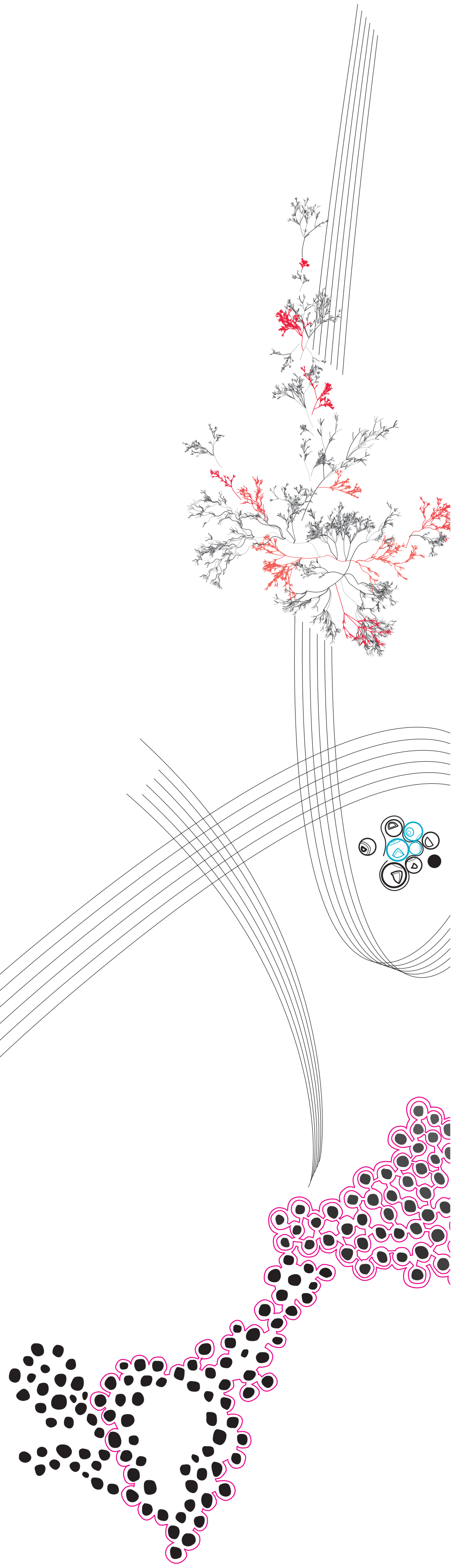
MSc Interaction Technology

Master thesis



Improving spontaneous speech using a pronunciation training game

Emma Tas

Supervisors:
dr. Mariët Theune
dr. Johannes Steinrücke

November 4, 2024

Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Abstract

Language-learning applications have become increasingly popular over the past decade. However, these applications, are not sufficient for language learners to practice the skills necessary for natural speech, such as spontaneous pronunciation. This research explores the use of a serious game to target the spontaneous pronunciation knowledge of Dutch language learners. For this, we developed two versions of a serious game, one with a controlled pronunciation task (control game), and one with a spontaneous pronunciation task (experimental game). We compared the two versions of the game in a user study (n=23), during which participants were randomly assigned to play either the control game (n=11) or the experimental game (n=12) for one week. We compared the effects of the experimental game on the participants' controlled pronunciation, spontaneous pronunciation, playing style, and user experience. Our results show that participants in the experimental group played more consistently and rated their game as more challenging than participants in the control group. This suggests that serious games can support Dutch language learners by increasing the challenge of the task, thereby enhancing engagement. While the control group showed improvements in controlled knowledge, the experimental group did not. Neither game improved participants' spontaneous pronunciation knowledge, likely due to their low Dutch proficiency and the challenges of mastering pronunciation rules and game mechanics simultaneously. Therefore, we recommend using serious games for more advanced learners who could benefit from the added challenge of complex game mechanics.

Chapter 1

Introduction

In today's globalized world, mastering more than one language has become an increasingly valuable skill. In Europe alone, 98% of all secondary school students are learning at least one foreign language (European Commission, Eurostat, 2024). These students learn the majority of this language in a classroom, which comes with disadvantages: a typical foreign language student is exposed to the language only a few hours per week, and the type of language they learn tends to be more formal than the language used in everyday social interactions (Lightbown & Spada, 2013).

Classroom-based learning is not the only way to study a foreign language. The use of language-learning technology has become increasingly popular, and many language learners gravitate towards mobile technologies to either substitute or supplement traditional classroom-based instructions (Reinders & Benson, 2017). The fact that the most downloaded educational application, with over 500 million users worldwide, is a language-learning application (*Duolingo*[1]) is proof of this popularity (Blanco, 2022). These language-learning applications offer learners the benefit of practicing their target language anytime, anywhere, and at their own pace (Reinders & Benson, 2017).

However, relying solely on mobile applications to study a foreign language has its downsides, especially when it is the learner's only study method. Because the main focus of most language-learning applications is on vocabulary acquisition, students who practice with these applications often improve in written vocabulary and grammar, but not on their speaking skills, including pronunciation (García Botero et al., 2019; Loewen et al., 2019, 2020). This is problematic because a learner's pronunciation has influences beyond general comprehensibility. Language learners who struggle with pronunciation often feel less confident than their peers with better pronunciation (Zielinski, 2012), and heavily accented speech has a negative effect on credibility (Lev-Ari & Keysar, 2010).

It is thus important for foreign language learners to have opportunities to practice their pronunciation, but opportunities to do so are restricted when relying solely on (mobile) language-learning applications. The limited speaking exercises that these applications *do* offer are typically drill-based exercises, which benefit only novice learners with minimal prior knowledge of the language, as they transition from having limited speaking skills to acquiring some proficiency (Loewen et al., 2020). Once learners move past the novice stage of language learning, they require

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

exercises that allow them to practice the skills necessary for natural, spontaneous speech. This research looks into the use of serious games to target this type of speech, by answering the following research question:

RQ: How does the inclusion of a spontaneous pronunciation exercise in a serious language-learning game influence Dutch pronunciation learning?

Currently, no language-learning applications or games exist that target a language learner's spontaneous pronunciation knowledge. Therefore, the first aim of this research is to answer the following sub-question:

SQ1: How can a Computer-Assisted Pronunciation Training (CAPT) game be designed to effectively target spontaneous pronunciation knowledge of Dutch language learners?

To answer this sub-question, two versions of a CAPT game were developed. These versions differ only in the type of pronunciation exercise they offer. The first version, referred to as the *control game*, has an exercise that is often included in existing CAPT applications. The second version has an exercise that adheres to the requirements of a spontaneous speech exercise and is referred to as the *experimental game*.

We compared the two versions of the game in terms of their effectiveness in improving Dutch pronunciation of Dutch language learners. We also tested how the experimental game affects the playing style and overall user experience of participants, compared to the control version of the game. To address these objectives, we formulated the following sub-questions:

SQ2: What is the effect of the experimental game on learners' controlled pronunciation knowledge, compared to the control game?

SQ3: What is the effect of the experimental game on learners' spontaneous pronunciation knowledge, compared to the control game?

SQ4: How do the two versions of the game influence the playing style of the participants?

SQ5: How does the perceived user experience differ between participants of the control and experimental group, across the subjective metrics of competence, flow, tension/annoyance, challenge, negative affect, and positive affect?

To answer these sub-questions, a user study was conducted. The user study followed a pretest-posttest control group design, during which participants were randomly assigned to play either the control game (n=11) or the experimental game (n=12) for one week. During the pretest and posttest, both types of pronunciation knowledge were elicited. Additionally, participants completed a posttest questionnaire about their game experience, and data on their interactions with the game were tracked.

Our results show a trend for participants in the experimental group to play more consistently throughout the experiment and rate the experimental game as more challenging, compared to the control group. Furthermore, results show an improvement in controlled knowledge for the control group, but not for the experimental group. Neither the experimental game nor the control game improved the participants' spontaneous knowledge. The effectiveness of serious games for improving spontaneous pronunciation seems limited for the participants of this study, possibly due to their low Dutch proficiency. The combined difficulty of learning the pronunciation rules of the Dutch language, as well as mastering the game mechanics might have limited their progress. Therefore, we suggest the use of serious games for more advanced learners, who would benefit from the additional challenge of more complex game mechanics that go beyond what is typically offered in current CAPT applications.

The remainder of this research is structured into six chapters. Firstly, Chapter 2 contains a literature review on Skill Acquisition Theory, pronunciation exercises, computer-assisted pronunciation training, and serious games. Chapter 3 discusses the development of the game, based on insights from the literature and an expert interview. This chapter also includes the methodology and results of a usability test, which shaped the final version of the game. Chapter 4 outlines the methodology for the user study, and Chapter 5 presents the findings of this study. Chapter 6 discusses these results and offers design suggestions for future games. This chapter also addresses the limitations of the user study and provides directions for future research. Lastly, Chapter 7 provides the conclusion of this research.

Chapter 2

Related work

This chapter reviews prior research relevant to this study, beginning with the role of pronunciation in developing language proficiency. It then examines *Skill Acquisition Theory* to provide a theoretical basis for the stages of developing a new skill. Pronunciation exercises targeting different types of pronunciation knowledge are discussed next, followed by a review of computer-assisted pronunciation training tools. The chapter concludes with an overview of the elements that make serious games effective, and their current role in supporting pronunciation learning.

2.1 Role of pronunciation in second language acquisition

Second language learning is commonly described as being composed of four basic skills: speaking, listening, reading, and writing (Blake, 2016). These four skills can be classified into receptive and productive skills, also known as active and passive skills (Sreena & Ilankumaran, 2018). Reading and listening are receptive skills, whereas writing and speaking are productive skills. Speaking is inherently linked to pronunciation: to speak, one must pronounce words (Pawlak et al., 2011).

Current pronunciation instruction methods are mainly focused on improving intelligibility and comprehensibility (Pennington, 2021). Intelligibility can be defined as the degree to which a listener's understanding matches with the speaker's intended message (Tergujeff, 2021). The two main categories of pronunciation features that influence intelligibility are segmentals (e.g. vowels and consonants) and suprasegmentals (e.g. stress, rhythm, and intonation) (Wang, 2022). Intelligibility is measured objectively, for example by having listeners transcribe speech in standard orthography (Tergujeff, 2021). Comprehensibility is a judgment from the listener on how difficult the speaker is to understand and is usually measured subjectively using listener ratings (Tergujeff, 2021). Comprehensibility can also be affected by factors such as vocabulary choice and grammar.

2.2 Skill Acquisition Theory

Learning a new language involves acquiring all of the aforementioned skills. The stages in which people progress when learning a new skill can be explained using Skill Acquisition Theory (DeKeyser, 2020). Three stages of development are recognized in Skill Acquisition Theory: *declarative*, *procedural*, and *automatic*.

Most beginner learners of any skill will first obtain *declarative knowledge*: knowledge of events and facts (DeKeyser et al., 2017). Declarative knowledge, also referred to as *knowing-that*, can be acquired through observation and/or instructions, for instance when a teacher demonstrates a certain skill to a student (DeKeyser, 2020). For example, a beginner guitarist can learn declarative knowledge about a new chord by watching their teacher demonstrate how to play the chord. Learners with declarative knowledge of a certain skill have learned what they should do and when they should do it. However, they are unable to act upon this knowledge unless they are completely focused on form and have enough time (DeKeyser et al., 2017)

Once a learner begins practicing the skill, they start a process of proceduralization (DeKeyser, 2020). Procedural knowledge is also referred to as *knowing-how*, and is knowledge that can only be performed. The learner develops *procedural knowledge* relatively quickly during the proceduralization phase, by engaging in tasks that draw on their declarative knowledge (DeKeyser et al., 2017). As long as the learner uses the necessary declarative knowledge while executing the task, proceduralization can be completed after only a few attempts (DeKeyser, 2020). For example, once the beginner guitarist starts practicing the new chord, their declarative knowledge about this chord is used to form procedural knowledge about playing the chord.

Procedural knowledge alone is not sufficient for the learner to execute the skills fluently, without making any errors. However, with continued practice, the time required to execute the task, the error rate, and the amount of attention required will decrease (DeKeyser, 2020). This process, called *automatization*, requires a considerable amount of time and and for most skills never truly finishes. The guitarist in the previous examples will automatize their knowledge of playing the chord after some months of consistent practice. They can now play the chord without thinking about finger placement, and use the chord seamlessly while playing songs.

Practice plays an important role in skill acquisition, especially to get the learner to a level of automatization in which they can apply the skill at a normal speed with a high degree of accuracy (DeKeyser et al., 2017). The type of practice that is appropriate for the learner is dependent on a few factors, such as the type and amount of prior knowledge, the time and resources available, and the type of skill that is desired as a result of the practice (DeKeyser et al., 2017). When the goal of the practice is to gain declarative knowledge, practice should be focused on providing more understanding to the learner. When the goal is to advance proceduralization or automatization, the respective focus should be on applying the understanding or doing so faster and with less effort.

2.2.1 Acquisition of pronunciation knowledge

Second language learners often first acquire declarative pronunciation knowledge through instruction and demonstration. They learn the specific pronunciation rules for their target language, and when to apply them. This declarative knowledge can be further consolidated with controlled and repetitive practice, such as pronunciation drills with minimal pairs (pairs of words that differ by only one sound) (DeKeyser et al., 2017).

Once the learner has obtained the necessary declarative knowledge, they can move to the procedural stage of pronunciation skill development. This is done with practice that focuses on applying the knowledge they obtained (DeKeyser et al., 2017). These exercises are less controlled and force the learner to prioritize meaning over linguistic accuracy, such as during an object naming task (Saito & Plonsky, 2019). Once the learner has started the process of automatization, the same practice tasks can be used, but with an increased focus on speaking smoothly, without many pauses or hesitations, at a normal speed (DeKeyser et al., 2017). Some forms of practice have a strong focus on increasing automatization, such as participating in an immersion classroom or studying abroad.

To assess the effectiveness of pronunciation instruction and practice methods, it is necessary to elicit the learner's knowledge. Saito and Plonsky (2019) argue that two types of pronunciation knowledge need to be assessed separately: controlled pronunciation knowledge and spontaneous pronunciation knowledge. These types of knowledge are based on the stages of Skill Acquisition Theory. Controlled knowledge is synonymous with declarative knowledge and spontaneous knowledge comprises the knowledge obtained during the procedural and automatic stages of pronunciation skill development.

Controlled pronunciation knowledge can be elicited through controlled speech tasks (Saito & Plonsky, 2019). These tasks only show the learner's controlled pronunciation knowledge when three conditions are met (Krashen, 2013). Firstly, the learner must consciously know the pronunciation rules that the task is targeting. Secondly, the learner must have enough time to think about the pronunciation rules. Lastly, the learner must be able to actively think about the correctness of their pronunciation (focus on form). Usually, these tasks are highly structured, such as fill-in-the-blank exercises and word reading tasks (Saito & Plonsky, 2019).

Similarly, spontaneous pronunciation knowledge can be elicited through spontaneous speech tasks, which are typically characterized by three features (Saito & Plonsky, 2019). Firstly, unlike controlled speech tasks, spontaneous speech tasks should be focused on function rather than form (e.g. describing a picture while using the appropriate tense markers). Next, the task should include time pressure so that learners do not have as much time to access their declarative pronunciation knowledge. Finally, spontaneous tasks are often semi-structured, to ensure that the learner can form their own sentences, while using the specific phonological features that need to be assessed.

We expect that the differences between the different types of pronunciation knowledge also apply when language learners practice their pronunciation through CAPT tools. Learners using a CAPT application that focuses on obtaining declarative knowledge will improve their controlled pronunciation knowledge, but not their spontaneous knowledge. Conversely, language learners using a CAPT application that focuses on increasing automatization will improve their spontaneous pronunciation knowledge, but not their controlled knowledge. This leads us to formulate the following hypotheses for SQ2 (*What is the effect of the experimental game on learners' controlled pronunciation knowledge, compared to the control game?*) and SQ3 (*What is the effect of the experimental game on learners' spontaneous pronunciation knowledge, compared to the control game?*):

H1: Learners who play the experimental game will not show any improvement in their controlled pronunciation knowledge, whereas learners who use the control game will improve their controlled pronunciation knowledge.

H2: Learners who play the experimental game will improve their spontaneous knowledge, while learners who play the control game will not improve their spontaneous knowledge.

2.3 Pronunciation exercises

When a second language learner starts learning the pronunciation of a language, they first acquire declarative pronunciation knowledge through explicit instructions, which is either articulatory-based or auditory-based (Saito & Plonsky, 2019). Articulatory-based instructions are focused on showing students the manner and place of articulation, often using visual materials such as diagrams. Auditory-based instructions are focused on showing learners the differences and similarities between their first language and the language they are learning, e.g. by demonstrating a certain sound by various speakers in various contexts (Saito & Plonsky, 2019).

Pronunciation exercises, like pronunciation instruction, generally involve two main types: perception (auditory) and production (articulatory) (Tejedor García et al., 2020). Perception exercises help learners become more aware of both segmental (individual sounds) and suprasegmental (intonation, stress) elements of language. Common perception exercises include *identification*, *discrimination* and *oddity* tasks (Nagle, 2018; Tejedor García et al., 2020). In identification tasks, learners match a spoken word to its corresponding image or written form. Discrimination involves determining whether two heard words are the same or different. In oddity exercises, learners hear multiple sounds and must identify which one is different from the others.

Production exercises are designed to improve either controlled pronunciation knowledge or spontaneous knowledge (Saito & Plonsky, 2019). Controlled pronunciation exercises focus purely on pronunciation accuracy and include tasks like reading words or sentences aloud, as well as repetition activities where learners repeat words or sentences after hearing them (Nagle, 2018; Saito & Plonsky, 2019). These tasks help learners practice specific sounds or phrases in a highly controlled environment. In contrast, spontaneous production tasks are more concerned with using language naturally for communication, and they involve activities like describing pictures, naming objects or images, and narrating a story (Nagle, 2018; Saito & Plonsky, 2019). These exercises encourage learners to focus on using language for meaning rather than form.

2.4 Computer-Assisted Pronunciation Training

Computer-Assisted Language Learning (CALL) can be broadly defined as any type of digital technology that is used in formal or informal language learning, both inside or outside language classrooms (Chen et al., 2021). CALL applications, such as *Duolingo*[1], *Rosetta Stone*[2] and *Babbel*[3], typically offer exercises for all four basic skills (speaking, listening, reading and writing), and often also include exercises focused on pronunciation. Systems that are specifically designed to target only a learner's pronunciation, are called Computer-Assisted Pronunciation Training (CAPT) systems (Fouz-González, 2020). Examples of these types of applications include *Elsa Speak*[4], Clash of

[1] https://www.duolingo.com/mobile

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

[3] https://www.babbel.com/mobile

[4] https://elsaspeak.com/en/

Pronunciations (Tejedor García et al., 2020) and English File Pronunciation (Fouz-González, 2020).

2.4.1 Strengths of CAPT applications

CALL and CAPT applications can have significant pronunciation improvement results, for both the production and perception of target features (Fouz-González, 2020; Martinelli, 2016; Tejedor-Garcia et al., 2020). The motivation and engagement of the users strongly influence these outcomes: users who practice more and repeat lessons are more likely to improve (Martinelli, 2016; Tejedor-Garcia et al., 2020).

CAPT tools gain more benefits when they can be accessed through a learner's mobile phone. Mobile applications are convenient, easily accessible, and allow learners to practice at their own convenience (Pennington, 2021; Rogerson-Revell, 2021). By independently practicing their pronunciation in settings where they feel comfortable, learners can also increase their confidence. This, in turn, can reduce foreign language anxiety (Pennington, 2021).

Another, more recent, benefit of CAPT applications is that they allow for personalized, immediate feedback. Using speech recognition software, learners can receive individualized feedback on their speech recordings, as is done by applications such as *Duolingo* and *Elsa Speak* (Pennington, 2021). This feedback has been shown to have a positive impact on learners' pronunciation (Rogerson-Revell, 2021).

2.4.2 Weaknesses of CAPT applications

CAPT resources, whether integrated into CALL applications or used independently, typically offer only a small variety of tasks (Rogerson-Revell, 2021). While most applications offer both perception and production tasks, these tasks almost exclusively target controlled knowledge. For example, the exercises included in *Duolingo* that target a user's pronunciation all consist of controlled tasks (*word identification*, *word and sentence reading*, and *word and sentence repetition*).

Although CAPT applications have the potential to give personalized feedback based on the input of the user, most applications give out very generalized feedback, that is limited to simple right/wrong indications (Pennington, 2021; Rogerson-Revell, 2021). It is also not uncommon for the feedback to be incorrect, which is frustrating and demotivating for learners (Rogerson-Revell, 2021). CAPT systems that provide more extensive feedback, such as spectrograms or waveforms, are often not user-friendly. They require some expertise to be interpreted, and often do not give learners enough information to pinpoint the cause of their errors (Rogerson-Revell, 2021).

Lastly, many CAPT applications prioritize technological innovation (e.g., including artificial intelligence), over exploring new ways to innovate pedagogically (Pennington, 2021; Rogerson-Revell, 2021). This is especially noticeable due to the overreliance on controlled pronunciation tasks by most CAPT applications, as those are not sufficient to improve all aspects of pronunciation knowledge of a learner (DeKeyser et al., 2017). Additionally, many CAPT applications still promote themselves as teaching native-like pronunciation to their users (Rogerson-Revell, 2021). This is in contradiction with the trend in current pronunciation instruction methods, which aims to improve intelligibility and comprehensibility instead of achieving native-like pronunciation (Pennington, 2021).

