Word Saber: An Effective and Fun VR Vocabulary Learning Game

Judith Hartfill
Universität Hamburg
Germany
hartfill@informatik.uni-hamburg.de

Jenny Gabel
Universität Hamburg
Germany
jenny.gabel@studium.uni-hamburg.de

Daniel Neves-Coelho
Universität Hamburg
Germany
daniel.neves.coelho@studium.uni-hamburg.de

Daniel Vogel
Universität Hamburg
Germany
daniel.vogel@studium.uni-hamburg.de

Fabian Räthel
Universität Hamburg
Germany
fabian.raethel@studium.uni-hamburg.de

Simon Tiede
Universität Hamburg
Germany
simon.tiede@studium.uni-hamburg.de

Oscar Ariza
Universität Hamburg
Germany
ariza@informatik.uni-hamburg.de

Frank Steinicke
Universität Hamburg
Germany
steinicke@informatik.uni-hamburg.de

Figure 1: Experiment setup. Subject wearing the HMD with controllers (left), and a picture of the corresponding gameplay.

ABSTRACT

In this paper, we introduce *Word Saber*, a virtual reality (VR) edu-game for vocabulary learning. The game design and game mechanics are inspired by *Beat Saber*, where players have to cut virtual cubes from predefined directions by using bi-manual interactions with two lightsabers. *Word Saber* uses a similar concept, but players have to cut the corresponding object to the presented word. We evaluated the effects of *Word Saber* on learning efficiency and enjoyment in comparison to a traditional vocabulary learning method. The results show that *Word Saber* is effective for vocabulary learning. However, as *Word Saber* presented significantly lower scores in terms of recognition and recall rate compared to a flashcard method, it has lower efficiency. Regarding subjective preferences, our findings suggest that VR edu-games can be fun and useful for language learning in the long run as participants reported to be more motivated to use *Word Saber* afterward.

MuC 2020, September 06–09, 2020, Magdeburg, Germany

CCS CONCEPTS

• **Human-centered computing → Virtual reality**; **Empirical studies in interaction design**; • **Applied computing → Interactive learning environments**.

KEYWORDS

virtual reality, edu-game, vocabulary learning, CALL, VRALL, gamification, digital game-based language learning

ACM Reference Format:
Judith Hartfill, Jenny Gabel, Daniel Neves-Coelho, Daniel Vogel, Fabian Räthel, Simon Tiede, Oscar Ariza, and Frank Steinicke. 2020. *Word Saber*: An Effective and Fun VR Vocabulary Learning Game. In *Proceedings of Mensch und Computer 2020 (MuC 2020).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404983.3405517

1 INTRODUCTION

In a globalized world, multilingualism is of particular importance. Foreign language knowledge is essential, for instance, for business usage, online communication as well as for private traveling. Several studies showed that digital and gamified teaching methods could improve the learners' efficiency and motivation [26, 33].

In 2015 a meta-analysis on virtual reality for foreign language education was conducted by Solak et al. [28]. They found that "the capacities of virtual reality have not been utilized adequately in foreign language teaching and learning facilities". They concluded that more research is required to explore the possibilities for VR-assisted language learning (VRALL).

In this paper, we introduce the VR educational game *Word Saber*, see Figure 1, which aims at helping foreign language learners with vocabulary acquisition and training. It is universally applicable and independent of the target language. Furthermore, *Word Saber* does not rely on the learners' first language, as no translation in the first language is provided. This way, users can learn words they do not know in their first language. *Word Saber* follows a multimodal approach as the vocabulary is presented written in the target language, as a 3D object, as well as auditory.

The idea of *Word Saber* is to provide an edutainment game, that should be both effective at vocabulary learning as well as highly motivating for the users. From the game design perspective, it is obvious that the gameplay should motivate players by providing a good user experience while also mentally challenging the user during the learning process.

The joy of use is of particular importance so that players are willing to play repetitively, as learning, in general, is known to be most effective if it is done regularly. Early research in human learning by Ebbinghaus from 1885 has introduced the forgetting curve [9] in this context. Hence, we assume that if the users enjoy playing *Word Saber*, their willingness to play the game repeatably (and learn vocabularies) would be high, and so the game could be very effective in the long run.

Vocabulary acquisition in *Word Saber* is based on deliberate, paired-associate word learning, which was found to be effective by Elgort in 2011 [11]. Learning from word lists or flashcards are conventional methods based on deliberate learning. For *Word Saber*, we transferred the idea into VR.

In 2012 van der Ploeg et al. found that prolonged sitting is associated with higher all-cause mortality risk [31]. Moreover, Labonté-LeMoyne et al. in 2015 conducted a study with a treadmill desk setup, where participants either sat or walked on a treadmill while reading a text. The participants who walked had a short-term increase in memory and attention [17]. So we designed *Word Saber* to be a VR experience that requires the player to stand and use the whole body to fulfill the task. It offers the possibility to decrease sitting time while learning and might support the learning process with a light full-body motion.

Word Saber focuses on vocabulary acquisition and training for words that can be visualized, like many nouns. The immersive nature of a head-mounted display (HMD) increases the sense of presence, and the distraction from the real world can be minimized [24]. Furthermore, the game environment is designed in a simple and minimalist form and feel to support the focus on the learning task, as illustrated in Figures 2-4. The game mechanics are inspired by *Beat Saber*. *Beat Saber* is one of the most successful VR games originally developed by Beat Games, which have been acquired by Facebook and integrated into Oculus Studios in 2019 [3]. However, instead of the beat, the focus of *Word Saber* is on the vocabulary in the target language. In the following, the procedure of the game is described. In the first phase of the game, i. e., the learning phase, new words are introduced to the player by presenting the word as text, auditory, and the corresponding 3D object. Afterward, in the recall phase, the user's task is to remember the word for each presented object. Finally, in the gaming phase, the task is to hit the correct 3D object for a presented word with one of the two virtual 3D sabers. A high-score system gives immediate feedback to the user's performance.

