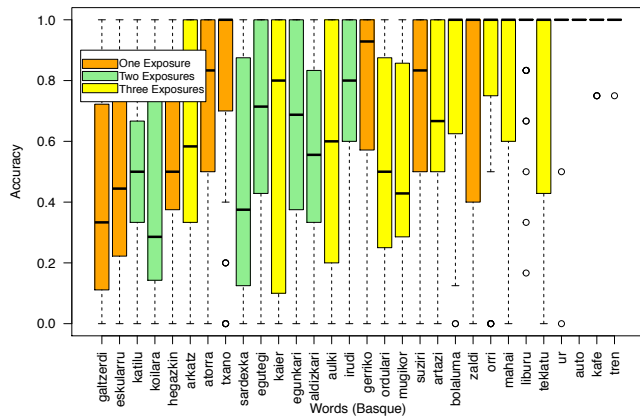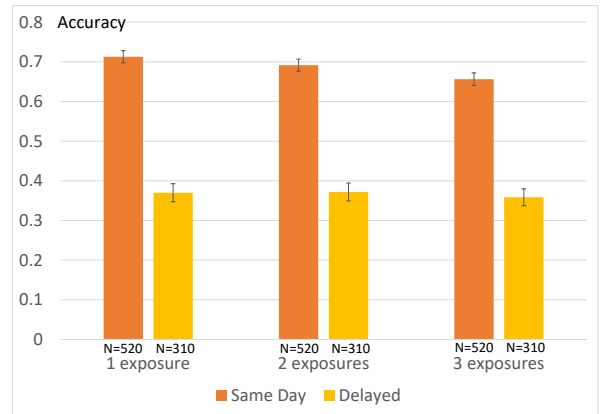

Figure 9. Mean accuracy in productive recall for each exposure group. Side-by-side bars show same the result for the same-day and delayed recall tests. Error bars show standard error here.

while yellow appeared in three learning phases for each participant. Here, accuracy is also computed as the score function defined in Eq. 2. Both graphs reveal that the terms are fairly evenly distributed across the range of scores, with an obvious relationship between performance and word length. For example, 'galtzerdi' and 'eskularru' are the longest words in the set and received the lowest scores. As expected, the three highest scoring words in both tests (auto, kafe and tren) are English cognates. Again, these were carefully distributed across the term groups.

Repeat Exposure

To recap, participants had three learning phases per modality, with sets of 5, 10 and 15 terms in the first, second and third, respectively. In each phase, one subgroup of 5 terms were new, meaning that the first subgroup was seen in three phases, and the last subgroup of 5 was seen only once. Figure 9 shows grouped barplots representing mean accuracy for exposure groups. Mean accuracy is shown for same-day and delayed productive recall tests. Here, we see that in the delayed tests, repeat exposure does not seem to have any effect on recall. Surprisingly, there was no significant effect between terms that had one exposure and terms that had three exposures in the same-day test. To further investigate this finding, a discussion of eye-tracking and head orientation data for flashcards and AR is provided below.

Receptive Recall

Receptive recall was analyzed in the same manner as productive recall, however a histogram of response accuracy revealed a ceiling effect in the data, where many participants provided fully correct responses. The mean receptive recall score was 0.89 for the same-day test in both modalities and 0.84 in the 4-day delayed test again for both modalities. In the delayed test, the productive recall was presented first to avoid learning

Figure 8. User performance on same-day productive recall tests, ranked by term. Colors show exposure groups. The accuracy score on the y-axis is computed from the mean of the normalized Levenshtein distance between the participant's spelling and the correct spelling.

effects from viewing multiple choice options. There was no significant difference between the modalities in this test.

Attention Metrics

Gaze data was gathered for both modalities as described above. For each of the terms in the three different exposure groups we computed the average time that participants' attention was focused on that item. This was performed primarily to examine why repeated exposure to terms did not produce an observed improvement in accuracy. For the first group, the mean was 13.5 seconds (SD 6.3 seconds), for the second, the mean was 10.8 seconds (SD 7.2 seconds), and third group had a mean attention time of 7.2 seconds (SD 4.5 seconds). The differences in attention times not being significant for each group may imply that during the learning phases, participants focused mainly on the new items, or that users chose to focus on different words on average. This is a possible explanation for the lack of accuracy improvement for repeated-exposure items.