2.5 Serious games

A serious game is a game that is not designed with entertainment as its primary purpose, instead aiming to achieve a specific, non-entertainment goal (Casañ-Pitarch, 2018; Caserman et al., 2020). This goal, also known as a *characterizing goal*, can be educational, health-related, or focused on training (Caserman et al., 2020; Krath et al., 2021). To help players achieve this goal, the game's content should be accurate and relevant, and the game should provide appropriate feedback (Caserman et al., 2020). An effective serious game is able to balance its serious aspects and its game aspects. In other words, it needs to target the characterizing goal, while also offering an interesting gameplay experience (Caserman et al., 2020).

2.5.1 Core elements of effective serious games

Ideally, a player reaches a state of flow while playing a serious game, during which they are deeply immersed and focused (Calvillo-Gámez et al., 2015). A game can encourage players to reach this state by including several elements. Firstly, the game should maintain the player's concentration, by having an appropriate workload, as well as immersing the player in a world with interesting details and a captivating storyline (Desurvire & Wiberg, 2009; Sweetser & Wyeth, 2005). The game should also be challenging in a way that is appropriate for the skill level of the player (Desurvire & Wiberg, 2009; Hamari et al., 2016). Clear goals are crucial, both short-term and long-term, and should be presented to the player in an engaging way, such as through cut scenes or mission briefings (Calvillo-Gámez et al., 2015; Desurvire & Wiberg, 2009). Additionally, the game should have an intuitive interface, clear controls, and appropriate feedback mechanisms (Desurvire & Wiberg, 2009). Through this feedback, the player can track their progress, which in turn can help keep them motivated.

Serious games with these elements have positive effects on both students' behavior, as well as their affect (Krath et al., 2021). Behavioral outcomes include increased participation and engagement, as well as improvements in the performance on academic and work tasks (Krath et al., 2021; Landers et al., 2017). Similarly, serious gaming increases the learner's feelings of competence more than students who learned through classroom instructions (Bakhuys Roozeboom et al., 2017). In terms of affective outcomes, serious games increase motivation, engagement, and enjoyment of an activity (Bakhanova et al., 2020; Krath et al., 2021). In turn, these factors increase the player's interest to continue the game, as well as the likelihood of them returning to the game and recommending the game to others (de Almeida & dos Santos Machado, 2021).

2.5.2 Serious games used for pronunciation training

In computer-assisted pronunciation training, most applications are not designed as serious games, but instead incorporate gamification elements. These elements include avatars, achievements, leaderboards, and performance graphs (Tejedor-Garcia et al., 2020). These social competition elements (such as leaderboards) positively influence motivation and encourage learners to play more regularly. (Tejedor-Garcia et al., 2020).

Serious games focused on pronunciation are relatively uncommon; however, both *LINGO Online* and *Spaceteam ESL* have been effective tools for improving English pronunciation training (Berry, 2021; Trooster et al., 2017). Dutch primary school students who played Lingo Online, improved their English pronunciation notably more compared to students who did not play the game. Additionally, the game group was also more motivated to learn (Trooster et al., 2017). Similarly, students who played Spaceteam ESL, a collaborative mobile game in which students have to give each other instructions, outperformed their peers who relied on traditional paper-based methods (Berry, 2021). The engaging nature of the video game also made the learning process more enjoyable for the students who played the mobile game.

We expect that a CAPT game incorporating a spontaneous knowledge exercise will allow users to experience the benefits of a serious game more strongly than one using a controlled knowledge task. This expectation stems from the fact that a spontaneous knowledge exercise aligns more closely with typical game mechanics, as players have to focus on additional (gameplay) elements, as well as perform under time pressure. This leads us to formulate the following hypotheses for SQ4 (*How do the two versions of the game influence the playing style of the participants?*) and SQ5 (*How does the perceived user experience differ between participants of the control and experimental group, across the subjective metrics of competence, flow, tension/annoyance, challenge, negative affect, and positive affect?*):

H3: Participants in the experimental condition will spend more time playing the game compared to participants in the control condition.

H4a: Participants will respond more positively to the experimental game than to the control game. This will improve the user experience metrics *challenge*, *flow*, *positive affect* and *competence*.

H4b: Participants will respond more positively to the experimental game than to the control game. This will decrease the user experience metrics *negative affect* and *tension*.

2.6 Contributions

Current CAPT applications typically only offer exercises that target a language learner's controlled knowledge. These applications thus fall short for learners who have already acquired declarative knowledge about the pronunciation of their target language. Instead, these learners would benefit from exercises with which they can advance their procedural knowledge. This research aims to evaluate a CAPT game that uses a spontaneous pronunciation exercise for Dutch language learners. This is done by developing two versions of a CAPT game: one including the spontaneous task and one including a controlled task. Through a user study, we evaluate the effectiveness of both versions on learners' controlled and spontaneous knowledge, as well as the user experience of the players of both games.

Chapter 3

Game Development

This chapter describes the game development process in three main phases: Concept Development, Prototype Development, and Game Refinement. During the Concept Development phase, we decided upon the game's core elements based on earlier literature and an expert interview. The section on Prototype Development describes the different elements that make up the prototype of the game. The section concludes with the methodology and results of a usability test using this prototype. Lastly, the Game Refinement section describes the changes that were made based on the usability test in order to create the final version of the game.

3.1 Concept Development

There are multiple core elements that are necessary to create the basis of the CAPT game. Firstly, since this game targets Dutch learners' spontaneous pronunciation knowledge, the game must include an exercise that specifically targets this knowledge. Another key aspect is the technical framework, which includes the game engine for development, the database for storing user data, and the Automatic Speech Recognition System to assess the user's pronunciation. Lastly, the game must adhere to specific design requirements relevant to a language learning game, which we determined using an expert interview with a Dutch teacher.

3.1.1 Initial concept and requirements

Developing an effective CAPT application that learners can use to advance their spontaneous knowledge requires designing an exercise that specifically targets this type of knowledge. Such an exercise needs to fulfill two requirements. Firstly, the learner must be able to use the exercise to apply their existing declarative pronunciation knowledge (proceduralization), as well as learn to apply this knowledge faster and with less effort (automatization) (DeKeyser et al., 2017). Secondly, the exercise needs to elicit the learner's spontaneous pronunciation knowledge, so that it can be assessed. Spontaneous knowledge is elicited when three requirements are met: 1) *the focus of the task is on function rather than form*, 2) *the task includes time pressure*, and 3) *the task is semi-structured to allow for autonomy of the user* (Saito & Plonsky, 2019). Undertaking an exercise that elicits spontaneous knowledge can also be used to advance proceduralization and automatization, particularly when a learner can do the exercise multiple times as practice. Thus, for the design of the spontaneous pronunciation exercises, the three requirements for eliciting spontaneous speech will be followed.

A ‘perfect’ spontaneous pronunciation exercise would consist of free speech with a (native) speaker in the target language, allowing the learner to convey their message autonomously while experiencing the time pressure of a real conversation. However, it is not desirable to implement a task mimicking free speech into an application that is purely focused on pronunciation, as it requires a comprehensive language proficiency that extends beyond just pronunciation. Instead, gamification will be used as a basis for the spontaneous pronunciation exercise. A game compels the learner to focus on in-game elements and strategic decision-making, instead of being able to solely focus on the correctness of their pronunciation. Many games include some form of time pressure, and players of a game typically have the autonomy to make their own choices throughout a game (Deen, 2015). Thus, a game can be an effective basis for a spontaneous pronunciation exercise, while also demanding fewer additional language skills, such as grammar.

To measure the effects of including a spontaneous pronunciation exercise in a CAPT application, a second application is needed that is identical in all aspects, except for the pronunciation exercise. This second application will be referred to as the *control game*, as opposed to the *experimental game*, which includes the spontaneous pronunciation task. The control game will use a *word reading task* instead of the spontaneous pronunciation task, as this is a widely used pronunciation exercise in current CALL applications. Both versions of the game are specifically designed to be played on mobile phones, as the quality of voice recordings made using mobile phones is generally better than using laptops and other devices (Vogel et al., 2015). Designing the game for mobile phones also makes the game more accessible, as it allows users to play at their convenience, anytime and anywhere.

3.1.2 Technical Framework

The game’s technical framework consists of three main elements: the game engine in which the game is made, the database that collects user data, and the Automatic Speech Recognition system used to assess the users’ pronunciation.

Game Engine

To create the two versions of the game, the Godot Engine[1] is used. This is an open-source game engine that has experienced rapid growth in popularity, especially in the indie game industry (Holfeld, 2023). The Godot Engine is a cross-platform game engine that can be used to create games for PC, consoles, mobile phones, and web browsers. The game engine is optimized for 2D game development, making it an appropriate choice for the development of our game.

[1] https://godotengine.org/

Data collection

While playing the game, user data is collected to 1) ensure the smooth functioning and management of the game and 2) to measure user activity for analysis during the user study. The user data is saved locally on the user's device, and also sent to a database hosted on Google Firebase[2], using the real-time database and authentication services offered by Google Firebase[3]. The operational data that is saved includes login details, as well as the progress of the user in the game. This data is also stored locally, to prevent potential issues with sending data to Firebase from impacting the user experience.

Automatic Speech Recognition

In order to assess the pronunciation of participants during the pronunciation exercises, the intelligibility of their speech is measured using a readily available Automatic Speech Recognition (ASR) system. The capacity of ASR systems to accurately transcribe non-native speech matches human performance, and thus can be used to replace human annotators for this task (Ivanov et al., 2016; Mulholland et al., 2016). The ASR service of Deepgram[4] is used to assess the speech of the learner while doing the exercise. Although Deepgram has a slightly lower accuracy than other publicly available ASR systems, the processing speed of this system is much faster than other systems (Kuhn et al., 2024). Processing speed is especially important in the context of a game, as it reduces latency (the delay between a player's action and the system's response). High latency reduces the user experience and player performance while playing a game (Halbhuber, 2022).

3.1.3 Expert interview

To gain insights into additional requirements and functionalities for the game, we conducted a semi-structured interview with a Dutch teacher from the University of Twente Language Centre [5], who specializes in teaching Dutch to non-native speakers. He teaches pronunciation and grammar classes, and conducts intake sessions to determine students' Dutch proficiency levels. The interview took place through a video call and took approximately one hour. The participant filled out an informed consent form before the start of the interview. The entire interview was recorded for later transcription and analysis. The expert interview was approved by the Ethical Review Committee of the University of Twente.

An interview guide was created to conduct the semi-structured interview, which can be found in Appendix A. The guide organizes the interview into three main objectives: 1) identifying common pronunciation difficulties, 2) determining current methods for teaching pronunciation, and 3) gaining expert insights on the design of the CAPT application. Additionally, the expert was shown a demo video showing the most recent version of the application at the time of the interview. Three questions of the third objective (Q3.2, Q3.3, and Q3.4), ask the expert about his opinion on the application shown in this demo video.

[2] https://firebase.google.com/

[3] All data collection and storage was done according to GDPR compliance.

[4] https://deepgram.com/

[5] https://www.utwente.nl/en/language-centre/

Expert interview results

Regarding the first objective, the expert indicated that although there is a wide variety of sounds that Dutch learners have difficulties with, a common difficulty is the pronunciation of vowels (*"Usually it's the vowels that are really difficult. This is because Dutch has, depending on how you look at it, thirteen vowels. Most other languages have about five."*, Q1.1). Especially the pronunciation of diphthongs, such as 'ui' (/œy/) are problematic (*"Especially diphthongs are very difficult. The 'ui' is one that almost always goes wrong."*, Q1.2.1). Dutch learners also often struggle with consonant clusters, especially when these are not pronounced the way they are written (*"Most people know that 'ch' should be read as a /x/ sound, but when you read the word 'geschreven' this sound almost completely disappears. You only say an 's' and an 'r'."*, Q1.2.3).

Correcting a student's pronunciation is not the main focus of Dutch lessons, unlike grammar and vocabulary, unless it hinders communication (*"The first priority is communication. Pronunciation can hinder this, but it doesn't have to."*, Q.2.1.2). The interviewee indicated that there are multiple methods he uses to teach pronunciation, such as by using visuals (e.g. diagrams of the vocal tract) to show where different sounds are made. More importantly, however, is that a student needs to be able to hear the sound correctly before they can produce the sound correctly (*"In principle, you cannot produce a sound correctly until you can hear the sound correctly."*, Q2.1).

An ideal pronunciation application would, according to the expert, be based on the user's language background (*"If everything was possible, I would personalise the application based on language background. So, let people indicate which languages they know best, so that the app can somehow take that into account."*, Q3.1). Another important aspect of a pronunciation application would be that it includes example words with audio (*"It is very important that people can hear the difference between their pronunciation and the model pronunciation"*, Q3.1).

Implications for the application design

The expert interview influenced several key features of the CAPT game design. Firstly, the main focus of the game is placed on the Dutch vowel sounds, as these are the sounds that Dutch learners have the most difficulties with. Consequently, the game is structured with different levels, one for each vowel sound. Secondly, instead of having a rigid level structure that forces users to complete one level before moving to the next, all levels are available to the user. This allows users to personalize their learning by focusing on the sounds they struggle with. Lastly, the interview made clear that it is important that users of the game also have access to example audio so that they can hear the proper way to pronounce the vowel sounds. These example audio and general pronunciation instructions are included in a pronunciation guide, which is added to the game. A more detailed description of these features and their design is given in Section 3.2.2.

3.2 Prototype Development

In this section we describe the prototype of the game, starting with an explanation of the lexicon that serves as the basis of the pronunciation exercises. This section also provides a detailed description of the various components of the game, including the *tutorial*, *pronunciation notebook*, *battles*, *levels*, and *general functionalities*. Finally, with the use of a usability test, we identify issues and areas for improvement of the prototype.

3.2.1 Lexicon

Based on the findings from the expert interview, the game focuses on the Dutch vowel sounds. These findings are in line with earlier research by Neri et al. (2006), in which vowels (both diphthongs and monophthongs) were identified as more problematic than consonants for foreign Dutch learners. For this research, we focus on 13 vowel sounds, an overview of which can be found in Table 3.1. These sounds consist of the most problematic vowel sounds identified by Doremalen et al. (2013) and Neri et al. (2006), as well as their most common incorrect realizations (Cucchiarini et al., 2009; Doremalen et al., 2013; Neri et al., 2006).

Target phoneme	Example	Incorrect realization	Example
/ɑ/	man	/aː/	maan
/aː/	maan	/ɑ/	man
/ɛ/	leg	/eː/	leeg
/eː/	leeg	/ɛ/	leg
/ɪ/	lip	/iː/	liep
/iː/	liep	/ɪ/	lip
/ɔ/	bot	/oː/	boot
/oː/	boot	/ɔ/	bot
/ʏ/	bus	/u/, /y/	boek, buur
/y/	buur	/u/	boer
/øː/	deur	/u/, /y/, /oː/	doek, duur, door
/œy/	huid	/ʌu/	houd
/ɛi/	wijs	/eː/	wees

Table 3.1: Frequent Dutch pronunciation errors and their most common incorrect realizations.

The game is divided into levels based on the different vowel sounds. However, for certain vowel sounds, it is most effective to practice them alongside the sound they are frequently confused with (Neri et al., 2006). This is specifically the case for the tense and lax vowels /ɑ/-/aː/, /ɔ/-/oː/, /ɪ/ - /iː/ and /ɛ/ - /eː/ as they are frequently confused, as well as /ʏ/ and /y/, which both often result in /u/. For these vowels, the Dutch spelling can also cause confusion regarding pronunciation, as vowels spelled with an identical letter can be pronounced in a different way (e.g. 'o' in 'kom' is pronounced /ɔ/, whereas 'o' in 'komen' is pronounced /oː/) (Nunn, 2006). This inconsistency also makes it more beneficial for Dutch learners to practice these sounds together within the same level. As a result, a total of eight levels were created: Level Aa (/ɑ/-/aː/), Level Ee (/ɛ/ - /eː/), Level Ei (/ɛi/), Level Eu (/øː/), Level Ie (/ɪ/ - /iː/), Level Ui (/œy/), Level Oo (/ɔ/-/oː/) and Level Uu (/ʏ/-/y/).

A different word list is created for each of the eight different levels. Every word list consists of 40 different words (for a total of 320 words), which can be divided into three categories: 1) words without additional difficulties, 2) words with a minimal pair, and 3) words with additional

Category 1	Category 2	Category 3
neus	beuk (pair with boek)	spreuk
heuvel	keuken (pair with koken)	deuntje
kleur	veulen (pair with voelen)	keukendeur
jeuk	heup (pair with hoop)	goedkeuring
leuk	reuzen (pair with rozen)	augustus

Table 3.2: Example words for the three categories, for the /øː/ sound.

difficulties. Examples of words for each of these categories can be found in Table 3.2. The first category consists of 20 one- or two-syllable words that do not have any other features that foreign Dutch learners struggle with, such as consonant clusters or the inclusion of the /x/ sound (Neri et al., 2006). The second category consists of 10 words that form a minimal pair with a word with the incorrect realization of the target vowel. These words are included, because errors that lead to the realization of a completely different word are likely to hinder communication (Cucchiarini et al., 2009). For example, the incorrect realization of /oː/ in 'boom' (tree) as /ɔ/ in 'bom' (bomb) can greatly change the meaning of a sentence. For the vowel sound /œy/ not enough minimal pairs exist to create a 10-word list, and thus this category is supplemented with words from the first category. The last category consists of words that contain additional difficulties, such as three or more syllables, consonant clusters, and the inclusion of the /x/ sound. The complete word list used for the final version of the game can be found in Appendix B.1.

3.2.2 Description of the prototype

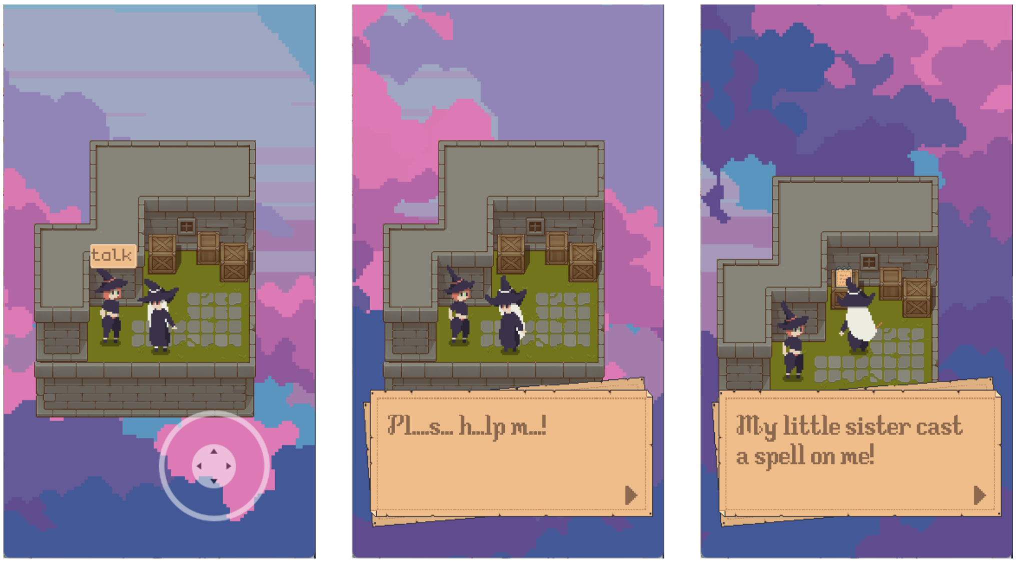
The game is structured around multiple key components. The first is the *tutorial* scene, in which the storyline is introduced, and the mechanics of the game are explained to the user. The second component is the *pronunciation notebook* that contains explanations of the Dutch vowel sounds, as well as example audio of different words with these vowels. Next are the eight different *levels* that the user can explore. The fourth, and main component of the game is the *battles*, with which the user can practice their Dutch pronunciation. Lastly, the game has some *general functionalities*, such as the login and registration features.

Tutorial

After creating an account, the player first enters a tutorial scene. Adding a tutorial to a serious game is beneficial, as it allows players to learn and get used to the game's interface (Ravyse et al., 2017). In this tutorial, the player has to move their character (the *White Witch*) to get to the different elements of the tutorial. This allows the player to get comfortable with moving their character around the game. The three items in the tutorial (the *handwritten note*, the *battle instruction book*, and the *map*) appear one after another (e.g. the battle instruction book only appears after reading the handwritten note). This ensures that the player does not skip over any information in the tutorial.

The tutorial starts when the player talks to the first non-playable character, the *Red Witch*.

The dialogue with this character introduces the storyline of the game. This storyline is included, because stories and plots in serious games can increase the engagement and motivation of learners (Couceiro et al., 2013; Hämäläinen, 2011). The storyline starts with the Red Witch asking the player for help, because her little sister cast a spell on her that made her unable to say any vowel sounds (see Figure 3.1). The Red Witch explains to the player that they might have to battle her sister to get her to undo the spell. When the player finishes the tutorial, and enters one of the eight levels, they meet the sister (the *Blue Witch*), whom they have to battle by doing the pronunciation exercise. The full transcript of the dialogue with the Red Witch and the Red Witch's note can be found in Table B.1 in Appendix B.2.