We performed a user study to test the effectiveness and efficiency of vocabulary learning as well as the users' enjoyment of *Word Saber* in a between-subjects design. One group played *Word Saber*, and the control group used conventional flashcards, which are a typical approach for deliberate word learning. We assumed the following hypotheses:

- Hypothesis 1: Playing *Word Saber* leads to better recall and recognition of the vocabulary than vocabulary learning with the flashcard method.
- Hypothesis 2: Playing *Word Saber* results in more fun than vocabulary learning with the flashcard method.

The remainder of this paper is structured as follows. Section 2 resumes work related to our approach. Then, Section 3 introduces a detailed overview of *Word Saber*. Section 4 presents an evaluation that we conducted to analyze *Word Saber*. Section 5 summarizes the results of the user study. Section 6 discusses those results. Finally, Section 7 concludes the paper and gives an overview of future work.

2 RELATED WORK

The idea of game-based language learning is not new. In 1965 Lee published "Language Teaching Games and Contest", a collection of playful activities, with a whole chapter on vocabulary learning [18]. Since then, several studies have further investigated game-based language learning and have found positive effects for learners [13], also for vocabulary learning specifically [20].

Computer-assisted language learning (CALL) approaches have been around since the '60s, and many of them have tried to combine learning and gaming. However, often, especially in early approaches, the results were neither effective nor fun resulting in "chocolate-dipped broccoli" as Bruckman [5] puts it.

Some more recent approaches for computer-assisted vocabulary learning make use of two-dimensional displays, as desktop monitors or mobile phones. For instance, in 2017 Zhu et al. developed *ViVo*, a "Video-augmented dictionary for vocabulary learning" [33]. Their tool adds subtitles with target words to existing short video clips. They conducted a study that showed that learners could recall significantly more words with their method than with a standard dictionary. The difference was significant for mid-term retention testing, in particular, several days after the learning phase. The participants preferred using *ViVo* and reported to feel more engagement and enjoyment. This approach does not include any gamification aspects.

Palomo et al. developed a "Mobile Game-Based Learning Environment" for foreign language learning [23] in 2016. It is set into a classroom context and is designed to encourage students to practice their vocabulary knowledge outside the class. The user has to translate a word from the foreign language to the first language by reading and typing and may listen to the correct pronunciation of the word. The effectiveness or enjoyment of this approach remains unclear, as no user study was conducted.

In 2013, Edge et al. developed an language learning edu-game with body motion, *SpatialEase* [10]. It is an application for the Microsoft Kinect camera, where the user is advised to perform specific actions, addressing the kinesthetic intelligence. Despite the small sample size in the corresponding study, they found learning gains of *SpatialEase* comparable to the Rosetta Stone[1] software. The study did not evaluate the users' enjoyment of the game. Furthermore, vocabulary learning with *SpatialEase* is limited to expressions for body movements.

Silva et al. conducted a study on the motivational impact of VR on game-based learning [26] in 2017. Their findings lead them to the proposition that "an immersive version of a game is more effective in the aspect of attention when compared to a non-immersive one" and that the presented content seems to be perceived as more relevant to the participants when it is presented in virtual reality. Additionally, they found that the feeling of confidence is much higher in VR.

With the rise of VR technologies, language learning techniques were transferred to more immersive setups, including gamification aspects. Many of them are based on role-playing scenarios, where the user has to communicate in a particular situation, like [7] and [32] or utilize multiuser virtual environments (MUVEs) for more free learning techniques like [1] and [8].

VR approaches that specifically aim at vocabulary learning are often set into real-world inspired virtual environments, like a household [12] or a zoo [19] and require the user to find a particular object.

All these game mechanics are mainly oriented on tasks that could very similarly be done in the real world. The interaction-based tasks could be done with real-life partners, and the search tasks, at least those set in a household environment, could be done with sticky notes at home, for example. There is little research to develop more diverse game mechanics for language, and specifically, vocabulary learning in VR and the potential of immersive setups with HMDs and handheld controllers is not nearly fully utilized in this context.

[1]https://www.rosettastone.com

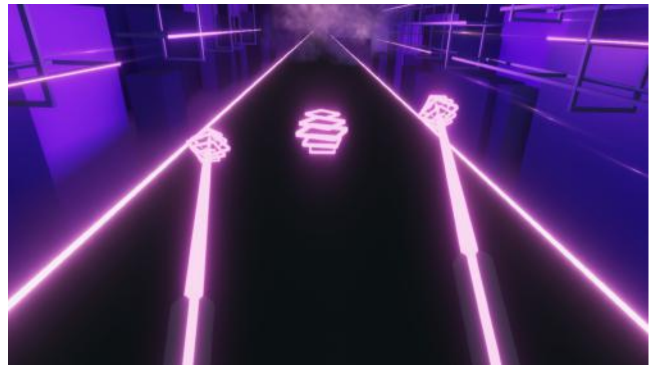

Figure 2: Environment of *Word Saber*.

3 *WORD SABER*

In *Word Saber*, vocabularies of any language can be learned with object-target language pairs additionally supported via audio. This section describes the game design and hardware setup used for the game.

3.1 Game Design

At the beginning of the game, the user stands in an abstract surrounding, equipped with the two controllers, which are visualized as short lightsabers, as illustrated in Figure 2. From an indicated area, various objects fly towards the user. In the starting area, the current word is displayed in the target language. The user is supposed to hit the object, which corresponds to the currently displayed word with one of the virtual 3D sabers.

