UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

**MSc in Computer Science**

# Automated Sentence Generation for a Spaced Repetition Software

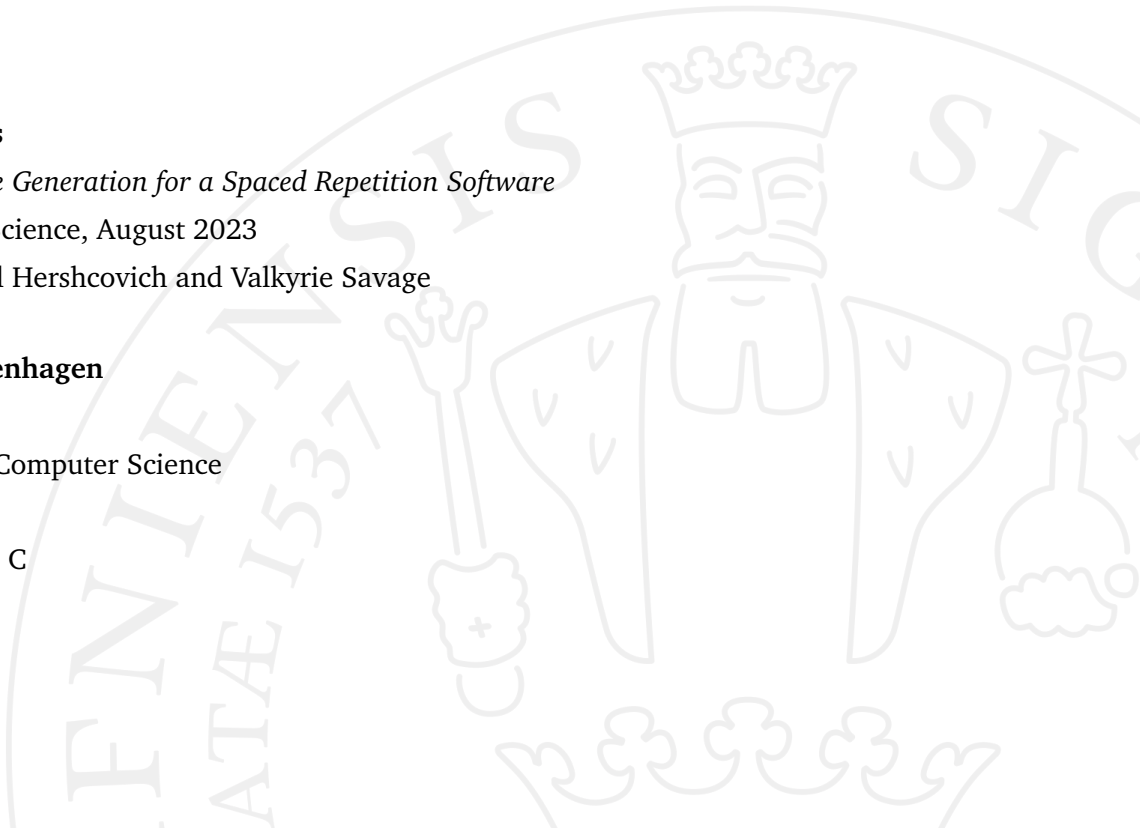
Benjamin Paddags

Supervised by Daniel Hershcovich and Valkyrie Savage

August 2023

**Benjamin Paddags**
*Automated Sentence Generation for a Spaced Repetition Software*
MSc in Computer Science, August 2023
Supervisors: Daniel Hershcovich and Valkyrie Savage

**University of Copenhagen**
*Faculty of Science*
Masters Degree in Computer Science
Bülowsvej 17
1870 Frederiksberg C



# Acknowledgements

I would like to express my sincere gratitude to my supervisors Daniel Hershcovich and Valkyrie Savage for their invaluable support along the way while writing this dissertation. I appreciated Daniel's creativity and expertise a lot in suggestions on how to overcome obstacles and optimize my methodology. Valkyrie provided helpful guidance on how to design a user study, something I had very little prior experience with. Especially their prompt feedback and holding me up to academic writing standards, have greatly improved the quality of this work.

I would also like to thank all those who volunteered to participate in the user study. This research would not have been possible without you and many of you even provided constructive ideas on how to keep improving the system beyond this dissertation's scope.

Finally, I would like to acknowledge that this thesis was written using the UCPH thesis template by Mikkel Roald-Arbøl to define the layout, which is available under the LaTeX Project Public License.

# Abstract

This dissertation proposes and user-tests AllAI, an app that utilizes state-of-the-art NLP technology to assist second language acquisition through spaced repetition, a procedure that spaces out exposure to each vocabulary item and thereby improves long-term recall. Other than current approaches, where words are either repeated solo and out of context, or fixed sentences are repeated, the proposed approach still schedules words independently but combines several words that are due for repetition into a dynamically chosen or generated sentence so that they are still learned in context. First, different NLP paradigms are investigated for their suitability to generate correct sentences that optimize the spaced repetition timing and it is found that retrieval using a BM25 ranking from a Wikipedia-based corpus, as well as a few-shot prompting approach both are suitable. Then a user study is carried out, comparing learning outcomes and user engagement in users using these two methods to the conventional approach of having a fixed sentence associated with each word. It was found that the use of the proposed sentence-based spaced repetition significantly increased learning outcomes (four- to six-fold) compared to the conventional approach, primarily by increasing efficiency and vocabulary growth by showing more words more quickly, without decreasing the fraction of words remembered by learners. In the retrieval group, a significantly higher enjoyment was observed, possibly due to the higher efficiency, hinting at a higher user engagement.

# Contents

# 1 Introduction

## 1.1 Background

Second language acquisition (L2-acquisition) is a challenging process that billions of people go through in formal or informal education. Technology has been used in this process for decades to speed it up and make it more enjoyable (see section 2.1). As one of these technologies, Natural Language Processing (NLP) technology has significantly advanced in recent years and has become able to create text of human-like quality (see section 2.2). However, most language teaching content is still created by humans. Spaced repetition is a well-known learning technique that involves repeated exposure to learning material, usually at increasing intervals, which has been shown to enhance long-term retention (see section 2.1.1). Usually, spaced repetition in language learning is done by repeating single words or whole sentences which have previously been curated by humans. Developing a software system that automatically generates sentences for spaced repetition has the potential to provide learners with a more efficient learning experience by generating sentences with many words that are due for repetition, with more personalized and versatile tasks which make studying more enjoyable and engaging. Furthermore, it could free up human language teachers to focus on in-person teaching instead of writing example sentences.

## 1.2 Research Questions and Scope

The aim of this thesis is to design and develop an app that utilizes state-of-the-art NLP technology to assist L2-acquisition through spaced repetition. The app would keep track of the vocabulary of the user and generate sensible sentences (spaced repetition "tasks") from only the subset of words of a language that the user knows and currently needs to repeat, with some minor amount of new words that make sense to learn. The user can then calibrate the spaced

repetition of each word by answering which of the words in the sentence they correctly remembered. This work mainly consists of two parts – the design of a language learning system with the schedule-constrained sentence generation as a heart piece, and a user study to assess whether this solution can outperform current solutions. As such, the main research questions are the following:

1. Which NLP paradigm and which configuration of the former can optimize spaced repetition timing and best avoid out-of-user-vocabulary words, while retaining as high as possible correctness of the generated sentences?

2. How does sentence-based spaced repetition using the best-performing options from the first question influence user engagement and learning outcomes among language learners, compared to conventional approaches?

From the previously described motivation and research questions, we can define the scope of this work. The goal is to reach a minimum viable product (MVP) from a product development perspective, so that it can be tested with actual users, but keeping in mind what is achievable in the two months of development period between the start of research and the beginning of the user study.
Since spaced repetition systems are usually used for vocabulary retention, this will also be the main goal of the system developed here. The practice of speaking, writing, or listening comprehension is thus out of scope. The main learning outcome for the user is thus reading comprehension in the target language, which, although never directly practiced on larger texts, is arguably an attainable goal, since larger texts would be comprised of the sentences that the system shows to the user, and the sentence-based nature of the system gets the user into contact with the semantics of different words and their forms, as well as grammatical and syntactical structures, which are the main building blocks of texts. In fact, vocabulary learning can be considered paramount to L2 acquisition, as suggested by (Richards 1976). (Nation 1982) calls lack of vocabulary "one of the main obstacles to progress in the receptive skills of listening and reading."
The MVP would thus start with a minimal set of as fundamental words as possible in the language, which are the first words assumed to be in the user

vocabulary, or for beginners the first new words they will learn. New words are always set to be due on the current day and thereafter scheduled according to a spaced repetition scheduling algorithm.
The user path is very similar to classical spaced repetition systems, a loop where the user is shown a task, solves it, sees the solution, and reports which parts (words) of the task they remembered (this part is different from traditional spaced repetition systems, where the task is rated as a whole), hereafter they move on to the next task until no words are due anymore. The key contribution of this work is putting the due words into context by automating the forming of sentences with them. This has the advantage that it does honor the minimum information principle (see 2.1.1 for an explanation and the benefits), as the words can be scheduled independently, while the words are at the same time presented in a sentence context. Furthermore, new words are naturally introduced in the context of existing words whenever it is not possible to form a sentence only with the existing words, this is how the user's vocabulary grows over time.
For this sentence forming, a maximum length of 10 words is defined, to avoid overwhelming the user with too long sentences, while still being able to provide enough context.

#### Learners with previous experience

Due to the combination of spaced repetition and sentence generation, it is arguably not necessary to know the user's exact vocabulary with absolute certainty to generate useful sentences that teach the user new words, meaning that users with existing knowledge can be treated like beginners. The system will still generate more advanced words that are new even to advanced learners, where they fit the sentence well, even though it might be optimized to prefer simpler words. This is just a hypothesis and has to be confirmed in the user study.

#### Language

The system will be implemented and tested in Danish, as the author is associated with the University of Copenhagen and it was thus easiest to source

test users willing to learn this language. The idea of the system is, however, applicable to any language in which the sentences are made up of words, and the software system around the NLP solution (front-end, back-end) should be developed in a way that allowed to teach a different language if the NLP component is swapped out, e.g. by translating the prompts of a prompting-based solution to a different language.
The choice of Danish does bring some drawbacks, such as reduced availability of language models and of metrics for comparability, since most NLP research happens in English. However, even in English no previous research was to be found publishing metrics and scores for spaced repetition scheduling accuracy, so comparability is impacted by that fact in the first place. For the correctness of the sentences, very few metrics results, such as perplexity, also exist for Danish. For more details on these computed metrics, see section 3.4.

## 1.3 Contributions

As previously mentioned, the main contribution of this work is putting the current and soon-to-be due words of a spaced repetition system into context by investigating different methods of automating the forming of sentences with them.
Secondly, a metric for calculating the scheduling accuracy is developed and other metrics are selected to assess the quality of the output sentences for the task.
A range of candidate methods and configurations that managed to return sensible sentences containing target words are compared with regard to these metrics.
An application is developed consisting of a front-end for the user to interact with the generated tasks and a back-end to do the spaced repetition scheduling and house the developed methods for sentence generation.
Finally, the real-world usefulness of two of the best-performing methods, a retrieval-based method and a GPT-3.5-based method using few-shot prompting, is tested in a user study, assessing learning outcomes, indicators of user engagement and technology acceptance, against a baseline similar to current spaced repetition practices.
The proposed system combines the following potential advantages over the conventional spaced repetition approaches mentioned in 2.1.1:

1. It honors the minimum information principle,

2. Shows words in context for a less artificial learning situation and the possibility to infer meaning,

3. Is able to generate a variety of tasks for high novelty value,

4. Sentence generation could be optimized for additional objectives, such as possibly entertainment value (e.g. subsequent sentences could form a story), variety of grammar, or others.

## 1.4 Overview

This dissertation consists of two parts, corresponding to the two main research questions. After this introduction and an initial background chapter on previous research related to the topic, part one begins, which answers the first research question. It is made up of two chapters. First, the methodology is explained in chapter 3, including the objectives to be optimized, the paradigms being considered, and the metrics to compare them. Chapter 4 presents and discusses the results of that assessment and concludes which two methods to continue with in the user study.
Part two answers the second research question by conducting a user study. The method is explained in chapter 5, which goes into detail about the baseline the two methods are compared to, the test system's design and the experimental procedure, including sourcing of the participants and descriptive statistics about possible confounding variables. Chapter 6 then presents the results, which are then discussed in chapter 7.
Finally, a short conclusion is drawn in chapter 8.

# 2 Literature Review

This chapter dives into the theoretical background related to the research questions to identify promising approaches and distinguish this work's contribution from that of the most similar works.

## 2.1 Vocabulary Learning and Technology

There has been a plethora of research on the impact of the use of technology on L2-acquisition and vocabulary learning specifically. Much of the research focuses on the use of computer-assisted language learning (CALL) and mobile-assisted language learning (MALL), since these two are among the most accessible and thus most used technologies for language learning. Apart from CALL and MALL, specific concepts employed in the technological support of L2-learning include gamification (Al-Dosakee and Ozdamli 2021) and spaced repetition (see 2.1.1).
A meta-analysis by Hao, Wang, and Ardasheva (2021) found a large effect of technology-assisted language learning on vocabulary learning, compared with traditional instructional methods in preschool-to-college learners of English as a foreign language. Larger effects were observed for MALL than for CALL, for non-game-based technologies, for enabling students to learn outside of the classroom, and for productive test formats or a combination of productive and receptive ones.

### 2.1.1 Spaced Repetitions Systems and Their Use in L2-Acquisition

Spaced repetition means reviewing information that one wants to remember repeatedly and with temporal spacing in between each exposure to the same information. A review usually involves the learner being prompted, making an

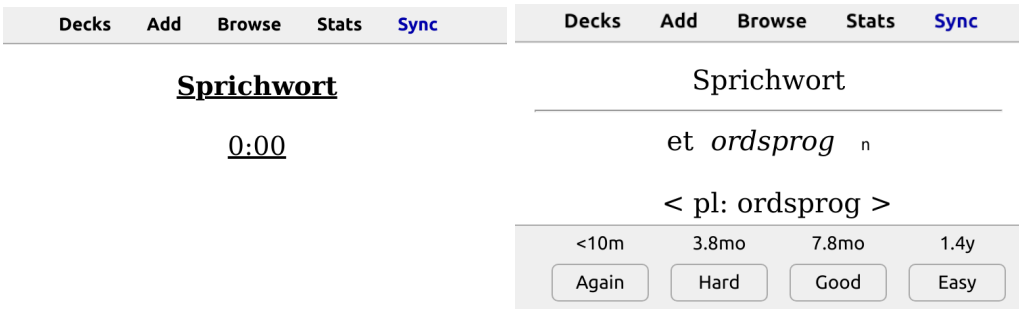



**Figure 2.1.:** The Anki spaced repetition system, step by step: a task is presented (left), the solution (translation) is shown and the user is prompted to rate how well they remembered (right)

effort to recall, and then getting feedback. It has (independently of the specific spacing intervals) been shown to produce better learning than immediate repetition without spacing, e.g. in this meta-analysis by Carpenter et al. (2012) for spacing in general.

A low-tech way to implement spaced repetition is to use physical flashcards where the information to be remembered is split into a prompt and an answer, which are written on the two sides of the card, which is then reviewed at increasing intervals (Leitner 1972). Based on this idea, most spaced repetition software (e.g. Anki (Elmes, n.d.), which is shown as an example in figure 2.1, Mnemosyne (Çakmak, Namaziandost, and Kumar 2021), Supermemo (Wozniak, n.d.)) usually show a memory recall task to the user and expect the user to try to solve it. Thereafter, the solution is shown to the user, and the user rates how well they were able to recall the solution of the task. The system uses the recall quality to calculate the length of the spacing until the task should be presented to the user again, which should ideally be right before the user is likely to forget it. Compared to manual spaced repetition (i.e. repeating paper flashcards regularly), this automates the process of card scheduling, so users are not tempted to do large cramming sessions, but can only study those digital flashcards that are actually due to be reviewed (Kornell 2009).