Pl...s... h...lp m...!

My little sister cast a spell on me!

(a) Start of the tutorial scene. Only the Red Witch is visible.

(b) Part of the dialogue with the Red Witch.

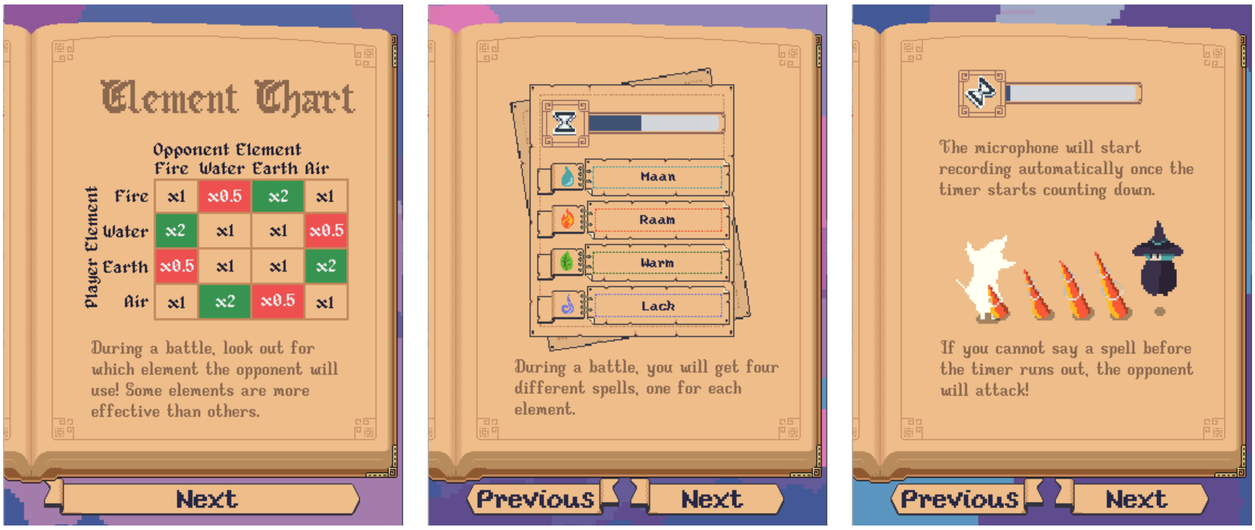
(c) Part of the Red Witch's note. The note is visible after completing the dialogue.

Figure 3.1: Three screenshots from the tutorial scene[6]

Aside from introducing the player to the storyline, the tutorial also contains a book with battle instructions. This book is different for both versions of the game. Figure 3.2 shows three pages of the battle instruction book for the experimental version of the game. The complete battle instruction book for the control version and the experimental version of the game can be found in Appendix B.3.

After the player finishes reading the book with battle instructions, a map appears (see Appendix B.5, Figure B.5). By clicking on this map, the player exits the tutorial scene, and gets access to the eight different levels. The book with battle instructions remains accessible in the levels, allowing players to refer back to these instructions at any time during the game.

[6] To improve readability, text shown in the screenshots uses the font altered after the usability tests (Section 3.2.3).



Element Chart

Opponent Element

Player Element	Fire	Water	Earth	Air
Fire	x1	x0.5	x2	x1
Water	x2	x1	x1	x0.5
Earth	x0.5	x1	x1	x2
Air	x1	x2	x0.5	x1

During a battle, look out for which element the opponent will use! Some elements are more effective than others.

During a battle, you will get four different spells, one for each element.

The microphone will start recording automatically once the timer starts counting down.

If you cannot say a spell before the timer runs out, the opponent will attack!

(a) Page with a chart of the element system.

(b) Page showing an explanation for part of the battle interface

(c) Page explaining when a player gets attacked.

Figure 3.2: Three pages from the battle-instruction book of the experimental version of the game[6]

Pronunciation Notebook

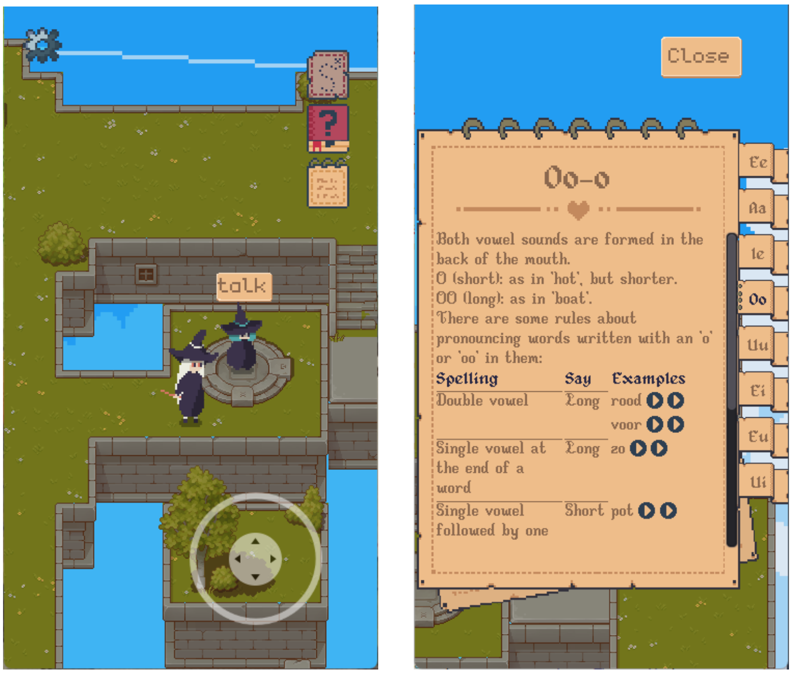
Based on the input of the Dutch teacher during the expert interview, a pronunciation notebook was added to the game, which includes tips on pronouncing the vowel sounds correctly, as well as example audio (Appendix B.4). Once the player first enters one of the eight levels, they get access to this notebook. To open the notebook, the player has to press the icon, as can be seen in Figure 3.3.

The pronunciation notebook consists of explanations on how to pronounce the different vowel sounds, as well as audio recordings of example words. The explanations are based on those found in the Routledge Intensive Dutch Course textbook (Quist et al., 2015). These explanations were supplemented with example words found online[7].

The example audio was created using Narakeet[8] with a male and female voice. The audio fragments were judged by a native Dutch speaker on naturalness. Words that did not sound natural were replaced with more natural-sounding alternatives. For every level, 9 example words were added, pronounced by the two voices. In total 144 audio fragments were added to the game.

[7] https://www.heardutchhere.net/pronunciation_overview.html#vowels

[8] https://www.narakeet.com/



Oo-o

Both vowel sounds are formed in the back of the mouth.
O (short): as in 'hot', but shorter.
OO (long): as in 'boat'.
There are some rules about pronouncing words written with an 'o' or 'oo' in them:

Spelling	Say	Examples
Double vowel	Long	rood voor
Single vowel at the end of a word	Long	zo
Single vowel followed by one	Short	pot

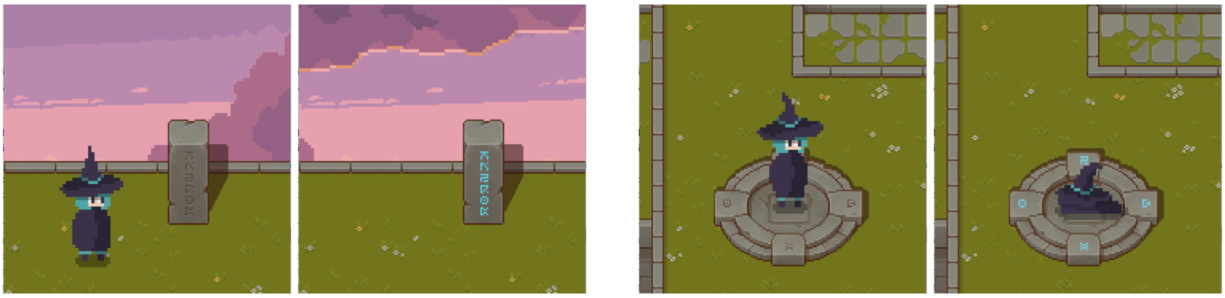
(a) The button to open the pronunciation notebook is located in the top right corner, next to the buttons for the map and the battle instruction book.

(b) A page with explanations on how to pronounce the /oː/ and /ɔ/ sounds.

Figure 3.3: Screenshots the pronunciation book, which can be accessed in any of the eight levels.

Levels

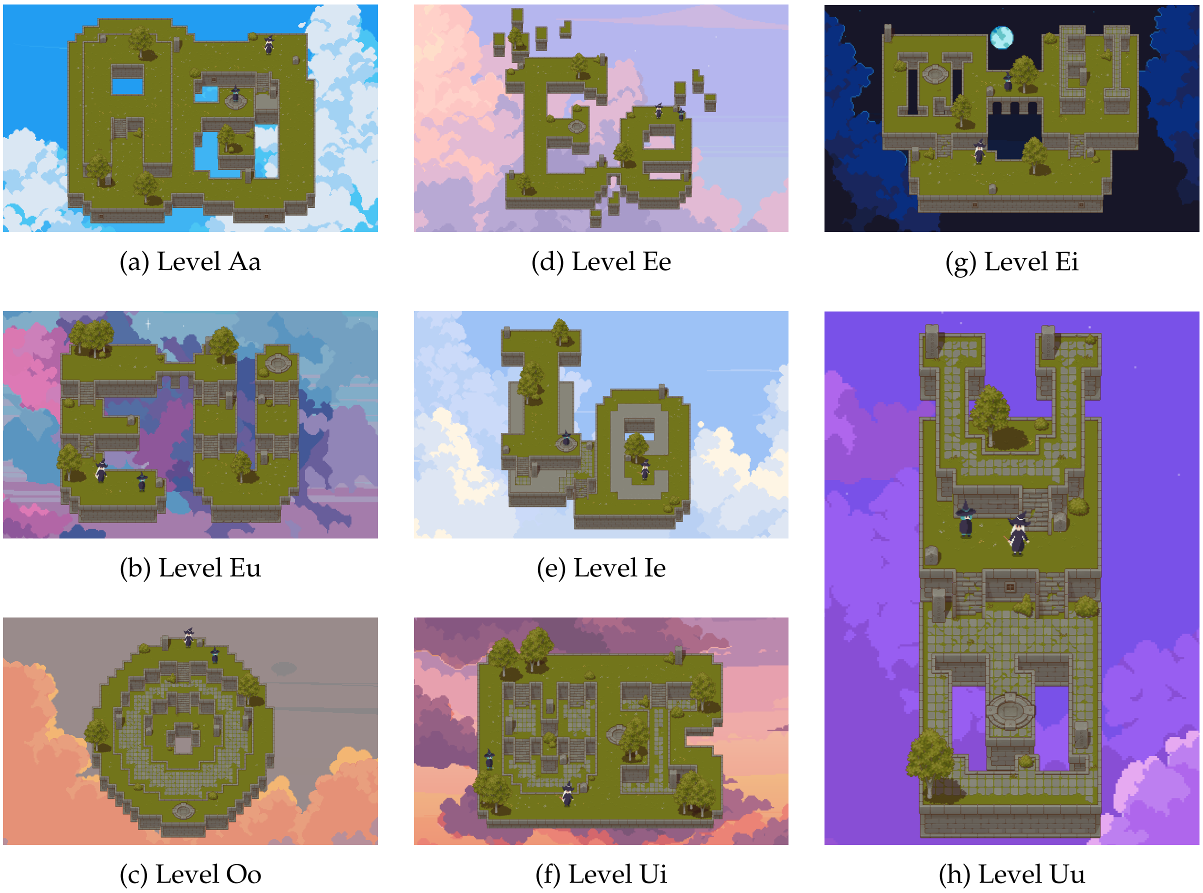
Eight different levels were made, one for each of the target vowel sounds described in Section 3.2.1. The eight different levels each have their own layout and design, as can be seen in Figure 3.5. This was done to encourage exploration and create a sense of novelty, which has a positive effect on player satisfaction (Anolli et al., 2010). The player can access the battle instruction book, the pronunciation guide, and the map from every level. Each level also contains the Blue Witch, whom they can talk with to start a battle.



(a) A stone with ruins before and after winning a battle for this stage of the level.

(b) The stone stage before and after winning the final battle of a level.

Figure 3.4: Two types of stones inside the different levels, that indicate whether or not a battle has already been won. Note that the Blue Witch remains on the stone stage to allow the player to replay the battle for the final stage.



(a) Level Aa

(d) Level Ee

(g) Level Ei

(b) Level Eu

(e) Level Ie

(c) Level Oo

(f) Level Ui

(h) Level Uu

Figure 3.5: Level layouts for the eight different levels. For each level, the target vowel sound is integrated into the level design.

Each level consists of seven *stages*. When a player wins a battle, the stage of the level increases. The stage of the level determines which words are shown during the battle. In the first two stages, the player practices category 1 words. Battles during stages 3 and 4 have category 2 words, while stages 5 and 6 feature category 3 words. The final stage (stage 7) includes words from all three categories. The stage of the level also determines the dialogue of the Blue Witch in which she challenges the player to a battle (see Appendix B.2, Table B.2 for a complete overview).

To further encourage exploration, the location of the Blue Witch changes depending on the stage. This is implemented by making the Blue Witch run away from the player when she loses a battle and having her stop at the next location. Additionally, when a player wins a battle, the *stone with runes* (see Figure 3.4a), starts glowing. This helps the player identify the next location of the Blue Witch, as the glowing stones indicate the battle at that location is already won. For the final stage, the Blue Witch is located on the *stone stage*. After finishing the battle in this stage, the stone stage also starts glowing, and the 'defeated' animation of the Blue Witch plays (as can be seen in Figure 3.4b). The Blue Witch does not change locations after this stage, and the player can replay the final battle to keep practicing.

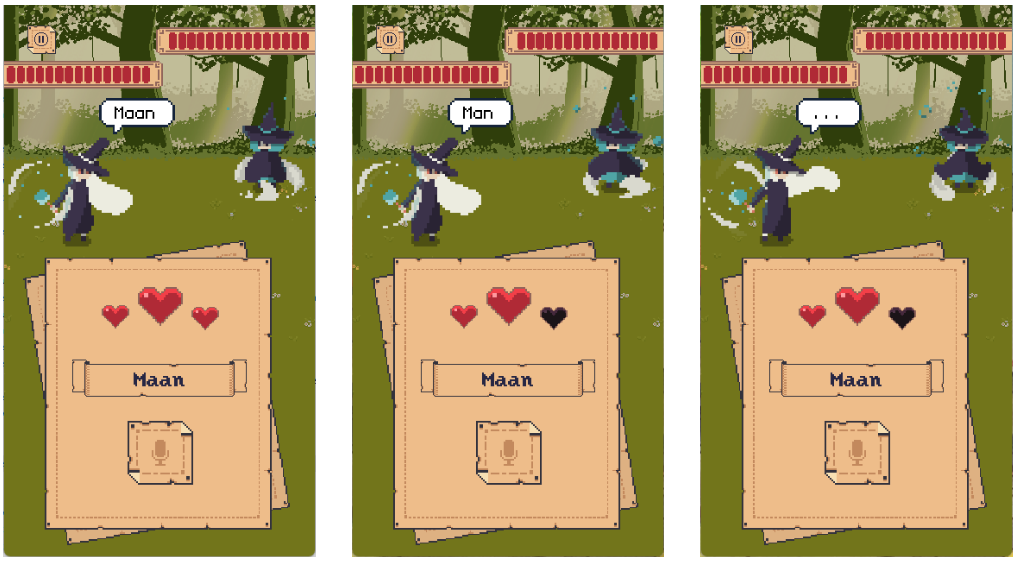
Battle

The two versions of the game are completely identical, except for how the battle system is designed. Some elements of the battles are the same for both versions.

Firstly, for both versions of the game, the course of the battle is determined by the amount of the health of the player and the opponent. Both characters start with 100 health points (HP). When the player attacks (by saying a word correctly), the health of the opponent decreases, and vice versa. When either the player or the opponent loses all their health, the battle ends.

Furthermore, the appearance of the battle is also kept as similar as possible. This is done by, e.g., using the same background, the same interface elements, and by showing almost the same attack animations in both versions of the game.

Lastly, the feedback that a player gets on their speech is the same in both versions, which is shown in a speech bubble next to the White Witch (see Figure 3.6). When the player's speech is not correctly recognized, the speech bubble either shows the incorrect word that was recognized, or it shows "...", when no alternative word was recognized. When the correct word is recognized, this word is shown in the speech bubble.



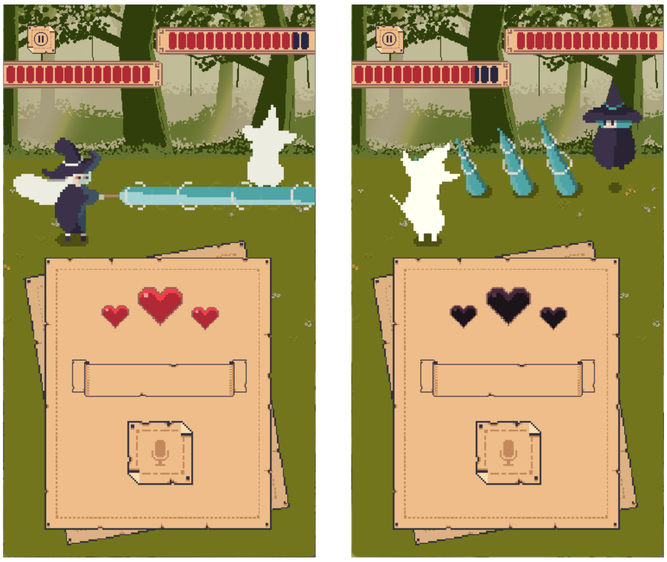
(a) Speech bubble shown when the user pronounces the target word correctly.

(b) Speech bubble with the incorrectly recognized word said by the user.

(c) Speech bubble indicating that no word was recognized.

Figure 3.6: Screenshots of a battle in the control version of the game, showing three different ways in which the user receives feedback on their pronunciation.

Control version In the battle of the the control version of the game, the player practices their pronunciation using a simple word reading task. One word is shown, and the player can decide when to start recording this word by pressing the record button. The player attacks the opponent by saying the target word correctly (see Figure 3.7a). The amount of damage the attack does to the opponent is determined by the amount of hearts the user has. The user can lose a heart by saying a word incorrectly or unintelligibly. If the user still has all three hearts, the attack does 8



(a) The player attacks after saying a word correctly.

(b) The opponent attacks when the players loses three hearts.

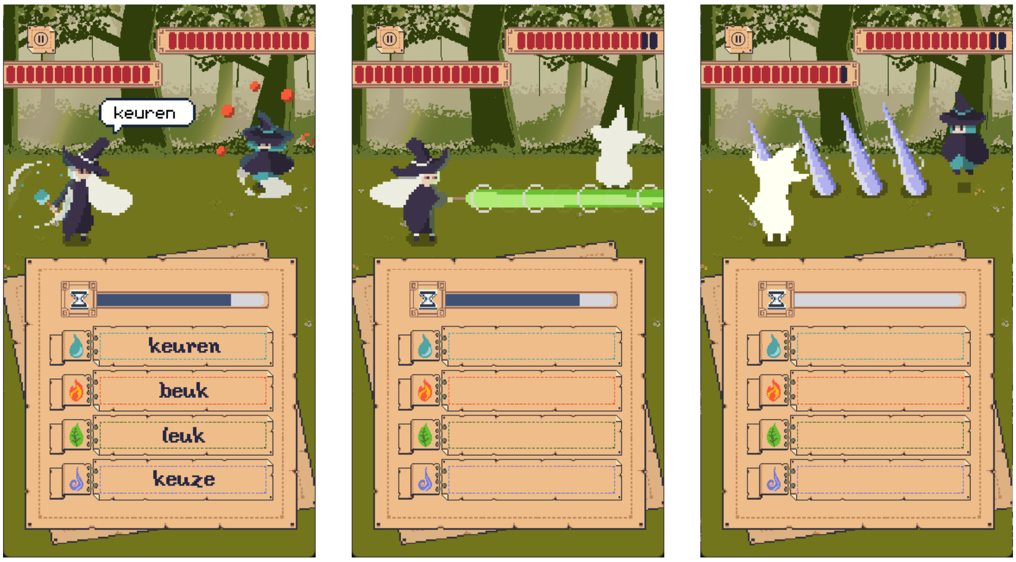
Figure 3.7: Screenshots of a battle in the control version of the game, showing an attack from the user, as well as an attack from the opponent.

HP worth of damage. Each time the player loses a heart, the strength of their attack is reduced by 2 HP. When all hearts are gone, the opponent attacks, and does 5 HP worth of damage to the player (see Figure 3.7b). The hearts refill when the next word appears after an attack.

