Automated Sentence Generation for a Spaced Repetition Software

Benjamin Paddags **Daniel Hershcovich** **Valkyrie Savage**
Department of Computer Science, University of Copenhagen
{bepa, dh, vasa}@di.ku.dk

Abstract

This paper presents and tests AllAI, an app that utilizes state-of-the-art NLP technology to assist second language acquisition through a novel method of sentence-based spaced repetition. Diverging from current single word or fixed sentence repetition, AllAI dynamically combines words due for repetition into sentences, enabling learning words in context while scheduling them independently. This research explores various suitable NLP paradigms and finds a few-shot prompting approach and retrieval of existing sentences from a corpus to yield the best correctness and scheduling accuracy. Subsequently, it evaluates these methods on 26 learners of Danish, finding a four-fold increase in the speed at which new words are learned, compared to conventional spaced repetition. Users of the retrieval method also reported significantly higher enjoyment, hinting at a higher user engagement.

1 Introduction

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). Usually, spaced repetition in language learning is done by repeating single words or whole sentences curated by humans. Already a decade ago, the potential of computational linguistics for vocabulary learning was identified by Zock et al. (2014, p. iii): "There is so much more we could do these days by using corpora and computational linguistics know-how, to extract the to-be learned words from text and to display them with their context. Hence, rather than having the user repeat single words (or word pairs) we could display them in various contexts (e.g. sentences), thereby making sure that the chosen ones correspond to the learners' level and interests.." 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 that 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.

This work introduces AllAI (Automated Language Learning with AI), an application utilizing NLP to create such a sentence-based approach to spaced repetition. The app keeps track of the user's vocabulary and generates 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. We then investigate the learning outcomes of using such a system compared to current solutions. As such, the main research questions are the following:

1. Which NLP paradigm and configuration can optimize spaced repetition timing and best avoid out-of-user-vocabulary words, while retaining high 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?

The proposed system combines the following potential advantages over the conventional spaced repetition approaches mentioned in 2:

1. It honors the minimum information principle.

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 351–364
June 20, 2024 ©2024 Association for Computational Linguistics

2. It shows words in context for a less artificial learning situation and the possibility to infer meaning.
3. It can generate a variety of tasks for high novelty value.
4. It could be optimized for additional objectives, such as entertainment value (e.g. subsequent sentences could form a story), variety of grammar, or others.

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. We also develop a metric for calculating the scheduling accuracy and select other metrics to assess the quality of the output sentences for the task. We compare a range of candidate methods and configurations that managed to return sensible sentences containing target words with regard to these metrics. We develop an application 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, we test 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, in a user study, assessing learning outcomes and indicators of user engagement against a baseline similar to current spaced repetition practices.

We implement and test the system in Danish. Still, it applies to any language in which the sentences are made up of words and is developed in such a way that it could 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.

2 Background and Related Work

2.1 Spaced repetition

Previous research has found a large beneficial effect of computer-assisted language learning (CALL) on vocabulary learning (Hao et al., 2021). One possible CALL technique is spaced repetition. Spaced repetition means reviewing information that one wants to remember repeatedly and with temporal spacing between each exposure to the same information. A review usually involves the learner being prompted, trying to recall, and then getting feedback. It has 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. Based on the idea of physical flashcards with a prompt on one side and the correct answer on the other, that are reviewed at increasing intervals (Leitner, 1972), most spaced repetition software (e.g. Anki (Elmes), shown as an example in figure 1, Mnemosyne (Çakmak et al., 2021), SuperMemo (Wozniak)) usually show a memory recall task to the user and expect the user to try to solve it. Thereafter, the solution is shown, and the user rates how well they could recall it. The system uses the recall quality to calculate the spacing until the task is presented to the user again, which should ideally be right before the user is likely to forget it.

In the context of language learning, spaced repetition can be used for the parts of L2 acquisition that require memorization, such as vocabulary learning. 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 [1]. 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), allowing for independent scheduling of each of the bits of knowledge. On the other hand, language is naturally used in context, where words learned in the context of a sentence reinforce each other, strengthening thus learning and recall, meaning that remembering words out of context is a very artificial task and much harder than if related words are present which can give hints about the meaning (Ramos and Dario, 2015). This work sets itself apart from the existing literature on spaced repetition by examining the effects of integrating a sentence generation component that generates sentences for single use on demand, which makes it possible to keep scheduling single words and adhering to the minimum information principle while showing words in context.

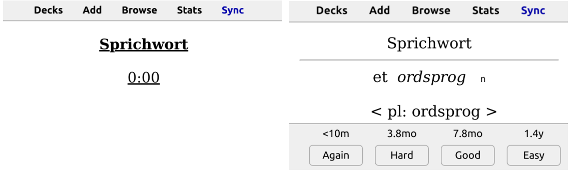
[1] "Shared Decks" https://ankiweb.net/shared/decks/danish

2.2 Language models, Text Generation and Language Teaching

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, 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).