In the context of language learning, spaced repetition focuses on the parts of L2 acquisition that require memorization, such as vocabulary learning or grammar rules, and leaves out any parts that require active practice, for example speaking, listening comprehension or text production.

There are thus three common approaches for vocabulary retention using spaced repetition systems, as evidenced by the kinds of card decks users have published for the Anki app ("Danish - AnkiWeb" n.d.). The first one is to use single pieces of vocabulary as the task, the second one is to use whole sentences

or text snippets, and the third one is to use single words, but with one or more example sentences also provided on either the solution side or both sides of the flashcard. The main argument for the first practice is the minimum information principle; Each task should be as minimal as possible, ideally one piece of information (Jankowski 1999). This allows for more accurate scheduling. If the task contains multiple bits of knowledge, some easy parts of it might be scheduled more often than they would need to, because the harder parts require more frequent review. On the other hand, language is naturally used in context, meaning that remembering words out of context is not just a very artificial task, but also much harder than if related words are present which can give hints about the meaning (Ramos and Dario 2015). Thus using sentences or text snippets as the task can be assumed to lead to easier recall and thus less time spent reviewing each word. Ramos and Dario (2015) do, however, conclude that the context has to be high-quality, meaning with a high percentage of previously known words, to be able to infer meaning.
This dissertation sets itself apart from the existing literature on spaced repetition by examining the effects of the integration of a sentence generation component, which makes it possible to keep scheduling single words and adhering to the minimum information principle while showing words in context.

## 2.2 Language Models

A central concept in NLP is the language model (LM): A statistical model that assigns a probability to any possible sequence of tokens (Jurafsky and Martin 2023). This probability distribution can be sampled from, thereby generating text. The ability of language models to generate fluent text has significantly advanced in recent years, to the point where they can create text of human-like quality (Fatima et al. 2022).
It is possible to directly modify the probability distribution of the output tokens and thereby influence the likelihood that certain words are generated, as is successfully attempted by Chen et al. (2022), who talk about how to satisfy language constraints, such as not mentioning specific categories of words. The approach thus also seems promising if a certain set of words (such as the due words in a spaced repetition schedule) shall be more likely to be generated. While Chen et al. (2022) focus on avoiding certain words given as a category,

this dissertation tries to do the opposite and encourage certain words that are explicitly listed.

While the traditional approach for using LMs was to train them from scratch for the task at hand, in recent years, the strong performance of transformer-based (Vaswani et al. 2017) pre-trained models (PLMs), such as BERT (Devlin et al. 2019) and GPT-3 (Brown et al. 2020) which are pre-trained on large amounts of text data, usually web-crawls, have lead to two paradigms to NLP tasks becoming very popular: Fine-tuning, as well as prompting of these PLMs. These two paradigms profit from the excellent general understanding of the semantics and syntax of language that large pre-trained models can develop through the training on these large and diverse text corpora.

In fine-tuning, a pre-trained neural network is taken as the starting point and then the whole network or parts of it are further trained on the specific task. Usually, this involves labeled training data. Due to the already existing general understanding, this can reduce the amount of training data and training repetitions needed (Han et al. 2021). This training data-based approach uses a loss function that is based on a distance measure between the LM's output and the correct solution. A loss function, or its opposite, a reward function can also be formulated without specific training data, and the LM can be directly optimized for that reward function using its own outputs. This is a key aspect of reinforcement learning, which is the term for when an agent (the LM) learns a policy (which outputs to generate at which conditions) from its own interactions with a reward function (Uc-Cetina et al. 2023).

Opposed to fine-tuning and reinforcement learning, prompting does not alter the pre-trained model's weights. Instead, it relies solely on the inputs given to the language model to alter its outputs, by combining the inputs with a prompt. The prompt is a description of the task that the language model is supposed to perform. While this so-called zero-shot prompting itself performs quite well for some tasks, due to the general language understanding of the language model, accuracy can often be improved by few-shot prompting, which is when a small number of examples are also provided in the prompt. These examples demonstrate which outputs should be generated for exemplary inputs. Since language models assign probabilities to a token to be generated based on previous tokens, they are often capable of continuing the pattern observed in the examples into the final output. Recent research has demonstrated that especially for more recent, very large LMs, prompting approaches can reach similar results to fine-tuning-based approaches on many NLP tasks, or even outperform them (Wei et al. 2022; Brown et al. 2020). Still, these papers also

show that output quality can be very dependent on the specific formulation of the prompt and the examples, which is why it is important to tweak these, in a process called prompt engineering.

## 2.3 Use of NLP in L2-acquisition

Even before the advent of modern language models, Brown, Frishkoff, and Eskenazi (2005) used a corpus of words with example sentences to generate cloze questions with a keyword missing, which the user has to fill in, to assess language learners' level. This is similar to the task this dissertation tries to achieve, of generating sentences based on multiple words that should be contained. However, they only use one input word which in their database is already associated with sample sentences, so the exact approach cannot be copied for multiple input words. The approach of using a corpus of example sentences still seems worth considering. The field of text retrieval is concerned with retrieving documents from a corpus which are relevant to a specific query which can consist of multiple words. One of the simplest and most popular and reliable document scoring formulas is BM25 (Robertson and Zaragoza 2009).

When it comes to using LMs in second language teaching, Kasneci et al. (2023) identify many promising aspects that language models could assist with and which most of the current research effort is focused on, such as using chatbots to mimic L2 conversation or assist with explanations (Jeon 2021), or to give feedback, score (Mizumoto and Eguchi 2023) and suggest improvements to learner's written or oral output. Okano et al. (2023) try a reinforcement learning approach, as well as a few-shot prompting approach to make large language models output sentences containing specific grammatical structures and find that both approaches are feasible. Their research was published after the experiments in this dissertation were already finished, so it could not be used for inspiration. And while they focus on generating sentences with specific grammatical structures, this work instead tries to achieve the use of specific words in the sentence, which is easier in the sense that instead of transferring implicit grammatical patterns, the model just needs to use the same words already given in the input, but harder in the sense that there are thousands of words that might need to be generated, while Okano et al. (2023) only had 20 grammatical structures to optimize for.

## 2.4 Conclusion

Despite the identified opportunities from the previous paragraph, it seems that only a minuscule fraction of the published research on language models is on applying them to more specific real-world problems within L2-acquisition. Neither have there been many papers with actual test users of language models which compared their results with traditional approaches to solving the tasks the LM is supposed to assist in. This leaves a research gap that this dissertation would like to contribute to filling.

Even though the specific combination has not been tested, the research suggests that the application should, for ideal learning outcomes, take the form of a mobile application that can be accessed on the go, is non-game based, and has users not only receive input but also produce sentences in the language. The last property is, however, out of scope for this work since verification of user-produced sentences is a whole research area by itself, for which there is not enough time to implement.

As for the core of the system, the sentence generation, research suggests that it is beneficial to maximize the content of known words in the sentence so that unknown ones can be inferred from context and remembered in context.

Different paradigms for how to apply LMs to downstream tasks have been discussed and seem like possible solutions to the task at hand. Modifying the probability distribution of a PLM directly, few-shot prompting and reinforcement learning have already been shown to be suitable to achieve constrained generation, which is why all of these were explored in section 3.3.

# Part I

Choice of NLP paradigm

# 3 Method

This chapter defines the method used to answer the first research question: Which NLP paradigm and which configuration of the former can optimize spaced repetition timing and best avoid out-of-user-vocabulary words, while retaining as high as possible correctness of the generated sentences?
First, the objective to be optimized is defined, then the experimental process and with it the paradigms, models and configurations (all of them together referred to as sentence generation "methods"), as well as comparison metrics.

## 3.1 Objective Definition

The objective of the system is to suggest sentences ("tasks") for the user to review, with the goal of practicing their vocabulary and thus reading understanding of the language they are learning. For optimal efficiency, the words of which each task consists should be scheduled, as much as possible, according to their due dates coming from the spaced repetition scheduler. This results in the following three main objectives imposed by the first research question:

1. **Correctness** of the sentence, meaning that the language is used in the same way a native speaker would, using proper grammar, orthography and style.

2. **Amount of known words in the output, prioritized by due dates**, meaning that the sentence should ideally consist of words that the user knows and that are due to be practiced soon, with words being due today having the highest priority to be in the task and the longer in the future, the lower the priority.

3. **Avoid long sentences over ten words** to not overwhelm users. The ten word boundary was chosen by showing Danish sentences of different

lengths to three test users recruited for the user study (see 5.5.1) and asking them to choose the length that is okay without overwhelming them.

There is a fourth objective which is somewhat more subtle and thus intentionally left out of the research question: **Proper difficulty of new words**, meaning that when new words are introduced because they were necessary for the sentence or the user wants to learn new words, start from a difficulty that is appropriate to the user, to not bore them with words they already known nor overwhelm them with very advanced words. For beginner users that means as low a difficulty as possible and increasing difficulty as slowly as possible. For advanced users, this objective involves finding the right difficulty for the user, which is challenging to accurately determine, which is why the focus of this work is mostly on the three main objectives, but a limited effort will be done to not completely neglect this objective so that the lack of optimization for it does not destroy the efforts done to optimize the other three.

### 3.1.1 How is it evaluated how advanced a word is?

The above mentions more and less difficult/advanced words. While words might have intrinsic difficulty, like longer words being more difficult to recall, this is subjective and varies between persons. Arguably, a better measure of how advanced a word is is how frequently used it is, since this determines how quickly a learner would learn it by exposure to the language. Frequency lists are available for most languages, such as contained in the Python library WordFreq (Speer 2022) used in the implementation of the system.

### 3.1.2 Measuring the Loss

From objectives one and two, we can define a loss formula which should be optimized. $l = -a * p_{correct} + b * \frac{t_{wasted}}{sec}$ where a and b are weights for the two sub-terms.

$t_{wasted}$ is a heuristic for the time wasted by non-optimal scheduling.

$$t_{wasted} = \begin{cases} t_{rev\_word} * \frac{max(t_{due} - t_{now}, 0)}{t_{due} - t_{last\_seen}} & \text{if word in user vocabulary} \\ 1 & \text{otherwise} \end{cases}$$

Out-of-user-vocabulary words get a score of one, since the model should use them as rarely as possible to avoid exponential growth of the vocabulary and thus the due words on each day. If the model puts more unknown words in each task than words that are due today, and the user studies until all due words are reviewed, that would lead to exponential growth, since these words would then be added to the user vocabulary and be due to be reviewed on the coming days.
To give two examples of the losses, if we are reviewing an item that we just reviewed successfully immediately again, the fraction is 1 and thus we just wasted the time it takes to review the word. If we are already reviewing tomorrow's due word, which had a review interval of 2 days, we have wasted half the review interval, and thus half the time that the review took was wasted.

**How to estimate the review duration** $t_{rev\_word}$

If we had actual users, over time, the actual review durations for different sentences can be measured empirically, and from that the contribution of each word in the sentence could be factored out. For testing purposes, the same review duration was assumed for every word and thus the review duration could be removed from the loss by setting $t_{rev\_word} = 1$.

## 3.2 Experimental Procedure

Different methods of generating sentences were implemented for testing purposes. For each method, a preliminary assessment was done, consisting in determining whether it generated at least more than 50% correct sentences that contained at least of one the words it was given as inputs. If a method reached these criteria, it was moved on to the next stage where then for each

of the promising methods, different configurations were subjected to a range of metrics to determine how suitable they are to be used in the system.

## 3.3 Methods

The objective defined in section 3.1 can be approached by many different approaches, which are presented in this section. The approaches considered were, as identified in chapter 2, retrieval of suitable sentences from a corpus, modifying the probability distribution of a PLM directly, few-shot prompting, and reinforcement learning with a static reward function that was the inverse of the loss from section . The reinforcement learning approach was not at all able to generate coherent sentences in preliminary tests, while the others were, which is why only they will be described further.

### 3.3.1 Finding the Best Sentence in a Corpus

In information retrieval, the objective is to retrieve relevant documents to a query consisting of multiple words. If the documents only consist of one sentence, this task becomes very similar to the objective of this work.
If the corpus is chosen well, the correctness of the sentences will be high, which optimizes the first part of the objective.
Sentences containing more advanced words can be filtered out from the corpus, which optimizes the fourth part of the objective, at least for beginner learners.
Different from LM-based approaches, the amount of distinct possible outputs is limited to the number of sentences in the corpus, and thus it is limited how well the second objective can be optimized. There is thus a trade-off between excluding sentences with advanced words and including them to have more options among which there might be sentences that contain more due words.
The BM25 retrieval algorithm (Robertson and Zaragoza 2009) was taken as a starting point. It is suitable insofar as it ranks the sentences based on how many of the query words they contain and gives reduced importance the more common a query word is.
Intuitively, this means that the algorithm is less likely to miss the opportunity to suggest a sentence containing multiple rare words, e.g. if the words "Java", "big" and "debugging" are due, it will not make the mistake of picking a

sentence like "Java is a big island" and leaving "debugging", for which it is hard to find good sentences in the corpus. It will instead pick a sentence like "Java debugging explained" and leave "big" for a later task, for which there is a plethora of sentences containing it.
BM25 was modified to add query word weights. Query word weights allow to give a higher importance to words that are due earlier (e.g. a word due today gets a higher weight than a word due tomorrow). Query words were discounted with exponential decay the longer in the future they were due.
The following formula was used to rank the sentences:

$$\text{BM25}(\text{query, sent}) = \sum_{\text{w} \in \text{query}} \left( \text{idf}_\text{w} \frac{(\text{k1}+1) \cdot \text{q\_freq}_\text{w}}{\text{q\_freq}_\text{w} + \text{k1}(1 - \text{b} + \text{b}\frac{\text{sent\_len}}{\text{avgsl}})(\text{days\_til\_due}_\text{w} + 1)} \right)$$

Where:

$\text{idf}_\text{w}$ is the inverse document frequency of the word in the collection.
$\text{q\_freq}_\text{w}$ is the frequency of the word in the query.
sent_len is the length of the sentence (number of words).
avgsl is the average sentence length in the collection.
k1 is 1.5
b is 0.75

Like in all other methods, same-day repetitions of a task were disallowed, for reasons further explained in 3.4.2. This was implemented by finding the best-ranking sentence that had not been previously shown.
In addition to this standard version described above, a version was tested which instead of selecting the best-ranked task, selected the task with the best scheduling score among the 25 best-ranking tasks.

#### The wiki40b Corpus

The Wiki-40B Corpus (Guo et al. 2020) was chosen as the source of the sentences since it is one of the biggest corpora for Danish (and 40+ languages in total, allowing for easy adaption) with ca. 200MB worth of Danish sentences

and, as it is sourced from Wikipedia articles, contains mostly correct use of the language.
The corpus was processed to remove any annotations added, such as _START_ARTICLE_, it was split into sentences using the SpaCy (Montani et al. 2023) sentencizer, and sentences containing certain punctuation (()"»«›‹„""‚‘’[]{}) were removed, since they often only make sense in the context of another sentence. All text was converted to lowercase. Duplicate sentences were removed, as well as sentences consisting of just a single word (since they are not really sentences) or that were longer than 10 words (to optimize the third objective) or which contained words that were not among the 25000 most frequent in the language, since such sentences would not be suitable for learners. This number 25000 was chosen quite arbitrarily and is a trade-off between allowing very niche words and reducing the size of the usable corpus so much that it would for some queries exclude the best match. Ideally, different options should be considered and compared. However, since this dissertation was more focused on how language models can be used to generate the tasks, I decided to leave this optimization as a future research opportunity.
After the filtering, the resulting corpus contained 64259 sentences, of which the average length was 5.9 words.