For each hit, immediate feedback is given on the correctness, indicated by red crosses or green ticks around the object, as can be seen in Figure 3. The visual feedback is complemented by auditory feedback with specific sounds indicating success and failure. Additionally, when the user hits an item, the corresponding word is provided as the written text is displayed next to the object and auditory, using speech synthesis.

After a certain amount of time, if the user did not hit the correct target item, help is provided by an appearing text box that displays the correct word and the corresponding object.

The entire game is divided into three phases. In addition to the actual game phase described above, two more phases were added to the overall gameplay, which complements the vocabulary learning process. In the first phase, new unknown words are introduced to the user. In the pilot phase of the development, early feedback from testers indicated that five words per round seemed to be a good number, which is in accordance with findings on chunking in working memory by Miller in 1956 [21] and Baddeley and Hitch in 1974 [2]. The new item is presented to the user with the written word displayed next to it (see Figure 4). Besides, is provided audi-

Figure 3: Hitting an object, that does not correspond to the currently displayed one. Red crosses provide visual feedback on the error.

Figure 4: Presentation of the word *Ninjin* (carrot). Additionally, the word is provided auditory.

tory using speech synthesis, and the word-object pair flies towards the user. Afterward, different objects, including the new one–but without the corresponding word–fly towards the user. The new word is displayed on the background panel, and the user is supposed to hit the correct item with one of the lightsabers.

Then, in the second game phase, the user is instructed to repeat the vocabulary for themselves. Again, each of the five new objects introduced in the first game phase flies towards the user, but without the written word or pronunciation provided. Now, the users' task is to remember the spelling of these words.

The third game phase is the actual gaming phase, where the user has to hit the correct items as described above. In this phase, a high score is calculated, and the user can score an extra point by doing "combos" by hitting the correct item multiple times in a row.

The virtual environment (VE) has a simple and minimalist design to keep the user focused on the task. In the VE, the user stands at the end of a corridor with abstract shelves to the sides and storage boxes in the front. The user's sense of presence, in particular the plausibility illusion, is further increased by 3D objects flying out of the storage boxes towards the player. The surrounding elements in the VE further enhance this illusion, as they form a path on which the objects move. The colors in the VE are chosen from a violet color palette to support the abstract and minimalist design and provide a cohesive look and feel. Additionally, the relatively dark background color offers a contrast to the multicolored objects.

Similarly, the 3D objects were chosen and designed to be easily recognizable. The models were either taken from the Unity Asset Store (and modified to fit into the VE) or manually built.

Following the design of *Beat Saber*, the environment is illuminated by neon lights and glowing volumetric particles. A light bloom effect was added through the use of Unity's Post-Processing Stack v2 to achieve a more realistic soft glowing effect.

To invoke and enhance the sense of motion, and the feeling of flow during gameplay, a "warp drive" effect was used where particles are continuously moving towards the user. Moreover, particle effects were used to guide the users' focus of attention. For example, a bright particle effect guides the user towards the central UI panel every time a new word is displayed. Some particle effects and animations were specially designed to give the user information about the game state as well as direct and unambiguous feedback on the user's interaction within the game. Sound effects further support the particle effects. Moreover, this ensures that the user notices changes in the game even if the objects are currently not in the field of view. Additionally, the particle and the sound effects for hitting correct objects were designed as gamification techniques to visually and auditory reward the user.

Altogether, the combination of visual and auditory cues was implemented to provide a pleasant user experience. To maintain the user's place illusion in the VE [27], all UI elements that are visible to the user are directly integrated as neon panels into the 3D world. Apart from displaying the current word, the UI elements also show instructions and scores.

In general, the development of *Word Saber* was oriented along the eight game-flow criteria proposed by Sweetser et al. [29] in 2005 to ensure a high level of player enjoyment. In the following, the implementation of each criterion in *Word Saber* is explained.

- **Concentration** According to [29] to achieve a high level of player enjoyment, a game should require concentration, but should also enable the player to concentrate on the game. This is ensured by the regularity in which the items appear and the minimalist and structured design of the VE.
- **Challenge** Games should be challenging but not overstraining. This was implemented with the different phases of the game: the learning phase for the new words, the repetition phase, and the final game phase. By going through these phases with more and more new words, the level of challenge increases.
- **Player Skills** A game should be playable without prior knowledge and enhance the players' skill level. In *Word Saber*, this was achieved by using natural and intuitive interaction methods (based on the movement of controllers, not on buttons), and the tasks were explained in the game.
- **Control** Players should have control over their actions in the game. The action of hitting an item was optimized so that the users had as much control as possible.
- **Clear Goals** Games should have clear goals, which need to be presented at appropriate times. At the beginning of each game phase, the goal was displayed to the user.

Feedback Players must receive appropriate feedback at appropriate times. Feedback was provided immediately after each hit, as well as after each round.

Immersion Players should experience deep but effortless involvement in the game. Through the use of a head-mounted display, a high degree of immersion was achieved. The impression was supported by the objects flying towards the user.

Social Interaction Games should support and create opportunities for social interaction. With the high score, a possibility to easily add social interaction was implemented.

3.2 Hardware and Infrastructure

The game was implemented using the Unity 2018.2 game engine[2]. The game runs on an Intel Core i7-6900K CPU @ 3.20 GHz with 16 GB RAM and an NVIDIA GeForce GTX 1080 graphics card on a Windows©10 64-bit system. Players use the HTC VIVE Pro[3] with two controllers.