With the strong performance of transformer-based pre-trained models (PLMs), such as GPT-3 (Brown et al., 2020), zero- or few-shot prompting of these PLMs have gained popularity, profiting from the excellent general understanding of the semantics and syntax of language that they can develop through pre-training on large and diverse text corpora. Recent research has demonstrated that especially for 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).

Even before the advent of modern language models, Brown et al. (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 work tries to achieve: 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. However, using a retrieval system on a corpus of example sentences can be a viable approach since queries can consist of multiple words. When it comes to using LMs in second language teaching, 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 this paper's experiments were finished, so it could not be used for inspiration. 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. There have also been successful attempts at creating flashcards for spaced repetition systems using LLMs, such as Gossmann (2024), Cruz (2023) and Velde (2023). Gossmann and Cruz focus on summarizing knowledge from articles into flashcards while Velde is applying their approach to vocabulary learning. Differently from what we are attempting, their flashcards are static, so they will still always show each word in context of the same information, which is equivalent to the third existing approach mentioned in section 2.1.



Sprichwort

0:00

Sprichwort

et *ordsprog* n

< pl: ordsprog >

Figure 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)

3 Comparing candidates methods for the sentence generation component

This section describes our simulated study to narrow down the methods and configurations that could optimize the system's objective to two that can be tested in the user study.

3.1 System objectives

The system's objective is to suggest sentences ("tasks") for the user to review, while following as closely as possible the due dates of the contained words coming from the spaced repetition scheduler. This results in the following three main objectives imposed by the first research question:

1. Maximize the correctness of the sentence
2. Maximize the amount of due and future due words contained, prioritize by upcoming due dates
3. Avoid sentences exceeding ten words (which was the maximum length that three test users reported not finding overwhelming)

3.2 Simulated Metrics

To automatically evaluate the different methods, we simulated their use over 20 days by a user who remembers any word with an 85% chance and then calculated the following automated metrics:

1. A scheduling score measuring how well the spaced repetition scheduling is adhered to and only due and future due vocabulary is used (for more details on the scheduler, see 4.1)
2. Too long sentences, to measure the fraction of sentences that are longer than the ten word limit from the third objective

We defined the scheduling score as the average fraction of the scheduling intervals wasted by scheduling words before they are due or 1 when a new word is introduced without the user asking for it, to discourage exponential vocabulary growth. It can be between zero and one and should be minimized.

$$S = \frac{1}{n_{tasks}} \sum_{tasks} \frac{1}{n_{taskwords}} \sum_{taskwords} s_{word}$$

$$s_{word} = \begin{cases} \frac{max(t_{due} - t_{now},\ 0)}{t_{due} - t_{last\_seen}} & \text{if in user vocab} \\ 0 & \text{user requested new word} \\ 1 & \text{new word, not requested} \end{cases}$$

Additionally, the correctness of the sample sentences was rated by a human evaluator and GPT-3.5-turbo-0301. While the human saw 20 samples per method, the LM saw 1000. They agreed fairly (Cohen's Kappa = 0.35), indicating that the LM's ratings can be useful when based on larger samples, but should not solely be relied upon.

3.3 Sentence Generation Methods

We implemented a variety of methods for generating or selecting sentences for testing purposes. Reinforcement learning with a static reward function (scheduling score) and modifying the probability distribution of a PLM directly (GPT-2 and OPT-1.3B) were briefly explored but were not able to generate at least 50% correct sentences that contained at least one of the words it was given as inputs. Meanwhile, retrieval of suitable sentences from a corpus and few-shot prompting did pass and they were thus moved on to the next stage where we subjected different configurations to the previously listed metrics.

The BM25 retrieval algorithm (Robertson and Zaragoza, 2009) was taken as a starting point for the retrieval method. 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. We modified BM25 to add query word weights 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). We discount query words with exponential decay the longer in the future they were due. The following formula was used to rank the sentences: BM25(query, sentence) =

$$\sum_{\mathrm{w} \in \mathrm{query}} \left(\mathrm{idf_w} \frac{(\mathrm{k1} + 1) \cdot \mathrm{q\_freq_w}}{\mathrm{q\_freq_w} + \mathrm{k1}(1 - \mathrm{b} + \mathrm{b}\frac{\mathrm{sent\_len}}{\mathrm{avgsl}})(\mathrm{dtd_w} + 1)} \right)$$

Where $\mathrm{idf_w}$, $\mathrm{q\_freq}_w$, sent_len, avgsl as in BM25,
dtd means days until the word is due for repetition,
$\mathrm{k1} = 1.5$ and $\mathrm{b} = 0.75$