Click data was recorded for the flashcard application to help identify potential learning patterns. Recall that the flashcards had two sides and required a click to turn from text to image and back again (Figure 2). The click patterns showed that people tended to click more often towards the end of the study. 18 of the participants had a pattern of clicking the same flashcard over 5 times in a row, perhaps indicating a desire to see both image and text at the same time, or testing themselves during the learning phase. Both possibilities are supported by users reporting in the post-study interview that they enjoyed the ability to see the object and the word simultaneously in AR, while others mentioned making use of the flashcards' two-sided nature to self-test.

Confidence

For each question in the recall tests, participants reported their confidence level in the provided answer. Figure 10 shows the distribution of those scores for the same-day productive recall test. The scores follow a U shaped distribution, showing that participants tended to be sure they were wrong, or sure they were right about their responses. To assess how correct these judgments were, an analysis of mean score on productive recall was performed for each confidence level. Figure 11 shows a breakdown of median accuracy for each reported confidence level. This informs us that participants were good predictors of their performance in the productive recall tests. Confidence scores were also analyzed by modality to understand if the learning method had an impact on participants confidence in their own answers. Despite the fact that significant effects were shown on accuracy metrics across the modalities (delayed and same-day), and that accuracy and confidence were strongly correlated (results shown in Figure 11), there was no significant effect observed for the confidence metric in between the modalities.

Perception

Participants were asked about their experience using AR and flashcards, and their subjective ratings correspond with their learning performance. In terms of what was fastest for learning words, 54% found AR fastest, compared to 46% who
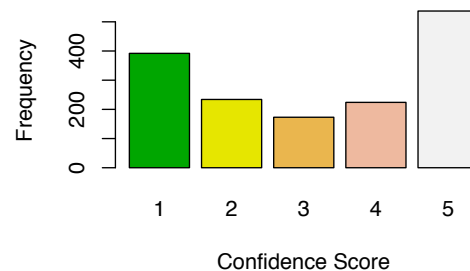

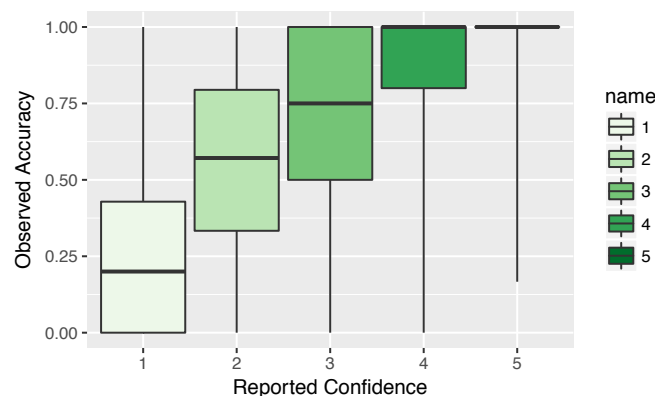Figure 10. Distribution of reported confidence scores for the productive recall test.

Figure 11. Relation between participant confidence and actual performance in productive recall tasks.

found flashcards fastest. As a side note, 13 among the delayed test population had reported preferring the flashcards, as opposed to 18 for AR. As for the learning experience, 75% of participants rated AR "good" or "excellent", while 63% rated flashcards "good" or "excellent".

Figure 12 shows that when asked about the effectiveness of each platform for learning words, 88% of participants "somewhat agreed" or "strongly agreed" that the AR headset was effective, while 79% "somewhat agreed" or "strongly agreed" that the flashcards were effective.

Participants' comments comparing the two platforms revealed that about 20% (10 of 52) felt AR and flashcards were equally effective for learning because of the visual imagery both provide. 14 of 52 specifically mentioned that they found AR better because they saw the word and object at the same time. Almost 20% (10 of 52) stated that AR was better because it was more interactive, immersive, and showed objects in real time and space (e.g., "The flashcards are classic and I have experience learning from them but the AR headset was more immersive" and "The headset was more interactive because it was right in front of you with physical objects rather than through a computer screen"). Only 13% of the participants commented that flashcards were better, due to their familiarity and being confined spatially to the tablet.