4 EVALUATION

Japanese was chosen as the language for this study, as it has few similarities with commonly spoken languages in professional or academic environments. Moreover, Japanese itself is a rarely spoken language. Only *romaji* (Latin script) was used to display the Japanese words in the game since all participants had no prior knowledge of the Japanese writing system.

29 people participated in the study (18 male, 11 female), aged from 19 to 56 (M = 25.83, SD = 8.18). Participants were recruited via email invitation from the students and employees from our university but were not limited to it. They were told that participation required to have no advanced knowledge of Japanese.

4.1 Study

We conducted a study with a between-subject design. One participant at a time was tested in a 45 minutes time slot. First of all, participants were asked to read and sign a consent form. They were told that we were evaluating different vocabulary learning methods against each other. Then they were randomly assigned to one of the two conditions: *Word Saber* or flashcards. The procedure of both is explained in detail in the following.

Flashcard condition In the flashcard condition, the participants were asked to fill out a preliminary questionnaire on demographics, including a self-report on Japanese skill level. Then they were handed 20 flashcards with images of food items on the one side and the corresponding romanized Japanese words on the other. An example image used for the flashcards is given in Figure 5. The images were generated from the 3D models used in *Word Saber*.

The participants were asked to learn as many of the words as possible to prepare for a vocabulary test. They were further told that the test would require recall as well as recognition of the words. They were given 20 minutes to learn the vocabulary, but they were not told about the amount of time they had and were asked not to look at a clock while learning. Additionally, it was pointed out to the participants that questions regarding the images could not be answered. If they were not able to identify an image, they were asked to remember the image so that they would be able to describe it afterward.

After the learning time, the participants were asked to play the game 2048[4] for three minutes as an interference task to remove the words from the working memory. Then they were asked to fill out the test on recall and recognition of the vocabulary on a computer. The test was implemented in Unity using 3D models of the images used on the cards. The virtual environment was comparable to that in the game. Finally, the participants were asked to fill a post questionnaire containing the NASA Task Load Index (TLX) [15] questionnaire, the System Usability Scale (SUS) [4] and questions on perception of time, fun and motivation.

***Word Saber* condition** In the *Word Saber* condition, the participants were also asked to complete the questionnaire on demographics, including the self-report on their Japanese knowledge level. Additionally, information on VR and gaming experience was gathered in this condition, and the Simulator Sickness (pre-) Questionnaire (SSQ) [16] was run. Then, the participants were told that they were going to play a virtual reality game to learn Japanese vocabulary. They were asked to try to remember as many words as possible to prepare for a vocabulary test. The participants were told that the test would require recall as well as recognition. The basic game concepts and the three game-phases were explained

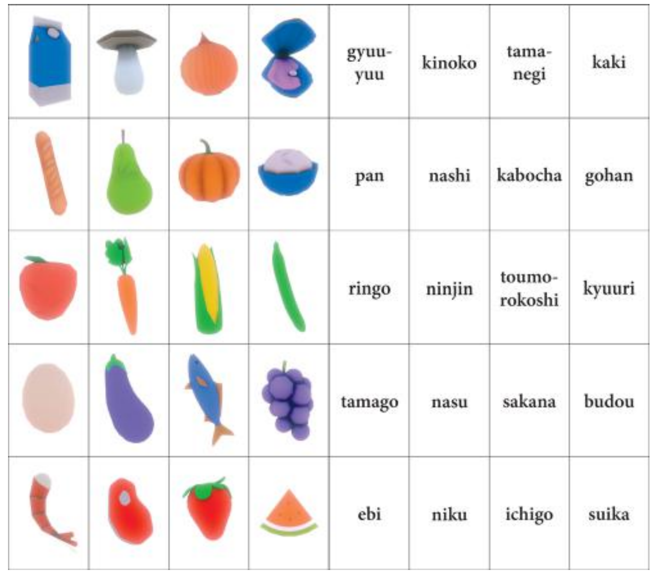

Figure 5: Overview of the flashcard images and corresponding words. For example, the card showing a carrot had the word *Ninjin* written on the back of the card.

[2] https://unity.com/
[3] https://www.vive.com/us/product/vive-pro/
[4] 2048game.com

shortly, and they were free to ask questions. They were not told how long the game would take.
Afterward, the participants were asked to put on the VIVE headset and were handed the controllers and told to start the game when they are ready. The game took about 20 minutes and contained four rounds of the game. In the first round, five words were introduced to the player, and the round was completed with these five words. In the next round, five new words were introduced in the first phase and repeated in the second phase. In the third phase of the second round, all ten words introduced by now appeared in a quasi-random order. This procedure was repeated in the third round of the game. So in the fourth and last round, the last five new words were introduced, and in the gaming phase, all 20 words appeared. The game phases in the following rounds were each a little longer than the previous one, to ensure that the new words appeared to the same amount in each round.
After they finished the game and took off the headset, the participants were asked to play the game 2048 for three minutes on a computer. The vocabulary test was the same as for the flashcard condition. The order of the images and words was different from the order of appearance in the game. Finally, the participants were asked to fill out the post questionnaire, which contained the SSQ post-test, the NASA TLX questionnaire, the SUS questionnaire, and a post-questionnaire on the perception of time, fun, and motivation. For the Simulator Sickness Questionnaire, they were asked to think of the time when they just took off the headset.

4.2 Measures

In this study, different kinds of measures were taken, which will be discussed in detail in the following.

Scoring Participants were asked to fill out two different vocabulary tests afterward, one for recall and one for recognition.

- Recall: For the recall test, they were asked to fill the blanks next to each object image with the correct word or as much of the word as they could remember.
- Recognition: In this test, the Japanese words were given, and the participants were asked to translate them to German or any language of their choice. If they did not know the translation, they were asked to describe the object as precisely as possible.