Same-day repetitions of a task are disallowed by finding the best-ranking sentence that had not been previously shown. In addition to this standard version described above, we test a version that selects the task with the best scheduling score among the 25 best-ranking tasks. We chose the Wiki-40B Corpus (Guo et al., 2020) as the source of the sentences since it is one of the largest corpora for Danish (and 40+ languages in total, allowing for easy adaption, even though the BM25 would have to be re-tuned for some languages' features, e.g. different tokenization) with ca. 200MB worth of Danish sentences and, as it is sourced from Wikipedia articles, contains mostly correct use of the language. We removed sentences with rare words (not in the 25000 most frequent from the language), shorter than two, or longer than 10 words. After the filtering, the resulting corpus contained 64259 sentences, of which the average length was 5.9 words.

For the prompting approach, we chose GPT-3.5-turbo-0301 as the language model since it was the largest model that was partly trained on Danish data (0.1%, 220 million words in Danish (Brown et al., 2020)) at the time of writing, trained to be helpful with answering prompts containing instructions and relatively cheap to use. We explore different zero and few-shot prompts, with the best performing one given in appendix A and used for all further experiments. Input words are taken from the words scheduled for the current day and upcoming ones if fewer words were due on the day than the method takes as input. We also test two different system messages given to the model before the prompt, instructing it to generate a maximum of 5 words in the first and 10 words in a correct and meaningful sentence in the second. We also explore two temperature settings (0.2 and 0.8), five versus ten

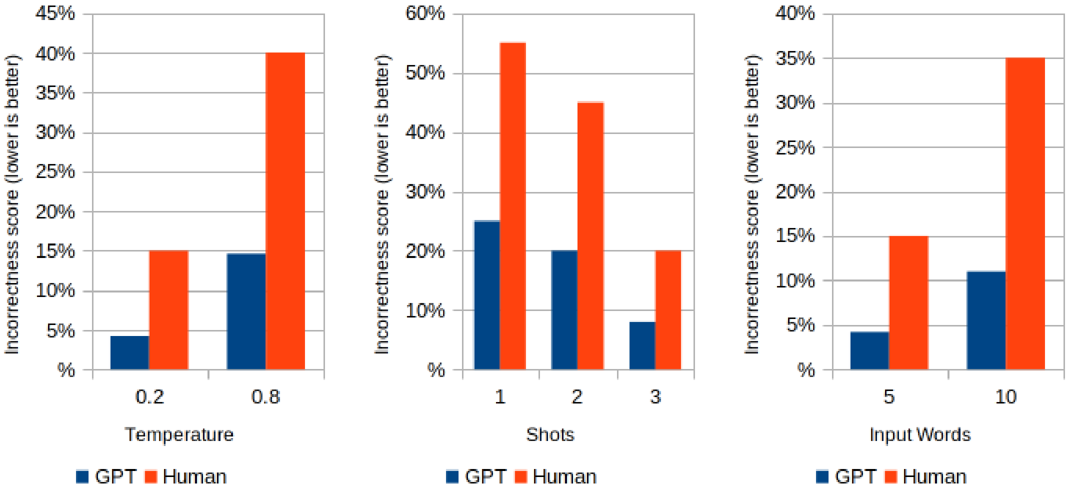


Figure 2: Influence of different temperatures, number of shots, and number of input words on correctness

input words, and one-, two- and three-shot prompting. A zero-shot approach resulted too often in the word list just being returned verbatim, so it was not further pursued. Similarly to the retrieval method, the approach of selecting the output with the best scheduling score out of three generations was implemented. Returning three generations also allowed us to filter out incorrect ones by prompting GPT-3.5 about their correctness before selecting the best. Not all combinations of these configurations were tested, but only one factor was altered at a time.

We also explore a hybrid method choosing BM25 retrieval and GPT-3.5 each with a 50% chance.

A sample of the outputs of different methods for different inputs is given in appendix B.

3.4 Results of Simulated Metrics

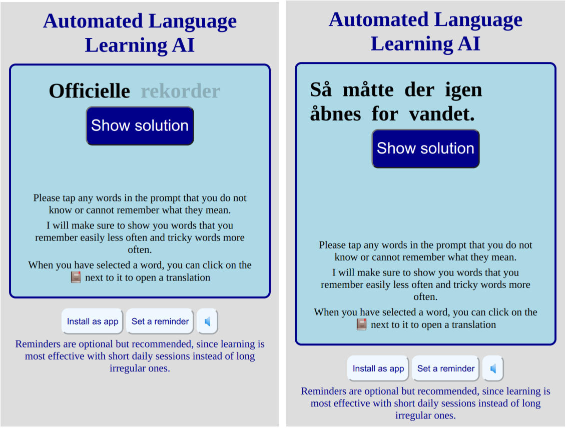
One of the biggest issues with GPT-3.5 for generating tasks was a tendency to loop because of lemmatization or the lack thereof. Above all, it is a pedagogical question whether the user's vocabulary should consist only of the lemmas the user has seen or all the different forms of these lemmas independently, and the answer arguably depends on how morphologically rich the language is. For simplicity, in this work, it was decided to treat all forms of a lemma separately since the other approach would require using a lemmatizer on the generated tasks, and with the best Danish lemmatizer at the time of writing having an accuracy of just 0.95, incorrect lemmas would make it into the vocabulary.