### 3.3.2 Prompting GPT-3.5

While a retrieval-based system is limited to only a certain set of possible outputs, a language model can theoretically output any sequence of tokens and thus adapt more precisely to the inputs.
While fine-tuning approaches were also considered, preliminary tests revealed that only the prompting approach gave meaningful results (see section 2.2 for background on both paradigms). Opposed to all other approaches, with prompting it was possible to use the fairly recent GPT-3 and GPT-3.5 models and the cost of using the OpenAI API is far lower than that of hosting a server with enough RAM to run comparable (or even smaller) LLMs. The whole experiment cost less than $10 due to the use of prompting, while a server with a GPU cluster capable of running comparable models would likely cost hundreds of dollars per month.

### Choice of Language Model

Different language models, as well as formulations of the prompt and hyper-parameters for the generation, were tested. Firstly, three different language models available through the OpenAI API were tested: GPT-3-curie, GPT-3-davinci and GPT-3.5-turbo-0301. GPT-3-babbage and GPT-3-ada had also been looked at but found to be generating nonsensical or non-Danish text too often when replying to the prompts. GPT-3-curie and GPT-3-davinci fared better, sometimes outputting very good sentences containing the input words, but upon further investigation, they were still found to often just output the input words in verbatim instead of a sentence, even when different variations of the prompt were tried. This left GPT-3.5, which is also the cheapest one to use.
The GPT-3 models have been trained on quality-filtered web crawls and book and Wikipedia corpora. They contain 93% English text and 0.1% (220 million words) in Danish (Brown et al. 2020). The GPT-3 models had been shown by Brown et al. (2020) to have multilingual capabilities. It was chosen to use GPT-3.5 over a Danish-specific model because at the time of writing, the only other remotely comparably-sized PLM specifically trained on a greater portion of Danish text is GPT-SW3 (Ekgren et al. 2023) with 40B parameters instead of up to 175B for GPT-3, and GPT-SW3 was only available on a very restrictive license where each test user in the user study would have had to agree to the license and the experimenter would have had to monitor their use, which was deemed too impractical.
GPT-3.5 has, in addition to the above, been trained with reinforcement learning with human feedback to be helpful with answering prompts containing instructions (Ouyang et al. 2022). Thus, it is indeed more suitable for the prompting approach to be used in this section than the standard GPT-3 models.

### Prompt Design

For this one, a greater variety of prompts were tried. Zero-shot prompting was the first variant tried, with the prompts tried being shown in figure 3.1 and an inverted version shown in figure 3.2.
While valid sentences containing the input words were returned most of the time, there were still instances where the word list was just returned verbatim or the model refused to perform the task, returning the Danish equivalent of

```
"[List of 5 due words] - Lav en sætning med de givne ord

###

"
```

**Figure 3.1.:** Zero shot prompt 1 (Between the quotes. Danish instruction translates to "Make a sentence with the given words")

```
"Lav en sætning med de givne ord: [List of 5 due words]

###

"
```

**Figure 3.2.:** Zero shot prompt 2 (Between the quotes. Danish instruction translates to "Make a sentence with the given words")

either "It is not possible to form a sentence with these words." or "I am not sure what you would like me to do". Showing such a sentence to the user once might not be a big problem, but if it happens repeatedly, it would be an annoyance to see these very specific sentences regularly, and it would reduce the scheduling score. This problem was mainly addressed by switching to few-shot prompting. The prompt was extended to give one, two or three examples (always the same ones) as shown in figure 3.3. The three examples were chosen to be diverse in their grammatical structure (two normal sentences and one exclamation, two pure main clauses and one with a subordinate clause) and in the words used to demonstrate to the LM that any kind of sentence may be generated and it might encounter any kind of word in the input words.
Prompting GPT-3.5 through the API has the format of a chat where multiple messages can be provided as input. It allows to assign messages different roles, such as "user" or "system", where the user message is the instruction to which a reply should be given and the system sets more of a general setting or scenario for the conversation (OpenAI n.d.). So in addition to the previously described user prompt, it was possible to provide a system prompt and I experimented with two different ones and with not having a system message at all. The two system messages were:

```
"Lav en korrekt sætning med de givne ord.

###

ord: en har at sådan; sætning: Vi har ønsket, at der var en løsning.
ord: nyhed for god rimmelig; sætning: Det er en god nyhed for os!
ord: rigtigt se hellere i udenfor københavn; sætning: Jeg vil hellere
     kunne se rigtigt udenfor.
ord: [List of 5 due words]; sætning:"
```

**Figure 3.3.:** Three shot prompt (Between the quotes. First line translates to "Make a sentence with the given words". "ord" translates to "words", "sætning" to "sentence".)
The one and two shot version only used the first or first two of these examples.

1. "Du er conciseGPT, dine svar er meget korte, maks 5 ord." – meaning "You are conciseGPT, your answers are very short, max 5 words.",

2. "Du er conciseGPT, dine svar er meget korte, maks 10 ord, men korrekte og giver mening." – meaning "You are conciseGPT, your answers are very short, max 5 words, but correct and making sense."

### Other Parameters Tuned

Another approach to counteract the problem of the model sometimes refusing to return valid sentences was to use the API's option to return multiple replies for the same prompt. Since usage is billed by tokens and most of the tokens are used to formulate the prompt containing up to three examples, returning three possible output sentences instead of just one did not add much cost compared to the cost of sending the prompt. It was thus decided to always request three outputs and choose the one with the best scheduling score since outputs like "I cannot do this" will probably have very bad scheduling scores unless the words in them are by coincidence due soon.
Having three output sentences also allowed to take only those into account that were grammatically correct: For each generated sentence, the language model was again asked "Only reply with yes or no: Is "{sentence}" a correct sentence in {language_name}?" (the same question is asked to get the correctness

| Generated task | Due words on the day |
|---|---|
| Jeg føler mig følelsesmæssigt drænet. | [så, de, se, følels, følelsesmæssig, over, sig, tæt] |
| Jeg føler mig følelsesmæssigt påvirket. | [så, de, se, følels, følelsesmæssig, over, sig, tæt] |
| Jeg føler mig følelsesmæssigt drænet. | [så, de, se, følels, følelsesmæssig, over, sig, tæt] |

**Table 3.1.:** Chronological account of tasks and vocabulary when the system was looping because of incorrect lemmatization. "følelsesmæssig" remains in the due words despite the forms "følelsesmæssigt" being generated and marked correct by the user. There might be the same issue with "mig" being generated for the lemma "sig" by GPT-3.5, but "mig" being lemmatized by the lemmatizer to "jeg". "følels" should never have entered the vocabulary in the first place, since it is not a correct lemma.

metric, see section 3.5.1).

If only one or two sentences out of the three were labeled as correct, only they were then ranked by scheduling score and the best one used. For this correctness prompt, no system message was used, and the temperature, which determines how likely the model is to sample less likely tokens, was set to 0.0 to get the most likely answer.

Finally, the last parameter altered was the sampling temperature during the task generation with 0.2 being suggested in the API reference as a value that makes the output quite deterministic and 0.8 making it more random (OpenAI n.d.). Thus, these two values were tested.

Due to the costs associated with using the OpenAI API, not all combinations were assessed, only the most promising ones. That means when modifying a parameter in a certain direction yielded considerably worse results, this modification was not tested in other combinations. Please refer to chapter 4 for a table with the combinations that were assessed.

#### Lemmatization and Issues about Looping Tasks

One of the biggest issues observed with the use of GPT-3.5 for generating tasks was lemmatization. Above all, it is a pedagogical question whether the user's vocabulary should consist only of the lemmas the user has seen or all the different forms of these lemmas. To give an example from the Danish language, if the word "god" (good) has been reviewed by the user today, it might not make sense to show the inflected form "godt" of the same lemma to the user for a while. It can be argued that both of these forms should be

| Generated task | Due words on the day |
|---|---|
| Det var godt for os. | [god, er, sundt, vi, dette] |
| Det var godt for os. | [god, er, sundt, vi, dette] |

**Table 3.2.:** Chronological account of tasks and vocabulary when the system was looping because of not using lemmatization. "god", "er" and "dette" remain in the due words despite the forms "var", "godt" and "det" being generated and marked correct by the user. Since the sentence uses many of the due words, the model assigns it a very high probability and it will be generated over and over.

scheduled together and then when the lemma is due, one of them will be generated, ideally be chance to make sure the user still sees a variety of forms. On the other hand, if a language has many inflected forms with different meanings, it can be argued that these meanings should be introduced only gradually, following how commonly they are used. Danish has only three verb inflections to represent time (present, perfect and imperfect), but still, it might be pedagogically valuable to show only present forms up to a certain point to not overload the user with verb forms, which could be achieved by not having a lemmatized vocabulary, since then only a present form could be used in the generated tasks if a present form from the vocab is due to be reviewed.
So with both approaches having their pros and cons for the Danish language, I decided to base my choice on the technical difficulty of the approaches. With the previously chosen prompt and parameters, GPT-3.5 has a tendency to generate the word form related to the input word, which best fits the grammar of the sentence. Meaning, that if the form "godt" is in the due input words, it might use the form "god" in the sentence. I tried whether this tendency could be overcome by adding a sentence "Generate the exact words forms given" to the prompt or system message, but the LM still had that tendency.
This tendency should in theory be perfect for the lemmatized approach, where all input words would just be the lemma and the LM would generate any of the related forms. However, the biggest difficulty to a lemmatized approach is posed by the imperfection of machine lemmatization. The spacy lemmatizer for Danish has an accuracy of 0.95 (Explosion 2016), meaning that it is not always possible to do the reverse association from the single words in the sentence, for which the user rates how well they remembered them, and the input words from the existing vocabulary. For an example that actually occurred in the evaluation runs, see table 3.1: When the user reported they correctly remembered the word "følelsesmæssigt" in the sentence, it was incorrectly

lemmatized to "følelsesmæssig**t**" and thus the actual lemma "følelsesmæssig" was not be rescheduled and remained in the due words. Then for the next task, it was one of the input words again and this is how the LM ended up generating the same task containing many words that are incorrectly lemmatized over and over because the lemma is never removed from the due words.
Loops also occurred if lemmatization was not used. If the model generated a task that contains a different form of each of the input words' lemmas, only these inflected forms were rescheduled when the user marked them as known or unknown, leaving the actual input words to still be due, which in turn means that the language model would keep generating the same sentence again, see table 3.2. Sooner or later the random sampling would result in a different sentence, but especially at temperature 0.2, this can take a long time if the LM has a very strong preference for this token sequence.
Four different measures were taken to alleviate the issue. Firstly, if there were more than five due words, the input words actually shown to the LM were sampled randomly from all due words, and their order was always shuffled. This partly mitigates the problem of loops due to missing lemmatization discussed in the previous paragraph, since the model would get different input words (if more than five are due) or at least a shuffled version at each generation time and would thus be far less likely to generate the exact same sentence again.
Secondly, same-day repetitions of the exact same task were avoided if possible, just like in the other methods. They were avoided by filtering out any tasks that had previously been generated on the day from the three options generated by the LM, even before incorrect ones are filtered out and the one with the best scheduling score is selected.
Thirdly, lemmatization was discarded not to be implemented in the scope of this dissertation, since the problem of words getting stuck in the due state due to inaccurate lemmatization is not solved by the previous two measures.

### 3.3.3 Hybrid Model

The hybrid method combines a BM25 retrieval method with a GPT-3.5 model. For each task to be generated, either of the two methods is chosen with a chance of 50%. The proportion of 50% was chosen as a trade-off between using the LM as much as possible, since determining how well-suited an LM is for task generation is one of the main purposes of this dissertation, and using

retrieval often enough to catch words that are stuck in the due words list, as explained in 3.3.2. Also, a half-half split will arguably maximize variability and thereby make the tasks less boring to users.
If the LM method was not able to generate a previously unseen task (because all three options generated were previously seen), the retrieval method was queried instead, but this happened less than one percent of time in the simulated user runs.
The aim of experimenting with this hybrid model was that it should not suffer from the lemmatization problems that the LM-only method has, since a hybrid model which makes intermittent use of the retrieval method will sooner or later remove any stuck words from the due words list since the retrieval method only looks for exact matches for words and not related words that share the same lemma.

### 3.3.4 Modifying the Probability Distribution for each Token

Using a pre-trained model and modifying the probability distribution of the output tokens was identified in section 2.2 as a promising approach used in previous literature. Preliminary tests were done using GPT-2 (Radford et al., 2018) where the next token was selected in a greedy way as $max(\frac{p_{token}}{(rank_{token})^{\alpha}})$ favoring the first tokens of more common words in the English language (rank of the first token of the most common word is 1, down to 50000, tokens not in the 50000 most common words are disallowed, spaces and punctuation ranked one). $\alpha$ is a scaling factor and it was experimented with different ones, where 0.075 turned out to be the sweet spot where a small influence on the generation was observed but higher values lead to only the most common word or punctuation being generated. In general, the method was deemed to be too unstable and the tuning of the factor to be too much guesswork to move on in the implementation.

## 3.4 Computed Metrics

From the objective from 3.1 and observations during the development process, metrics are defined to assess any possible solutions. Metrics are divided into automated metrics and human/LM-based metrics.

1. Perplexity, to measure cohesiveness and correctness

2. Scheduling score, to measure how well the spaced repetition scheduling is adhered to and only known vocabulary is used (for more details on the scheduler, see 5.2.2)

3. Too long sentences, to measure the fraction of sentences that are longer than the ten word limit from the third objective

### 3.4.1 Perplexity

The Perplexity (Jelinek et al., 1977) of a large language model, which we know has a good representation of the language, can be calculated on the tasks that each method outputs, to get an estimate of the cohesiveness and correctness, like has been done using GPT-2 as a censor in (Guo et al. 2022) and (Chen et al. 2022). GPT-SW3 (Ekgren et al. 2023) serves as the censor in this work since it is the only other big general-purpose LM where the training data included Danish corpora. Similarly to the standard GPT-3 models, it was trained on web crawls and book and Wikipedia corpora, and there is likely some overlap with GPT-3 training data, but it should have lower bias compared to the alternative which would be using GPT-3.5 as a censor for GPT-3.5 outputs.
The output tasks on which the perplexity is calculated are obtained by simulating a set of potential users and the tasks they would be shown if the method under assessment was used. For the simulation, five users were simulated to keep requesting new tasks from the retrieval-based algorithm and review the words contained in them, with an 85% chance of remembering them. After each task, they had a 10% chance of quitting (and coming back the next day, with new words due), unless they had reviewed the last task the system had to offer them, in which case they quit in half of the cases and requested to learn previously unseen words in the other 50% of cases.

### 3.4.2 Scheduling Score

This metric should measure the second objective defined in section 3.1. In accordance with the loss defined in that section, the "sched score" (equation 3.1) is calculated as the fraction of the scheduling interval that is wasted by scheduling words before they are due. It can be between zero and one and should be minimized. Out of vocabulary words that are new to the user get a score of one, since the model should use them as rarely as possible unless the user requested new words to be added to their vocabulary, in which case these specific selected new words get a score of zero, since they are due today.

$$S = \frac{1}{n_{tasks}} \sum_{tasks} \frac{1}{n_{words\_in\_task}} \sum_{words\_in\_task} s_{word} \tag{3.1}$$

$$s_{word} = \begin{cases} 0 & \text{if not seen before, user-requested new word} \\ \frac{max(t_{due} - t_{now}, 0)}{t_{due} - t_{last\_seen}} & \text{otherwise, if word in user vocabulary} \\ 1 & \text{not seen before, no new words requested} \end{cases} \tag{3.2}$$

Here again, the tasks are obtained by simulating a set of potential users and the tasks they would be shown.
I considered giving a higher $s_{word}$ than 1 for words that are not in the correct language, however, decided against it, since it would permit scheduling scores greater than one, which would reduce interpretability. Also, this case should be covered by the perplexity metric, if the censor model assigns low probabilities to mid-sentence language changes, as well as by the correctness metric (see 3.5.1).
It should be noted, that often it would possible to achieve good scheduling scores just by repeating the same sentences over and over, especially if the user marked most of the words in the sentence as unknown. This would avoid incrementing the sum by 1 for having scheduled previously unseen words that were not user requested and thus bring down the scheduling score. This did in fact happen in testing and would reduce the variety of contexts the user sees, which is why repetitions of the same sentence on the same day were disallowed when generating tasks. An alternative option to also discourage repetitions across days would have been having a term in the scheduling score that increases the score when a previously scheduled sentence is scheduled. However, this means that all previously scheduled sentences would have to be

saved and compared at each generation, and seeing some sentences again on the other days is arguably not as big of a problem as seeing them again on the same day, so this option was not chosen.