A stark/striking difference was found in participants' opinions about which platform was enjoyable for learning. Figure 12 shows that 92% of participants "somewhat agreed" or "strongly agreed" that using the AR headset was enjoyable for learning words, compared to only 29% for using the flashcards. Open-ended comments from the participants pointed to the not unexpected novelty effect of AR (21 of 52 or 40%), "The AR Headset because it was an incredibly futuristic experience." In addition, 16 of 52 participants (31%) commented explicitly on how AR is more interactive, engaging, hands-on, natural, and allowed for physical movement (e.g., "The AR headset was more interactive and required movement which engaged my mind more" and "The AR Headset was more fun because it's more fun to be able to move around and see things in actual space than on a computer screen" or "The AR headset was more enjoyable because it allowed for you to interact with the objects that you are learning about. It felt more realistic and applicable to real life, plus I had the visual image that helped me remember the words"). Only 8 of 52 participants (15%) indicated that flashcards were more enjoyable because they were familiar, practical, and straightforward.

As we noted earlier in the discussion of productive recall results (Section 6.1.1, several participants commented in interviews or left text feedback related to the spatial aspect of the AR condition, generally saying that it helped give them an extra dimension to aid in learning. For example, one participant reported that: *The AR headset but me in contact with the objects as well as had me move around to find words. I was able to recall what words meant by referencing their position in the room or proximity to other objects as well. ***Seeing the object at the same time as the word strengthened the association for me greatly***.* Another participant said *"the AR seems like it would work better with friends or family trying to learn together, while the flashcards seem to work on an individual level."*. The latter comment points towards a social or interactive aspect of AR-based learning which we have not focused on in this study, but is nonetheless of potential interest to system designers and language learning researchers. The potential for social interaction and learning that this participant mentioned is likely linked to the availability of an interactive learning space.

Another possible benefit to learning in the AR condition is that it can facilitate the so called "memory palace" technique, frequently depicted in popular TV by Sherlock Holmes. It has been shown to be useful when applied to learn the vocabulary of a foreign language. The method is described in Anthony Metivier's book "How to learn and memorize German vocabulary..." [17]. The author suggests to begin by creating a memory palace for each letter of the German alphabet by associating it with a location in an imaginary physical space. According to Metivier, each memory palace then shall include a number of loci where an entry (a word or a phrase) can be stored and recalled whenever it is needed. One of our participants made a comment about this learning method after learning in the AR condition: "I use memory palaces, so I really enjoyed AR as it felt somewhat familiar and made it easier for me to use the technique than the flashcards".

LIMITATIONS AND FUTURE WORK

This paper has described a proof-of-concept experiment that shows that AR can produce better results on the learning of foreign-language nouns in a controlled lab-based user study. However, the study has several limitations. First, learning itself occurred in a controlled experimental context, in which subjects were paid an incentive. This can not be assumed to be representative of real-world learning, and it is possible that our results may vary in real learning context. Second, and related, it is likely that novelty effects had some impact on the study given that the HoloLens remains in the category of new and exciting technology. Our design included a long acclimatization phase with the device, but it is difficult to be sure that our qualitative results have not been impacted by novelty effects. Third, our design choice for the flashcard application mirrored the traditional flashcard design for self-testing. That is, an image on one side and foreign text on the other. A small number of participants noted that they preferred the AR condition's inherent ability to view the object label and the object at the same time. Others clearly made use of the self-testing feature. Last, our receptive recall tests, while carefully controlled based on informal pre-studies and performance information from existing literature, showed ceiling effects with a large number of participants. No ceiling or floor effects were observed for the productive recall test. In follow-up experiments, we will increase the difficulty of the receptive recall tests.