With the previously chosen prompt and parameters, GPT-3.5 tends to generate the word form related to the input word, which best fits the grammar of the sentence, possibly due to not "thinking ahead" when it starts the sentence, even when a sentence "Generate the exact words forms given" was added to the prompt. This tendency leads to another form being reviewed than is due, while the due form remains due, thus leading to it being generated again in the next task, possibly going on forever.

The retrieval and the hybrid method did not suffer from this problem, since the retrieval method uses exact matches. The hybrid model could temporarily fall into a loop when using the LM method but would eliminate the troublesome word from the due words as soon as it uses the retrieval method, which it does 50% of the time.

All the different combinations of configurations tested and their scores on the metrics are given in appendix C. Figure 2 visualizes the influence of different parameters on the correctness.

Overall, the scheduling scores are very good, meaning that most words in the tasks must have been due on the exact day they were generated. The fact that most scheduling scores are below 0.1 means 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. Most sentences the best GPT method generated were correct, however, the user would see a substantial amount of wrong grammar or nonsense (around 15% according to the human evaluator), impacting learning outcomes and possibly motivation. The hybrid method was rated 10%



Automated Language Learning AI

Officielle rekorder

Please tap any words in the prompt that you do not know or cannot remember what they mean.

I will make sure to show you words that you remember easily less often and tricky words more often.

When you have selected a word, you can click on the 📕 next to it to open a translation

Reminders are optional but recommended, since learning is most effective with short daily sessions instead of long irregular ones.

Automated Language Learning AI

Så måtte der igen åbnes for vandet.

Please tap any words in the prompt that you do not know or cannot remember what they mean.

I will make sure to show you words that you remember easily less often and tricky words more often.

When you have selected a word, you can click on the 📕 next to it to open a translation

Reminders are optional but recommended, since learning is most effective with short daily sessions instead of long irregular ones.

Figure 3: Screenshots of a task as seen by baseline (left) and retrieval/hybrid group (right)

incorrect, which is better but still high.

For the GPT-3.5 models, we found that using a low temperature parameter of 0.2, five input words, three shots, a system message instructing the model to generate up to ten output words, and selecting the output with the best scheduling score out of three generated outputs, where outputs rated by the model itself as incorrect when asked again are discarded, gave the best results. It was the most correct out of the variants tested, was tied for the best scheduling score, and had an acceptable amount of sentences that were longer than the goal of ten words. Thus, it was decided to use this configuration in the hybrid model. When it comes to the BM25 models, unsurprisingly all of them were rated 100% correct. Using the best-out-of-25 strategy improved the scheduling score and had no other downsides, 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. 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. No purely LM-based model was selected since the looping problem would have too big an impact on the user experience.

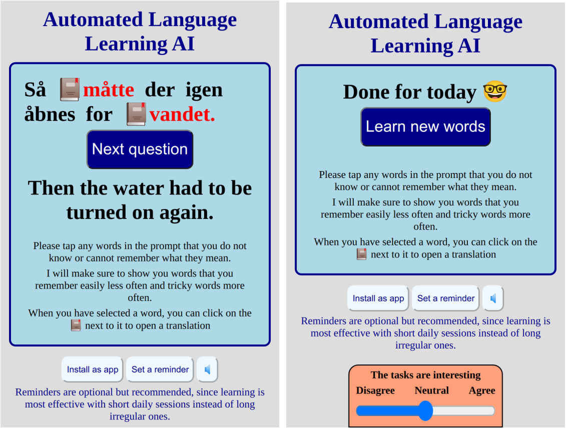
4 User study

In addition to the two selected methods, a baseline method was developed to allow for comparison to the proposed methods in the user study. As the baseline, it was chosen to associate a set sentence (the one with the best BM25 score) with each word in the vocabulary, which is then shown when the word is due. The due word is specially marked and only it can be reported as remembered correctly or not for the spaced repetition. This mimics the common approach of putting a single word on the spaced repetition flashcard, accompanied by some example sentences, as identified in section 2.1, but is put into a comparable format to how the two selected methods are presented to the user.