## 3.5 Human / LM-based Evaluated Metrics

The following metrics about correctness and variability cannot automatically be calculated, so they had to be evaluated manually by humans. However, as Chiang and Lee (2023) note, prompting a large language model can be used as an alternative to human evaluation and the scores are likely to be correlated with human judgment for many tasks. Thus, human evaluation of a subset of outputs was used, along with a more extensive LM-based evaluation of all outputs. The scores were checked for correlation between the human and LM to judge their aptness.

### 3.5.1 Correctness

In addition to the perplexity metric, the correctness of the outputs of a task generation method was assessed by asking a human and a language model the same question, "Only reply with yes or no: Is "{sentence}" a correct sentence in {language_name}?"
A human evaluator was shown twenty randomly sampled output sentences of each generation method that should be assessed. The language model (GPT-3.5-turbo-0301) was shown all of the ca 1000 outputs. Both saw them one at a time. The mean of the human and the mean of the LM scores are both reported.
To determine whether the LM had the same understanding of correctness as the human evaluator, agreement and Cohen's Kappa between the two was calculated on those samples that were shown to both the human evaluator and the LM (220 in total). LM and human evaluator agreed in 84% of cases and Cohen's Kappa was 0.35, which means that there is moderate agreement between the two, surpassing what would happen by random chance.
Thus, it can be argued that the LM's score might be more accurate since it is based on 1000 instead of just 20 samples per method, and should somewhat align with human judgment. The human score is additionally also reported.

```
"Sentences:
[Sample of 20 sentences, one per line]

---

How varied do you find the topics of these Danish sentences (on a
scale of 1-5, with 1 being the lowest)?"
```

**Figure 3.4.:** Zero shot prompt 1 (Between the quotes. Danish instruction translates to "Make a sentence with the given words")

### 3.5.2 Variability

Even though not directly defined as an objective in 3.1, in the introduction 1.3 it was mentioned that using a language model to generate sentences might be able to generate a great variety of sentences. Thus, a metric for this was added to assess this.

Both the human and LM evaluator were also asked to rate variability, with the LM prompt shown in figure 3.4. The human evaluator was asked the same question directly after rating the correctness of each method's batch of 20 output samples, while the LM got to rate 20 randomly sampled sentences, 200 times for each method, with the average of these 200 runs being the final score. Here, there was not much agreement between LM and human rater on the ratings for the eleven methods, in fact, their ratings were slightly negatively correlated (Pearson's $r$ -0.28). Thus, both human and LM ratings are reported, but not much weight was given to this metric when selecting the methods to be tested in the user study since the human evaluator only saw a fraction of samples and the LM seems not to share the same criteria for variability.

# 4 Results and Discussion

This chapter presents and discusses the results of the experiments defined in the previous chapter. Table 4.1 shows the results of computing the metrics mentioned in section 3.4 on those methods and their configurations that were implemented and not immediately discarded.

### Sentence Length

While the retrieval model never outputs sentences longer than ten words, since they have been removed from the corpus, the language model has a tendency to output more than the desired ten words sometimes. This is even more pronounced when providing no system message, or a system message that instructs the model to output a maximum of ten instead of five words, but even more when providing ten input words out of which to form sentences, instead of five. The fewer examples ("shots") were provided in the prompt, the fewer of the outputs violated the ten word maximum.
Predictably, the hybrid model was around half as likely as the GPT model that it was using (line 6 in the table) to output too long sentences since the retrieval was used in the other half of cases, which never returns too long sentences.

### Correctness

Perplexity turned out not to be a very good indicator of correctness. Because of this, I only relied on the correctness metric (labeled "incorrect" in the table since lower is better) to draw conclusions about it. As was to be expected, the retrieval models only returned correct sentences, since they are human-written and Wikipedia has a community curation process that mostly prevents bad language.
Figure 4.1 visualizes the main trends among the different GPT-3.5 configura-

tions from the results table. Out of the GPT-3.5 models, the least incorrectness was achieved by the sixth configuration, while increasing the temperature parameter, as well as providing more input words negatively influenced correctness. It makes sense that higher temperature would lead to the model being more likely to generate incorrect sentences. It also makes sense that the model might de-prioritize correctness when more inputs are provided, just trying to squeeze in as many of the then longer list of input words into the sentence. However, this was not investigated further to confirm it, since the combination of lower correctness and tendency to generate too long sentences was already reason enough to not give more than five input words.
Also, having less than three shots seems to reduce the correctness of the generated sentences, which could also be a sign that one or two examples do not make it clear enough that the outputs should be correct Danish sentences.
As was to be expected, correctness is also increased by around five percent by filtering out sentences which GPT-3.5 would label as incorrect ("prefer correct" criterion in the table), as is evident when comparing lines five and six of the results.
Once again, the correctness of the hybrid model was in between the two methods it switches between, at least for the human ratings.

**Variability**

None of the GPT-3.5 outputs was rated varied at all by human evaluators. There was a slight difference in the LM's evaluations, it found the models which got more input words or used a higher temperature slightly more varied. However, this small possible advantage cannot offset the fact that these models generated far more incorrect sentences and thus they were not considered further.
The retrieval models were preferred by the human evaluators getting a slightly better score of 2, but rated as the least varied by the LM. Human evaluators had a clear preference for the hybrid model, which the LM rated as more or less equally varied as the GPT outputs.
The differences between human and LM ratings were not further investigated, since these results are not used to select the methods to move on with, as explained in 3.5.2.

| Model | Tempe-rature | Input Words | Shots | System Message | Best out of n, critera | Sched score | Ppl | >10 words | Incorrect (GPT \| Human) | Varied (GPT \| Human) |
|---|---|---|---|---|---|---|---|---|---|---|
| gpt3.5 | 0.2 | 5 | 3 | none | 3, best sched score | 0.068 | 150 | 18.7% | 8.5% \| 50% | 3.0 \| 1 |
| gpt3.5 | 0.2 | 5 | 3 | 1 | 3, best sched score | 0.124 | 170 | 5.4% | 11.5% \| 25% | 2.6 \| 1 |
| gpt3.5 | 0.2 | 5 | 1 | 2 | 3, best sched score | 0.094 | 123 | 7.0% | 25.3% \| 55% | 2.9 \| 1 |
| gpt3.5 | 0.2 | 5 | 2 | 2 | 3, best sched score | 0.068 | 150 | 12.7% | 20.2% \| 45% | 3.0 \| 1 |
| gpt3.5 | 0.2 | 5 | 3 | 2 | 3, best sched score | 0.070 | 133 | 19.1% | 8.0% \| 20% | 2.6 \| 1 |
| gpt3.5 | 0.2 | 5 | 3 | 2 | 3, prefer correct ->best sched score | 0.068 | 155 | 19.6% | 4.2% \| 15% | 2.4 \| 1 |
| gpt3.5 | 0.8 | 5 | 3 | 2 | 3, prefer correct ->best sched score | 0.082 | 125 | 13.1% | 14.6% \| 40% | 3.0 \| 1 |
| gpt3.5 | 0.2 | 10 | 3 | 2 | 3, prefer correct ->best sched score | 0.077 | 152 | 44.1% | 11.0% \| 35% | 3.2 \| 1 |
| BM25 | - | 25 | - | - | 1 | 0.113 | 121 | 9.9% | 0% \| 0% | 1.5 \| 2 |
| BM25 | - | 25 | - | - | 25, best sched score | 0.098 | 102 | 8.5% | 0% \| 0% | 1.8 \| 2 |
| Hybrid | 0.2 | 5 (LM) / 25 (BM25) | 3 | 2 | 3 (LM) / 25 (BM25), prefer correct -> best sched score | 0.078 | 203 | 11.2% | 4.5% \| 10% | 2.8 \| 4 |

**Table 4.1.:** Comparison of the considered models and parameters with computed, LM-prompted and human scores (for varied and wrong metrics only).
System messages:
1: "Du er conciseGPT, dine svar er meget korte, maks 5 ord.",
2: "Du er conciseGPT, dine svar er meget korte, maks 10 ord, men korrekte og giver mening."
The column "Best out of n, criteria" describes how many outputs were generated by the method and the criteria by which the best of them was selected as the final output. "Prefer correct" means that out of the n results, only the correct ones (determined by prompting GPT-3.5) were considered for the next criterion. If none was correct, all were be considered. For more details see section 3.3.2

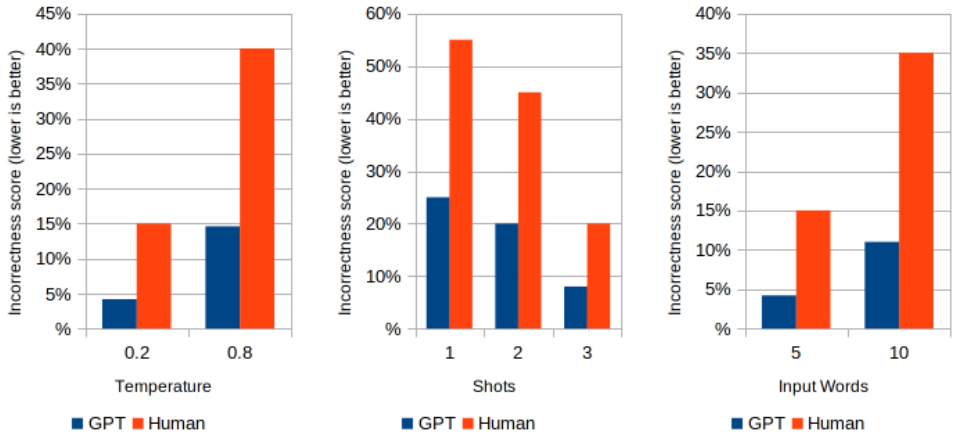



**Figure 4.1.:** Influence of difference temperatures, number of shots, and number of input words and correctness

### Scheduling Score

The scheduling score was quite similar for most of the GPT-3.5 models, around 0.07-0.08. Only the model which was instructed to generate shorter sentences of up to five words had a far worse scheduling score. This might be due to the division by task length in the formula, e.g. if one new word is used in a task of length five and one new word in a task of length ten (and all the other words are due on the day), the first one would get a score of 0.2 and the second one a score of 0.1. Since all the alternatives produced decent scheduling scores, the fine differences here were not primarily used to choose which configuration to continue with. One-shot-prompting resulted in a worse scheduling score, which could mean that it was not clear enough to the language model what to do with the input words from just one example.

The retrieval models had slightly worse scheduling scores. Predictably, choosing the sentence with the best scheduling score out of the 25 best-ranked sentences turned out to improve the scheduling score.

The hybrid model had a scheduling score in between the retrieval model and the LM it is based on.

All in all, the scheduling scores are very high, meaning that most words in the tasks must have been due on the exact day they were generated. The scheduling scores below 0.1 mean that on average, less than one in ten words in the tasks were out-of-user-vocabulary, and less than one in five was not due on the day the sentence was generated.

**Looping**

All LM-only methods suffered from problems with looping explained in 3.3.2, despite the three countermeasures mentioned there decreasing their likelihood. The retrieval and the hybrid method did not suffer from this problem.

## 4.1 Conclusion

Out of the GPT-3.5 models, the sixth one was the most correct, was tied for the best scheduling score, and had an acceptable amount of sentences that were longer than the goal of ten words. It was the least varied according to LM-based evaluation, however, this possible tradeoff was acceptable, especially since humans rated non of the models very favorably with regards to variability and the LM-based variability might not be very accurate in the first place, as discussed in 3.5.2. Thus, it was decided to use this model in the hybrid model, but not to risk users getting stuck in loops in the actual user study.
When it comes to the BM25 models, using the best-out-of-25 strategy improved the scheduling score and had no other downsides, as documented in table 4.1 and was thus chosen as the retrieval method to test in the user study and to be part of the hybrid model. As was to be expected with the hybrid model using two models 50% of the time each, most metrics come in right between the used GPT-3.5 model and the used BM25 model. Variability, however, was higher than that of its parts. Thus, solving the looping problems and performing decently in the metrics, it was decided that the hybrid model is adequate to be the way how LM generated tasks are tested in the user study.

# Part II

User Study

# 5 Method

This chapter defines the method used to answer the second research question: How does sentence-based spaced repetition using the best-performing options from the first question influence user engagement and learning outcomes among language learners, compared to conventional approaches?In addition to the two proposed sentence forming methods, which have been selected in the previous part, a baseline method was developed for them to be compared to, which is presented here. Then, I designed the user study by defining the metrics to be collected from the participants, as well as the sourcing and procedure for the division of participants into experimental groups. Concurrently, the test system was designed and implemented, and the details about its design are documented in section 5.2.

## 5.1 Baseline ("Single Word Group")

Four candidates for baselines were considered, with the last one being chosen:

1. ignoring the spaced repetition scheduling and retrieving sentences with any words from the user vocabulary

2. only showing single words at a time instead of sentences

3. use sentences as the scheduling object, instead of single words

4. having a set sentence for each word, show that sentence when the word is due and only focus on that word

These are similar to approaches that learners are already using to learn a language. The last three had already been identified as common practices in section 2.1.1. Ignoring the spaced repetition scheduling is similar to reading

texts of approximately the difficulty that is suitable to the learner (since even though the exact scheduling is ignored, only sentences with mostly words in the user vocabulary are preferably retrieved).
The drawback with showing only a single word at a time is that users could become aware that they are in the control group if they notice that they only get single words while other users get whole sentences, and become demotivated by that fact.
A common practice is also to fill spaced repetition flashcards with sentences, as mentioned in section 2.1.1, instead of each word independently. The intention is to provide context for either a specific word or to learn all the words in the sentence together. Something similar could be used as the baseline, either by also scheduling whole sentences or by having a set sentence that is associated with each word and shown when that word is due, while the due word is highlighted and the others grayed out, see the left screenshot in figure 5.1.
I opted for the latter since not scheduling words independently from each other would make it more difficult to accurately assess the user's vocabulary growth (number of words). The drawback of highlighting a single word in the latter baseline is that users could notice differences between each other and conclude that they are in different groups since the other two groups don't focus on a specific word in the sentence, but all of them. But since the users are not provided with specific information about the different groups, how many there are, or what the differences are, there should not be any major unwanted effects on motivation.

## 5.2 Test System Design

A mobile app was developed as a front-end for the user to interact with the generated tasks. The portability aspect was deemed important, since MALL had been found to deliver the best learning outcomes in section 2.1, at the time of writing this, mobile is the most popular platform to access web apps from ("Global Mobile Traffic 2022" n.d.) and due to the nature of spaced repetition, which relies on one or a few short review sessions every day, instead of more infrequent longer sessions, which a desktop would be more for.

### 5.2.1 UI

Due to the limited time and since the focus of this work was on the task generation methods, it was decided to keep the UI of the language learning app very simple and focus on the implementation of absolutely necessary components. The goal was to make the UI intuitively understandable and provide all necessary instructions in the UI, to minimize the need for any user training.

It was decided that only one UI screen would be necessary which presents the tasks and their solutions to the user. Account creation was handled by the experimenter directly in the back-end, and authentication was through a personalized link that the experimenter then sent to the participants, so that the first thing the user would see when opening the app was the first task and a button to show the solution, along with some minimal instructions and two buttons to install the app to their device's home screen and activate a daily notification. This state is shown in figure 5.1

The user would then, after thinking about a translation to the task, click the button to show the solution and compare it with their translation. They would then mark all words in the task that they did not remember correctly (or had never seen before). Through seeing a solution and the option to click a dictionary icon next to the words they marked, they could learn the meaning of new words, and refresh their memory of old ones. This is shown in figure 5.2 on the left.