There are several avenues to continue our research on the AR-bis Pictus system, most notably, by taking the system beyond the controlled learning environment that was described in this paper and applying it to real-world learning tasks. As an initial step towards this, we have implemented a first-draft of real time object labeling with HoloLens and YOLO [22] and are also in the process of working with students and course administrators to develop personalized language learning plans that could be deployed in the system. Evaluating the performance of a real-world AR personalized-learning system is clearly a non-trivial task that will require complex longitudinal studies with many learners to account for differences in user experiences brought about by uncontrolled data in real-world environments. In terms of education and learning theory, it may be possible for our results to contribute to expanding the existing and established theories of CTML –but this would also benefit from running larger studies with more participants in real world settings. Finally, there is a wealth of interaction data that we gathered from this study through eye-tracking, click-interaction and HoloLens interactions that may contain interesting learning patterns that can be related back to performance.

CONCLUSION

This paper has described a novel system for language learning in augmented reality, and a 2x2 within-subjects experimental evaluation (N=52) with the system to assess the effect of AR on learning of foreign-language nouns compared to a traditional flashcard approach. Key research questions were proposed, related to quantitative performance in immediate and delayed recall tests, and user experience with the learning
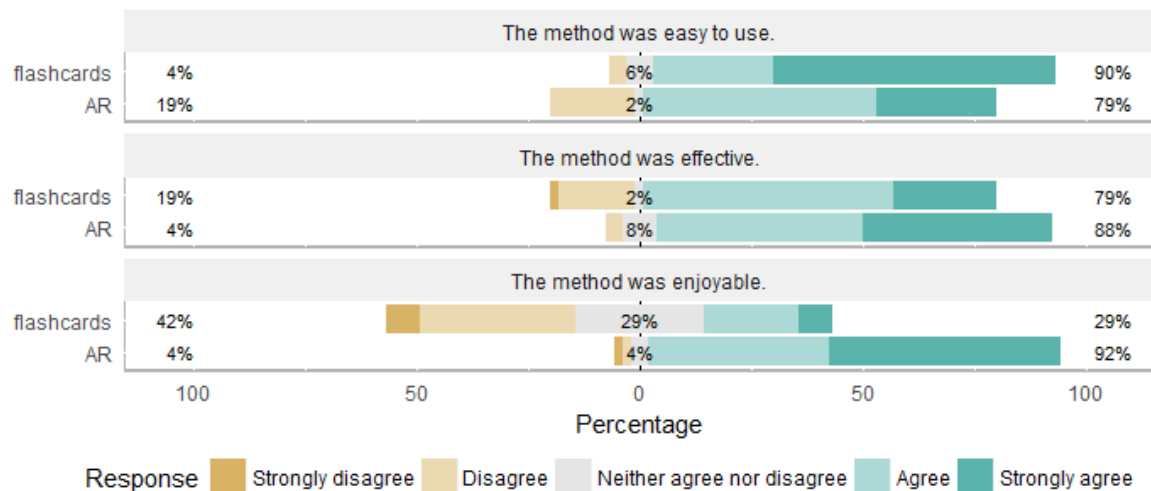
Figure 12. Qualitative feedback for the 52 participants

modality (qualitative data). Results show that 1.) AR outperforms flashcards on productive recall tests administered same-day by 7% (Wilcoxon Signed-rank p=0.011), and this difference increases to 21% (p=0.001) in productive recall tests administered 4 days later. 2.) Participants reported that the AR learning experience was both more effective and more enjoyable than the flashcard approach. We believe that this is a good indication that AR can be beneficial for language learning, and we hope it may inspire HCI and education researchers to conduct comparative studies.

ACKNOWLEDGMENTS

REFERENCES

1. —— 2017. Suppressed for double-blind review. (2017). ——- Technical Report, —, —.

2. Carrie J Cai, Philip J Guo, James Glass, and Robert C Miller. 2014. Wait-learning: leveraging conversational dead time for second language education. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2239–2244.

3. Dorothy M Chun and Jan L Plass. 1996. Effects of multimedia annotations on vocabulary acquisition. *The modern language journal* 80, 2 (1996), 183–198.

4. J.A. Comenius. 2014. *Orbis Pictus*. Literary Licensing LLC. `https://books.google.com/books?id=InbKoAEACAAJ`

5. Maria F Costabile, Antonella De Angeli, Rosa Lanzilotti, Carmelo Ardito, Paolo Buono, and Thomas Pederson. 2008. Explore! possibilities and challenges of mobile learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 145–154.