Standard Questionnaires The participants' workload was measured with the NASA TLX score, a self-report on perceived demands, overall performance, effort, and frustration level. Based on findings by Bustamante et al. in 2008 [6], only the first part of the original NASA TLX score was used, as the second part, the individual weighting of the items can be "meaningless and misleading".
The overall usability of the game was measured with the System Usability Scale[4] developed by Brooke et al. in 1996. It contains questions on the users' general feelings towards the system and the perceived complexity.
Moreover, the simulator sickness caused by *Word Saber* was measured with the Simulator Sickness Questionnaire [16]. It contains three subscales: Nausea, Oculomotor issues, and Disorientation, which are measured with 16 symptoms. These symptoms are weighted according to an importance factor, which makes the overall SSQ score.

VR & Gaming Experience In the pre-questionnaire basic demographics of the participants were gathered. Moreover, they were asked to rate their Japanese knowledge level on a five-level Likert scale from *None* to *Native speaker*. Participants in the *Word Saber* condition were also asked about their VR and Gaming experience. Questions in this section included former participation in VR studies, average hours of playing per week as well as questions on experience with stereoscopic displays and experience with 3D computer games (the last two as five-level Likert scale from *Very much* to *Not at all*).

Quantitative and Qualitative Feedback The questionnaires contained a quantitative question on fun, motivation, perceived difficulty of the vocabulary, perceived success, and perceived duration. A seven-level Likert scale was used for the following questions:

- Q2: How much fun did you have while learning?
- Q3: Would you have liked to continue learning with the given learning method?
- Q5: How would you rate your motivation to learn the vocabulary?
- Q6: Could you imagine to continue learning vocabulary with the learning method outside of the experiment given that you had access to it?
- Q7: Do you think learning the vocabulary in the given time was too difficult?
- Q8: How successful do you rate yourself in having learned the vocabulary?
- Q9: How would you rate your perceived duration of the experiment?

For each question, the possible answers were indicated with appropriate adjectives in the extremes of the scale. For example, for Q2, one side of the scale was labeled *Very little fun*, and the other one was labeled *Very much fun*.
Moreover, two free text questions regarding time were included:

- Q1: How much time do you think you were given to learn the vocabulary? (in minutes)
- Q4: For how many more minutes could you learn with this method if the experiment would not have been stopped? (Imagine you were given more words and could decide whether to stop.)

Free text answering was chosen here to avoid anchoring effects.
Three qualitative questions were included in the questionnaire. The participants were asked about learning strategies they used, whether they had ideas to improve the vocabulary learning technique, and they could leave additional comments.

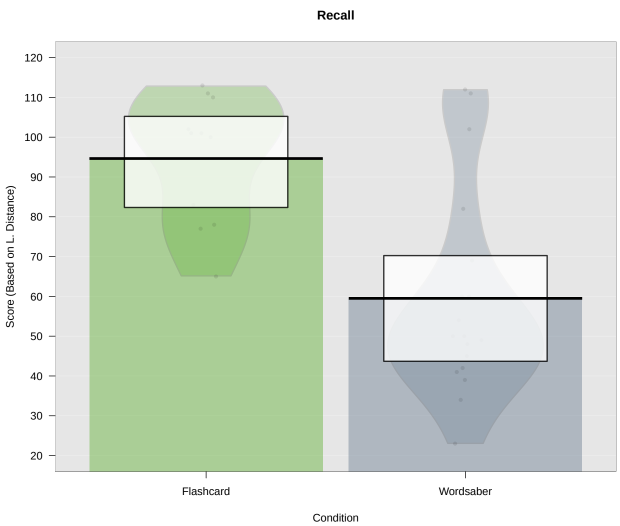

Figure 6: Comparison of the recall values between conditions.

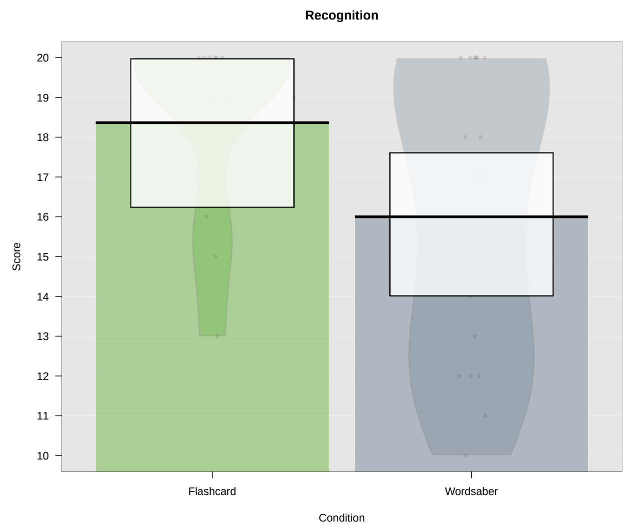

Figure 7: Comparison of the recognition values between conditions.

5 RESULTS

In this section, we present the results of the evaluation. First, we describe the pre-processing of the data. Then, the results for each measure are presented.

5.1 Data Pre-Processing

For the answers gathered in the recall condition, we calculated the Levenshtein distance to the correct answer [25]. The Levenshtein distance measures how similar two words are. It is based on the number of insertions, deletions, or substitutions that are required to transform one word into the other [25].
The data gathered in the recognition questionnaires were manually checked for correctness. Answers were classified as correct in one of the two cases: (i) The correct translation was either identifiable, regardless of the correct spelling, or (ii) a description of the image was given, which clearly showed that the participant had remembered the correct item.