After selecting all unknown words, they would then press the main button again to be shown the next task, and so on, until they either wanted to stop, or they had reviewed all words that, according to the spaced repetition system, were due on the day. At that time, a "done for today" screen was shown, as seen in figure 5.2 on the right. This was intended as a natural stopping point for users, however, if they were motivated enough to spend more time, they were given the option to add five new words to the vocabulary and the system would generate tasks containing these words and show them immediately. This option could be used repetitively, so the user could study for as long as they wanted.


**Figure 5.1.:** Screenshots of a task as seen by baseline (left) and retrieval/hybrid group (right)

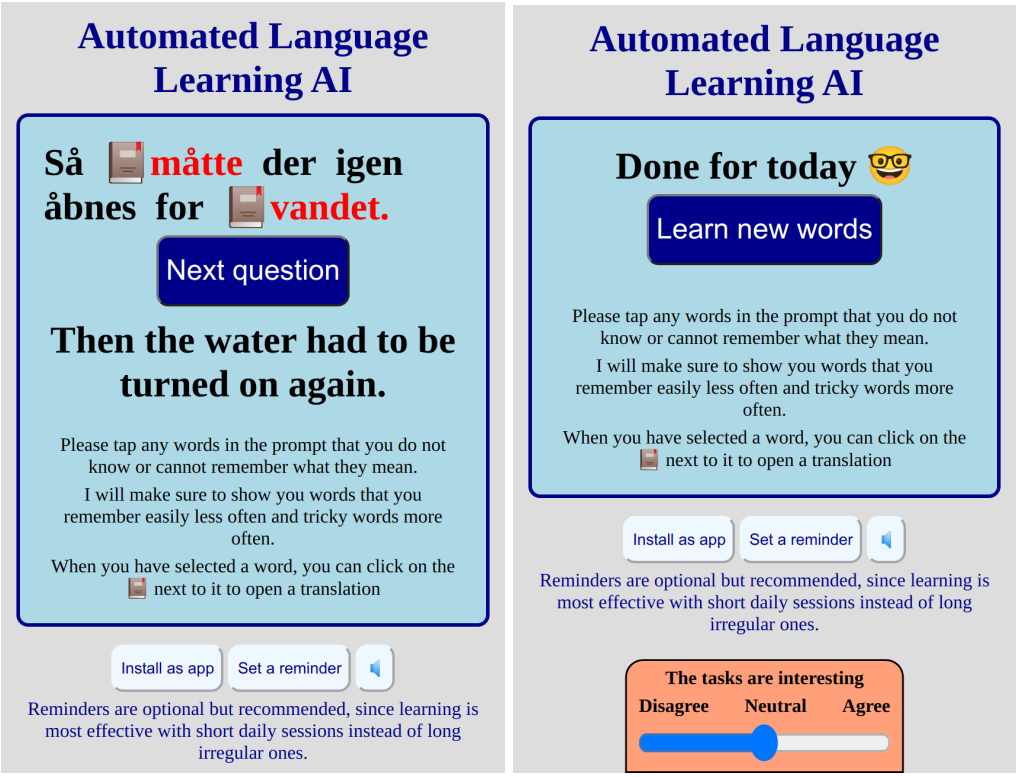



**Figure 5.2.:** Screenshots of solution being shown with two words selected as unknown (left), and "done for today" screen (right) with the interestingness prompt being shown, as described in section 5.3.6

### 5.2.2 Spaced Repetition Algorithm

Each word in a user's vocabulary acts as a spaced repetition item. The user would start out with the ten most common words of the target language in their vocabulary, and each time the "More words" button was pressed, five new words of appropriate difficulty (see section 5.2.2 for how this was determined) were added to the vocabulary. Since we have no practical way of knowing, if and how well a user already knows a word when they first see it, the assumption is made that all the words are completely unfamiliar when they first appear, and the spaced repetition intervals thus start from zero for any word encountered.
The spaced repetition algorithm implemented was the popular SM-2 algorithm (Wozniak 1990), a variation of which is for example used by Anki (Elmes, n.d.), one of the most widely used spaced repetition programs.
One simplifying modification was made: While the SM-2 algorithm grades responses on a six-point scale, assuming three different severities of not remembering a word and three different levels of correctly remembering a word (e.g. "perfect response" and "correct response after a hesitation"), here it was decided to only use a two-point scale, one for remembering and one for not remembering, doing away with the need for measuring the time a user hesitates and making assumptions, such as where to set the time limits and that hesitation could not be caused by for example exterior distractions. The grades used in this system correspond to grades 1 (not recalled) and 4 (recalled correctly) in the original SM-2 algorithm.

#### User Requesting more Words

As previously mentioned, the user could request to have new words added to their vocabulary when done with all tasks on a day, to allow them to be fully flexible with the amount of time they would like to spend on the app on a given day. For beginner users, new words are just added from the simplest words available on the frequency list. For more advanced users, there is some skipping ahead if the user has marked very few words as unknown up to the point when they requested more words. If the user had marked less than five words as unknown, the 500 easiest new words were skipped on the frequency list, if it was less than nine words, the 200 easiest new words, otherwise the

five easiest new words were added to the user vocabulary.
For an impression of the hike in difficulty, among the initial words in the user vocabulary are words like "jeg" ("I") or "på" ("on"). At difficulty 210 we have words like "gange" ("times") and "gennem" ("through") and if we skip 500 words to 510, we have words like "stod" ("stood") and "venstre" ("left"). This means that very advanced learners in the single-word group will probably have to request new words multiple times to find new words.

### 5.2.3 Architecture

The app was chosen to be a progressive web app (PWA), since these are platform-independent, reducing development time since no separate apps had to be developed for different mobile operating systems. They do provide most features that normal apps do, like the option to install and send notifications that have both been implemented.
A client-server architecture was employed to run spaced repetition scheduler and sentence generation method on a Python server and communicate via web APIs with the front-end app.
For simplicity, only one server and worker thread were implemented, which did sometimes result in up to a few seconds of delay when multiple users were using the app concurrently.

## 5.3 Metrics

To answer the question of how user engagement and learning outcomes are affected by the different methods, the following metrics were assessed for each user group using a different method (see the following subsections for how they are connected to engagement and learning outcomes):

1. User vocabulary growth

2. Time efficiency

3. Word effectiveness

4. Quantitative proxies for user engagement

5. Subjective interestingness, enjoyment, perceived learning, challengingness, and confusion while learning

6. Perceived usefulness and perceived ease of use before and after the trial period

All of the metrics were assessed either from usage data or questionnaires and each user received the same weight. The data was then analyzed for correlations between all the metrics and demographical data, in case these uncover some major confounding factors, and these are reported if the absolute value of the Pearson coefficient is at least 0.5. Primarily, the differences between the participant groups for each metric were reported and the significances of these differences were reported to answer the question of whether the use of retrieval-based methods and LM-based methods increases user vocabulary growth, learning efficiency, user engagement, and perceived usefulness and ease of use.
For the significance testing, the one-sided Mann–Whitney U test (Mann and Whitney 1947) was used to determine the significance of the differences between the groups with regards to the metrics (for the exact hypotheses formulation, please refer to A.3. It tests whether a probability distribution is greater than the other and does not assume normally distributed data. This is important since for these metrics it cannot be expected that data follows a normal distribution. E.g. for learning and engagement metrics, it is likely that data is skewed to the left because most users only learn very little or interact with the app very little, while some few very motivated participants would learn/engage a lot. For the 1-5 scaled subjective metrics, a normal distribution cannot be assumed unless the mean lies in the middle at 3. Results were considered significant if the p-value was smaller than 5%. P-values above 10% will not be reported, in that case, it will only be reported that the h1-hypothesis is refuted.

### 5.3.1 User Vocabulary Growth

The research question is to assess differences in learning outcomes. This can be operationalized mainly as vocabulary growth since that is the main goal of

a vocabulary teaching system.
While the most common and most accurate way to test vocabulary growth is conducting a vocabulary size test (Olmos 2009) before and after the period when the system is being used, this approach requires extra time investment by the students and investigator.
Thus, it was decided to extract vocabulary growth directly from usage data that is logged in the background while users interact with the app. Any word initially marked as unknown but then declared known when the user last sees it was counted as growth.
It is important to note that word in this case refers to any possible word form, e.g. counting "help" and "helps" as two different words, even though they share the same lemma. More on considerations about lemmatization in 3.3.2.

### 5.3.2 Time Efficiency

Another way to see the purpose of a vocabulary teaching system is that it should teach vocabulary as efficiently as possible, which could also be deemed an aspect of "learning outcomes" mentioned in the research question. Measuring both time spent using the app and vocabulary growth will allow to explain whether any improvement in vocabulary growth was due to more efficient learning, or due to more time spent learning. The learning efficiency was calculated as vocabulary growth divided by the total time spent.

### 5.3.3 Word Effectiveness

A slightly different aspect of learning outcomes is effectiveness. The word effectiveness was defined as vocabulary growth divided by number of words seen. This makes it possible to draw conclusions on whether any method is didactically advantageous by making users remember a larger share of the words seen.

### 5.3.4 Number of Distinct Words Seen

The number of distinct words seen by each user was measured. While it is not directly related to learning outcomes nor engagement, but to the variedness of

the methods, it is logged to calculate other metrics, such as word effectiveness, and to analyze the reasons for differences seen in other metrics.

### 5.3.5 Proxies for User Engagement

Even though we can measure the vocabulary growth directly, it is a good idea to assess the engagement as a proxy, since this might be a good indicator of long-term learning outcomes - even if users have good learning outcomes during a short ten-day user study, if they don't feel engaged, they might not be able to keep up that performance in the long term.
User engagement cannot be measured directly, but generally, the more engaged the user is, the more time will they spend using the system. Consequently, the total time using the system and the number of sessions were measured.

### 5.3.6 Subjective Ratings

I suspected that more interesting/enjoyable tasks would tend to improve engagement and learning success, while confusing or overly challenging tasks would be detrimental. These were thus also assessed by prompting users during each usage session. The subjective amount of learning was also prompted. The point of time when to ask was randomized with a random question appearing with a chance of 20% after each task the user completed and when the user was done for the day. Of course, the user can stop the session before that, which means that for some sessions, no data might be obtained for certain metrics. The statements shown to the users were:

- interestingness: "The tasks are interesting"
- enjoyment: "I am enjoying this"
- learning: "I am learning a lot"
- challengingness: "This is challenging"
- confusion: "I am confused"

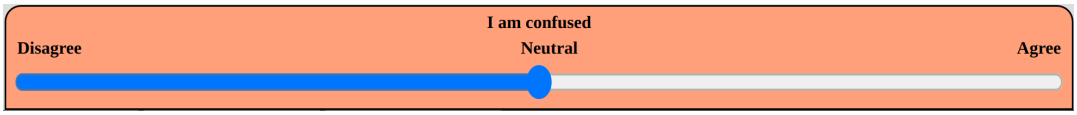



**Figure 5.3.:** Prompt to the user to give a subjective rating

They were presented at the bottom of the page and the user could drag a slider to mark how much they agreed with the statement, see figure 5.3. Before the start of the trial, the users were instructed in writing to provide this feedback. The users were also prompted once to give these subjective ratings again in the final questionnaire, which also included an optional free text comment field. The questions in that questionnaire are listed in appendix A.1.

### 5.3.7 Perceived Usefulness and Perceived Ease of Use

Following the Technology Acceptance Model (Davis 1989), perceived usefulness and perceived ease of use are the key determinants of the adoption of new technologies. It seems plausible, that increased user engagement and learning success would also increase user's willingness to adopt the new technology proposed in this dissertation. Thus, users were asked to rate these two, once after seeing initial information about the system's functionality and purpose, including two screenshots, and once after the trial period. The exact questions asked were: "From the info you have, do you think the app will be useful for you?" and "From the info you have, do you think the app will be easy to use?" before they had used the app, and in the final questionnaire "How useful did you find the app?" and "How easy was the app to use? Please rate only the design and ignore any technical problems you might have encountered.", rated on a five-point scale.

## 5.4 Pilot Study

To rehearse the actual user study and confirm that all the necessary data can be obtained, a three-day-long pilot study was conducted ca. 1 month before the actual user study, using a development version of the system and two of the final methods; the retrieval method and the single word method, albeit

the single word method at that time only showed the word, without showing any low-opacity context. A third method that ignored spaced repetition timing was also tested. For each method, one test user participated. One user was a beginner, and the other two were at different intermediate levels.
It was generally concluded that the methodology is suitable to determine the effectiveness of the concept and compare different methods, and all the necessary data can be collected.
Some other takeaways from the pilot study were, that daily reminders were necessary to keep users from forgetting to study, and they were consequently implemented. Also, a previously considered "I am in the flow" subjective metric was dropped for being too unspecific. Having a final questionnaire after the trial period was also a result of the pilot study, since some of the randomly prompted subjective metrics were missing data from two of the users.
Two of the users from the pilot study also participated in the final user study. However, by then the system had changed considerably so that any familiarity the users had gained with it was negligible. Also, as to not give them an advantage in initial vocabulary size, the pilot study was conducted with French as the target language to be learned.

## 5.5 Users

### 5.5.1 Sourcing

After the pilot study, 26 test users were recruited for the actual user study, mainly through social media from the researcher's acquaintances. The only exclusion criterion used was that the user should not be completely fluent in Danish. The test users were thus not representative of the general population. Participants were shown an initial questionnaire (see appendix A.1) before the beginning of the user study, collecting demographical information as well as their background in language learning and initial motivation, which were treated as potential confounding variables. Participants were aged 19 to 56 (mean 28.9, std 11.1). 9 were female and 17 male. Users had 15 different native languages (7 German, 3 Spanish, 3 Chinese, 2 English, 1 each had Albanian, Bengali, Dutch, Hungarian, Italian, Konkani, Odia, Persian, Polish, Portuguese, and Slovak as native languages). 17 were living in Denmark and 9 had never lived there. Those who lived in Denmark had lived there from

ten months up to 6 years (mean 2.5 years, std 1.4 years). 14 had learned Danish before and out of them, 10 of these had used the language outside of a class context. 23 had previously used other language-learning apps. Users reported an average motivation of 3.1 on a 1-5 scale, std 1.0) and mainly career prospects, curiosity, and social life as the motivating factors.
The recruitment through acquaintances could affect the subjective metrics mentioned in 5.3.6 through the social desirability bias, making participants more likely to give more favorable ratings. This has been partly mitigated by putting emphasis on the anonymity of the participants' answers, but cannot fully be avoided. However, it affects all test groups equally, since users did not know which intervention they had been assigned to, so the results still remain comparable between the groups.

### 5.5.2 Grouping

Each user was assigned a participant ID consisting of five random letters and numbers. The participants were allocated randomly into the three intervention groups. Since previous knowledge of the Danish language was suspected to be the biggest confounding factor, users were first divided in two blocks, depending on whether they had answered in the initial questionnaire that they had previously learned Danish. Then, blocked randomization was used to allocate the participants from each block to each treatment group. Lots were drawn without replacement to select an equal number of participants for each group.
The study was double-blind: Users were not told which group they had been assigned to, only that multiple interventions were being compared. As explained in 5.1, for two of the groups the UI looked exactly alike, while the baseline method was slightly different by showing all context words with low opacity. The experimenter was equally unaware of the treatment group assignments during all communication with the participants.

### 5.5.3 During the User Study

Users were allowed to choose freely, how much time they would like to spend using the app, to make it possible to measure engagement in the way described in 5.3.5, and to allow for the long study duration of 10 days with volunteer

testers without getting users frustrated. Users were also allowed to drop out of the study at any time, with their results until that point being considered. They were still asked to fill out the final questionnaire, which 22 out of the 26 did, minimizing non-response bias since any questions asked during the sessions were asked once again in the final questionnaire.

The ten-day trial period ran from Monday to Wednesday the next week for all participants. According to their availability, participants were allowed to choose their start date out of two possible Mondays.