Experimental version In the experimental version of the game, the player is presented with four different words they can say during the battle. These words each have a label for one of four elements: water, fire, earth, and air, as can be seen in Figure 3.8a. The opponent also randomly displays one of these four elements. When the player says one of the four words correctly, the player attacks the opponent (see Figure 3.8b). The amount of damage that is done to the opponent depends on the element of the word the player said. A regular attack deals 6 HP damage to the opponent. However, when the player chooses the element that is very effective, the amount of damage is doubled. Similarly, when the player chooses the element that is not effective, the amount damage is halved. An overview of the element interactions can be found in the battle instructions book in Appendix B.3, Figure B.2.

Additionally to the inclusion of the different elements, the battle for the experimental version of the game also includes a timer. This timer starts when the four words are shown to the user, and counts down for 10 seconds. When the timer has run out, the opponent attacks and does 5 HP damage to the player (see Figure 3.8c). After an attack by either the opponent or player, four new words are shown, the timer resets, and a new element is randomly selected for the opponent.

As stated in Section 3.1.1, the battle in the experimental version of the game is designed based on the three requirements of a spontaneous speech exercise. Firstly, the battle focuses on function over form: the player has to focus on the opponent's element and choose the word with the opposing element in order to do the most damage. Additionally, the learner experiences time pressure, as the player has to say a word correctly before the timer runs out in order to avoid getting attacked by the opponent. Lastly, the battle gives some autonomy to the player by giving them the freedom to choose which words to say out of the four words that are shown.



(a) The player says the word that is effective against the opponents current element (*water* against *fire*)

(b) The player attacks the opponent after saying a word labelled with the earth element.

(c) The opponent attacks using the air element.

Figure 3.8: Three screenshots of different parts of a battle in the experimental version of the game.

General Functionalities

The game has a few general functionalities that help it run smoothly. Firstly, there are the registration and login functionalities. Although the game was originally designed to be downloaded onto a user's mobile phone, issues with the microphone did not allow this. As an alternative, the game was altered to be played on a web browser. Because of this change, a registration and login system was created to ensure that player data could be effectively sent to the database. This system also allows data to be retrieved when a user wants to continue playing at a later time. The built-in authentication feature of Firebase was used, and a registration and login page was created for this feature, allowing users to sign up and log in with an email address and password. The registration and login pages can be seen Section 3.3, Figure 3.9a and 3.9c respectively. Lastly, a pause menu was added to the battles that allows players to stop the battle and go back to the level.

3.2.3 Usability test

As preparation for the user study (Chapter 4), a usability test was conducted to identify issues and areas for improvement of the prototype. The usability test was approved by the Ethical Review Committee of the University of Twente. This usability test has two main purposes. Firstly, we tested the general user experience of the game, including the different game features and the aesthetics of the game. Secondly, we tested the performance of the Speech Recognition System during the pronunciation exercises. The Speech Recognition System might be biased towards certain voices or accents. For example, it might recognize speech less accurately for deeper voices compared to higher-pitched ones (Feng et al., 2021). To address this, the game was tested by both male and female native Dutch speakers to identify and replace words in the lexicon that are recognized by one voice type but not the other. Additionally, to account for accent-related biases, the system was tested by non-native Dutch speakers with different accents. While it is not feasible to test every possible accent, words that are consistently not recognized, even though the speech of the participant is understandable, were replaced as well.

Participants Demographics

The participants were recruited through convenience sampling. The sample consisted of two foreign Dutch learners and two native Dutch speakers of various ages. For both groups, one female and one male participant was recruited. An overview of the participant demographics can be found in Table 3.3. All participants filled out an informed consent form before the start of the usability test. The test was conducted in Dutch for the native Dutch speakers, and in English for the foreign Dutch learners. All participants had good knowledge of English and were able to understand the instructions and dialogue within the game without issues.

ID	Age	Gender	Dutch level	Native Language
P1	18-24	Male	beginner	English
P2	18-24	Female	intermediate	German
P3	55-64	Female	native	Dutch
P4	25-34	Male	native	Dutch

Table 3.3: Participant information. *Dutch level* is self-assessed by the participants (beginner, intermediate, advanced, native)

Test procedure

The usability test took place in a reserved room on the University of Twente campus (P1, P2, P4) or a reserved room in a public library (P3). These locations were chosen because they are quiet and will thus not interfere with the speech recording while playing the game. Each usability test took approximately one hour. The usability test made use of a think-aloud protocol. Participants were instructed to freely speak their minds while interacting with the game. Participants were encouraged to share their opinions on the applications, and were occasionally asked follow-up questions based on what they shared while playing the game.

The participants were first randomly given access to either the control version or the experimental version of the game. If after 15 minutes a participant had not accessed a certain feature of the game, they were prompted by the researcher to interact with this feature. This ensured that all participants experienced and interacted with the entire game. Participants were required to complete at least two exercises for two unique levels. Additionally, they were instructed to go over both the pronunciation notebook and the battle instruction book. After interacting with all features in the control version, the participants were given access to the other version of the game. For this version of the game, the same instructions were given.

Data collection

During the usability test, observational notes were taken. These notes include comments made by the participants, as well as any behavior that is noticeable or unexpected while playing the game. During the battles, notes were made on the words that were spoken and whether or not they were accurately assessed by the Speech Recognition. For the foreign Dutch Learners, the time it took to finish a battle was also written down.

Results

The participants made multiple suggestions about the general functionalities of both versions of the game. Firstly, two participants mentioned that they thought the background music would get repetitive, and thus annoying after playing the game for some time (P1, P4). They suggested adding an option to turn off the music while playing the game. Secondly, one participant had difficulties with finding the opponent after winning a battle, and mentioned that it would be better if the game was more zoomed out (P1).

The participants also noticed a few issues with the prototype. These issues included some spelling errors (P1), as well as some bugs (P1, P3, P4). Three minor bugs were identified in total: the pause button remained visible after winning a battle, the music restarted every time the participant said a word, and the player could walk through a wall in level 'Aa'. One major issue was identified during the usability test with participant P3. The entire game zoomed in when the phone's built-in keyboard was opened. The participant continued the usability study using the phone of the researcher, as this issue could not be solved and made the game unplayable. This issue seemed to only be present on iPhones, as the other participants who used Android phones did not have this issue when logging into the game.

Participants also had some comments on the aesthetic features of the game. Firstly, three participants mentioned that they found the font difficult to read (P1, P2, P3). Secondly, multiple participants mentioned that they expected more sound effects during the battles. Two participants stated that they would like to hear a sound effect to indicate when a word is recognized (P1, P3). Another participant stated that they expected a sound effect when the game started recording during the control version of the game (P2). Lastly, three participants expressed that they would like to hear sound effects during the attacks (P1, P2, P4).

Generally, the speech recognition system worked well for both versions of the game. One participant noted that in the experimental version, sometimes a different word from the list was recognized instead of the word she was trying to say (P2). However, she mentioned that she did not find this frustrating, and that it did not hinder the game experience. Another participant found the words included in the battles during the fifth and sixth stages of a level (category 3 words) too difficult, and mentioned preferring to have some of the easier words (category 1 words) included as well (P1). A total of 21 words were not recognized when spoken by the native Dutch speakers.

Occasionally this was caused by the ASR recognizing a homophone instead of the target word (e.g. 'wij' instead of 'wei') (P3).

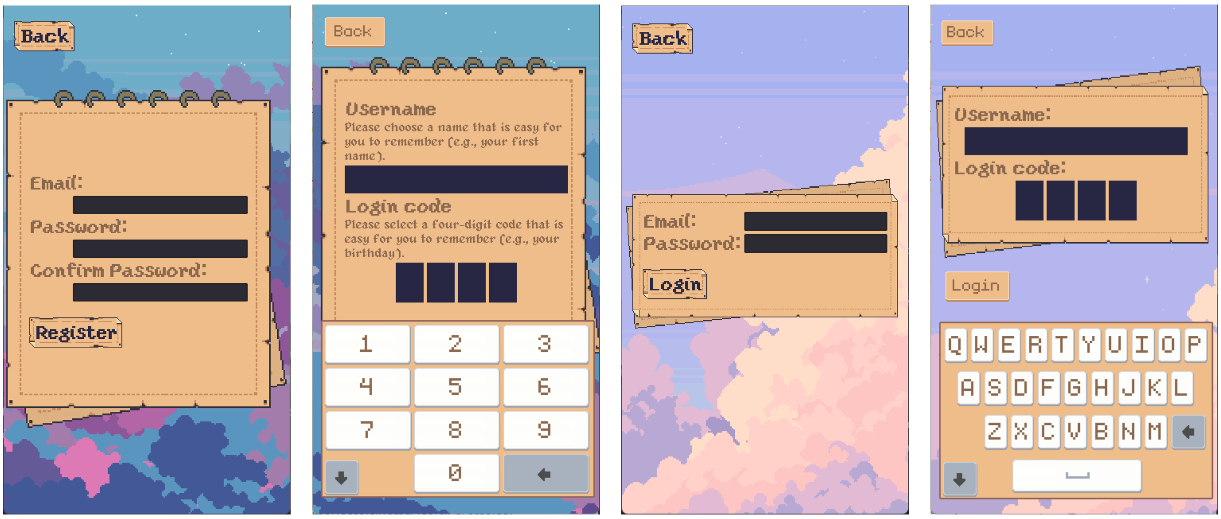
The amount of time it took for the two Dutch learners (P1 and P2) to complete a battle was roughly equal for both versions of the game. In the control version, it took the participant P1 and P2 respectively 3:10 minutes and 5:00 minutes on average to finish a battle. Conversely, it took respectively 2:30 minutes and 5:30 minutes on average to finish a battle in the experimental version.

Control version There were a few issues identified specifically for the control version of the game. Firstly, for two participants it was not clear that losing hearts would mean that the attack does less damage (P1, P4). Both Dutch learners (P1 and P2) became frustrated when the word they were trying to pronounce was not recognized multiple times in a row. One of these participants suggested adding a skip button to move on to the next word (P1). One participant consistently started speaking before the phone microphone was turned on (P3). Another participant also noted that there should be a more clear signal to indicate when the recording starts (P2).

Experimental version Participants noticed more problems with the experimental version of the game. Firstly, multiple participants noted that the instructions on the battle system for this version of the game were confusing (P2, P3, P4). One participant mentioned that the texts explaining the different elements were too long (P4). Two participants would like to see added that the recording starts immediately for this version (P2, P3). During the battle, two participants noted that they expected feedback about the effectiveness of a move (P1, P3). One participant mentioned that the difference between the 'air' and 'water' color of the opponent is hard to see, making the battle confusing (P2).

3.3 Game Refinement

Based on the usability test, a few aspects of the game were changed. Firstly, the registration and login system was changed. Instead of using the phone's built-in keyboard, a keyboard was added to the game. Additionally, to avoid having to build a very complex keyboard, the authentication was changed from email and password to username and code, as can be seen in Figure 3.9. An additional benefit to this method is that it made the login process quicker, as the players had to type out fewer characters.



Please choose a name that is easy for you to remember (e.g., your first name).

Please select a four-digit code that is easy for you to remember (e.g., your birthday).

(a) Registration page designed for the prototype. The registration page makes use of the phone's built-in keyboard.

(b) New registration page. A keyboard with only numbers allows users to fill out a code.

(c) Login page designed for the prototype.

(d) New login page. A keyboard with letters allows users to fill out their username.

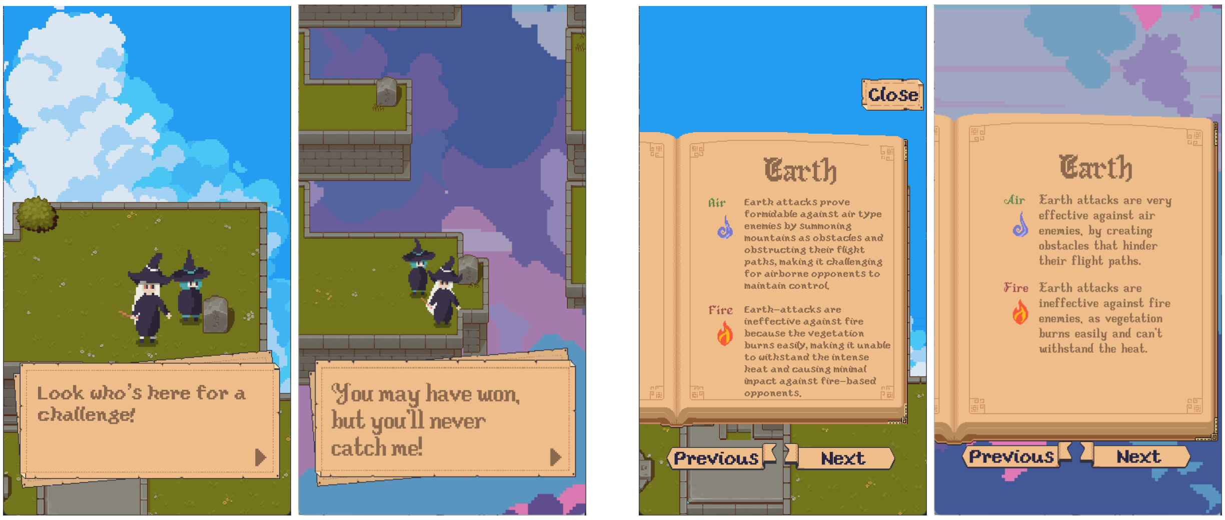
Figure 3.9: Changes made to the registration and login pages of the game. Changes were made based on the usability tests.

The second change was made to the words included in the battles. As was suggested by one of the participants in the usability test, instead of having exclusively one category of words per battle, the later stages of the level also have the words from category 1 included. Additionally, the 21 words that were not recognized by the ASR system were replaced with alternatives. The final version of the word list can be found in Appendix B.1.

The fonts used in the game were changed to fonts that are more easily readable. Examples of the changed fonts can be found in Figure 3.10. Additionally, the texts in the battle instruction book were shortened and altered to be more clear and concise.

A change was made to both versions of the battle. A skip button was added to the control version of the game, which allows players to move to the next word without doing or taking damage (see Figure B.4a in Appendix B.5). For the experimental version of the game, a label was added next to the opponent, displaying the current enemy element (see Figure B.4b in Appendix B.5). This makes it easier to quickly recognize what the opponent's element is. It also makes the game more accessible to participants who are color blind, as they can use the symbols to make decisions on which words to say. Lastly, sound effects were added to the battles. For both versions, sound effects are played when the word is correctly or incorrectly recognized. Additionally, two different sound effects are added to the attacks of the White Witch and the Blue Witch. For the control version, short sound effects are added to indicate when the microphone is turned on and turned off.

Lastly, a settings menu was added to the game, which is accessible in the levels as well as during the battles. In this menu, the player is able to close the game or log out. In this menu, an option to turn off the music and sound effects was also added. Additionally, for the menu inside



Look who's here for a challenge!

You may have won, but you'll never catch me!

Earth

Air Earth attacks prove formidable against air type enemies by summoning mountains as obstacles and obstructing their flight paths, making it challenging for airborne opponents to maintain control.

Fire Earth-attacks are ineffective against fire because the vegetation burns easily, making it unable to withstand the intense heat and causing minimal impact against fire-based opponents.

Earth

Air Earth attacks are very effective against air enemies, by creating obstacles that hinder their flight paths.

Fire Earth attacks are ineffective against fire enemies, as vegetation burns easily and can't withstand the heat.

(a) Dialogue of the Blue Witch, before and after changing the font.

(b) Page of the battle instruction book of the experimental version of the game. Not only was the font changed, the text was also shortened.

Figure 3.10: Screenshots that show the difference between the font in the prototype compared the the font in the final version of the game.

a battle, the user also has the option to quit a battle and return to the level. This menu can be seen in Figure B.3 in Appendix B.5.

Chapter 4

User Study

We conducted a user study to answer our research question, employing a pretest-posttest control group design. The study consisted of a pretest, a week-long intervention in which participants played the game, and a posttest. Both the pretest and posttest were conducted via video calls and took about 20 minutes.

The pretest consisted of a language assessment questionnaire and three pronunciation exercises. After completing the pretest, participants were given access to one of two games, depending on their assigned group. Participants were instructed to engage with the game for at least 15 minutes per day over the course of one week.

The posttest took place about one week after the pretest. This test involved the same pronunciation exercises as in the pretest, as well as a game experience questionnaire. Upon completing the posttest, the participants were granted access to both versions of the game as part of their reward for participating.

4.1 Participant demographics

A total of 26 participants were initially recruited, 13 for each condition. Participants were randomly assigned to one of two conditions: a control group and an experimental group. Recruitment was carried out through personal connections, university group chats, and a participant recruitment page. One participant was recruited via the university's psychology test subject pool and received academic credit for their involvement. Out of the 26 participants, 23 completed the full study: 12 from the experimental group and 11 from the control group. Three participants withdrew from the study during the intervention phase.

Among these 23 participants, 9 identified as women, 13 as men, and 1 as non-binary. The majority of participants were between 25 and 34 years old. The participants came from diverse linguistic backgrounds, with a large variety of native languages and dominant languages (the language the participant uses most frequently), as can be seen in Table C.1 and Table C.2 in Appendix C.1. An overview of the descriptives for both conditions can be found in Table 4.1.

Figure 4.1 shows the self-reported levels of Dutch proficiency for the categories: speaking, understanding spoken language, reading, and native-like accent. Overall, the participants scored their language skills relatively low across all skills, with mean scores below 5 out of 10. On average,

	Age (in years)			Gender			Native languages
	17-24	25-34	35-44	Female	Male	Non-binary	
Control	4	6	1	4	6	1	German, Hindi, Javanese Russian, Sourashtra, Spanish, Tamil
Experimental	2	9	1	5	7	0	Chinese, English, Greek, Indonesian, Italian, Russian, Spanish, Tamil, Vietnamese

Table 4.1: Participant demographics per condition

the participants from the control group reported their own Dutch language skills as slightly better than the participants from the experimental group. This difference is small, however, and not statistically significantly different between the two groups.

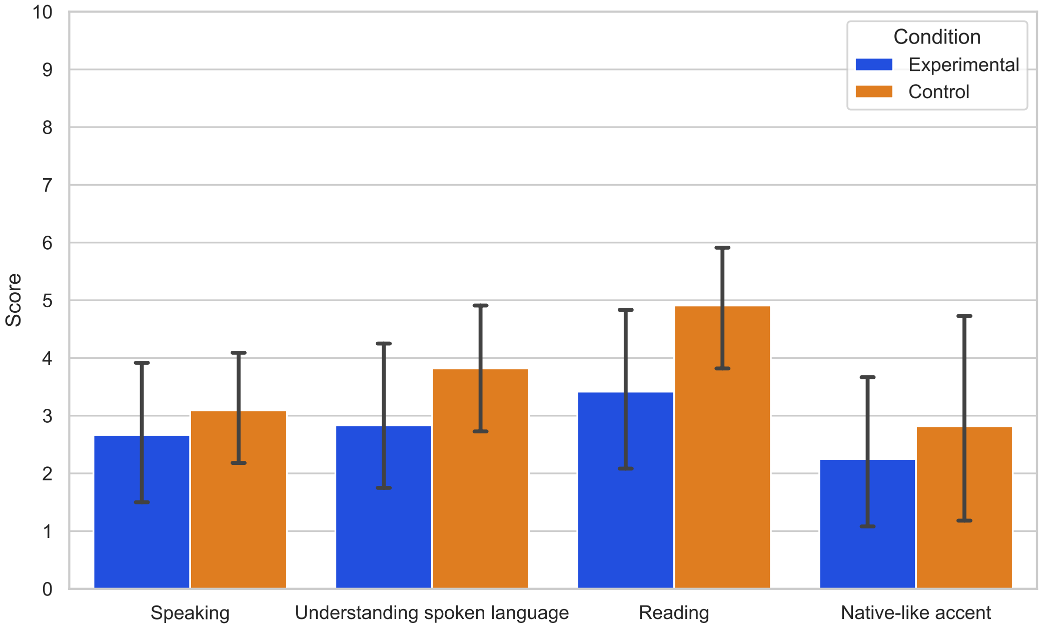


Figure 4.1: Self-reported level of Dutch proficiency, based on responses from the pretest questionnaire.

4.2 In-game data collection

To collect information required for later analysis, data on participants' interaction with the game was sent to the online database. This data includes information on the daily activity of the participants. More specifically, the number of minutes spent playing the game per day, as well as the specific days they logged into the game were recorded.

4.3 Pretest and posttest

The pre- and posttest consist of two questionnaires and two audio-recorded pronunciation assessment tasks: an untimed *word list reading task* and a timed *passage reading task*. The word list reading task involves participants reading a series of isolated words aloud, while the passage reading task requires them to read a continuous text passage aloud. The word list reading task is used to measure the controlled pronunciation knowledge of the target phonemes. The passage reading task is used to measure spontaneous knowledge of pronunciation, and also takes into account the pronunciation of non-target features. Lastly, we use questionnaires to collect general information about the participants and to evaluate the participants' user experience when playing the game.