5.2 Recall and Recognition Score

For the recall condition, we performed a Welch's t-test and found a significant effect for learning method (t(27)=.49, p <.05, Cohen's d=1.54, CI=[0.8, 2.6]) with flashcards ($M = 94.6$, $SD = 16.1$) outperforming *Word Saber* ($M = 59.5$, $SD = 26.0$). A boxplot with the mean values is illustrated in Figure 6.

For the recognition condition, we performed a Welch's t-test and found a significant effect for learning method (t(27)=.06, p <.05, Cohen's d=0.72, CI=[0.0, 1.6]) with flashcards ($M = 18.4$, $SD = 2.5$) outperforming *Word Saber* ($M = 16.0$, $SD = 1.56$). A boxplot with the mean values is illustrated in Figure 7.

	Flashcard	*Word Saber*
Mental Demand	25.5	60.0
Physical Demand	21.9	28.9
Temporal Demand	18.8	57.8
Overall Performance	16.1	64.4
Effort	69.0	60.0
Frustration level	27.4	30.0
Overall workload	29.8	50.2

Table 1: NASA TLX scores for both conditions.

5.3 Questionnaires

- **System Usability Score**. Both conditions had similar, very high SUS scores, with 85 for the flashcard condition and 89.5 for the *Word Saber* condition.
- **NASA Task Load Index**. The results for the NASA TLX can be seen in Table 1.
- **Simulator Sickness Questionnaire**. For the SSQ, we compared the results with a Wilcoxon signed-rank test, and there was no significant difference between the pre-questionnaire and the post-questionnaire. They scored with 16.68, and 16.97 for the flashcard and the *Word Saber* condition respectively.

5.4 VR & Gaming Experience

61.5% of the participants in the *Word Saber* condition had participated in a VR-study before. 77% of the participants in that group reported to have experience with 3D computer games, and only 7% had little to no experience with stereoscopic displays. The average gaming time per week in the *Word Saber* condition was 4.7 hours, with 46% reporting no playing time at all.

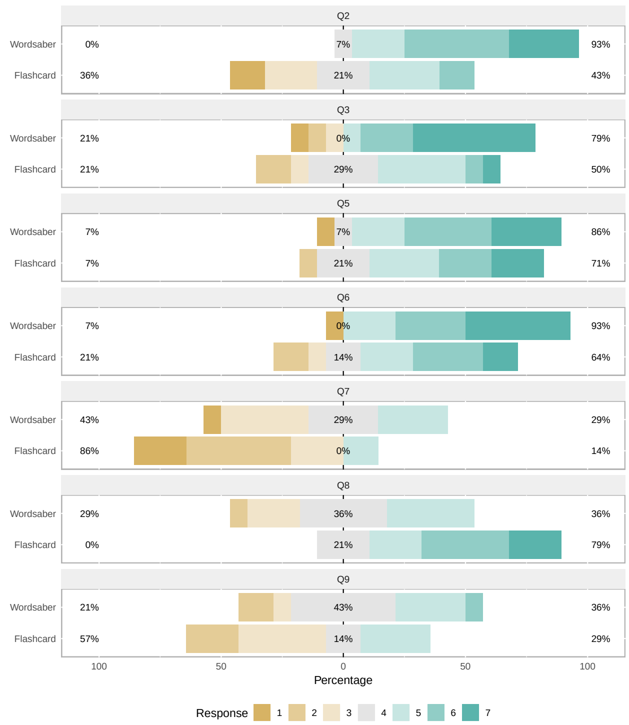

Figure 8: The results of question 2, 3 and 7 were significantly better for the *Word Saber* condition, while for question 8 the results of the flashcards condition were significantly better.

5.5 Feedback from Questionnaires

As the collected Likert ratings by condition were ordered with no spacing assumed and were not normally distributed, we analyzed the data with an ordinal logistic regression [14]. We found significant differences in the following questions:

- Q2 ($p < .001$), with the *Word Saber* condition ($M = 4.93$, $SD = 0.92$) offering more fun while learning than the flashcard condition ($M = 3.07$, $SD = 1.33$).
- Q3 ($p < .05$), with the *Word Saber* condition ($M = 5.57$, $SD = 2.06$) preferred to continue learning afterward than the flashcard condition ($M = 4.36$, $SD = 1.39$).
- Q7 ($p < .01$), with the *Word Saber* condition ($M = 3.71$, $SD = 1.14$) perceived as more difficult in the given time than the flashcard condition ($M = 2.43$, $SD = 1.28$).
- Q8 ($p < .01$), with the flashcard condition ($M = 4.57$, $SD = 1.09$) subjects rating themselves as more successful than the subjects with the *Word Saber* condition ($M = 3.0$, $SD = 0.96$).

We did not find significant differences in the remaining questions. Figure 8 presents the results for all questions.

For the quantitative open question Q1 on the perceived duration presented no significant differences with the *Word Saber* condition ($M = 17.86$, $SD = 4.50$) having similar times than the flashcard condition ($M = 18.0$, $SD = 10.79$). In the same way, the quantitative open question Q4 showed no significant differences regarding the intention of the subjects to continue learning more minutes if they could, with the *Word Saber* condition ($M = 16.57$, $SD = 16.89$) having similar times than the flashcard condition ($M = 15.71$, $SD = 15.55$).

6 DISCUSSION

The results show that *Word Saber* is both effective for vocabulary learning as well as fun, even though it is not as efficient as learning with flashcards. Although, the first hypothesis ("Playing *Word Saber* leads to improved recall and recognition of vocabularies compared to vocabulary learning with the flashcard method") needs to be rejected, the second hypothesis ("Playing *Word Saber* results in more fun than vocabulary learning with the flashcard method") could be confirmed in this study.