During the user study, some technical problems surfaced, e.g. problems with different operating systems, or timeouts when querying the OpenAi API. Most of them were fixed within a few hours after becoming apparent, and their impact should thus be minor. Still, for some users, this could have caused some frustration and worse usability. The problem of the application not being installable on iOS and iOS users not having the option of receiving notifications could not be fixed due to iOS's limitations, however, the remaining functionalities of the app was still available to these users through a web browser.

# 6 Results

This section discusses the results of the user study. As previously stated, I report the differences between the participant groups for each metric and the significance of these differences, as determined by the one-sided Mann–Whitney U test. Correlations between all the metrics and demographic data are also reported if the absolute value of the Pearson coefficient is at least 0.5 or if there is unexpectedly no correlation where one could be expected.

## 6.1 Presentation of results

### 6.1.1 User vocabulary growth

Figure 6.1 shows the distribution of the vocabulary growth in the different groups. Overall, users' vocabulary grew by 7 words in the median or 11.5 words in the mean. The standard deviation was, however, high at 19.3. The group which only learned a single word per task (labeled "single") had the lowest growth (median 1.5, mean 3.4, std 4.1) while the hybrid group (median 6.0, mean 18.8, std 31.0) and the retrieval method (median 10.0, mean 11.4, std 7.7) had stronger growth. While the retrieval group's median is the highest, the hybrid group's mean is by far the highest due to two outliers with exceptionally high vocabulary growth.

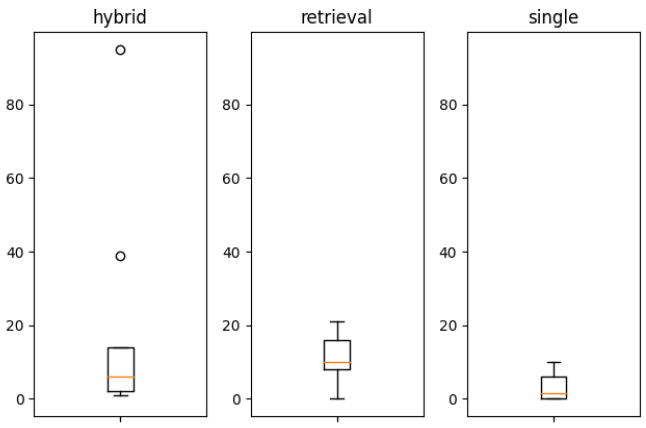



**Figure 6.1.:** Box plot of the vocabulary growth in the different groups

The difference between the retrieval group and the single-word group was

significant (p=1.7%), while the difference between hybrid and single-word group did not reach significance by a small margin (p=5.6%). Between the two intervention groups, there was no significant difference. Vocabulary growth was correlated with the number of words seen (Pearson 0.7), number of sessions (0.6), time spent (0.8), efficiency (0.5), and effectiveness (0.6). There were only minor correlations, however, between the actual vocabulary growth and enjoyment, interestingness, and age (0.0, 0.1, and -0.2), even though perceived learning was strongly correlated with these variables (see 6.1.4). Neither was there a strong correlation between vocabulary growth and initial motivation nor with prior perceived usefulness (in fact both -0.3). There was a slight negative correlation (-0.4) between having previously learned the language and vocabulary growth.

### 6.1.2 Time Efficiency

Figure 6.2 shows the distribution of the time efficiency in the different groups. Overall, users learned 0.38 words per minute in the median and 0.43 in the mean. Standard deviation was 0.35. The single-word group fared far worse (median 0.10, mean 0.14, std 0.16) while the hybrid group (median 0.38, mean 0.54, std 0.42) was much more efficient and the retrieval method even better than that (median 0.59, mean 0.60, std 0.24).

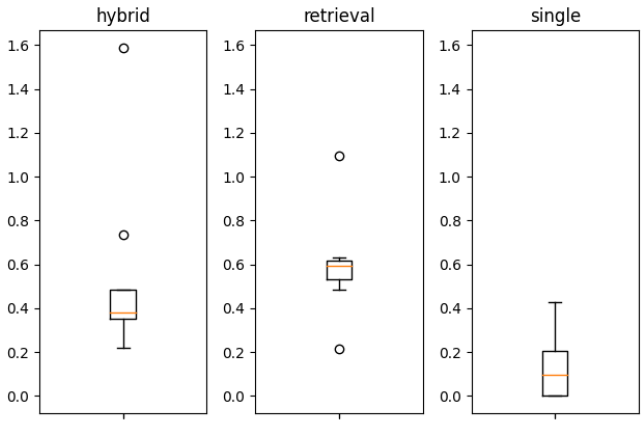



**Figure 6.2.:** Box plot of the efficiency (vocabulary growth per minute) in the different groups

For both intervention groups, the difference to the single word baseline was significant (p=0.3% that hybrid $\leq$ single, p=0.1% that retrieval $\leq$ single), but the difference between retrieval and hybrid did not reach significance (p=8.9% that retrieval $\leq$ hybrid).

There was a very slight negative correlation (-0.2) between having previously learned the language and efficiency.

### 6.1.3 Word Effectiveness

Figure 6.3 shows the distribution of the word effectiveness in the different groups. Overall, users learned 0.12 words per minute in the median and 0.15 in the mean. Standard deviation was 0.13. The single-word group had the lowest effectiveness (median 0.05, mean 0.12, std 0.15) while the hybrid group (median 0.12, mean 0.16, std 0.14) was more effective and the retrieval method even slightly better than that (median 0.17, mean 0.18, std 0.12). None of the differences were significant though.

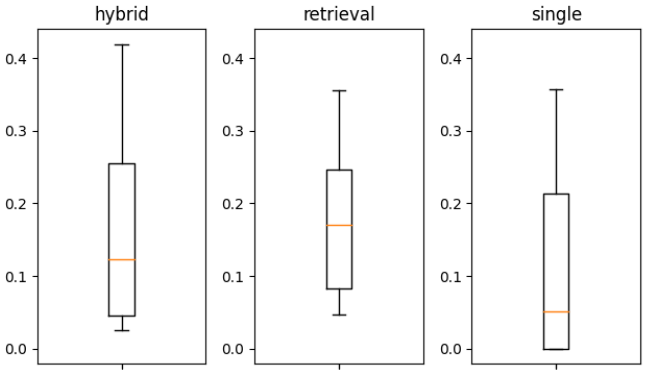



**Figure 6.3.:** Box plot of the effectiveness (vocabulary growth per distinct word seen) in the different groups

There was a pronounced negative correlation with prior Danish knowledge (Pearson -0.7), and moderate correlations with number of sessions (0.5) and with efficiency (0.6).

### 6.1.4 Proxies for User Engagement

**Words seen**

Figure 6.4 shows the distribution of the number of words seen in the different groups. Overall, each user saw 46.5 words in the median or 65.3 words in the mean. The standard deviation was 82.0. The group which only learned a single word per task had the low-

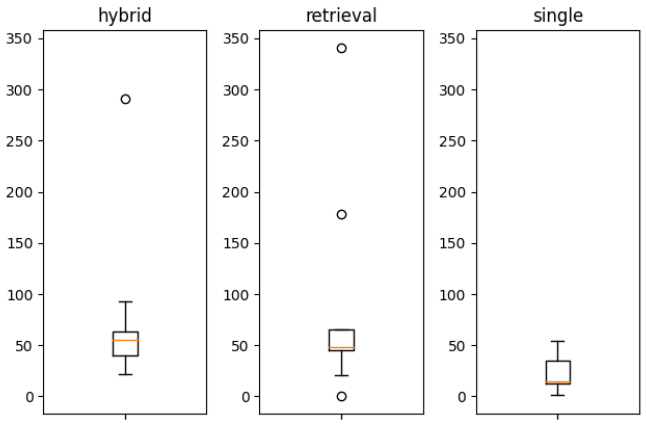



**Figure 6.4.:** Box plot of the number of words seen in the different groups

est number of words seen (median 15.0, mean 24.0, std 19.5) while the hybrid group (median 55.0, mean 78.0, std 82.4) and the retrieval method (median 48.0, mean 89.4, std 106.6) saw far more words. While the hybrid group's median is the highest, the retrieval group's mean is the highest due to two high outliers while the hybrid method only had one outlier.
For both intervention groups, the difference to the single word baseline was significant (p=0.5% that hybrid $\leq$ single, p=3.4% that retrieval $\leq$ single), between them there was no significant difference.

### Total Time Spent

Figure 6.5 shows the distribution of the total time spent using the app in the different groups. Overall, the average user spent 17.4 minutes in the median or 23.7 minutes in the mean. Standard deviation was 27.5. The groups were quite similar: The single-word group had spent a median of 16.4 minutes, mean of 21.9 minutes (std 25.0) while the hybrid group spent a median of 17.1, mean of 27.3 minutes (std 39.3) and the retrieval method spent a median of 26.2, mean of 21.7 minutes (std 15.9). None of the differences were significant. Time spent was strongly correlated with the number of sessions (Pearson 0.8).

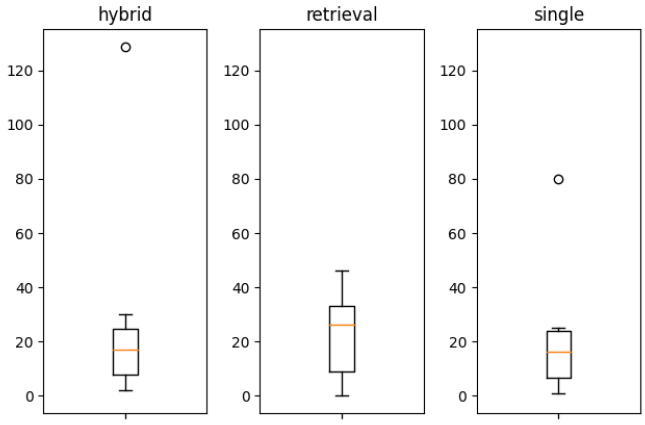



**Figure 6.5.:** Box plot of the total time spent using the app (minutes) in the different groups

### Number of Sessions

The user with the least sessions only had one and the user with the most had 16. The differences between groups weren't significant either for the number

of sessions users had: median 6, mean 7.2 in hybrid, 5 and 6.3 in retrieval, and 6 and 5.9 in the single-word group.

### Subjective Interestingness, Enjoyment, Perceived learning, Challengingness, and Confusion while Learning

Figure 6.6 shows the different groups' responses to the interestingness question and figure 6.7 to the enjoyment question. In fact, there was a strong positive correlation between interestingness and enjoyment (Pearson 0.8). For both metrics, the single-word group rated worst and the retrieval group best. The retrieval group's rating of enjoyment was significantly higher than both the hybrid group (p=2.8%) and the single word baseline (p=4.2%).
Enjoyment was also positively correlated with time efficiency (Pearson 0.5).

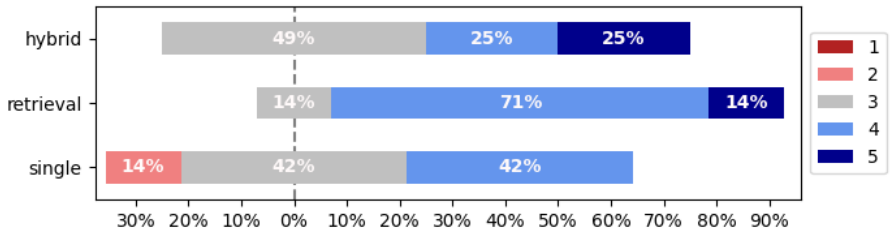



**Figure 6.6.:** User ratings of "This is interesting" across the different groups (1 = disagree, 5 = agree)

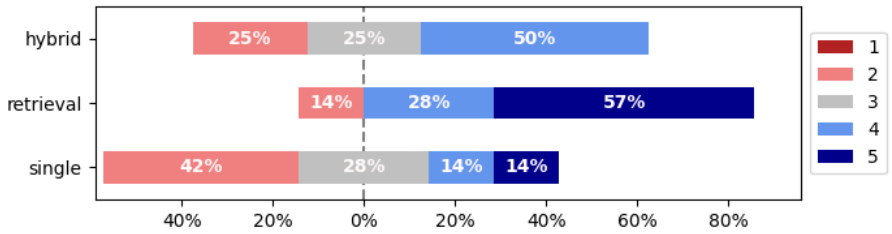



**Figure 6.7.:** User ratings of "I am enjoying this" across the different groups (1 = disagree, 5 = agree)

Figure 6.8 shows the different group's perceived learning. In all of the groups, the majority agreed that they were learning a lot, but in the single-word group, it was a smaller portion, but the difference was not significant. Perceived learning was positively correlated with enjoyment (Pearson 0.7) and interestingness

(0.8) and negatively correlated with age (-0.5). Actual vocabulary growth, however, was almost uncorrelated to the perceived one at Pearson 0.1.

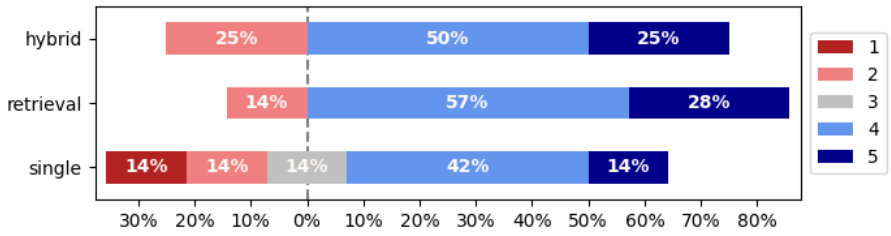



**Figure 6.8.:** User ratings of "I am learning a lot" across the different groups (1 = disagree, 5 = agree)

Figure 6.8 shows the different groups' responses to the challengingness question. While the single-word group was exactly split, most in both of the intervention groups reported they felt challenged. None of the differences were significant.

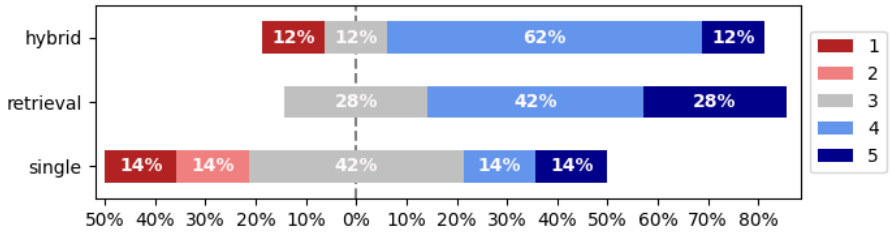



**Figure 6.9.:** User ratings of "This is challenging" across the different groups (1 = disagree, 5 = agree)

Finally, figure 6.8 shows the different groups' responses to the confusion question. While the intervention groups were mostly split, the single-word group leaned more towards being confused.

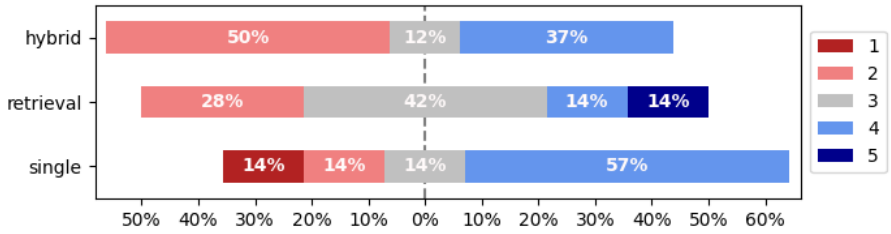



**Figure 6.10.:** User ratings of "I am confused" across the different groups (1 = disagree, 5 = agree)

With all of the subjective metrics, there were largely no trends in how they would change over time, for some users they decreased, and for some, they increase or remained the same.