4.3.1 Word list reading task

A word list reading task is the most commonly used type of assessment for pronunciation training (Mahdi & Al Khateeb, 2019). This task typically uses lists of minimal pairs, as demonstrated by Ghorbani et al. (2016) and Guskaroska (2020). However, the number of minimal pairs containing the target phonemes is limited, and most are used in the game. Including the same words in the game and posttest is undesirable, as that could allow participants to memorize the pronunciation of those specific words, which can incorrectly lead to a higher score on the word list reading task in the posttest. Instead, a list of consonant-vowel-consonant (CVC) words is used for the word list reading task. These words are deemed appropriate for this task, as they are short and do not contain extra difficulties (such as consonant clusters), which allows participants to focus on the correct pronunciation of the vowels. Therefore, if a word is unintelligible, it is very likely due to the mispronunciation of the vowel sound.

los	buur	zuur	mok	tik
muur	pit	ruis	laat	vuil
bok	zaal	sein	leun	luid
mus	top	nut	hut	mes
kuur	das	mep	kijk	raak
pijp	vaas	net	rat	ken
reis	wat	deur	dus	dik
zit	duif	leuk	keus	lap

(a) Pretest word list

buis	zes	keur	heus	zon
reuk	duik	bal	taal	put
buur	vaak	fit	sok	bijt
zeil	pen	wit	muur	vat
sip	ruk	ruil	kom	zus
duim	sap	raad	zeur	zuur
dek	lik	leid	bel	rijm
tong	mat	kuur	maat	bus

(b) Posttest word list

Figure 4.2: Two unique word lists containing words from the Thomas More Lists.

For each of the 10 target phonemes, eight CVC-words were randomly selected from the Thomas More Lists: a corpus of 16 lists each containing 25 Dutch CVC-words (Vanpoucke et al., 2022). Words that also appear in the game were replaced with other words from the Thomas More Lists. For each target phoneme, four words were randomly assigned to the pretest list, and four words were assigned to the posttest list. An exception was made for the target phoneme /y/. For this phoneme only four CVC-words were present in the Thomas More Lists, resulting in their presence in both the pretest and posttest. The order of the words in each list was randomized. This process resulted in two unique lists for the pretest and posttest, each consisting of 40 words. Previous work utilizing similar approaches shows that this is a sufficient number for assessing controlled pronunciation (Ghorbani et al., 2016). The two word lists can be found in Figure 4.2.

4.3.2 Passage reading task

Although spontaneous speech is best elicited through free speech, a passage reading task is used to give an indication of the spontaneous pronunciation knowledge of the participants. The use of a standardized text is preferred, as it prevents errors that could arise due to the participants' word choice (lexical errors) or sentence structure and grammar (morphosyntactical errors) (Cucchiarini et al., 2009). Such errors affect the accuracy and reliability of the assessment, potentially leading to the misclassification of a word or sentence as unintelligible due to factors unrelated to pronunciation. Additionally, a standardized passage reading task allows for comparison within and between participants, since all participants are tested on the exact same sentences.

The texts chosen for the passage reading task are phonetically balanced. This means that they contain phonemes in proportions that reflect their frequency and distribution in natural speech (Radová & Vopálka, 1999). Because these texts mimic the natural occurrence of phonemes, it allows for a more realistic assessment of natural speech production.

Two phonetically balanced texts are used: 'Papa en Marloes' (Van de Weijer & Slis, 1991) and 'De auto' (Martens et al., 2010). These texts comprise a total of 22 sentences (8 and 14 sentences, respectively), which previous research has shown to be a sufficient number of sentences for assessing pronunciation (Franklin & McDaniel, 2016). Both texts are also commonly used in Dutch speech and language therapy to assess the intelligibility of an individual's speech (Beijer et al., 2014; Middag, 2012; Xue et al., 2023). Some low-frequency sounds are not represented in the texts, including the target phoneme /ø/ (Martens et al., 2010). To ensure that all target phonemes are present in the passage reading task, one additional word containing /ø/ (*leuke*) was added to the text 'De auto'. The two texts can be found in Figure 4.3.

Papa en Marloes staan op het station. Ze wachten op de trein. Eerst hebben ze een kaartje gekocht. Er stond een hele lange rij, dus dat duurde wel even. Nu wachten ze tot de trein eraan komt. Het is al vijf over drie, dus het duurt nog vier minuten. Er staan nog veel meer mensen te wachten. Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.

(a) Papa en Marloes

Er was eens een man uit Finland. Hij had veel geld gespaard. Dat was voor de auto van zijn dromen. Hij nam de trein om de *leuke* auto te gaan kopen. Maar de man was bang voor dieven. Hij bewaarde het geld in zijn onderbroek. Hij droomde al van de eerste rit in de nieuwe wagen. Plots moest hij naar het toilet. De man dacht niet meer aan het geld. Het zakje met geld viel recht in de pot. En de man spoelde door. Daar ging zijn fraaie plan! Gelukkig was de politie in de buurt. Die vond het zakje terug op de sporen.

(b) De auto

Figure 4.3: Two phonetically balanced texts used for the passage reading text. The word 'leuke' (in italics) has been added.

4.3.3 Questionnaires

Before completing the two pronunciation assessment tasks during the pretest, participants are asked to fill out a questionnaire (Appendix C.2). This pretest questionnaire collects participants' demographic information and details about their Dutch proficiency and pronunciation skills, using questions from the Language Experience and Proficiency Questionnaire (LEAP-Q) (Marian et al., 2007). Additionally, two open-ended questions about the participants' previous experience learning Dutch were added.

The participants are asked to fill out a second questionnaire during the posttest (Appendix C.3). The core module of the Game Experience Questionnaire is used, which consists of 33 questions to measure the participants' experiences with the game (IJsselsteijn et al., 2013). The questionnaire allows us to make a distinction between the user experience of the control and experimental version of the game, and allows us to better understand the findings from the pronunciation analysis. The questionnaire is used to measure the following six game-experience components on a scale from 0 (not at all) to 4 (extremely): competence, flow, tension/annoyance, challenge, negative affect, and positive affect. A Mann-Whitney U Test is used to compare the results of the control and experimental condition (independent variable) for the six components (dependent variables) ($\alpha = 0.05$).

4.4 Pronunciation analysis

In order to determine whether participants improved their Dutch pronunciation, the recordings of the different tasks from the pretest and posttest are analyzed. The first task (word list reading task) is used to elicit controlled pronunciation knowledge, and the number of correctly recognized words (by the ASR system) is used as a measure of this knowledge.

To measure the participants' spontaneous pronunciation knowledge, the recordings from the second and third tasks (passage reading tasks) are combined and analyzed together. As a general measure of intelligibility, the Word Error Rate (WER) is calculated using automatically generated transcripts of the recordings and comparing these to the original texts of the passage reading tasks. WER is calculated as:

$$\text{WER} = \frac{Insertions + Deletions + Substitutions}{\textit{Number of words in the reference}} \tag{4.1}$$

In addition to the WER, three suprasegmentals are measured: *speech rate*, *pause frequency*, and *pause duration* (Trofimovich & Baker, 2006). Speech rate is measured as the number of spoken syllables, divided by the total duration of the speech (including pauses). Pause frequency and the average pause duration were computed for each participant, using a silent pause threshold of 250 ms (De Jong, 2016). These suprasegmentals are chosen because they influence a listener's judgment on comprehensibility and accentedness (Trofimovich & Baker, 2006).

A paired samples t-test is used for both the control and experimental conditions, to compare if the participants experienced a statistically significant improvement in their pronunciation after using either application ($\alpha = 0.05$). Because of the relatively small sample size for both groups, it is likely that the normality assumption is not met for every measure, in which case a Wilcoxon signed-rank test is used as an alternative. Similarly, the difference in gain scores (posttest – pretest) between the two conditions is tested with an independent samples t-test ($\alpha = 0.05$). In case of violations of the normality assumption or homogeneity of variances assumption, a Mann-Whitney U test is used.

Chapter 5

Results

This chapter presents the findings from the user study, analyzing data collected from both the pretest and posttest, as well as during gameplay. In the sections that follow, we examine the differences between the two conditions in terms of user experience, controlled pronunciation knowledge, and spontaneous pronunciation knowledge.

5.1 User experience

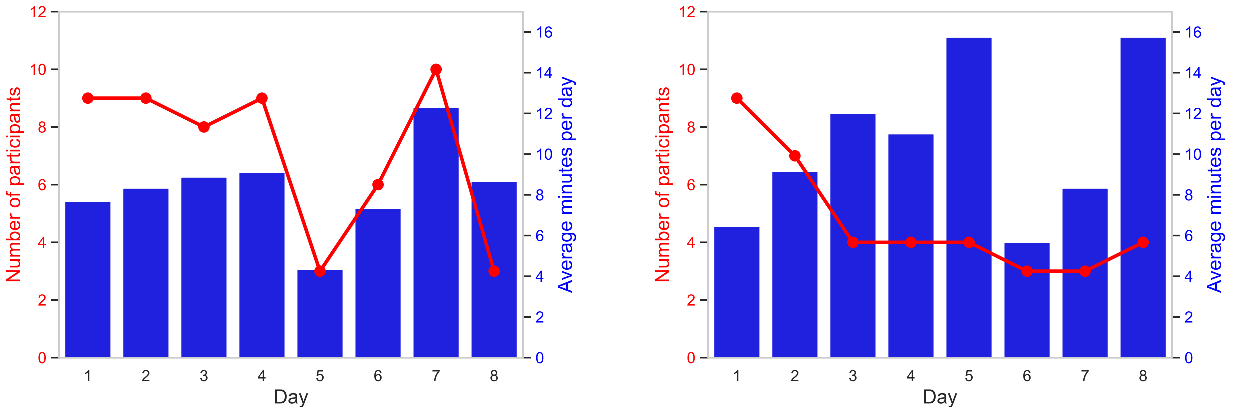
The user experiences of the two participant groups are compared across two factors. The first comparison is made by analyzing the differences in playing style between the participants in both conditions. Secondly, we analyze the answers to the posttest questionnaire, to see if there are any differences in the participants' game experience. The outcomes of the assumption tests for all statistical tests in this chapter can be found in Appendix D.1.

SQ4: How do the two versions of the game influence the playing style of the participants?

During the study, participants played the game for either 7 or 8 days, depending on when they scheduled their posttest. In the control group, 7 of the 11 participants had 7 days between the pretest and posttest. In the experimental group, 9 participants had a 7-day interval between the two tests, and 3 participants had 8 days between the pretest and posttest. An overview of the number of active participants (participants who played the game that day), as well as the average minutes these participants played that day can be found in Figure 5.1.

The number of active participants playing the game fluctuated differently between the experimental and control conditions over the course of the experiment. In the experimental condition, participation was high and consistent during the first four days, as can be seen in Figure 5.1a. In contrast, participation in the control group steadily declined after the first day, with minimal changes after the third day. For the experimental condition, the number of active participants increases during the final days of the experiment (days 7 and 8). This increase is much smaller for the control condition.

The average number of minutes played per day is quite similar for the two conditions. For

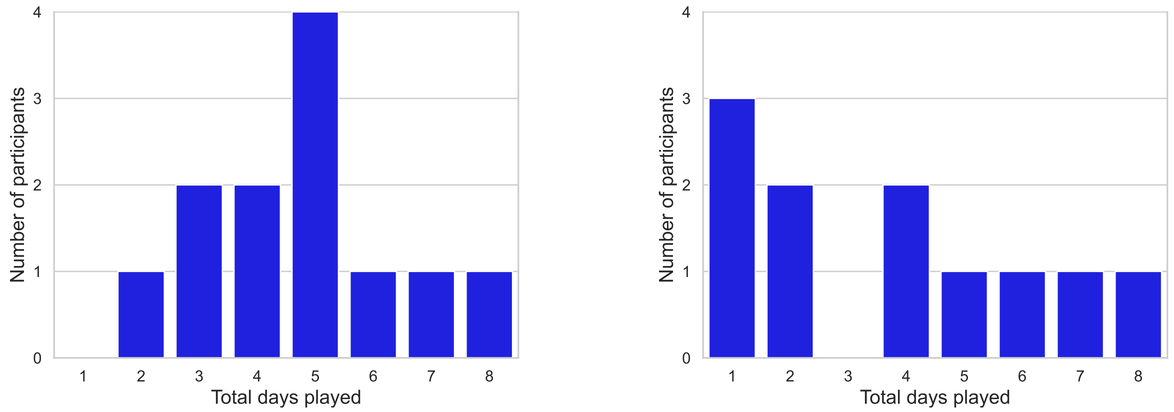


(a) Experimental condition

(b) Control condition

Figure 5.1: Number of active participants per day and average minutes played per day in the experimental and control conditions during the experiment. The average minutes per day were calculated by averaging the playtime across the number of active participants each day.

both conditions, there is an increase in average number of minutes played towards the end of the experiment. Noticeable is a high average playtime on day 5 of the control condition. This is likely because of one participant (P2) who played the game only once, on the fifth day of the experiment, for 46 minutes.



(a) Experimental condition

(b) Control condition

Figure 5.2: Number of participants per total number of days the game was played.

Figure 5.2 illustrates that participants in the experimental group played the game for more days than those in the control group. On average, the experimental group played the game for 4.75 days ($SD = 1.71$), while the control group averaged 3.73 days ($SD = 2.53$). Although the total number of days each participant played the game is different between the two versions, this difference is not statistically significant, $t(21) = 1.143, p = .133, d = 0.47$. The total number of minutes played over the span of the week also does not differ between the two groups. An Independent Samples T-test was conducted to compare the total number of minutes the participants from the experimental condition and the control condition spent playing the game. The test indicated no significant difference in minutes spent playing the game between the control group ($M = 34.06, SD = 7.56$) and the experimental group ($M = 41.38, SD = 4.44$), $t(21) = .851, p = .404, d = 0.36$.

SQ5: How does the perceived user experience differ between participants of the control and experimental group, across the subjective metrics of competence, flow, tension/annoyance, challenge, negative affect, and positive affect?

As shown in Table 5.1, there are no significant differences between the two conditions for any of the game experience components. However, a few trends are worth noting. Firstly, the experimental group had a slightly higher median rank than the control group for the challenge component ($Mdn = 1.90$ vs. $Mdn = 1.60$). Although this difference was not statistically significant ($p = .64, r = .29$), this suggests there might be a slight tendency for the experimental version of the game to be considered more of a challenge. Similarly, the participants of the experimental group gave their version of the game a slightly higher score on the negative affect component ($Mdn = 1.63$ vs. $Mdn = 1.00$).

Component	Control group Median	Experimental group Median	U value	p value	Effect size (r)
Competence	2.40	2.30	60	.710	.08
Flow	1.60	1.50	58.5	.642	.09
Tension/Annoyance	1.00	1.00	65	.950	.01
Challenge	1.60	1.90	43.5	.164	.29
Negative Affect	1.00	1.63	55	.516	.14
Positive Affect	2.60	2.60	64	.902	.03

Table 5.1: Summary of Mann-Whitney U test results for the various game experience components.

5.2 Controlled knowledge

During the pretest and posttest, the participants' controlled pronunciation knowledge was elicited by conducting a word list reading test. In this section, we analyze the number of words spoken correctly during these tasks to compare the two groups' controlled pronunciation knowledge after playing the game.

SQ2: What is the effect of the experimental game on learners' controlled pronunciation knowledge, compared to the control game?

Two participants were not included in the analyses of the controlled knowledge task, one for each condition. For these participants, the audio quality of the posttest recordings was too poor to obtain accurate transcriptions. The analyses were thus done for 10 participants of the control group and 11 participants of the experimental group.

Firstly, a paired samples t-test was conducted to see if there was a difference between the pretest and posttest scores for both conditions. For the participants in the control group, the t-test indicated that their scores after playing the game ($M = 26.70, SD = 5.72$) were significantly higher than pretest scores ($M = 22.40, SD = 7.32$), $t(9) = -4.872, p < .001, d = 1.54$. On the other hand, the difference between pretest scores ($M = 25.00, SD = 4.52$) and posttest scores ($M = 25.82, SD = 3.57$) was not statistically significant for the experimental condition, $t(10) = -0.962, p = .179, d = 0.27$. In line with these results, the independent samples t-test between the gain scores of the two conditions revealed a significant difference in gain scores between the control group ($M = 4.30, SD = 2.79$) and the experimental group ($M = 0.82, SD = 2.82$), $t(19) = 3.839, p = .011, d = 1.24$.

5.3 Spontaneous knowledge

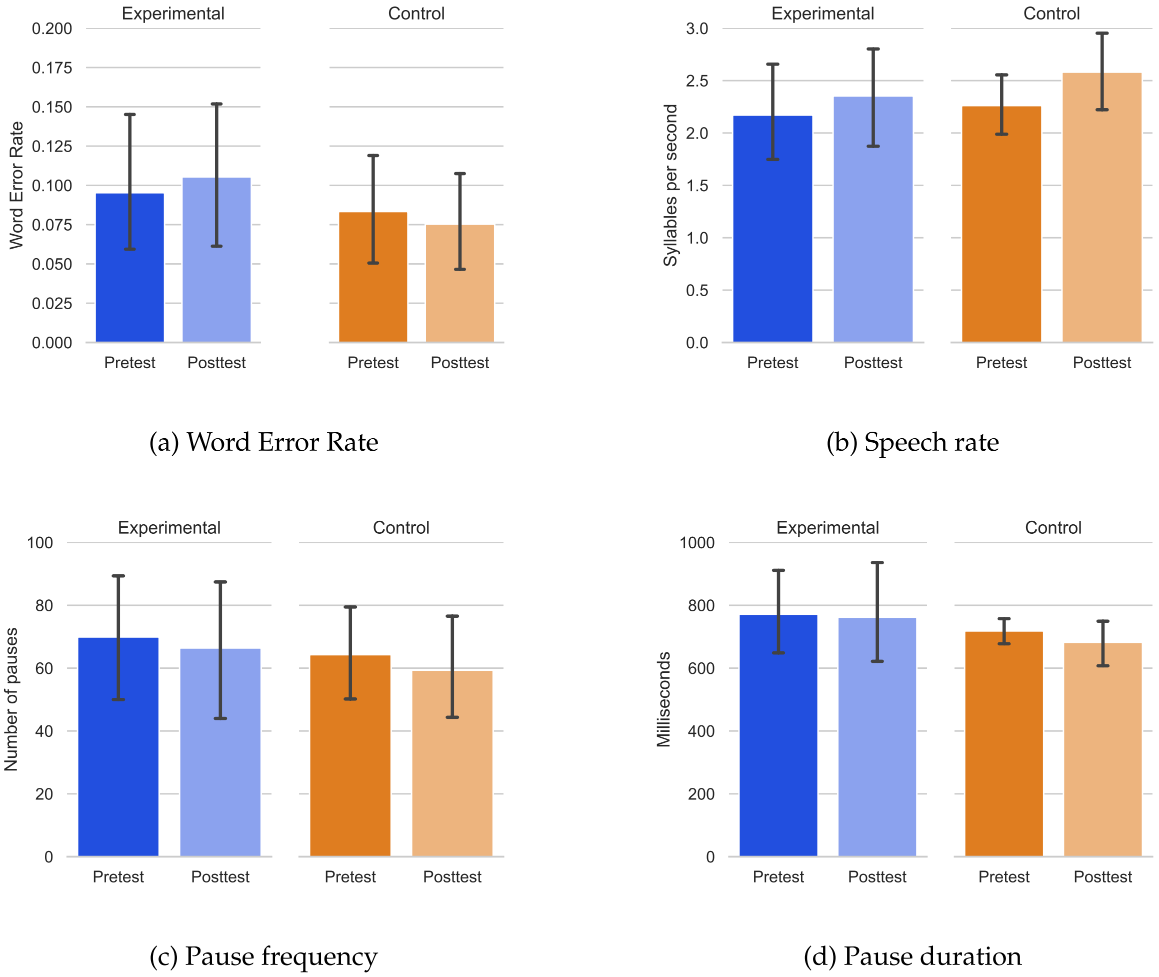
Lastly, the participants read two passages aloud during the pretest and posttest. Using the recordings from these tasks, we analyzed a total of four measures to show any changes in the participants' spontaneous speech knowledge. An overview of these four measures is presented in Figure 5.3.

SQ3: What is the effect of the experimental game on learners' spontaneous pronunciation knowledge, compared to the control game?

Word Error Rate For one participant in the control group, the WER could not be calculated due to the low audio quality of one of their recordings. The WER was thus calculated for 10 participants of the control group, and 12 participants of the experimental group.

As depicted in Figure 5.3a, the average WER decreased for the control group after playing the game for a week. However, a Wilcoxon signed-rank test revealed there was no significant difference between the pretest ranks ($Mdn = 0.07$) and posttest ranks ($Mdn = 0.06$) of this group, $p = .953, r = 0.12$. Similarly, there was no significant difference between the pretest ranks ($Mdn = 0.07$) and posttest ranks ($Mdn = 0.08$) of the experimental group, $p = .814, r = 0.49$. Lastly, the Mann-Whitney U test showed no significant difference between the gain scores of the control group ($Mdn = 0.00$) and the experimental group ($Mdn = 0.01$), $U = 58.5, p = .921, r = 0.02$. These results suggest that there was no significant change in Word Error Rate after playing either version of the game, nor was there a significant difference between the control and experimental conditions.