The flashcard condition outperformed the *Word Saber* condition in recall and recognition scores. However, the participants in the *Word Saber* condition reported having more fun while learning and would have liked to continue learning. This shows the motivating effect the game has on learners.

Even though the recall and recognition scores were lower than in the flashcard condition, the participants in the *Word Saber* condition could recall and recognize many words afterward. Although the game's efficiency is lower than traditional learning methods, its effectiveness could be shown.

The efficiency of recall and recognition of the flashcard condition might also be biased by the design of the study. The participants in the flashcard condition seemed to be highly motivated since they did not know how much time they were given to learn the vocabulary. The intention behind this was to gather data on the perception of time from both groups. However, it seemed to result in the participants of the flashcard group rushing through the flashcards as fast and highly concentrated as possible. In contrast, the players of *Word Saber* were also highly concentrated but could not adjust the speed of the game. This behavior in the flashcard condition is not representative for an average vocabulary learner.

Furthermore, the participants in the *Word Saber* condition had the additional task to understand the game mechanics and the rules of the game. As this only needs to be done when playing the game for the first time, the recall and recognition scores might be higher in the *Word Saber* condition in a longitudinal study. Consistent with this, the NASA TLX score for the overall workload was higher for *Word Saber* than for flashcards. Interestingly the mental workload and the physical workload were higher for *Word Saber* than for flashcards. In contrast, the effort score, which asks for mental and physical workload, was lower in the *Word Saber* condition. This indicates that although the workload was higher for the game, the users might have perceived it as less demanding.

Still, the higher workload of the game compared to flashcards could cause lower recall and recognition rates as the theory on cognitive load, proposed by Sweller in 1988 [30], suggests. Three different types of cognitive load are distinguished: Intrinsic, extraneous, and germane cognitive load.

While the intrinsic and the germane cognitive load should be the same across conditions as they depend on the content, the extraneous cognitive load, which depends on the way the material is presented, should be higher for *Word Saber*, because of the sound, game mechanics and speed of the game. Furthermore, findings by

Moreno et al. from 2000 on multimedia learning [22] show, that unrelated sounds can hinder learning.

Even with *Word Saber* being less efficient than flashcards, the motivating aspect of *Word Saber* is of high importance. The players of *Word Saber* reported having more fun and wanted to continue playing the game more often, which compensates for the game's lower efficiency. It might also be possible that if they could choose the playing time on their own, they could recall and recognize more words compared to traditional learning methods. Still, the efficiency would be lower, but combined with the perceived fun, efficiency would be less important, as players might have a feeling of spending their free time playing, not studying. This might not only lead to longer playing/learning times but also to a higher frequency, which might have a positive long-term effect on recall and recognition.

The high level of fun while learning vocabulary is of particular importance in this study. One could argue that something called a game should be more enjoyable than learning vocabulary with flashcards. However, this is not necessarily the case for edu-games, as they are often developed by teachers, with a strong focus on effectiveness and efficiency. So this result is very important for the concept of *Word Saber* to prove that effectiveness and fun of vocabulary learning can be combined in a VR game.

The *Word Saber* condition was perceived to be more difficult, which might be related to the overhead of the game mechanics and the rules mentioned above. However, it might also be related to the inflexible prototype design. In the flashcard condition, the participants were free to choose a learning style on their own. They reported having used various kinds of learning methods and approaches to cope with the considerable amount of flashcards. So they could choose whatever learning type they liked or used before. This was not possible in the *Word Saber* group.

A problem of a lower efficiency might be that players underestimate their progress. An interesting finding also is the participants reporting their feeling on success to be lower than in the *Word Saber* condition than in the flashcards condition, which is consistent with the recall and recognition scores in both conditions. This shows that players can estimate their success correctly while playing. So given a lower efficiency of learning; still, the players have a good feeling of how much progress they made.

The main design concept of *Word Saber* is its independence from the learner's first language. So in this study, native speakers of several languages reported successful participation. A drawback of this concept is that the learning success highly depends on the unambiguity of the 3D items and images. For example, the bowl of rice was confused with a cereal bowl by one participant. These mistakes need to be avoided to provide a reliable learning method.

7 CONCLUSION & FUTURE WORK

In this paper, *Word Saber*, a VR edu-game for vocabulary learning, has been introduced. It was evaluated whether learning efficiency and fun can be higher with such a game in comparison to a traditional vocabulary learning method. The findings suggest that games can be designed to be enjoyable for the player while being effective for vocabulary learning.

Overall, this approach should be emphasized more. It could be extended by a multiplayer mode so that players could compete against each other, which might be even more motivating. It could also be possible to add more kinds of words to the game, like adjectives or even verbs. An important point also mentioned by the participants is that more modalities, like writing and speaking, should be included in the game. Furthermore, the long time memory effects of the game should be evaluated. Two different scenarios seem to be interesting here: On the one hand, the long time memory effect of a single learning session and on the other one the development of recall and recognition values and the development of motivation over a more extended period of learning sessions, as discussed above. Participants mentioned that the game could be improved by adding more languages, and it should be possible to add new words to the game. Most importantly, the game should be adaptive to the users' progress. The item speed should be personalized as well as the number of new words, to make the game most efficient for the players.

REFERENCES

[1] Konstantinidis Andreas, Thrasyvoulos Tsiatsos, Theodouli Terzidou, and Andreas Pomportsis. 2010. Fostering collaborative learning in Second Life: Metaphors and affordances. *Computers & Education* 55, 2 (2010), 603–615.