### 6.1.5 Perceived Usefulness and Perceived Ease of Use Before and After the Trial Period

There were only negligible differences between the groups when it comes to their perception of the ease of use of the app, neither before nor after using it. The mean of the rating improved, however, from 4.1 to 4.4.
There were differences, though, between the groups when it comes to perceived usefulness. While the groups started out with similar overall very positive perceptions (means hybrid 3.9, retrieval 4.2, and single 4.1), they deteriorated for all groups, but most notably for the single word group, which fell to a mean of 3.0 while the others fell to 3.6. The difference was, however, not significant. Figure 6.11 shows the detailed distributions.

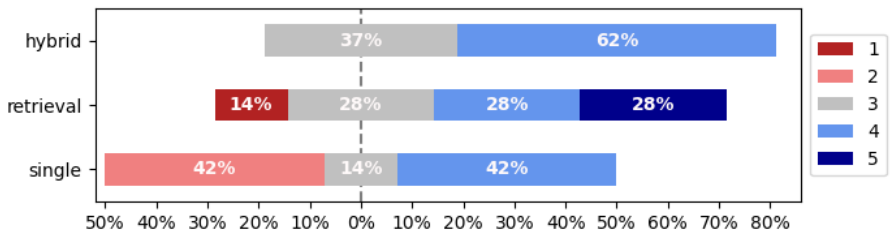



**Figure 6.11.:** User ratings of "How useful did you find the app?" across the different groups (1 = not at all, 5 = very)

Prior perceived usefulness was correlated with interestingness ratings and prior perceived ease of use (Pearson 0.5 each). Initial perceived usefulness was also correlated with initial motivation and inversely correlated with the number of sessions. Final perceived usefulness was correlated with interestingness, enjoyment, and learning (0.7 each) and inversely correlated with age (-0.5).

## 6.2 Supported and Refuted Hypotheses

Due to the previously reported differences and their significance levels, the hypotheses from section A.3 are supported/refuted as follows in table 6.1:

Please note that while the original hypotheses when comparing the hybrid and the retrieval group always were that the metric is greater in the hybrid group, this has been reversed for the enjoyment metric due to the results, so the supported hypothesis there was that enjoyment is greater in the retrieval group compared to the hybrid group.

| Metric | Comparison | Result |
|---|---|---|
| Vocabulary Growth | Retrieval vs. Single | Supported |
| | Hybrid vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |
| Words Seen | Hybrid vs. Single | Supported |
| | Retrieval vs. Single | Supported |
| | Hybrid vs. Retrieval | Refuted |
| Total Time Spent | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |
| Time Efficiency | Hybrid vs. Single | Supported |
| | Retrieval vs. Single | Supported |
| | Hybrid vs. Retrieval | Refuted |
| Enjoyment | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Supported |
| | Retrieval vs. Hybrid | Supported |
| Interestingness | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |
| Subjective Learning | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |
| Challengingness | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |
| Confusion | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |
| Perceived Usefulness | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |
| Perceived Ease of Use | Hybrid vs. Single | Refuted |
| | Retrieval vs. Single | Refuted |
| | Hybrid vs. Retrieval | Refuted |

**Table 6.1.:** Supported and refuted hypotheses where the h1-hypothesis is "{metric} is greater in the first out of the two following groups: {comparison}."

# 7 Discussion

This chapter discusses the main findings and relates them to each other. The implications of the results are analyzed by relating them to previous studies on vocabulary learning and limitations and future research opportunities are pointed out.

## 7.1 Summary of Findings

The results indicate that, compared to single-word spaced repetition with set assigned sentences, generating or selecting dynamic sentences based on multiple due words, can indeed increase learning outcomes and user engagement, as well as the user's perception of the usefulness of the system, even though that finding was not significant.
Both the group using a language model and the pure retrieval group achieved around four-fold greater efficiency of their vocabulary growth than the single-word group, while seeing three times more words and four-to-six times higher overall vocabulary growth, even though this was not significant for the hybrid group. The results are even stronger for the retrieval group than the hybrid group, since it also had significantly higher enjoyment ratings, which were even significantly higher than those of the hybrid group.

## 7.2 Interpretation of Results

The assessed metrics have intentionally been chosen to naturally be connected with each other, most obviously for efficiency, which is the quotient of vocabulary growth and time spent, but also for the other metrics, where it is plausible that the engagement-related metrics influence the learning-related metrics and

perceived usefulness. This allows for an analysis of which factors might have contributed most to the learning outcomes.

### 7.2.1 Drivers of Vocabulary Growth

This analysis should start from vocabulary growth, which I previously called the arguably most important metric since this is the tangible outcome users expect from using a language learning app. The first thing that should be noted here, is that there was a slight negative correlation between previous Danish experience and vocabulary growth. To me, this was surprising, since I had expected that having some familiarity with the language would help with acquiring new vocabulary. At first, I thought this was because of the single-word group, which might not be optimized for advanced learners since it only skips ahead to more advanced vocabulary if the user requests it and knew most of the ten most common words. However, upon further examination, this correlation remained equally strong if the single-word group was removed from the data. It thus seems that none of the methods was very suitable for advanced learners, leading to worse learning outcomes for them.

The results are still valid, since this being an advanced learner was eliminated as a confounding factor by the blocked randomization (see 5.5.2). Still, I re-analyzed the data without the results from advanced learners (which leaves n=4 for hybrid and single group, 5 for retrieval group) to find whether there were more pronounced effects for them specifically. Despite the low amount of remaining subjects, the main findings remain: hybrid and retrieval group still saw significantly more words and had significantly higher efficiency, retrieval still has significantly higher enjoyment, but for beginners, it is only the hybrid group that reached significance in its higher vocabulary growth.

Moving on, vocabulary growth had a moderate correlation with efficiency, and even more pronounced ones with effectiveness, words seen, time spent, and number of sessions. Considering that all groups spent similar amounts of time and sessions using the app, efficiency, effectiveness, and number of words seen must be the main drivers behind the differences in vocabulary growth observed between the groups.

Even though it could not be shown to be statistically significant, both of the intervention groups had higher effectiveness than the single-word group, which could thus explain a small part of their greater vocabulary growth and could mean that seeing each word in a variety of contexts did make a small

positive difference. Mainly, however, the vocabulary growth must be attributed to efficiency and the number of words seen. These two were also correlated with each other, which makes sense: the more distinct words the method presents to the user in the same amount of time, the more words can the user learn in that time, if effectiveness is factored out.

### 7.2.2 Drivers and Influence of User Engagement

Challengingness and confusion were neither significantly different between the groups nor strongly correlated with any other variables. The number of sessions and time spent were very strongly correlated and so were enjoyment and interestingness of the tasks. Consequently, for analyzing user engagement, I will focus on the number of sessions, enjoyment, and subjective learning.
The most objective measures of a user's engagement are arguably time spent using the app and the number of sessions since there is no subjective component. The number of sessions was moderately negatively correlated (-0.5) with being an advanced user, suggesting again, that the system might not have worked very well for them, which one advanced user also noted in the free text comment in the final questionnaire. It was also moderately negatively correlated (-0.5) with initially perceived usefulness, which is surprising. It can only be speculated about how this correlation arises. Maybe, the greater the initial expectations, the greater the disappointment when the system does not live up to these expectations, and that causes users to stop after only very few sessions.
The most interesting correlation was between number of sessions and learning effectiveness (Pearson 0.5). This could be interpreted in several ways: Either users become more effective over time as they get more used to the way of learning. Or users that are more effective see more benefits from the system and continue using it. But this reasoning is contradicted by there not being any correlation between effectiveness and enjoyment or perceived learning. The third interpretation is that this is due to more sessions increasing the spaced repetition effect and thereby leading to remembering more of the seen words. It could thus be evidence that the spaced repetition is functioning well in the implementation.
There were no significant overall differences between the groups when it comes to sessions and time spent. This could, however, also have been influenced by test subjects in all groups feeling obligated to do at least a few sessions as part

of social desirability bias, as discussed in 5.5.1.
When it comes to subjective learning, this was once again strongly correlated with enjoyment and interestingness, suggesting that it was a contributor to these, or that those who enjoyed it more also thought they were learning more. In any case, it can only have been a small contributor to enjoyment, since perceived learning is not significantly different between the groups, while enjoyment was.
The only subjective metric where there was a significant difference was enjoyment, and the retrieval group was the one to report the highest. Retrieval was moderately positively correlated with efficiency. Either higher enjoyment increases efficiency, or more plausibly, higher efficiency, learning more words in less time, leads to higher reported enjoyment.

### 7.2.3 Detailed Analysis of the Generated Tasks

Most of the hypotheses about user engagement-related metrics were refuted, meaning that any differences, if they existed, were too small to be detected with the limited sample size.
Analyzing the sentences that were shown to users in the three groups, the main difference was the sentence length and the number of distinct sentences seen among all users in the group. While the single-word group only saw 98 different tasks, the retrieval group saw 319, and the hybrid group had 400 distinct tasks. This is not very surprising, since users in the single-word group only saw new tasks beyond the first ten if they requested them, but it confirms that the goal of the implementation of showing users more diverse contexts was reached. The mean sentence lengths were also very different, 6.6 in the hybrid group, 4.5 in retrieval, and 3.2 in the single-word group. While all of these values fall well below the maximum of ten words that had been set, one could suspect that maybe the 4.5 words of the retrieval group are less overwhelming than the 6.6 words of the hybrid group, and that that led to the differences in learning and enjoyment between them. However, the hybrid group did not report higher challengingness or confusion ratings, which makes that possibility unlikely.
The topics of the sentences, identified by giving the whole list of generated tasks to GPT-3.5 and asking "Group the following sentences into five groups by topic area:", do not immediately yield an explanation for the differences in enjoyment that were observed, supporting the hypothesis that this difference

was mostly due to the increased speed of learning in the retrieval group. The topics the hybrid tasks were about included primarily music and entertainment, sports competitions, places and personal thoughts, emotions and actions. Meanwhile, in the retrieval group, the main topics were nature, historical figures and events, buildings and places, and culture. In the single-word group, topics included nature, music, entertainment and literature, as well as travel. Most of these topics are predictably what one would expect to find in a Wikipedia corpus, with only the "personal thoughts, emotions and actions" from the hybrid method being an exception. Including this type of topic is arguably important for language learners, since talking about personal topics is a very important aspect of speaking a language. In fact, looking at the free-text comment field from the final questionnaire (see A.2), four of the participants in the retrieval group complained about sentences being not very applicable to everyday life, while none of the participants from the other groups had that complaint. It is thus very surprising that this apparently did not increase the hybrid method's perceived usefulness over that of the retrieval group.

### 7.2.4 Concluding the Interpretation

Concluding the previous analysis, it seems likely that using sentence-based spaced repetition first and foremost manages to show users more new words to learn in less time, especially for beginners. This increases efficiency and vocabulary growth since users still retain the same fraction of words seen or even slightly more when they focus on several words in the sentence. The increased efficiency then probably leads to higher enjoyment.
Despite the positive effects on learning, no significant effect on perceived usefulness nor ease of use could be found, which means that the proposed learning method might not lead to better long-term adoption than current methods.
When it comes to the differences between the two intervention groups, these have mostly been minor, but they were significant for enjoyment and almost significant for efficiency, which could have led to the increased enjoyment. The difference in efficiency can, however, not be explained by one method showing more distinct words than the other. It can only be speculated that the difference can be explained by the higher effectiveness in the retrieval group (even though it was not significant).

## 7.3 Implications

The results mean first and foremost, that using a sentence-based spaced repetition scheme should be preferred over using single-word spaced repetition, even when the single word is shown in the context of an example sentence. This will show users more vocabulary in less time, increasing efficiency and thus enjoyment.
Even having tried a variety of different NLP paradigms, language models, and configurations, the more promising one, optimized for spaced repetition timing and correctness, was not able to deliver better efficiency or user engagement than the retrieval-based model, at least in the hybrid combination of LM and retrieval model. But since the hybrid configuration had slightly worse values in most user testing metrics than the retrieval model, the result can be extrapolated to that a pure LM model would probably not have performed better, even if the lemmatization issues were fixed.
Since a retrieval model is also by far a less costly option in terms of computing costs, it should be advised to be used instead of the LM-based option. Despite the fact that this specific prompting-based LM method and configuration did not prove worthwhile, with the current rapid advancements in LM size and tasks they can perform through prompting, other LMs e.g. GPT-4, which is the next generation of OpenAi's GPT language model and has substantially more parameters, could improve correctness and possibly scheduling timing.

### 7.3.1 How does the proposed sentence-based spaced repetition compare to conventional or other digital language learning methods?

While it is hard to find reliable estimates on how long it takes on average to learn a word, (Nation 1982) reviews different previous studies on learning lists of word pairs, and mentions an average of 34 words per hour for English-German word pairs, with 60% recall after 42 days for one study by (Thorndike 1908) and a range from 33 to 111 words per hour for another, but without mentioning any mean or recall. Calculating with the former, the word-pair method would average 0.57 words per minute, with 0.34 recalled words after

42 days. Since this dissertation only counted correctly recalled words after the span of a few days, if learning efficiency was equal, it should fall in between that range. Results are still not directly comparable, since the mentioned study did not have the problem of time being wasted on previously known words, which are not counted for vocabulary growth in this dissertation. However, considering these issues, it seems that hybrid and retrieval groups with mean of 0.54 and 0.6 words per minute had a higher efficiency than the results from Thorndike's study.

Another number for CALL comes from (Wozniak 2018), where he mentions that the "standard" for English vocabulary learning in the SuperMemo spaced repetition course is 40000 words in four years at 40 minutes per day, coming down to 0.68 words per minute. Like the previous one, this number seems not to account for previously known words and might thus be a bit high. Considering that, it seems that the intervention groups did comparably well to SuperMemo's word-based spaced repetition, even though it is not a very reliable comparison because of the very different timescales of 4 years compared to ten days.

The proposed system achieved a higher efficiency than a traditional way of vocabulary learning and was more or less on par with a fully matured spaced repetition system with years of development, a plethora of features, and a well-designed UI, while my system sent to the user study still had some minor bugs and very minimal UI and features. This means that the proposed system has shown some value and could potentially surpass the efficiency of traditional spaced repetition software if it gets refined further, for which some proposals are provided in 7.5.

## 7.4 Limitations

Convenience sampling has been employed to choose study participants, explained in further detail in 5.5.1. Participants were very diverse in some aspects such as native language, but very homogeneous in others, such as previous usage of language learning apps. This means that participants are not representative of the general population. While it can be reasonably assumed that learning works similarly in all humans, the evidence for the effect observed is strongest for people that are similar to the participants and might not be generalizable to persons with completely different backgrounds.

Furthermore, the sample size was small with 26 participants, looking at a population of hundreds of thousands of Danish learners or possibly billions of persons learning languages in general. This sample size might not have been big enough to detect some possible differences between the hybrid group and the control group or the retrieval group and the hybrid group. It was, however, big enough, to detect some of the most pronounced effects that this dissertation tried to assess.

The duration of the user study of ten days also only allows to draw direct conclusions for short-term use, but, as mentioned in 5.3.5 this was mitigated by measuring engagement as a possible predictor of long-term learning outcomes.

There have also been minor technical problems during the experiment affecting some participants. Firstly, a minority of the participants which was using iOS devices reported they were unable to install the app (only view it in the browser) and set remainders, which potentially reduced their engagement. Furthermore, some users on the first day were affected by an issue where they were shown a connection error when the OpenAI API took too long to respond, this issue was fixed by the second day of the experiment. Not implementing parallelization might have negatively affected engagement by increasing the delay for new tasks to show up to several seconds in the worst case. However, all groups were equally affected by these challenges, so the results are still valid, a perfectly working system from day one might just have allowed each method to show its advantages even more and potentially increase the significance of the differences for some metrics.

The choice of Danish as the language for the user study is also a slight limiting factor for generalizability. While it is reasonable to assume that learning happens in a similar way and is influenced by similar factors in most languages, details about the language such as its morphology, e.g. having many word forms for each lemma, could lead to reduced or increased suitability of the proposed approach and possibly increased importance of storing user vocabulary as lemmas instead of word forms.