Speech rate A paired samples t-test was conducted to compare the speech rate during the pretest and the posttest. This test indicated that the difference between pretest ($M = 2.26, SD = 0.50$) and posttest scores ($M = 2.58, SD = 0.67$) for the control group was statistically significant, $t(10) = 3.73, p = .004, d = 1.13$. Similarly, the paired samples t-test showed that participants in the experimental group spoke significantly faster during the posttest ($M = 2.35, SD = 0.91$) compared to pretest ($M = 2.17, SD = 0.82$), $t(11) = 3.91, p = .002, d = 1.13$. This increase in speech



(a) Word Error Rate

(b) Speech rate

(c) Pause frequency

(d) Pause duration

Figure 5.3: Comparison of measurements for spontaneous speech knowledge between the pretest and posttest.

rate for both conditions can also be seen in Figure 5.3b. There was, however, no significant difference between the gain scores of the control group $(M = 0.32, SD = 0.09)$ and the experimental group $(M = 0.18, SD = .05)$, $t(21) = -1.44, p = .166, d = 0.60$. These results indicate a significant improvement in speech rate for both groups, but neither condition performed better than the other.

Pause frequency Participants in both the control group, as well as the experimental group, paused less frequently during the posttest, as shown in Figure 5.3c. A Wilcoxon signed-rank test was conducted to assess this difference for both the control and experimental groups. For the control group, the difference between the number of pauses during the pretest $(Mdn = 56)$ and posttest $(Mdn = 48)$ was not statistically significant, $p = .075, r = 0.37$. Similarly, in the experimental group, no significant difference was found between the pretest $(Mdn = 59.5)$ and posttest $(Mdn = 54.5)$ pause frequencies, $p = .107, r = 0.34$. Additionally, a comparison of gain scores between the control group $(Mdn = -8)$ and the experimental group $(Mdn = -4.5)$ showed no significant difference, $U = 49.5, p = .309, r = 0.21$. These findings indicate that neither group experienced a significant change in pause frequency from pretest to posttest, and there was no meaningful difference in improvement between the conditions.

Pause duration A paired samples t-test was conducted to compare the average duration of pauses during the pretest and posttest tasks within each group. In the control group, there was no statistically significant difference between pretest pause duration ($M = 718.36, SD = 21.56$) and posttest pause duration ($M = 682.09, SD = 37.04$), $t(10) = -1.283, p = .229, d = .387$. Similarly, in the experimental group, no significant difference was found between the pretest pause duration ($M = 771.68, SD = 72.06$) and posttest pause duration ($M = 762.13, SD = 82.60$), $t(11) = -0.363, p = .728, d = .670$. Additionally, an independent samples t-test comparing the gain scores between the control group ($M = -36.27, SD = 28.28$) and the experimental group ($M = -9.55, SD = 26.31$) revealed no significant difference, $t(21) = -0.693, p = .496, d = .537$. These results suggest that neither condition showed significant changes from pretest to posttest when it comes to the average pause duration, and the gain scores were also not significantly different between the control and experimental conditions.

Chapter 6

Discussion

This chapter interprets and discusses the results of the user study in relation to the main research questions and their respective sub-questions. We explore the implications of these results and place these results in the context of previous studies. Additionally, this chapter reflects upon the game's design and provides suggestions for improvement. Lastly, we address the limitations of this study and suggest directions for future research.

6.1 User study

The results from the user study indicate that there are some differences in terms of pronunciation improvement and user experience between the two versions of the game. In this section, we discuss the possible causes and implications of these differences. Additionally, we assess our hypotheses formulated in Chapter 2.

6.1.1 Pronunciation improvement

In general, there were few significant improvements in pronunciation when comparing the pretest and posttest for either condition. Except for the word list reading task, the two versions of the game performed similarly when compared to each other. However, for almost all measures, there was a trend that participants performed better during the posttest than during the pretest. Although the results did not indicate a significant improvement for 4 out of 6 measures, they also did not indicate that participants' pronunciation became worse after playing the game. This is important, as it shows that playing the game (either version) will not cause any harm to Dutch learners' pronunciation.

Controlled pronunciation

Based on our findings, we accept Hypothesis H1 (*Learners who play the experimental game will not show any improvement in their controlled pronunciation knowledge, whereas learners who use the control game will improve their controlled pronunciation knowledge.*). The results show that the control group improved their controlled pronunciation, whereas the experimental group did not.

These results are in line with Skill Acquisition Theory, particularly regarding the requirements for an exercise designed to train declarative knowledge (DeKeyser et al., 2017). Participants in the control group had enough time to apply their knowledge of the different vowel sounds while playing the game. Additionally, they had no other distractions, and could thus completely focus on form. Conversely, these elements were not present in the experimental version of the game: participants had to complete the battle under time pressure and had other game elements to focus on.

Spontaneous pronunciation

Considering the outcomes of the passage reading tasks in the pretest and posttest, we did not find sufficient evidence to accept Hypothesis H2 (*Learners who play the experimental game will improve their spontaneous knowledge, while learners who play the control game will not improve their spontaneous knowledge.*). There were no differences between the control group and experimental group for the four different measures of spontaneous pronunciation. Neither condition improved in terms of word error rate, pause duration, and pause frequency. Both groups increased their speech rate, speaking faster during the posttest than during the pretest. This increase could be due to the participants' familiarity with the texts in the posttest, allowing them to read more quickly. This reasoning explains why the speech rate increased for both conditions, and not just the experimental condition.

The lack of improvement for the other three measurements of spontaneous speech could be linked to the Dutch level of the participants. As stated in Chapter 4, Section 4.1, the participants from both groups indicated that their Dutch proficiency was below average. As a result, their explicit (controlled) pronunciation knowledge might not have been able to turn into implicit (spontaneous) knowledge through practicing with the game, because they were developmentally not ready (Salaberry, 2018).

Lastly, the difference in familiarity with the game mechanics between the control group and the experimental group may have contributed to the lack of improvement. Prior experience with similar games positively influences the learning outcomes of a serious game (Orvis et al., 2008). The battles in the control game use a pronunciation exercise that many participants were familiar with, as it is commonly used in CALL applications like Duolingo. Most participants mentioned in their pretest questionnaire that they use Duolingo to practice Dutch (8 participants of the control group and 7 participants of the experimental group use Duolingo). Thus, the majority of participants in the control group did not need to learn a new skill to play the game, whereas those in the experimental group did. A lack of familiarity with the game mechanics could work as a threshold: having to learn how the battles work can act as a barrier to the participants' learning (Stapleton et al., 2012). During the time in which the experimental group participants were getting used to the game mechanics, they may have been unable to effectively practice their pronunciation.

6.1.2 User experience

Although there were no significant differences in the user experience measures, there were some distinctions between the two versions of the game. Firstly, the participants from the two conditions seemed to have different playing styles, as was shown by the data collected during gameplay. Additionally, the findings from the posttest questionnaire suggest that the experimental game is perceived as more challenging.

Playing style

In this study, we found no difference in the number of minutes the participants of the two conditions spent playing the game. As such, we did not find sufficient evidence for Hypothesis H3 (*Participants in the experimental condition will spend more time playing the game compared to participants in the control condition*). There was a slight trend that the participants in the experimental group played more, but this difference was not statistically significant. The distribution of active participants throughout the experiment did noticeably differ between the two conditions. In the experimental condition, participant numbers remained relatively stable, and quite high, during the first four days. In contrast, the control condition saw a sharp decline in participants during the first three days. This difference likely indicates that the experimental game kept participants' interest for a longer duration, while participants in the control group experienced a faster decline in motivation to play the game.

Interestingly, towards the end of the experiment, the number of active players in both groups increased again. This increase was especially notable for the experimental group, with day 7 having the highest number of active participants for this condition. The increase in activity is likely because their upcoming posttest reminded participants to play the game. This increase suggests that reminders (such as app notifications) could be an effective way to increase player activity. These findings are consistent with previous work showing that users are more likely to engage with an app within 24 hours when a notification is sent, compared to when it is not (Bidargaddi et al., 2018).

Although there were differences in the distribution of active players during the experiments, the average number of minutes played per day was similar for the two conditions. Similarly, the total number of days played by participants in the experimental group was higher than in the control group, but this did not influence the total number of minutes spent playing the game. This indicates that although the experimental condition might compel participants to play more often, it does not increase the duration of a game session. Additionally, these findings suggest that the participants in the experimental group may have a different playing style than those from the control group. Whereas the control group participants played fewer days with longer sessions, experimental group participants played more frequently, but for shorter durations. This playing style might have long-term benefits to the players of the experimental game, as research suggests that consistent practice over time leads to better retention and skill development compared to more intensive practice within a short period (Dunlosky et al., 2013).

Game experience measures

Based on our findings from the posttest questionnaire, we did not find sufficient evidence to support either H4a (*Participants will respond more positively to the experimental game than to the control game. This will improve the user experience metrics challenge, flow, positive affect and competence*) or H4b

(*Participants will respond more positively to the experimental game than to the control game. This will decrease the user experience metrics negative affect and tension*).

Overall, the participants had a moderate opinion of the game for all six game experience measures. The two versions of the game were rated similarly for most of the game experience components. The experimental group rated their version of the game as more challenging, with an average score of 1.90 compared to 1.60 in the control group. They also gave a higher rating on the 'negative affect' component, scoring 1.63, while the control group rated it at 1.00. However, neither of these differences were statistically significant.

Although the difference in challenge between the two versions was not statistically significant, it could still point to a positive aspect of the experimental game. Specifically, the added challenge is an intrinsic motivator, which can increase the engagement of learners (Laine & Lindberg, 2020). In the long term, this added engagement can positively influence learning outcomes, as well as increase enjoyment and player satisfaction (Hamari et al., 2016; Laine & Lindberg, 2020).

The fact that participants of the experimental group experienced more negative emotions while playing the game could be attributed to the game mechanics of this version. The experimental game included a pronunciation exercise that the participants were not familiar with. They had to focus on their Dutch pronunciation while simultaneously grasping the new mechanic, which could have been perceived as more difficult and frustrating.

6.2 Game design suggestions

We developed two versions of a CAPT game based on related research, as well as insights from an expert interview to answer the first research question (*How can a Computer-Assisted Pronunciation Training (CAPT) game be designed to effectively target spontaneous pronunciation knowledge in Dutch language learners?*). Based on the results from the user study, we propose several recommendations to improve the design of the game.

Firstly, we find that the battle system of the control game was designed adequately; conversely, our results indicate that the experimental game's battle system leaves room for improvement. The positive results for the control group in the word list reading task confirm that a simple word reading exercise is an appropriate exercise to improve controlled pronunciation knowledge. The lack of improvement in this task for the experimental group suggests that, as expected, the exercise in the experimental game did not target learners' controlled knowledge. However, the lack of improvement on the passage reading task does indicate that the exercise in the experimental version could be improved to more effectively target spontaneous pronunciation knowledge. One way this could be achieved while maintaining the same exercise format is to include phrases and sentences instead of singular words. This inclusion would make the task more similar to natural speech.

One key insight from the expert interview was that the inclusion of example audio in the game is important to allow students to hear the proper pronunciation of the sounds they are practicing (Chapter 3, Section 3.1.3). The current implementation of the game included example audio within a 'pronunciation notebook', alongside explanations on how to pronounce the different vowel sounds. We suggest extending this pronunciation notebook with more example words per vowel sound, as well as example audio of the words used within sentences. This allows learners to hear the word within the context of a sentence, which is how they would hear it in real-life situations. Additionally, it might be helpful for players to hear the correct pronunciation of the words they have to say in the battles. This would be especially beneficial for the words they could not pronounce correctly. This example audio could be included either at the end of a battle (as a list of words with their corresponding audio), or after each attack, before the next set of words is shown.

Which vowel sounds learners find difficult depends on factors such as their native language. For this reason, we gave players the freedom to choose which vowel they wanted to practice by unlocking all levels and allowing them to play in any order. However, a downside to this approach is that players might play the levels of vowels that they are already good at to make the game easier. To address this, adding a form of personalization would be an improvement to the game. This can be done by adapting the words shown in the exercises to include words the player has previously had difficulty with. Additionally, the order of the levels could be structured based on the sounds that the player is struggling with. This could be achieved by including a short test at the beginning of the game, for example in the tutorial scene. Based on the results of this test, the levels could be structured to fit the learner's needs.

Lastly, due to technical constraints, the current game was published on a website, which the participants could access on their phone's web browser. While this allowed the game to be accessed regardless of the type of phone the participants were using, it limited some features that could enhance user engagement. As mentioned earlier, sending reminders could help increase player activity, and this would be easier to implement if the game were developed as an app. An app would allow for push notifications, reminding players to return and continue their practice.

6.3 Limitations

This study faced several limitations that should be considered when interpreting the findings. The first limitation of the user study is the relatively short duration of the experiment. Most previous research that tests the effectiveness of CALL of CAPT applications uses spans multiple weeks, or even months (Fouz-González, 2020; Luo, 2016; Martinelli, 2016; Tejedor-Garcia et al., 2020). It is thus possible that a longer duration would reveal more differences between the two versions. However, a longer experiment could also decrease participants' willingness to participate (Ságvári et al., 2021). Within the scope of this study, ensuring an adequate number of participants was prioritized over the study duration.

Another notable limitation of this study is the relatively small sample size, consisting of 12 participants in the experimental group and 11 in the control group. This could limit the generalizability of the results, and is a possible explanation for the lack of significance of some results. Additionally, while the diversity of native languages among participants was a positive aspect, most of them spoke multiple languages fluently and all had learned a foreign language before, which is possibly not representative of the general population. Furthermore, the majority were recruited through a university setting. This may limit the generalizability of the findings, as the language learning experiences of this specific group might not accurately represent a broader population.

Thirdly, for the measures in which there was a significant improvement, it is possible that external factors also influenced these results. For example, some participants may have practiced Dutch with their friends during the course of the experiment, or have taken a Dutch class that focused on pronunciation. However, because, as stated in earlier sections, none of the results worsened, we can at least say that the game can be used as an additional tool for Dutch learners to improve these aspects of their pronunciation.

The last limitation is that the participants had, on average, a low proficiency in Dutch. The study aimed to test the impact of the experimental game on the automatization of procedural knowledge, which is the type of knowledge that intermediate and advanced learners already possess (DeKeyser et al., 2017). However, the low proficiency of the participants indicates that they might still be in the proceduralization stage of acquiring pronunciation knowledge. As a result, this group of participants was likely not suitable to show the influence of playing the experimental game on advancing spontaneous pronunciation knowledge.

6.4 Future work

Based on the aforementioned limitations, we now set out several suggestions for future work.

Firstly, the impact of the game should be tested using a larger group of participants, that have a higher proficiency of Dutch. Additionally, the duration of the experiment should also be increased to two or more weeks instead of one week. This would allow the experimental group more time to familiarize themselves with the game mechanics, and thus more time to improve their pronunciation. Increasing the sample size and extending the duration of the experiment would also allow future research to look into the effects of different playing styles on the effectiveness of pronunciation learning through CAPT games.

Secondly, future research can look at alternative spontaneous speech exercises that can be implemented in a CAPT application. The exercise included in this study is limited because it only lets the player practice individual words and only allows them to choose between four different options per turn. Future studies can look at ways to develop exercises that simulate real-life conversations more accurately, e.g., by giving the player more freedom to choose their responses.

The feedback on users' pronunciation in the game was quite minimal and did not include instruction on how learners could improve their pronunciation of a word. Current feedback mechanisms in CAPT applications are designed for controlled exercises, in which the user has no time pressure and can take their time to evaluate their feedback. Since time pressure is an important aspect of a spontaneous pronunciation exercise, future research needs to look into effective methods to give more extensive feedback on pronunciation in tasks that have time pressure.

Lastly, it would be valuable to explore alternative ways to elicit and assess spontaneous speech, that will still allow for comparison between groups. The passage reading text used in our study had the benefit of easy comparison between the two groups but is not as spontaneous as a task that involves free speech. Alternatively, picture description, naming, or narration tasks can be considered as they are more similar to free speech (Nagle, 2018; Saito & Plonsky, 2019). Additionally, alternative ways to measure the change in spontaneous speech knowledge should be considered, aside from the four measurements used in this research. For example, suprasegmentals such as stress, rhythm, and intonation could be included, as they also influence intelligibility (Wang, 2022).

Chapter 7

Conclusion

While language learning applications have become increasingly popular, they often fall short in teaching language learners the necessary skills for fluent speech. Specifically, current applications do not target a learner's spontaneous pronunciation knowledge. This study aimed to answer the following research question: *How does the inclusion of a spontaneous pronunciation exercise in a serious language learning game influence Dutch pronunciation learning?* To answer this question, we first developed a CAPT game that includes a spontaneous speech exercise, as well as a control version of this game that contains a controlled pronunciation task. We measured the controlled pronunciation, spontaneous pronunciation, and user experience of 12 participants who played the experimental game, and compared them to the measurements of the 11 participants who played the control game.

The results of the user study show that participants who played the control game significantly improved their controlled knowledge, while participants who played the experimental game did not improve this type of knowledge. These results were in line with expectations based on Skill Acquisition Theory, as the control game allowed participants to focus purely on the correctness of their pronunciation, while the experimental game did not. There was no statistically significant difference between the two versions of the game in improving the participants' spontaneous knowledge. Both groups of participants improved their speech rate, but none of the other measures yielded significant results. These results were not in line with the hypotheses and might have been due to the relatively low Dutch level of the participants, as well as their unfamiliarity with the experimental game's mechanics.

Additionally, results from the user study indicate that there was no difference in the total number of minutes the participants spent playing either version of the game, although the playing styles of participants for both versions of the game showed some differences. Participants in the experimental group played the game for more days compared to the participants from the control group. There also was no significant difference in the participants' user experience, although the experimental game was rated slightly more challenging. In the long term, the increased challenge can increase engagement and positively influence learning outcomes. However, to verify this claim, future research should extend the experiment duration to multiple weeks.

This research is the first to develop an exercise based on the requirements for eliciting spontaneous speech and integrating it into a CAPT game. Although there was no improvement in spontaneous speech for our participants after using this game, our research did provide relevant insights into the design of such a game and directions for future work. We recommend future research to test similar CAPT games and applications with larger participant groups that have higher language proficiency. Additionally, we identify opportunities for future research to explore different spontaneous speech exercises for CAPT games, as well as ways to improve the feedback included in these games. Lastly, we suggest that alternative methods to assess and elicit spontaneous speech (such as by using picture description, naming, or narration tasks) can be considered when testing the impact of CAPT applications on spontaneous speech.

Thus, to answer our main research question, we see a trend that the inclusion of a spontaneous pronunciation exercise in a serious game can support Dutch language learners by increasing the challenge of the task and thereby enhancing engagement. However, the effectiveness in terms of improving spontaneous pronunciation seems limited for beginner learners. For these learners, the combined difficulty of learning the pronunciation rules of their target language, as well as mastering the game mechanics may limit their progress. Therefore, we suggest the use of serious games for more advanced learners, who would likely benefit from the additional challenge of more complex game mechanics that go beyond what is typically offered in current CAPT applications.

Bibliography

Anolli, L., Mantovani, F., Confalonieri, L., Ascolese, A., & Peveri, L. (2010). Emotions in serious games: From experience to assessment. *International Journal of Emerging Technologies in Learning (iJET)*, *5*(2010).

Bakhanova, E., Garcia, J. A., Raffe, W. L., & Voinov, A. (2020). Targeting social learning and engagement: What serious games and gamification can offer to participatory modeling. *Environmental Modelling & Software*, *134*, 104846.

Bakhuys Roozeboom, M., Visschedijk, G., & Oprins, E. (2017). The effectiveness of three serious games measuring generic learning features. *British journal of educational technology*, *48*(1), 83–100.

Beijer, L., Rietveld, A., Ruiter, M., & Geurts, A. (2014). Preparing an e-learning-based speech therapy (est) efficacy study: Identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers. *Clinical Linguistics & Phonetics*, *28*(12), 927–950.

Berry, D. M. (2021). Level up your pronunciation: Impact of a mobile game. *Mextesol Journal*, *45*(1), n1.

Bidargaddi, N., Almirall, D., Murphy, S., Nahum-Shani, I., Kovalcik, M., Pituch, T., Maaieh, H., Strecher, V., et al. (2018). To prompt or not to prompt? a microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. *JMIR mHealth and uHealth*, *6*(11), e10123.

Blake, R. (2016). Technology and the four skills.

Blanco, C. (2022, December). 2022 duolingo language report. https://blog.duolingo.com/2022-duolingo-language-report/

Calvillo-Gámez, E. H., Cairns, P., & Cox, A. L. (2015). Assessing the core elements of the gaming experience. *Game user experience evaluation*, 37–62.

Casañ-Pitarch, R. (2018). An approach to digital game-based learning: Video-games principles and applications in foreign language learning. *Journal of Language Teaching and Research (Online)*, *9*(6), 1147–1159.

Caserman, P., Hoffmann, K., Müller, P., Schaub, M., Straßburg, K., Wiemeyer, J., Bruder, R., Göbel, S., et al. (2020). Quality criteria for serious games: Serious part, game part, and balance. *JMIR serious games*, *8*(3), e19037.

Chen, X., Zou, D., Xie, H. R., & Su, F. (2021). Twenty-five years of computer-assisted language learning: A topic modeling analysis.