[2] Alan D Baddeley and Graham Hitch. 1974. Working memory. In *Psychology of learning and motivation*. Vol. 8. Elsevier, 47–89.

[3] Beat Games. 2018. *Beat Saber*. Game [Steam Early Access and Oculus Store]. Beat Games, Prague, Czech Republic.

[4] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[5] Amy Bruckman. 1999. Can educational be fun. In *Game developers conference*, Vol. 99. 75–79.

[6] Ernesto A Bustamante and Randall D Spain. 2008. Measurement invariance of the Nasa TLX. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications Sage CA: Los Angeles, CA, 1522–1526.

[7] Alan Cheng, Lei Yang, and Erik Andersen. 2017. Teaching language and culture with a virtual reality game. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 541–549.

[8] Jessamine Cooke-Plagwitz. 2008. New directions in CALL: An objective introduction to Second Life. *CALICo Journal* 25, 3 (2008), 547.

[9] Hermann Ebbinghaus. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences* 20, 4 (2013), 155.

[10] Darren Edge, Kai-Yin Cheng, and Michael Whitney. 2013. SpatialEase: learning language through body motion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 13*. ACM Press, 469.

[11] Irina Elgort. 2011. Deliberate Learning and Vocabulary Acquisition in a Second Language. *Language Learning* 61, 2 (2011), 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

[12] Sarah Garcia, Denis Laesker, Derek Caprio, Ronald Kauer, Jason Nguyen, and Marvin Andujar. 2019. An Immersive Virtual Reality Experience for Learning Spanish. In *Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration (Lecture Notes in Computer Science)*, Panayiotis Zaphiris and Andri Ioannou (Eds.). Springer International Publishing, 151–161. https://doi.org/10.1007/978-3-030-21817-1_12

[13] Eleni Griva, Klio Semoglou, and Athina Geladari. 2010. Early foreign language learning: Implementation of a project in a game–based context. *Procedia-Social and Behavioral Sciences* 2, 2 (2010), 3700–3705.

[14] Frank E Harrell Jr. 2015. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[16] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.

[17] Élise Labonté-LeMoyne, Radhika Santhanam, Pierre-Majorique Léger, François Courtemanche, Marc Fredette, and Sylvain Sénécal. 2015. The delayed effect of treadmill desk usage on recall and attention. *Computers in Human Behavior* 46 (2015), 1–5.

[18] William Rowland Lee. 1965. *Language-Teaching Games and Contests*. Oxford: Oxford University Press.

[19] Jennifer Legault, Jiayan Zhao, Ying-An Chi, Weitao Chen, Alexander Klippel, and Ping Li. 2019. Immersive virtual reality as an effective tool for second language

vocabulary learning. *Languages* 4, 1 (2019), 13.
[20] Masoumeh Mehregan. 2014. Game-Based Tasks for Foreign Language Instruction: Perspectives on Young Learners' Vocabulary Acquisition. *IAFOR Journal of Language Learning* 1, 1 (2014), n1.
[21] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.
[22] Roxana Moreno and Richard E Mayer. 2000. A coherence effect in multimedia learning: The case for minimizing irrelevant sounds in the design of multimedia instructional messages. *Journal of Educational psychology* 92, 1 (2000), 117.
[23] Manuel Palomo-Duarte, Anke Berns, Alberto Cejas, Juan Manuel Dodero, Juan Antonio Caballero, and Iván Ruiz-Rube. 2016. Assessing Foreign Language Learning Through Mobile Game-Based Learning Environments:. *International Journal of Human Capital and Information Technology Professionals* 7, 2 (Apr 2016), 53–67.
[24] Maria V Sanchez-Vives and Mel Slater. 2005. From presence to consciousness through virtual reality. *Nature Reviews Neuroscience* 6, 4 (2005), 332–339.
[25] D Sankhoff and J Kruskal. 1983. Time warps, string edits, and macro-molecules: the theory and practice of sequence comparison. *Addison-Wesley Publication* (1983).
[26] Tassio de Souza Silva, Edwin Carlo Ribeiro Marinho, Giordano Ribeiro Eulalio Cabral, and Kiev Santos da Gama. 2017. Motivational Impact of Virtual Reality on Game-Based Learning: Comparative Study of Immersive and Non-Immersive Approaches. In *2017 19th Symposium on Virtual and Augmented Reality (SVR)*. IEEE, 155–158.
[27] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3549–3557.
[28] Ekrem SOLAK and Gamze ERDEM. 2015. A Content Analysis of Virtual Reality Studies in Foreign Language Education. *Participatory Educational Research* 2, 2 (2015), 21–26.
[29] Penelope Sweetser and Peta Wyeth. 2005. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment* 3, 3 (Jul 2005), 3. https://doi.org/10.1145/1077246.1077253
[30] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.
[31] Hidde P Van der Ploeg, Tien Chey, Rosemary J Korda, Emily Banks, and Adrian Bauman. 2012. Sitting time and all-cause mortality risk in 222 497 Australian adults. *Archives of internal medicine* 172, 6 (2012), 494–500.
[32] Ying Xie, Lanhui Ryder, and Yan Chen. 2019. Using Interactive Virtual Reality Tools in an Advanced Chinese Language Class: a Case Study. *TechTrends* 63, 3 (May 2019), 251–259. https://doi.org/10.1007/s11528-019-00389-z
[33] Yeshuang Zhu, Yuntao Wang, Chun Yu, Shaoyun Shi, Yankai Zhang, Shuang He, Peijun Zhao, Xiaojuan Ma, and Yuanchun Shi. 2017. ViVo: Video-Augmented Dictionary for Vocabulary Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI 17*. ACM Press, 5568–5579.