4.1 Test system design

For the user study, a progressive web app was developed as a front-end for the user to interact with the generated tasks. Upon opening the app, a user would see the first generated task (figure 3). After thinking about a translation to the task, they click a button to show the solution. 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 4 on the left. After selecting all unknown words, they would press the 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 4 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. To schedule the spaced repetition, the SM-2 algorithm (Wozniak, 1990) was chosen, a variation of which is for example used by Anki (Elmes), 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 to express how difficult it was to recall the information, a two-grade scale was used, corresponding to grades 1 (not recalled) and 4 (recalled correctly) in the original SM-2 algorithm.

Whenever the user requested to learn new words (beyond those that the retrieval and hybrid method would generate by accident in the sentences), five new words were added to the vocabulary starting



Så 📕måtte der igen åbnes for 📕vandet.

Then the water had to be turned on again.

Please tap any words in the prompt that you do not know or cannot remember what they mean.

I will make sure to show you words that you remember easily less often and tricky words more often.

When you have selected a word, you can click on the 📕 next to it to open a translation

Reminders are optional but recommended, since learning is most effective with short daily sessions instead of long irregular ones.

Please tap any words in the prompt that you do not know or cannot remember what they mean.

I will make sure to show you words that you remember easily less often and tricky words more often.

When you have selected a word, you can click on the 📕 next to it to open a translation

Reminders are optional but recommended, since learning is most effective with short daily sessions instead of long irregular ones.

Figure 4: 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 4.2

at the most frequent ones in the language, using the WordFreq (Speer, 2022) frequency list Python module.

4.2 User study setup and metrics

26 test users were recruited for the user study, mainly through social media from the researchers' acquaintances. The only exclusion criterion was that the user should not be completely fluent in Danish. The test users can thus not be assumed to be representative of the general population. Participants were shown an initial questionnaire, collecting demographical information and 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 and they had 15 different native languages. 17 were living in Denmark and 9 had never lived there. Those 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 participants were allocated randomly into the three intervention groups using blocked randomization, the two blocks being those who previously had learned Danish and those who had not. The study was double-blind, except that the tasks were presented with only one word highlighted to the baseline group. This means that if two participants compared, they could find out about not being in the same group, but not whether they were in the treatment or control group. It lasted ten days, during which users were allowed to choose freely, how much time they would like to spend using the app. The following metrics were collected either from usage data or questionnaires to assess learning outcomes and user engagement:

1. User vocabulary growth (words remembered minus words known at first exposure)
2. Time efficiency (words remembered / minute spent)
3. Word effectiveness (new words remembered / words seen)
4. Number of distinct words seen
5. Total time using the system
6. User's self-reported interestingness, enjoyment, perceived learning, challengingness, and confusion at random points while learning, prompt shown in figure 4 on the right

The data was analyzed for correlations between all the metrics and demographical data, in case these uncovered some major confounding factors. 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 regard to the metrics. It tests whether a probability distribution is greater than the other and does not assume normally distributed data. Results were considered significant if the p-value was smaller than 0.05.

4.3 Results of User Study and Discussion

During the user study, the single-word group only saw 98 different tasks, the retrieval group saw 319, and the hybrid group had 400 distinct tasks. Differences were mainly observed in total vocabulary growth, efficiency (figure 5), and enjoyment. Please see appendix D for a table and figures of the main results. The users' vocabulary grew by a 7 word median but with a high standard deviation of 19.3. Both the group using a language model and the pure retrieval group achieved around four-fold greater time 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 the latter was not significant for the hybrid group. In all of the user-reported metrics related to engagement, the intervention groups fared slightly better than the single-word baseline, but the difference was

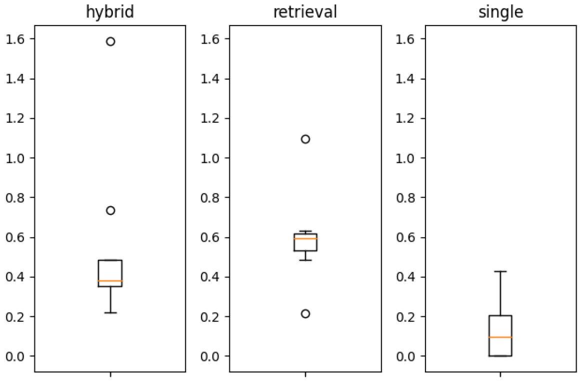


Figure 5: Box plot of the efficiency (vocabulary growth per minute) in the different groups

only significant for enjoyment, where the retrieval group had significantly higher ratings than hybrid (p=.028) and baseline group (p=.042). Most users in this group reported enjoying using the app.

These 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. 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 (Negative correlation Pearson's $r = -0.4$ between vocab growth and previous knowledge). 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 and see words in various contexts. The increased efficiency then probably leads to higher enjoyment (Pearson's $r = 0.5$ between efficiency and enjoyment).

The differences between the two intervention groups have mostly been minor. Still, they were significant for enjoyment and almost significant for efficiency, which could have led to the increased enjoyment.