Couceiro, R. M., Papastergiou, M., Kordaki, M., & Veloso, A. I. (2013). Design and evaluation of a computer game for the learning of information and communication technologies (ict) concepts by physical education and sport science students. *Education and Information Technologies*, *18*, 531–554.

Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral proficiency training in dutch l2: The contribution of asr-based corrective feedback. *Speech Communication*, *51*(10), 853–863.

de Almeida, J. L. F., & dos Santos Machado, L. (2021). Design requirements for educational serious games with focus on player enjoyment. *Entertainment Computing*, *38*, 100413.

De Jong, N. H. (2016). Predicting pauses in l1 and l2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113–132.

Deen, M. (2015). *Game, games autonomy motivation & education: How autonomy-supportive game design may improve motivation to learn* [Doctoral dissertation, Technische Universiteit Eindhoven].

DeKeyser, R. (2020). Skill acquisition theory. In *Theories in second language acquisition* (pp. 83–104). Routledge.

DeKeyser, R., Loewen, S., & Sato, M. (2017). Knowledge and skill in isla. *The Routledge handbook of instructed second language acquisition*, 15–32.

Desurvire, H., & Wiberg, C. (2009). Game usability heuristics (play) for evaluating and designing better games: The next iteration. *Online Communities and Social Computing: Third International Conference, OCSC 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings 3*, 557–566.

Doremalen, J. v., Cucchiarini, C., & Strik, H. (2013). Automatic pronunciation error detection in non-native speech: The case of vowel errors in dutch. *The Journal of the Acoustical Society of America*, *134*(2), 1336–1347.

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

European Commission, Eurostat. (2024). Pupils by education level, age and number of modern foreign languages studied - absolute numbers and % of pupils by number of languages studied [http://data.europa.eu/88u/dataset/jwawfiqzxvqn9iyjgo0la].

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Fouz-González, J. (2020). Using apps for pronunciation training: An empirical evaluation of the english file pronunciation app.

Franklin, A., & McDaniel, L. (2016). Exploring a phonological process approach to adult pronunciation training. *American Journal of Speech-Language Pathology*, *25*(2), 172–182.

García Botero, G., Questier, F., & Zhu, C. (2019). Self-directed language learning in a mobile-assisted, out-of-class context: Do students walk the talk? *Computer Assisted Language Learning*, *32*(1-2), 71–97.

Ghorbani, M. R., Neissari, M., & Kargozari, H. R. (2016). The effect of explicit pronunciation instruction on undergraduate efl learners' vowel perception. *Language and Literacy*, *18*(1), 57–70.

Guskaroska, A. (2020). Asr-dictation on smartphones for vowel pronunciation practice. *Journal of Contemporary Philology*, *3*(2), 45–61.

Halbhuber, D. (2022). To lag or not to lag: Understanding and compensating latency in video games. *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, 370–373.

Hämäläinen, R. (2011). Using a game environment to foster collaborative learning: A design-based study. *Technology, Pedagogy and Education*, *20*(1), 61–78.

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in human behavior*, *54*, 170–179.

Holfeld, J. (2023). On the relevance of the godot engine in the indie game development industry. *arXiv preprint arXiv:2401.01909*.

IJsselsteijn, W. A., De Kort, Y. A., & Poels, K. (2013). The game experience questionnaire.

Ivanov, A. V., Lange, P. L., Suendermann-Oeft, D., Ramanarayanan, V., Qian, Y., Yu, Z., & Tao, J. (2016). Speed vs. accuracy: Designing an optimal asr system for spontaneous non-native speech in a real-time application. *Proc. of the IWSDS, Saariselk, Finland*.

Krashen, S. (2013). The effect of direct instruction on pronunciation: Only evident when conditions for monitor use are met? *GiST: Education and Learning Research Journal*, (7), 271–275.

Krath, J., Schürmann, L., & Von Korflesch, H. F. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior*, *125*, 106963.

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, *16*(4), 1–23.

Laine, T. H., & Lindberg, R. S. (2020). Designing engaging games for education: A systematic literature review on game motivators and design principles. *IEEE Transactions on Learning Technologies*, *13*(4), 804–821.

Landers, R. N., Armstrong, M. B., & Collmus, A. B. (2017). How to use game elements to enhance learning: Applications of the theory of gamified learning. *Serious Games and Edutainment Applications: Volume II*, 457–483.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? the influence of accent on credibility. *Journal of experimental social psychology*, *46*(6), 1093–1096.

Lightbown, P. M., & Spada, N. (2013). *How languages are learned 4th edition-oxford handbooks for language teachers*. Oxford university press.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A duolingo case study. *ReCALL*, *31*(3), 293–311.

Loewen, S., Isbell, D. R., & Sporn, Z. (2020). The effectiveness of app-based language instruction for developing receptive linguistic knowledge and oral communicative ability. *Foreign Language Annals*, *53*(2), 209–233.

Luo, B. (2016). Evaluating a computer-assisted pronunciation training (capt) technique for efficient classroom instruction. *Computer assisted language learning*, *29*(3), 451–476.

Mahdi, H. S., & Al Khateeb, A. A. (2019). The effectiveness of computer-assisted pronunciation training: A meta-analysis. *Review of Education*, *7*(3), 733–753.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals.

Martens, H., Van Nuffelen, G., & De Bodt, M. (2010). De ontwikkeling van een fonetisch gebalanceerde standaardtekst [development of a phonetically balanced standard passage]. *Logopedie*, *23*(5), 31–36.

Martinelli, M. (2016). *Effectiveness of online language learning software (duolingo) on italian pronunciation features: A case study* [Doctoral dissertation, Oklahoma State University].

Middag, C. (2012). *Automatic analysis of pathological speech* [Doctoral dissertation, Ghent University].

Mulholland, M., Lopez, M., Evanini, K., Loukina, A., & Qian, Y. (2016). A comparison of asr and human errors for transcription of non-native spontaneous speech. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5855–5859.

Nagle, C. (2018). Perception, production, and perception-production: Research findings and implications for language pedagogy.

Neri, A., Cucchiarini, C., & Strik, H. (2006). Selecting segmental errors in non-native dutch for optimal pronunciation training.

Nunn, A. M. (2006). *Dutch orthography: A systematic investigation of the spelling of dutch words*. The Hague: Holland Academic Graphics.

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human behavior*, *24*(5), 2415–2433.

Pawlak, M., Waniek-Klimczak, E., & Majer, J. (2011). *Speaking and instructed foreign language acquisition* (Vol. 57). Multilingual Matters.

Pennington, M. C. (2021). Teaching pronunciation: The state of the art 2021. *RELC Journal*, *52*(1), 3–21.

Quist, G., Sas, C., & Strik, D. (2015). *Routledge intensive dutch course*. Routledge.

Radová, V., & Vopálka, P. (1999). Methods of sentences selection for read-speech corpus design. *International Workshop on Text, Speech and Dialogue*, 165–170.

Ravyse, W. S., Seugnet Blignaut, A., Leendertz, V., & Woolner, A. (2017). Success factors for serious games to enhance learning: A systematic review. *Virtual Reality*, *21*, 31–58.

Reinders, H., & Benson, P. (2017). Research agenda: Language learning beyond the classroom. *Language Teaching*, *50*(4), 561–578.

Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (capt): Current issues and future directions. *Relc Journal*, *52*(1), 189–205.

Ságvári, B., Gulyás, A., & Koltai, J. (2021). Attitudes towards participation in a passive data collection experiment. *Sensors*, *21*(18), 6085.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708.

Salaberry, M. R. (2018). Declarative versus procedural knowledge. *The TESOL Encyclopedia of English language teaching*, 1–7.

Sreena, S., & Ilankumaran, M. (2018). Developing productive skills through receptive skills–a cognitive approach. *International Journal of Engineering & Technology*, *7*(4.36), 669–673.

Stapleton, A., Costello, B., & Ryan, M. (2012). Threshold concepts: Implications for game design. *Asia Pacific simulation training conference & exhibition*.

Sweetser, P., & Wyeth, P. (2005). Gameflow: A model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, *3*(3), 3–3.

Tejedor García, C., et al. (2020). *Design and evaluation of mobile computer-assisted pronunciation training tools for second language learning* [Doctoral dissertation, Universidad de Valladolid].

Tejedor-Garcia, C., Escudero-Mancebo, D., Cardeñoso-Payo, V., & González-Ferreras, C. (2020). Using challenges to enhance a learning game for pronunciation training of english as a second language. *IEEE Access*, *8*, 74250–74266.

Tergujeff, E. (2021). Second language comprehensibility and accentedness across oral proficiency levels: A comparison of two l1s. *System*, *100*, 102567.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of l2 experience on prosody and fluency characteristics of l2 speech. *Studies in second language acquisition*, *28*(1), 1–30.

Trooster, W., Goei, S. L., Ticheloven, A., Oprins, E., van de Boer-Visschedijk, G., Corbalan, G., & Van Schaik, M. (2017). The effectiveness of the game lingo online: A serious game for english pronunciation. *Simulation and Serious Games for Education*, 125–136.

Van de Weijer, J., & Slis, I. (1991). Nasaliteitsmeting met de nasometer. *Logopedie en Foniatrie*, *63*(97), 101.

Vanpoucke, F., De Sloovere, M., & Plasmans, A. (2022). The thomas more lists: A phonemically balanced dutch monosyllabic speech audiometry test. *Audiology Research*, *12*(4), 404–413.

Vogel, A. P., Rosen, K. M., Morgan, A. T., & Reilly, S. (2015). Comparability of modern recording devices for speech analysis: Smartphone, landline, laptop, and hard disc recorder. *Folia phoniatrica et logopaedica*, *66*(6), 244–250.

Wang, X. (2022). Segmental versus suprasegmental: Which one is more important to teach? *RELC Journal*, *53*(1), 194–202.

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2023). Assessing speech intelligibility of pathological speech: Test types, ratings and transcription measures. *Clinical Linguistics & Phonetics*, *37*(1), 52–76.

Zielinski, B. (2012). The social impact of pronunciation difficulties: Confidence and willingness to speak. *Pronunciation in Second Language Learning and Teaching Proceedings*, *3*(1).

Appendix A Expert Interview Guide

Objective	Objective Description	Questions	Probing questions
1. Difficulties	To find out which word and sound characteristics students often have difficulties with.	1.1 What are the most common difficulties and challenges that foreign students experience when learning Dutch pronunciation? 1.2 What types of sounds or sound combinations do foreign students find difficult to pronounce correctly?	1.1.1 What causes these challenges? 1.2.1 What do the errors look like? for example, do they replace the target sound with another sound? 1.2.2 Do you see differences between students with different native languages? 1.2.3 Are there specific sounds that all students have difficulty with?
2. Current situation	To find out the current methods for improving pronunciation.	2.1 Which methods or techniques do you use during your lessons to improve pronunciation and speaking skills? 2.2 How do you assess the pronunciation of foreign students during your Dutch lessons? 2.3 To what extent, and in what way, do students practice their pronunciation outside the classroom?	2.1.1 How effective are these methods and are they equally effective for all students? 2.1.2 How quickly do you notice improvements in the pronunciation of students? 2.2.1 What are the characteristics of good Dutch pronunciation for you? 2.2.2 Are there specific criteria that you use? 2.3.1 What do you see as the advantages of this? 2.3.2 To what extent do you think it works or does not work?
3. Application	To gain expert insight about the implementation of the CAPT application.	3.1 In a world where everything is possible, what would the ideal app look like for you that students can use to practice their pronunciation? *[explanation of the application]* 3.2 If you had to divide the application into levels, how would you do this? 3.3 What challenges do you think students will face while using this application? 3.4 What would you like to see added or adjusted in the application?	3.1.1 What are the most important components of such an application for you and why? 3.2.1 Would you focus more on one sound, or multiple sounds?
4. General	To allow the expert space to provide additional information	4.1 Considering everything we have discussed today, what is the most important thing that I should take into account while developing the application? 4.2 Is there anything else you would like to share or anything that we have not discussed that you think is important to mention?	

Appendix B Game

Appendix B.1 In-game word list

	levelAa	levelEe	levelEi	levelEu	levelUi	LevelOo	levelUu	levelIe
Category 1	sla	nek	dijk	neus	bruin	hond	hut	mist
	arm	pen	ijs	heuvel	pruim	vork	tunnel	bitter
	naam	best	vijf	nerveus	duim	kort	bus	zitten
	bakker	zelf	lijst	kleur	suiker	spons	krul	blind
	kasteel	tent	partij	preuts	fruit	kopje	punt	hitte
	vast	mens	rijk	jeuk	zuid	mok	zusje	stil
	smal	veld	fijn	peuter	duizend	onder	kurk	kist
	varken	melk	smijten	meuk	bruid	honderd	tulp	dik
	kraan	ster	tijd	leuk	kruis	stop	hulp	hier
	kabel	engel	vijver	keuze	vuist	open	puur	diepte
	parels	beer	dweil	keuren	uil	foto	uniek	diep
	wapen	peer	eiland	leunt	kuif	troon	spuug	koffie
	maart	wereld	zeilen	jeugd	ruim	rozen	buurman	tien
	vader	meester	brein	kleuter	duim	woont	duur	ziek
	normaal	veer	trein	peuzel	tuin	oven	rups	rivier
	haat	peper	eik	sneu	vuil	kroon	smurf	manier
	baard	stelen	einde	meuk	duif	dood	uren	papier
	kamer	zee	wei	peuk	thuis	koken	minuut	idee
	baan	been	feit	sleutel	duif	brood	infuus	vier
	maken	kleed	prei	kleur	thuis	droom	textuur	ridder
Category 2	slap	mes	rijden	beuk	huid	rok	huur	bit
	man	beest	eind	keuken	huis	zoon	bukken	vis
	maand	wet	treinen	veulen	luid	bos	buren	vies
	tak	leeg	lijst	heup	muis	bomen	vuur	zin
	maat	spel	prijs	leunen	luis	poten	uur	riet
	zakken	spelen	lijn	breuk	uit	kop	stuur	lip
	bal	lessen	wijk	heus	kruimel	rot	rustig	fris
	latten	lezen	smijten	deuken	druipen	rood	pus	fris
	plaat	wetten	mijnen	reuzen	bruid	stoom	turen	liepen
	mannen	ver	lijk	speurt	fornuis	vlot	bruut	dip
Category 3	markt	herfst	dweiltje	serieus	uitkleden	monster	instrument	inkt
	afdruk	lengte	afwijkt	spreuk	luisteren	stro	instructie	verspilling
	alsnog	afspreken	splijten	deuntje	ziekenhuis	oploopt	urenlang	advies
	afstand	ernstig	bezeilen	milieu	uitzicht	persoonlijk	cultuur	kiespijn
	afspraak	extreme	schrijver	voorkeur	opruimen	ontstonden	formulier	spiegeltje
	achter	eenzelfde	grijns	keukendeur	verhuizen	bovendien	uurloon	iedereen
	agenda	scherp	gordijn	europa	uitsluiten	beoordeling	augustus	inmiddels
	aanvallers	eenvoudig	aardbeien	goedkeuring	huisnummer	programma	schaduw	subsidie
	ervaring	ergste	vijftig	deurknoppen	vuilniswagen	fotokopie	nummerbord	bibliotheek
	aangevraagd	eenentwintig	grijpen	augustus	duizelig	boodschappen	revolutie	enigszins

Appendix B.2 In-game dialogue

	Transcript
Red Witch's dialogue	H...ll...! Pl....s... h...lp m...! L......k ...t th... n...t...!
Red Witch's note	Hello stranger! Can you help me out please? My little sister cast a spell on me! I can't say any vowels anymore. Please get her to undo the spell! You might have to battle her. I wrote down the rules for a battle in the book! Thank you for helping me!

Table B.1: Red Witch's Dialogue and Note Transcript

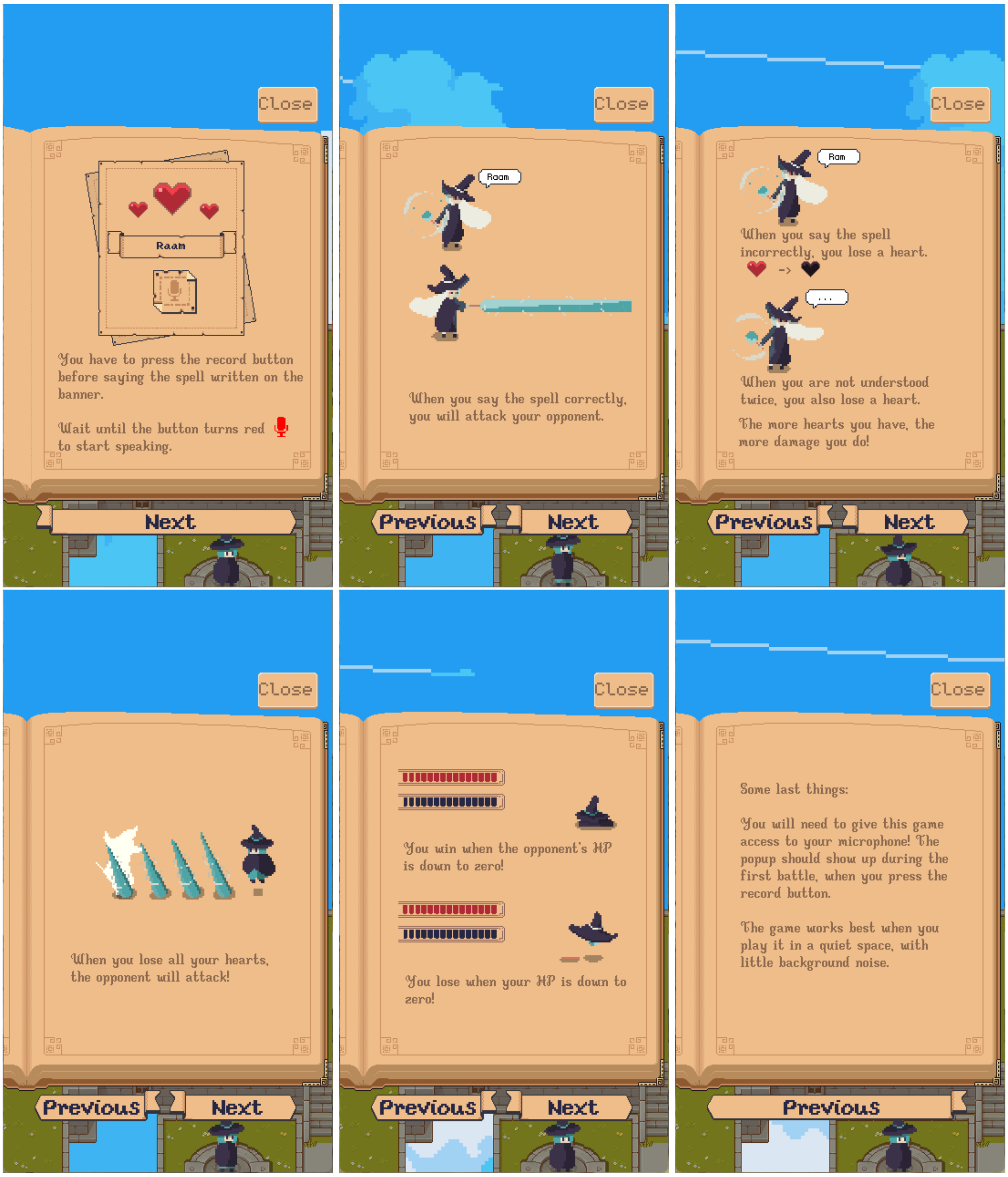
Level Name	**Level Stage**	**Dialogue**
levelAa	1	Look who's here for a challenge! But let me tell you, I own the 'AA'-sound. You'll never beat me!
	2	I hope you are ready for a real battle. Bring it on!
	3	Hey there! Back for another round? I won't go easy on you!
	4	Hey, hey! You're back! You'll never be able to defeat me!
	5	I've been practicing! Are you ready to be defeated?
	6	Oh, hey there again. I hope you are ready for a real battle!
	7	Are you ready to be defeated once again?
	complete	Do you want to keep practicing?
levelEe	1	Hey, you there! I am the master of the 'EE'-sound. You'll never be able to defeat me!
	2	I hope you're ready for a real battle. Bring it on!
	3	You won't be able to beat me this time!
	4	Back for more, huh? I've been practicing. Are you ready to be defeated?
	5	Oh it's you again. I won't go easy on you!
	6	Hey there! Back for another round?
	7	Back for more, huh? Let's see if you have improved!
	complete	Do you want to keep practicing?
levelEi	1	Hey there! Think you can outspell me? I am the undisputed ruler of the 'Ei'-sound! You don't stand a chance!
	2	Oh it's you again. I hope you're ready for a real battle! Bring it on!
	3	Back for more, huh? I've been practicing. Are you ready to be defeated?
	4	You won't be able to beat me this time!
	5	Hey there! Back for another round?
	6	Hey, hey! You're back! You'll never be able to defeat me!
	7	Oh it's you again. I won't go easy on you!