Finally, related to the small sample size, it is worth mentioning that most of the confounding variables were not directly controlled for, they were mostly addressed by randomization and blocked randomization, which works best at even larger scales, but should still be valid with 26 participants. It has been confirmed that there were no major imbalances among the groups for any of the confounding variables mentioned in 5.5.1.

## 7.5 Further Research Opportunities

During the development of the sentence generation methods, a preliminary attempt was made to implement reinforcement learning with a static reward function starting from a PLM, where the inverse of the loss from 3.3 was directly optimized. The attempt was unsuccessful, it led neither to natural sentences being produced nor the target words being contained. It should be noted that, due to limited access to computing resources, the implementation was only attempted on small LMs, such as GPT-2 and OPT-1.3B (Zhang et al. 2022). A proper implementation of reinforcement learning on a bigger and more current LM would be worth investigating since it can optimize the objective more directly than the prompting approach, which has been found to be quite limited for example in the number of due input words it can take without adverse consequences, while a reinforcement learning approach could potentially take an unlimited amount of input words if it is optimized for that, which could lead to better scheduling when a lot of words are due.

Other research opportunities lie in the challenges encountered in dealing with lemmas during the implementation of the methods in this work. If problems with looping due to lemmatization are solved, the user vocabulary could be stored and scheduled as lemmas instead of word forms, and it could be investigated whether this leads to better learning outcomes, especially in languages that have many different forms for each lemma. Even without solving the lemmatization problem, it would be valuable to test the current system in different languages with lots of forms of each lemma, such as conjugated forms, and investigate whether the effects of increased enjoyment and learning are still present.

Furthermore, as mentioned in 7.3, using more current and upcoming bigger PLMs like GPT-4 instead of GPT-3.5, it might be possible to achieve better results, for example for the correctness of the sentences, which was one of the main advantages of retrieval method over the LM method. If the looping problem is also solved, it will be interesting to test a purely LM-based method on actual test users and compare it to the results from this dissertation.

As a last research opportunity that should be mentioned, one of the potential advantages of generating sentences for spaced repetition mentioned in 1.3 was that the sentence generation could be optimized for additional objectives, such as entertainment value or variety of grammar. This could not be investigated in this work and thus remains a research opportunity.

# 8 Conclusion

The aims of this dissertation were first to identify NLP paradigms and configurations for sentence generation that can optimize spaced repetition timing and best avoid out-of-user-vocabulary words, while keeping the correctness of the generated sentences as high as possible, and then to quantify these methods' influence on user engagement and learning outcomes among language learners, compared to conventional approaches.

Two methods of achieving these goals were developed: one based on retrieval of suitable sentences from a corpus of high-quality sentences using many upcoming due words as queries, and the other was few-shot-prompting a PLM to generate sentences from a subset of the due words. Both of the methods were found to be able to form sentences that were mostly comprised of words from the user vocabulary, soon to be due and mostly correct, thereby reaching the objectives. While the retrieval method reached 100% correctness, the LM method optimized the spaced repetition scheduling even better but had worse correctness and has an unsolved problem with looping due to the treatment of lemmas, despite multiple countermeasures, which makes it unsuitable for deployment to users. A hybrid method switching between retrieval and LM generation could solve the looping problem while optimizing the research question's objectives.

Consequently, the hybrid and the retrieval method were compared to a baseline to answer the second research aim. It was found that the use of the proposed sentence-based spaced repetition significantly increased learning outcomes (four-to-six-fold) compared to the baseline, primarily by increasing efficiency and vocabulary growth by showing more words more quickly, without decreasing the fraction of words remembered by learners. In the retrieval group, a significantly higher enjoyment was observed, possibly due to the higher efficiency, hinting at a higher user engagement.

It was thus concluded that it is beneficial to use the proposed sentence-based spaced repetition over the conventional approach and that the retrieval approach is advisable over LM-based or hybrid approaches, but that further developments, such as fixing problems with lemmatization and looping and

higher correctness possibly achievable with newer language models could improve the results when using a more advanced LM based method in the future.

# 9 Bibliography

Al-Dosakee, Karwan, and Fezile Ozdamli. 2021. "Gamification in Teaching and Learning Languages: A Systematic Literature Review." Revista Romaneasca Pentru Educatie Multidimensionala 13 (August): 559–77. https://doi.org/10.18662/rrem/13.2/436.

Brown, Jonathan, Gwen Frishkoff, and Maxine Eskenazi. 2005. "Automatic Question Generation for Vocabulary Assessment." In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 819–26. Vancouver, British Columbia, Canada: Association for Computational Linguistics. https://aclanthology.org/H05-1103.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Çakmak, Fidel, Ehsan Namaziandost, and Tribhuwan Kumar. 2021. "CALL-Enhanced L2 Vocabulary Learning: Using Spaced Exposure through CALL to Enhance L2 Vocabulary Retention." Education Research International 2021 (September). https://doi.org/10.1155/2021/5848525.

Carpenter, Shana K., Nicholas J. Cepeda, Doug Rohrer, Sean H. K. Kang, and Harold Pashler. 2012. "Using Spacing to Enhance Diverse Forms of Learning: Review of Recent Research and Implications for Instruction." Educational Psychology Review 24 (3): 369–78. https://doi.org/10.1007/s10648-012-9205-z.

Chen, Howard, Huihan Li, Danqi Chen, and Karthik Narasimhan. 2022. "Controllable Text Generation with Language Constraints." arXiv. http://arxiv.org/abs/2212.10466.

Chiang, Cheng-Han, and Hung-yi Lee. 2023. "Can Large Language Mod-

els Be an Alternative to Human Evaluations?" arXiv. http://arxiv.org/abs/2305.01937.

"Danish - AnkiWeb." n.d. Accessed August 6, 2023. https://ankiweb.net/shared/decks/danish.

Davis, Fred D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." MIS Quarterly 13 (3): 319–40. https://doi.org/10.2307/249008.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Ekgren, Ariel, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. "GPT-SW3: An Autoregressive Language Model for the Nordic Languages." arXiv. http://arxiv.org/abs/2305.12987.

Elmes, Damien. n.d. "Anki." https://apps.ankiweb.net.

Explosion. 2016. "Danish · SpaCy Models Documentation." Danish. 2023 2016. https://spacy.io/models/da.

Fatima, Noureen, Ali Shariq Imran, Zenun Kastrati, Sher Muhammad Daudpota, and Abdullah Soomro. 2022. "A Systematic Literature Review on Text Generation Using Deep Neural Network Models." IEEE Access 10: 53490–503. https://doi.org/10.1109/ACCESS.2022.3174108.

"Global Mobile Traffic 2022." n.d. Statista. Accessed August 3, 2023. https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-device

Guo, Biyang, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. 2022. "GENIUS: Sketch-Based Language Model

Pre-Training via Extreme and Selective Masking for Text Generation and Augmentation." arXiv. http://arxiv.org/abs/2211.10330.

Guo, Mandy, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. "Wiki-40B: Multilingual Language Model Dataset." In Proceedings of the Twelfth Language Resources and Evaluation Conference, 2440–52. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.297.

Han, Xu, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, et al. 2021. "Pre-Trained Models: Past, Present and Future." arXiv. http://arxiv.org/abs/2106.07139.

Hao, Tao, Zhe Wang, and Yuliya Ardasheva. 2021. "Technology-Assisted Vocabulary Learning for EFL Learners: A Meta-Analysis." Journal of Research on Educational Effectiveness 14 (3): 645–67. https://doi.org/10.1080/19345747.2021.1917028.

Jankowski, Jakub. 1999. "Effective Learning: Twenty Rules of Formulating Knowledge." SuperMemo. December 6, 1999. https://www.supermemo.com/en/blog/twenty-rules-of-formulating-knowledge.

Jeon, Jaeho. 2021. "Chatbot-Assisted Dynamic Assessment (CA-DA) for L2 Vocabulary Learning and Diagnosis." Computer Assisted Language Learning 0 (0): 1–27. https://doi.org/10.1080/09588221.2021.1987272.

Jurafsky, Daniel, and James H. Martin. 2023. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." Learning and Individual Differences 103 (April): 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kornell, Nate. 2009. "Optimising Learning Using Flashcards: Spacing Is More Effective than Cramming." Applied Cognitive Psychology 23 (9): 1297–1317. https://doi.org/10.1002/acp.1537.

Mann, H. B., and D. R. Whitney. 1947. "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other." The Annals of Mathematical Statistics 18 (1): 50–60. https://doi.org/10.1214/aoms/1177730491.

Mizumoto, Atsushi, and Masaki Eguchi. 2023. "Exploring the Potential of Using an AI Language Model for Automated Essay Scoring." Research Methods in Applied Linguistics 2 (2): 100050. https://doi.org/10.1016/j.rmal.2023.100050.

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, Henning Peters, Paul O'Leary McCann, et al. 2023. "SpaCy: Industrial-Strength Natural Language Processing in Python." Zenodo. https://doi.org/10.5281/zenodo.8123552.

Nation, I.S.P. 1982. "Beginning to Learn Foreign Vocabulary: A Review of the Research." RELC Journal 13 (1): 14–36. https://doi.org/10.1177/003368828201300102.

Okano, Yuki, Kotaro Funakoshi, Ryo Nagata, and Manabu Okumura. 2023. "Generating Dialog Responses with Specified Grammatical Items for Second Language Learning." In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 184–94. Toronto, Canada: Association for Computational Linguistics. https://aclanthology.org/2023.bea-1.16.

Olmos, Carmen. 2009. "An Assessment of the Vocabulary Knowledge of Students in the Final Year of Secondary Education. Is Their Vocabulary Extensive Enough?"

OpenAI. n.d. "API Reference - OpenAI." Accessed August 6, 2023. https://platform.openai.com/docs/api-reference/introduction.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." arXiv. https://doi.org/10.48550/arXiv.2203.02155.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. “Language Models Are Unsupervised Multitask Learners.”

Ramos, Restrepo, and Falcon Dario. 2015. “Incidental Vocabulary Learning in Second Language Acquisition: A Literature Review.” Profile Issues in Teachers‘ Professional Development 17 (1): 157–66. https://doi.org/10.15446/profile.v17n1.43957.

Richards, Jack C. 1976. “The Role of Vocabulary Teaching.” TESOL Quarterly 10 (1): 77–89. https://doi.org/10.2307/3585941.

Robertson, Stephen, and Hugo Zaragoza. 2009. “The Probabilistic Relevance Framework: BM25 and Beyond.” Foundations and Trends in Information Retrieval 3 (January): 333–89. https://doi.org/10.1561/1500000019.

Speer, Robyn. 2022. “Rspeer/Wordfreq: V3.0.” Zenodo. https://doi.org/10.5281/zenodo.7199437.

Thorndike, Edward L. 1908. “Memory for Paired Associates.” Psychological Review 15 (2): 122–38. https://doi.org/10.1037/h0073570.

Uc-Cetina, Victor, Nicolas Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. 2023. “Survey on Reinforcement Learning for Language Processing.” Artificial Intelligence Review 56 (2): 1543–75. https://doi.org/10.1007/s10462-022-10205-5.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. “Emergent Abilities of Large Language Models.” arXiv. http://arxiv.org/abs/2206.07682.

Wozniak, P. A. 1990. “Optimization of Learning: A New Approach and Computer Application.” https://super-memory.com/english/ol.htm.

Wozniak, Piotr. 2018. “First Steps of SuperMemo - Supermemo.Guru.” 2018. https://supermemo.guru/wiki/First_steps_of_SuperMemo.

Wozniak, Piotr. n.d. “SuperMemo.” https://www.supermemo.wiki/en/supermemo.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, et al. 2022. “OPT: Open Pre-Trained Transformer Language Models.” arXiv. https://doi.org/10.48550/arXiv.2205.01068.

# A

# Appendix

## A.1 Sign-up and final questionnaire questions

This appendix lists the questions asked in the questionnaire asked before (figure A.1) and after the user study (figure A.2).

```
What is your native language?
How old are you?
How long have you lived in Denmark?
Have you learned Danish before?
Do you ever use it outside of class?
How strong would you say your motivation is to learn Danish?
What motivates you most to learn Danish?
Have you used language learning apps before?
From the info you have, do you think the app will be useful for you?
From the info you have, do you think the app will be easy to use?
```

**Figure A.1.:** Sign-up questions

```
The tasks were interesting
I was enjoying it
I was learning a lot
It was challenging
It was confusing
How useful did you find the app?
How easy was the app to use? Please rate only the design and (to the
extent possible) ignore any technical problems you might have encountered.
Any other comments / feedback
```

**Figure A.2.:** Final questions

## A.2 Free-text feedback from the final questionnaire

This appendix lists the free-form feedback from the last question of the final questionnaire in table A.1.

| Group | Feedback |
|---|---|
| single | Unfortunately, sometimes there is no connection between the selected word and the translator |
| single | I really like to participate in an AI development. I think it has a great potential to develop and certainly it will be useful for all student . |
| hybrid | the google translate version for the individual words wasn't the best solution I would say... |
| hybrid | I believe my biggest "issue" was that as I am familiar with a lot of words in Danish it took quite sometime to find an example that I would actually click on. In these cases, I would click 'Show Solution' to move on to a next sentence, but I wasn't sure if that was clear to the system. When I did click a word, it came back once again, but it was set in a very strange sentence, that I feel like it wouldn't be fitting to use it in that way. For example: "Det er nemt at læsse op stå et maling nu.", which seems to try to introduce the word "læsse" once again, but the sentence feels very nonsensical. I like the idea, but I feel like it would work best if I was a beginner or had been using this app for a longer period of time. |
| retrieval | The app wasn't difficult to use, but the translation suggestions from Google Translate often didn't match the suggested solutions, making learning difficult. It was good that you could look up every word. For someone without knowledge of Danish, it was often difficult sentences or long sentences with many difficult words. |
| retrieval | The new word I learned is not that common in daily life in my opinion. It would be more motivating if new words are tied with daily life. |
| retrieval | A few everyday and short sentences would have been good at the beginning to arouse interest and to enable small success stories right at the beginning |
| retrieval | I really enjoyed it and learnt some new words, my only comment is that many sentences or phrases were very random like phrases that might not be very used or common when learning a new language |
| retrieval | Sometimes it went in a Danish names loop |

**Table A.1.:** Free-text feedback from the final questionnaire

## A.3 Hypotheses tested in the user study

| Metric | Comparison |
|---|---|
| Vocabulary Growth | Retrieval vs. Single<br>Hybrid vs. Single<br>Hybrid vs. Retrieval |
| Words Seen | Hybrid vs. Single<br>Retrieval vs. Single<br>Hybrid vs. Retrieval |
| Total Time Spent | Hybrid vs. Single<br>Retrieval vs. Single<br>Hybrid vs. Retrieval |
| Time Efficiency | Hybrid vs. Single<br>Retrieval vs. Single<br>Hybrid vs. Retrieval |
| Enjoyability | Hybrid vs. Single<br>Retrieval vs. Single<br>Retrieval vs. Hybrid |
| Subjective Interestingness | Hybrid vs. Single<br>Retrieval vs. Single<br>Hybrid vs. Retrieval |
| Subjective Enjoyability | Hybrid vs. Single<br>Retrieval vs. Single<br>Hybrid vs. Retrieval |
| Perceived Learning | Hybrid vs. Single<br>Retrieval vs. Single<br>Hybrid vs. Retrieval |
| Challengingness | Single vs. Hybrid<br>Single vs. Retrieval<br>Retrieval vs. Hybrid |
| Confusion | Single vs. Hybrid<br>Single vs. Retrieval<br>Retrieval vs. Hybrid |
| Perceived Usefulness | Hybrid vs. Single<br>Retrieval vs. Single<br>Hybrid vs. Retrieval |

**Table A.2.:** List of hypotheses, where the h1-hypothesis is "{metric} is greater in the first out of the two following groups: {comparison}." and the h0-hypothesis is "{metric} is not greater in the first out of the two following groups: {comparison}."
Please note that confusion and challengingness are hypothesized to be lower in the intervention groups, while for the other metrics, it is the other way around