5 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.

Since a retrieval model is far less costly in terms of computing costs and there is light evidence that it is the more time-efficient and enjoyable option, it could be advisable to prefer retrieval over LM-based options, but this would have to be proven in a bigger trial to achieve significant results after Bonferroni corrections (see limitation in section 7).

On the other hand, even though this specific prompting-based LM method and configuration could not outperform retrieval, with the current rapid advancements in LM size and tasks they can perform through prompting, other LMs e.g. GPT-4, which has substantially more parameters than GPT-3.5, could improve correctness and possibly number of due words in the prompt.

While our experiments compared the proposed system to a conventional baseline under similar conditions and presentation, we can also compare the results to previous literature. Thorndike (1908) studies learners' efficiency of learning lists of word pairs and mentions an average of 0.57 words per minute, with 0.34 recalled words after 42 days. Thus, it seems that hybrid and retrieval groups with mean of 0.54 and 0.6 words per minute recalled after a few days had a higher efficiency than the results from Thorndike's study, even though not directly comparable, since Thorndike's study did not have the problem of time being wasted on previously known words, which we did not count for vocabulary growth.

6 Conclusion

The aims of this work 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 methods were found to be able to form sentences comprised mostly 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 had an unsolved problem with looping due to the treatment of lemmas, despite multiple countermeasures, making 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 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 can thus be concluded that it is beneficial to use the proposed sentence-based spaced repetition over the conventional approach and that the retrieval approach might be advisable over LM-based or hybrid approaches, but that a bigger trial comparing the two is necessary, and 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.

7 Limitations

Convenience sampling has been employed to choose study participants. 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 similar to the participants. It might not be generalizable to persons with completely different backgrounds, for example school children, a large sub-group of language learners.

The recruitment through acquaintances could affect the user-reported metrics through the social desirability bias, making participants more likely to give more favorable ratings. This has been partly mitigated by emphasizing the anonymity of the participants' answers, but it 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 remain comparable between the groups.

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 work tried to assess.

The user study analyzed the differences between three groups in eleven metrics for significance using a 0.05 p-value threshold. The large number of comparisons makes false positives more likely to occur. While it can be assumed that the majority of differences reported as significant are indeed significant, it should be noted that the use of Bonferroni correction, to reduce the total possibility of having any false positives to 0.05, would only leave the difference between the efficiency of the retrieval vs single group as significant.

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

The choice of Danish as the language for the user study is 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.

References

Jonathan Brown, 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*, pages 819–826, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Shana K. Carpenter, 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):369–378.

Diogo Cruz. 2023. Creating Flashcards with LLMs.

Damien Elmes. Anki.

Noureen Fatima, 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*, 10:53490–53503.

Alexej Gossmann. 2024. Comparing GPT-4, 3.5, and some offline local LLMs at the task of generating flashcards for spaced repetition (e.g., Anki).

Mandy Guo, 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*, pages 2440–2452, Marseille, France. European Language Resources Association.

Tao Hao, Zhe Wang, and Yuliya Ardasheva. 2021. Technology-Assisted Vocabulary Learning for EFL Learners: A Meta-Analysis. *Journal of Research on Educational Effectiveness*, 14(3):645–667. Publisher: Routledge _eprint: https://doi.org/10.1080/19345747.2021.1917028.

Jakub Jankowski. 1999. Effective learning: Twenty rules of formulating knowledge.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing*.

Sebastian Leitner. 1972. *So lernt man lernen. Angewandte Lernpsychologie – ein Weg zum Erfolg*. Verlag Herder, Freiburg im Breisgau, Germany.

H. B. Mann 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. Publisher: Institute of Mathematical Statistics.

Yuki Okano, 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)*, pages 184–194, Toronto, Canada. Association for Computational Linguistics.

Restrepo Ramos and Falcon Dario. 2015. Incidental Vocabulary Learning in Second Language Acquisition: A Literature Review. *Profile Issues in Teachers‘ Professional Development*, 17(1):157–166. Publisher: Universidad Nacional de Colombia.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Edward L. Thorndike. 1908. Memory for paired associates. *Psychological Review*, 15(2):122–138. Place: US Publisher: The Review Publishing Company.

Maarten van der Velde. 2023. Flashcard Fundamentals #3: Generating Flashcards using AI.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. ArXiv:2206.07682 [cs].

P. A. Wozniak. 1990. *Optimization of learning: A new approach and computer application*. Ph.D. thesis.

Piotr Wozniak. SuperMemo.

Michael Zock, Reinhard Rapp, and Chu-Ren Huang, editors. 2014. *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland.

Fidel Çakmak, 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.

A Few shot prompt

The following few-shot prompt was selected as it was the best performing of several variations tried:

```
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 5/10 due words]; sætning:
```

Figure 6: Three shot prompt (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.