6. Gabriel Culbertson, Shiyu Wang, Malte Jung, and Erik Andersen. 2016. Social Situational Language Learning through an Online 3D Game. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 957–968.

7. Matt Dunleavy. 2014. Design principles for augmented reality learning. *TechTrends* 58, 1 (2014), 28–34.

8. Matt Dunleavy and Chris Dede. 2014. Augmented reality teaching and learning. In *Handbook of research on educational communications and technology*. Springer, 735–745.

9. Robert Godwin-Jones. 2016. Augmented reality and language learning: From annotated vocabulary to place-based mobile games. (2016).

10. Raphael Grasset, Andreas Duenser, Hartmut Seichter, and Mark Billinghurst. 2007. The mixed reality book: a new multimedia reading experience. In *CHI'07 extended abstracts on Human factors in computing systems*. ACM, 1953–1958.

11. Hannes Kaufmann and Dieter Schmalstieg. 2003. Mathematics and geometry education with collaborative augmented reality. *Computers & graphics* 27, 3 (2003), 339–345.

12. Yang Liu, Daniel Holden, and Dongping Zheng. 2016. Analyzing studentsâĂŹ language learning experience in an augmented reality mobile game: an exploration of an emergent learning environment. *Procedia-Social and Behavioral Sciences* 228 (2016), 369–374.

13. Richard E Mayer. 2005. *The Cambridge handbook of multimedia learning*. Cambridge University Press.

14. Richard E. Mayer. 2009. *Multimedia Learning* (2 ed.). Cambridge University Press. `DOI: http://dx.doi.org/10.1017/CBO9780511811678`

15. Richard E Mayer. 2011. *Applying the science of learning*. Pearson/Allyn & Bacon Boston, MA.

16. Richard E Mayer and Valerie K Sims. 1994. For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of educational psychology* 86, 3 (1994), 389.

17. A. Metivier. 2012. *How to Learn and Memorize German Vocabulary: ... Using a Memory Palace Specifically Designed for the German Language (and Adaptable to Many Other Languages Too)*. CreateSpace Independent Publishing Platform.
`https://books.google.com/books?id=jTCuNAEACAAJ`

18. Sophio Moralishvili. 2015. Augmented Reality in Foreign Language Learning. (2015).

19. Roxana Moreno and Richard E Mayer. 1999. Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of educational psychology* 91, 2 (1999), 358.

20. Tatsuya Nakata. 2011. Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning* 24, 1 (2011), 17–38.

21. Jan L Plass, Dorothy M Chun, Richard E Mayer, and Detlev Leutner. 1998. Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of educational psychology* 90, 1 (1998), 25–36.

22. Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242* (2016).

23. Dieter Schmalstieg and Daniel Wagner. 2007. Experiences with handheld augmented reality. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 3–18.

24. Olga Scrivner, Julie Madewell, Cameron Buckley, and Nitocris Perez. 2016. Augmented reality digital technologies (ARDT) for foreign language teaching and learning. In *Future Technologies Conference (FTC)*. IEEE, 395–398.

25. Paul Seedhouse, Anne Preston, Patrick Olivier, Dan Jackson, Philip Heslop, Madeline Balaam, Ashur Rafiev, and Matthew Kipling. 2014. The European digital kitchen project. *Bellaterra journal of teaching and learning language and literature* 7, 1 (2014), 0001–16.

26. Ekrem Solak and Recep Cakir. 2015. Exploring the Effect of Materials Designed with Augmented Reality on Language Learners' Vocabulary Learning. *Journal of Educators Online* 12, 2 (2015), 50–72.

27. R.L. Trask. 1997. *The History of Basque*. Routledge.
`https://books.google.com/books?id=OiemTo_t5r8C`

28. Emine Türk and Gülcan Erçetin. 2014. Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. *Computer Assisted Language Learning* 27, 1 (2014), 1–25.

29. Nesra Yannier, Kenneth R Koedinger, and Scott E Hudson. 2015. Learning from Mixed-Reality Games: Is Shaking a Tablet as Effective as Physical Observation?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1045–1054.

30. Makoto Yoshii. 2006. L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language learning and technology* 10, 3 (2006), 85–101.