Level Name	Level Stage	Dialogue
	complete	Do you want to keep practicing?
levelEu	1	Ready to face the master of the 'Eu'-sound? Spoiler alert: it's me! You'll never defeat me!
	2	Hey there! Back for another round?
	3	You won't be able to beat me this time!
	4	Hey, hey! You're back! You'll never be able to defeat me!
	5	Let's see if you have improved!
	6	Back for more, huh? I've been practicing. Are you ready to be defeated?
	7	Oh, it's you again. I hope you're ready for a real battle. Bring it on!
	complete	Do you want to keep practicing?
levelIe	1	Hey there! Feeling brave today? Well, I hope you are because I am the master of the 'Ie'-sound. You'll never beat me!
	2	Oh it's you again. I won't go easy on you!
	3	Hey there! Back for another round? I won't go easy on you!
	4	Hey, hey! You're back! You'll never be able to defeat me!
	5	Oh, it's you again. I hope you're ready for a real battle. Bring it on!
	6	I've been practicing. Are you ready to be defeated this time?
	7	Hey there! Back for another round?
	complete	Do you want to keep practicing?
levelOo	1	Well, well, well, look who's here for a challenge! But let me tell you, I own the 'Oo'-sound. You won't stand a chance!
	2	Back for more, huh? I've been practicing! Are you ready to be defeated?
	3	Hey there! Back for another round?
	4	Oh it's you again. I won't go easy on you!
	5	Let's see if you improved!
	6	I've been practicing. Are you ready to be defeated this time?
	7	Hey there! Back for another round? I won't go easy on you!
	complete	Do you want to keep practicing?
levelUi	1	Hey you! The 'Ui'-sound is on the menu today. It's not going to be a walk in the park. Are you prepared for a battle?
	2	Oh, it's you again. I hope you're ready for a real battle. Bring it on!
	3	I hope you're ready for a real battle! Are you ready to be defeated?
	4	Hey there! Back for another round? I won't go easy on you!
	5	Let's see if you have improved!
	6	Oh it's you again. I won't go easy on you!
	7	Back for more, huh? I've been practicing! Are you ready to be defeated?

Level Name	Level Stage	Dialogue
	complete	Do you want to keep practicing?
levelUu	1	Hey, you there! You think you can beat me? I am the master of the 'UU'-sound. You'll never be able to defeat me!
	2	Let's see if you improved!
	3	Back for more, huh? I've been practicing. Are you ready to be defeated?
	4	Hey there! Back for another round? I won't go easy on you!
	5	Oh, it's you again. I hope you're ready for a real battle. Bring it on!
	6	Back for more, huh? I won't go easy on you!
	7	I've been practicing. Are you ready to be defeated this time?
	complete	Do you want to keep practicing?

Table B.2: Dialogue of the Blue Witch for each level and stage in the game

Appendix B.3 Battle Instruction Books



You have to press the record button before saying the spell written on the banner.

Wait until the button turns red 🎤 to start speaking.

When you say the spell correctly, you will attack your opponent.

When you say the spell incorrectly, you lose a heart.

When you are not understood twice, you also lose a heart.

The more hearts you have, the more damage you do!

When you lose all your hearts, the opponent will attack!

You win when the opponent's HP is down to zero!

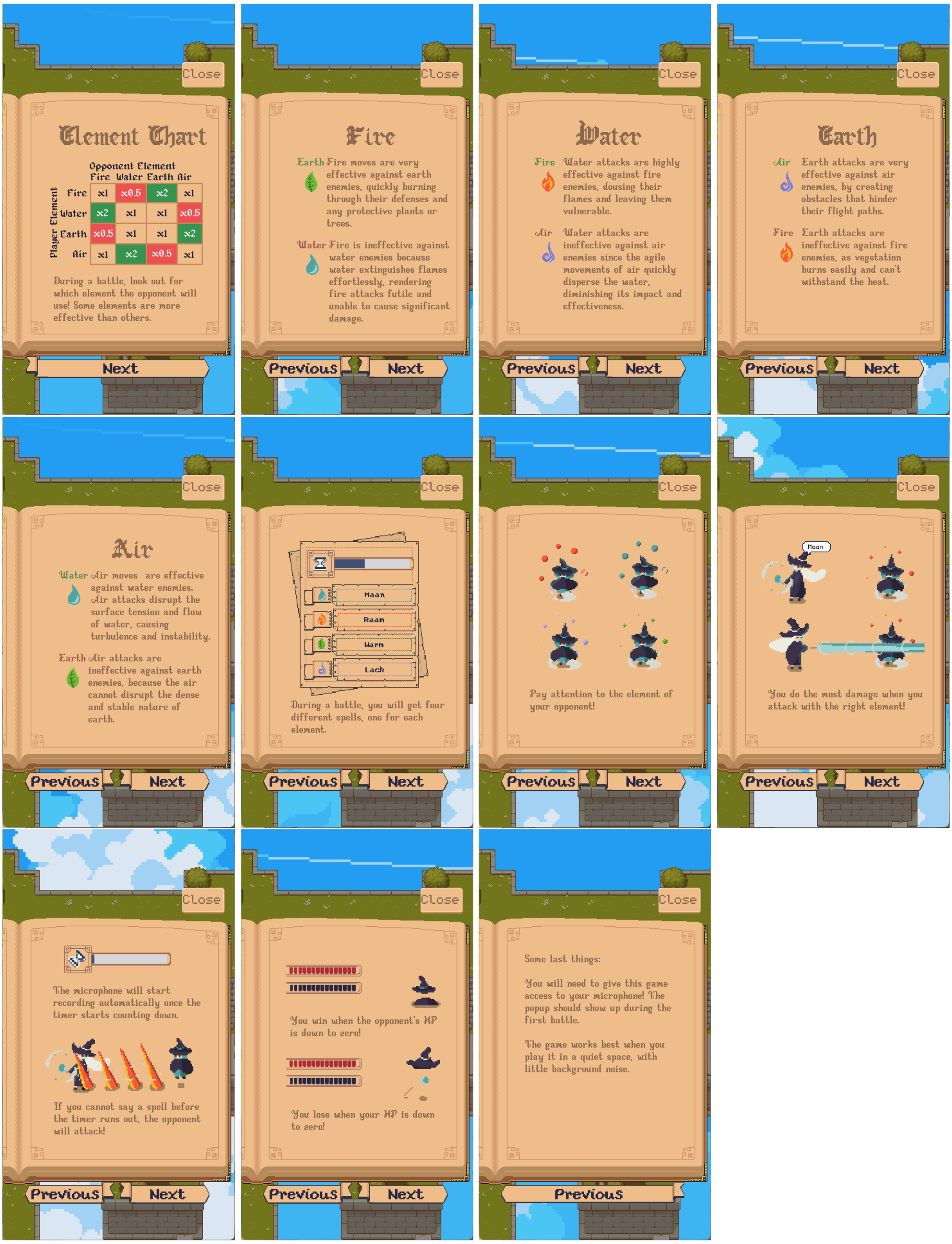
You lose when your HP is down to zero!

Some last things:

You will need to give this game access to your microphone! The popup should show up during the first battle, when you press the record button.

The game works best when you play it in a quiet space, with little background noise.

Figure B.1: Complete battle instruction book in the control version of the game.



Element Chart

	Opponent Element Fire	Water	Earth	Air
Player Element Fire	x1	x0.5	x2	x1
Water	x2	x1	x1	x0.5
Earth	x0.5	x1	x1	x2
Air	x1	x2	x0.5	x1

During a battle, look out for which element the opponent will use! Some elements are more effective than others.

Fire

Earth Fire moves are very effective against earth enemies, quickly burning through their defenses and any protective plants or trees.

Water Fire is ineffective against water enemies because water extinguishes flames effortlessly, rendering fire attacks futile and unable to cause significant damage.

Water

Fire Water attacks are highly effective against fire enemies, dousing their flames and leaving them vulnerable.

Air Water attacks are ineffective against air enemies since the agile movements of air quickly disperse the water, diminishing its impact and effectiveness.

Earth

Air Earth attacks are very effective against air enemies, by creating obstacles that hinder their flight paths.

Fire Earth attacks are ineffective against fire enemies, as vegetation burns easily and can't withstand the heat.

Air

Water Air moves are effective against water enemies. Air attacks disrupt the surface tension and flow of water, causing turbulence and instability.

Earth Air attacks are ineffective against earth enemies, because the air cannot disrupt the dense and stable nature of earth.

During a battle, you will get four different spells, one for each element.

Pay attention to the element of your opponent!

You do the most damage when you attack with the right element!

The microphone will start recording automatically once the timer starts counting down.

If you cannot say a spell before the timer runs out, the opponent will attack!

You win when the opponent's HP is down to zero!

You lose when your HP is down to zero!

Some last things:

You will need to give this game access to your microphone! The popup should show up during the first battle.

The game works best when you play it in a quiet space, with little background noise.

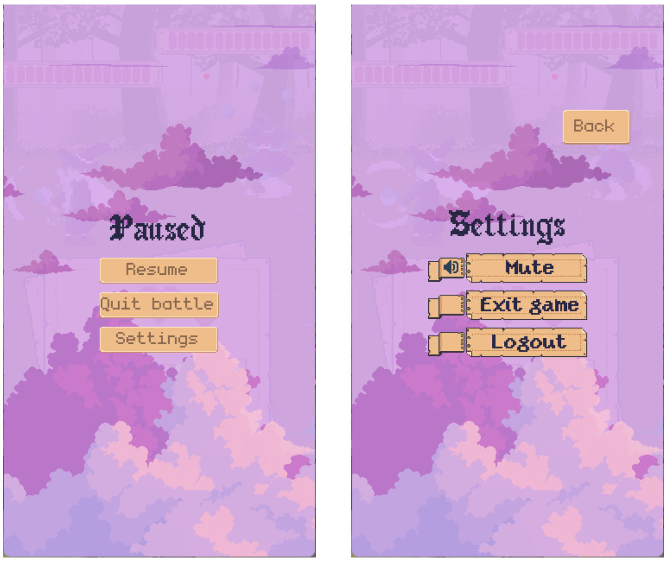
Figure B.2: Complete battle instruction book in the experimental version of the game.

Appendix B.4 Pronunciation Notebook

Sound	Explanation	Example Words
Ee	Both vowel sounds are formed in the front of the mouth. E (short): as in ’get’, but shorter. EE (long): as in ’gain’.	deel, breed, thee, mee, pet, gek, beren, brede, denken, echo
Aa	Both vowel sounds are formed in the back of the mouth. A (short): as in ’bath’, but shorter. AA (long): sounds like the A in ’Chicago’.	maan, slap, pa, sla, man, af, manen, ader, mannen, accent
Ie	Both vowel sounds are formed in the front of the mouth. I (short): as in ’fit’, but shorter. IE (long): as in ’cheat’. There is no double ’ii’ in Dutch; it’s always written as ’ie’.	vier, die, juli, pit, kennis, idee, zinken, beslissing
Oo	Both vowel sounds are formed in the back of the mouth. O (short): as in ’hot’, but shorter. OO (long): as in ’boat’.	rood, voor, zo, pot, mos, bonen, boren, bonken, belofte
Uu	Both vowel sounds are formed in the back of the mouth. U (short): sounds like ’dirt’, but shorter. UU (long): no English equivalent. Make a vowel sound as in ’feet’, while pursing your lips.	Ruud, vuur, nu, put, dus, buren, juni, bukken, dubbel
Ei	This vowel sound is formed in the front of the mouth. EI/IJ: between ’fate’ and ’fight’ or between ’mate’ and ’might’.	feit, meid, lijf, vijand, eiland, knijpen, paleis, blijven
Eu	This vowel sound is formed in the front of the mouth. EU: No English equivalent. Make a vowel sound as in ’dirt’ while pouting your lips tightly and pressing your tongue down.	geur, keus, steun, beugel, nerveus, meubels, monteur, neushoorn
Ui	This vowel sound is formed in the front of the mouth. UI: No English equivalent. Make a vowel sound as in ’house’ while pouting your lips tightly and pressing your tongue down.	ui, bui, trui, buigen, ruiken, buikpijn, buiten, huiswerk

Table B.3: The explanations for the different Dutch vowel sounds, and example words. The majority of explanations are from the Routledge Intensive Dutch Course textbook (Quist et al., 2015)

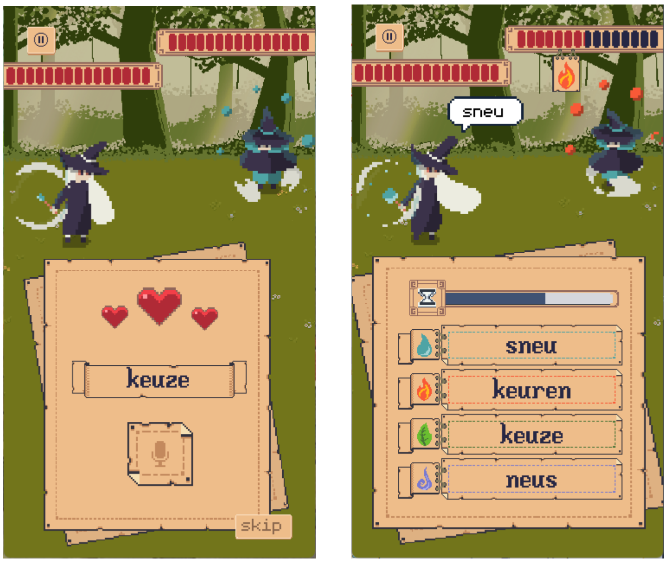
Appendix B.5 Additional screenshots



(a) The pause menu, opened from a battle.

(b) The settings menu.

Figure B.3: Screenshots of the pause and settings menu of the final version of the game.



(a) Control version.

(b) Experimental version

Figure B.4: Changes made to the two versions of the battle after the usability test.

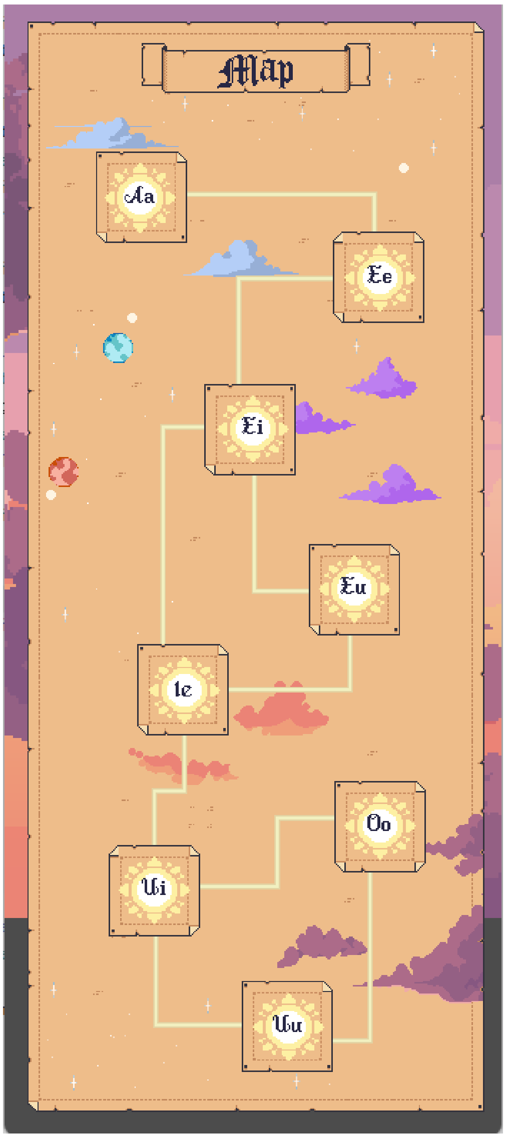


Figure B.5: Screenshot of the map, where players can scroll to select the level they wish to play.

Appendix C User test

Appendix C.1 Participant descriptives

	Control	Experimental	Total
Chinese	0	1	1
English	2	3	5
French	0	1	1
German	3	0	3
Greek	0	1	1
Hindi	1	0	1
Indonesia	0	1	1
Italian	0	1	1
Javanese	1	0	1
Sourashtra	1	0	1
Spanish	3	3	6
Tamil	0	1	1
Total	11	12	23

Table C.1: Most dominant language for the control and experimental group.

	Control	Experimental	Total
Chinese	0	1	1
English	0	2	2
German	3	0	3
Greek	0	1	1
Hindi	1	0	1
Indonesia	0	1	1
Italian	0	1	1
Javanese	1	0	1
Russian	1	1	2
Sourashtra	1	0	1
Spanish	3	3	6
Tamil	1	1	2
Vietnamese	0	1	1
Total	11	12	23

Table C.2: Native languages for the control and experimental group

Appendix C.2 Pretest Questionnaire

Demographic Information

1. **Age**:

[ ] Under 18

[ ] 18-24

[ ] 25-34

[ ] 35-44

[ ] 45-54

[ ] 55-64

[ ] 65 and over

2. **Gender**:

[ ] Male

[ ] Female

[ ] Non-binary

[ ] Other

[ ] Prefer not to say

3. Please list all the languages you know **in order of dominance.** *Make sure to include Dutch as one of the languages.*

1. 2. 3. 4. 5.

4. Please list all the languages you know **in order of acquisition** (you native language first). *Make sure to include Dutch as one of the languages.*

1. 2. 3. 4. 5.

All questions below refer to your knowledge of {language}

1. On a scale from 0 to 10, please select your *level of **proficiency*** in speaking, understanding, and reading.

Speaking Understanding spoken language Reading

2. On a scale from 0 to 10: In your perception, how much of a foreign accent do you have in {language}?

3. On a scale from 0 to 10: Please rate how frequently others identify you as a non-native speaker based on your *accent* in {language}?

All questions below refer to your experience learning Dutch

1. Please shortly describe the methods, resources, and activities you use or have used to practice Dutch.

2. What are your biggest challenges with Dutch pronunciation (e.g., specific sounds or patterns you find difficult)?

Appendix C.3 Posttest Questionnaire

Game Experience Questionnaire

	← not at all				extremely →
I felt content.	□	□	□	□	□
I felt skilful.	□	□	□	□	□
I was interested in the game's story.	□	□	□	□	□
I thought it was fun.	□	□	□	□	□
I was fully occupied with the game.	□	□	□	□	□
I felt happy.	□	□	□	□	□
It gave me a bad mood.	□	□	□	□	□
I thought about other things.	□	□	□	□	□
I found it tiresome.	□	□	□	□	□
I felt competent.	□	□	□	□	□
I thought it was hard.	□	□	□	□	□
It was aesthetically pleasing.	□	□	□	□	□
I forgot everything around me.	□	□	□	□	□
I felt good.	□	□	□	□	□
I was good at it.	□	□	□	□	□
I felt bored.	□	□	□	□	□
I felt successful.	□	□	□	□	□
I felt imaginative.	□	□	□	□	□
I felt that I could explore things.	□	□	□	□	□
I enjoyed it.	□	□	□	□	□
I was fast at reaching the game's targets.	□	□	□	□	□
I felt annoyed.	□	□	□	□	□
I felt pressured.	□	□	□	□	□
I felt irritable.	□	□	□	□	□

I lost track of time.	□	□	□	□	□
I felt challenged.	□	□	□	□	□
I found it impressive.	□	□	□	□	□
I was deeply concentrated in the game.	□	□	□	□	□
I felt frustrated.	□	□	□	□	□
It felt like a rich experience.	□	□	□	□	□
I lost connection with the outside world.	□	□	□	□	□
I felt time pressure.	□	□	□	□	□
I had to put a lot of effort into it.	□	□	□	□	□

Appendix D Results

Appendix D.1 Assumptions

	Measurement	F	Sig.
Playing style	Playtime	2.998	.098
Controlled knowledge	Correct words	0.011	.917
Spontaneous knowledge	Word Error Rate	2.454	.133
	Speech rate	2.351	.140
	Pause frequency	0.386	.541
	Pause duration	0.280	.602

Table D.1: Levene's test for equality of variances

Condition	Statistic	df	Sig.
Control	.920	11	.320
Experimental	.951	12	.647

Table D.2: Shapiro-Wilk test: average playtime

Condition		Statistic	df	Sig.
Control	Pretest	.934	10	.488
	Posttest	.942	10	.579
	Gain score	.949	10	.662
Experimental	Pretest	.953	11	.687
	Posttest	.940	11	.516
	Gain score	.941	11	.538

Table D.3: Shapiro-Wilk test: word list reading task

Measure	**Condition**		**Statistic**	**df**	**Sig.**
Word Error Rate	Control	Pretest	.909	10	.273
		Posttest	.902	10	.169
		Gain score	.843	10	.048
	Experimental	Pretest	.752	12	.003
		Posttest	.902	12	.169
		Gain score	.906	12	.188
Speech rate	Control	Pretest	.941	11	.537
		Posttest	.915	11	.279
		Gain score	.927	11	.385
	Experimental	Pretest	.917	12	.264
		Posttest	.926	12	.339
		Gain score	.958	12	.755
Pause freq	Control	Pretest	.889	11	.137
		Posttest	.890	11	.140
		Gain score	.810	11	.013
	Experimental	Pretest	.933	12	.414
		Posttest	.926	12	.338
		Gain score	.935	12	.435
Pause duration	Control	Pretest	.948	11	.614
		Posttest	.983	11	.979
		Gain score	.958	11	.741
	Experimental	Pretest	.945	12	.607
		Posttest	.860	12	.077
		Gain score	.991	12	.998

Table D.4: Shapiro-Wilk test: passage reading task