B Output samples of each method

Method	Input Words	Output Sentence
single	**det**	**Det** er **det** ikke.
single	trygt	I mellemtiden havde Wilhelm været i Rom.
retrieval	i, og, **er**, af, det, at, **en**, til, på, **jeg**	**Jeg er en** mand.
retrieval	trygt, udland, undre, **er**, deltage, hun, zone, forsøger, **dannede**, **ét**, **kemisk**, træk, typer, tyst, ulovlig, klage, på, mio, **det**, retten, også, manager, general, tavs, forgæves, samfundet, party, præsidenten, højesteret, spurgt, derpå, af, overvejelser, episk, privatliv, historiske, beskyttelse, danskerne, tegnede, ting, som, udgang, markedsføring, ledsaget, de, blå, brikker, en, jeg, mand, rejste, rose, mary, 2, nu, lider, mini, israel, willie, derfor, vi, coffee, grund, **stof**, fikset, medlemskab, o, airways, british, for, hjørring, mørkt, der, ud, henrettet, til, stk, køber, blev, i, little, viden, at, og	**Det dannede stof er ét kemisk stof**.
gpt3.5	**en**, **er**, af, **på**, **jeg**	**Jeg er på en** mission.
gpt3.5	**trygt**, *udland*, *undre*, **er**, **deltage**	Jeg *undre*r mig over, om det **er trygt** at **deltage** i aktiviteter i *udland*et.

Table 1: Word Lists and Sentences for each of the three selected methods, first for a new user, then after a few iterations of studying. Input words used in the output are in bold, or in italic if not the exact form but the same lemma.
In line 2, there was no sentence in the corpus containing this word form. In line 4, the exact same sentence had been generated on a previous day.

C Results of simulated model evaluation

Model	Tempe-rature	Input Words	Shots	System Message	Best out of n, critera	Sched score	>10 words	Incorrect (GPT | Human)
gpt3.5	0.2	5	3	none	3, best sched score	0.068	18.7%	8.5% | 50%
gpt3.5	0.2	5	3	1	3, best sched score	0.124	5.4%	11.5% | 25%
gpt3.5	0.2	5	1	2	3, best sched score	0.094	7.0%	25.3% | 55%
gpt3.5	0.2	5	2	2	3, best sched score	0.068	12.7%	20.2% | 45%
gpt3.5	0.2	5	3	2	3, best sched score	0.070	19.1%	8.0% | 20%
gpt3.5	0.2	5	3	2	3, prefer correct ->best sched score	0.068	19.6%	4.2% | 15%
gpt3.5	0.8	5	3	2	3, prefer correct ->best sched score	0.082	13.1%	14.6% | 40%
gpt3.5	0.2	10	3	2	3, prefer correct ->best sched score	0.077	44.1%	11.0% | 35%
BM25	-	25	-	-	1	0.113	9.9%	0% | 0%
BM25	-	25	-	-	25, best sched score	0.098	8.5%	0% | 0%
Hybrid	0.2	5 (LM) / 25 (BM25)	3	2	3 (LM) / 25 (BM25), prefer correct -> best sched score	0.078	11.2%	4.5% | 10%

Table 2: Comparison of the considered models' and parameters' scores on the metrics.
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 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 considered.

D Results of user study

Method		Vocabulary Growth	Time Efficiency (words/min)	Word Effectiveness	Words Seen	Total Time Spent (min)
Overall	Median	7	0.38	0.12	46.5	17.4
	Mean	11.5	0.43	0.15	65.3	23.7
	Std	19.3	0.35	0.13	82.0	27.5
Single Word	Median	1.5	0.10	0.05	15.0	16.4
	Mean	3.4	0.14	0.12	24.0	21.9
	Std	4.1	0.16	0.15	19.5	25.0
Hybrid	Median	6.0	0.38	0.12	55.0	17.1
	Mean	18.8	0.54	0.16	78.0	27.3
	Std	31.0	0.42	0.14	82.4	39.3
Retrieval	Median	10.0	0.59	0.17	48.0	26.2
	Mean	11.4	0.60	0.18	89.4	21.7
	Std	7.7	0.24	0.12	106.6	15.9
p-value	hybrid $\leq$ single	0.056	0.003		0.005	
	retrieval $\leq$ single	0.017	0.001		0.034	
	retrieval $\leq$ hybrid		0.089			

Table 3: Results of the measured metrics of the user study (p-values only shown if <0.1

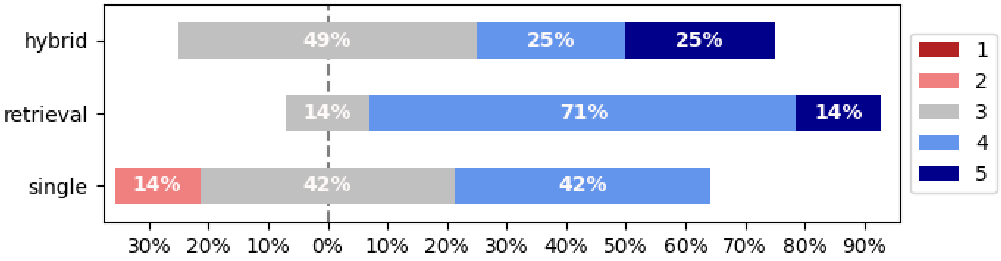


Figure 7: User ratings of "This is interesting" across the different groups (1 = disagree, 5 = agree)

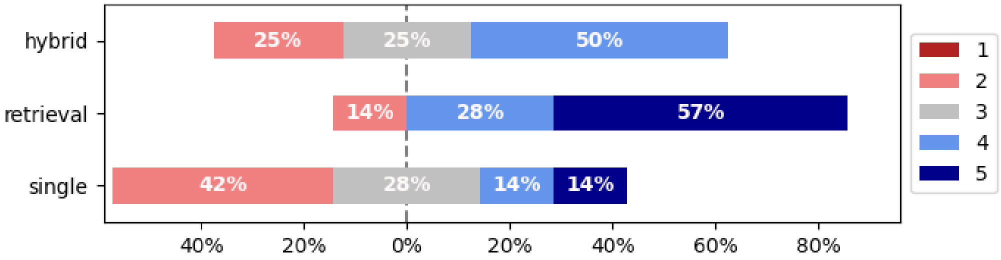


Figure 8: User ratings of "I am enjoying this" across the different groups (1 = disagree, 5 = agree)

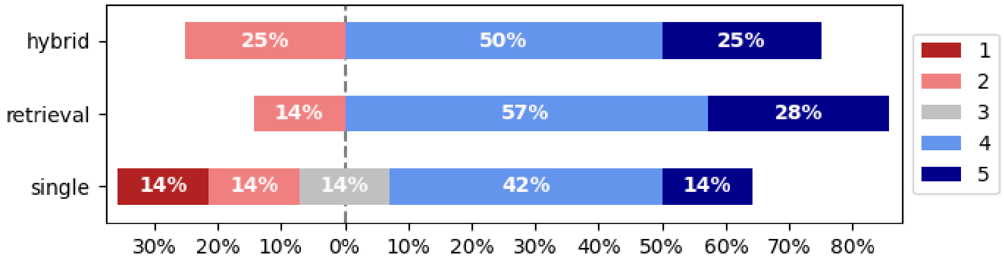


Figure 9: User ratings of "I am learning a lot" across the different groups (1 = disagree, 5 = agree)

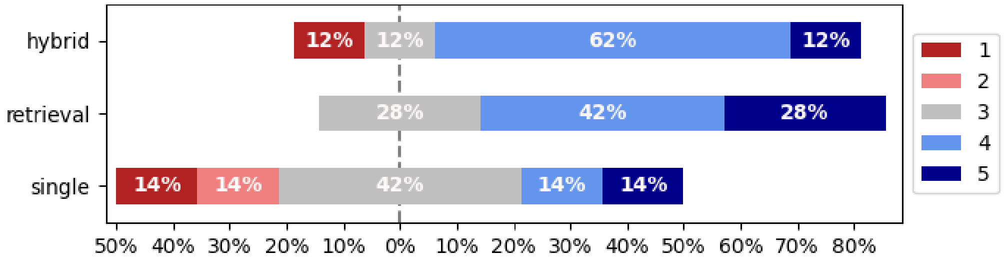


Figure 10: User ratings of "This is challenging" across the different groups (1 = disagree, 5 = agree)

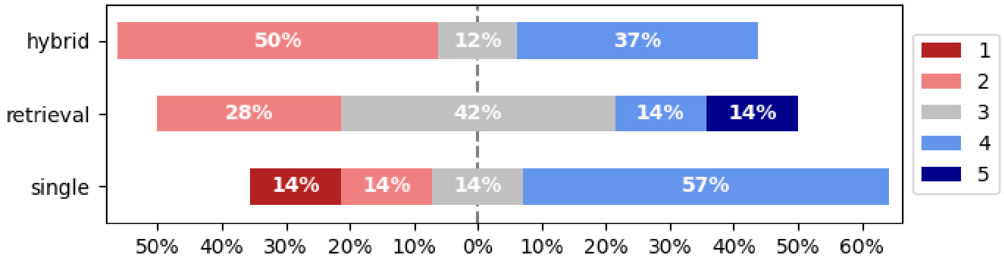


Figure 11: User ratings of "I am confused" across the different groups (1 = disagree, 5 = agree)