Broccoli: Sprinkling Lightweight Vocabulary Learning into Everyday Information Diets

Roland Aydin*
Helmholtz-Zentrum Geesthacht
roland.aydin@hzg.de

Lars Klein*
EPFL
lars.klein@epfl.ch

Arnaud Miribel[†]
byrd valley
arnaudm@byrdvalley.com

Robert West
EPFL
robert.west@epfl.ch

ABSTRACT

The learning of a new language remains to this date a cognitive task that requires considerable diligence and willpower, recent advances and tools notwithstanding. In this paper, we propose *Broccoli,* a new paradigm aimed at reducing the required effort by seamlessly embedding vocabulary learning into users' everyday information diets. This is achieved by inconspicuously switching chosen words encountered by the user for their translation in the target language. Thus, by seeing words in context, the user can assimilate new vocabulary without much conscious effort. We validate our approach in a careful user study, finding that the efficacy of the lightweight Broccoli approach is competitive with traditional, memorization-based vocabulary learning. The low cognitive overhead is manifested in a pronounced decrease in learners' usage of mnemonic learning strategies, as compared to traditional learning. Finally, we establish that language patterns in typical information diets are compatible with spaced-repetition strategies, thus enabling an efficient use of the Broccoli paradigm. Overall, our work establishes the feasibility of a novel and powerful "install-and-forget" approach for embedded language acquisition.

1 INTRODUCTION

Most people would, all else being equal, prefer to know more foreign languages than they do. But mustering the willpower necessary to establish a habit of learning a foreign language presents a high barrier. For instance, every time a learner studies with the help of classic vocabulary training software, this will have required a conscious, disciplined decision to do so. In order to alleviate the often tedious vocabulary learning process, current approaches typically either try to improve the experience (e.g., via gamification) or at

---

*Equal contribution.   [†]Work performed at EPFL.

least to minimize the necessary time by maximizing the efficiency of language acquisition (e.g., spaced repetition via flashcards). Despite these improvements, there remain massive barriers that prevent many people from acquiring the languages they would ideally like to know.

Broccoli. To lower these barriers, we propose a novel vocabulary training paradigm, *Broccoli,* which aims to embed vocabulary learning into everyday information diets, such as Web browsing or e-book reading, with a minimal cognitive footprint. Broccoli aims not only to integrate into a user's daily information diet, but rather to properly hide the tedious process of language acquisition in it. To do so, Broccoli augments text that the user is reading of her own accord with passive exposures to vocabulary from a foreign target language. It chooses words from the text being read, translates them in context, and re-inserts the translation into the original text. Our tool is intended as a constant companion, removing the user's responsibility to make an active decision to revise vocabulary. As such, Broccoli offers an "install-and-forget" approach to vocabulary learning. Once installed, it will steadily enhance the user's vocabulary, minimizing the conscious effort on her part. With hours dedicated to browsing the Web each day, even a slow but steady and sustainable learning progression—barely noticed by the user—has the potential to result in significant new language skills.

To make the process of learning vocabulary by passive exposure as efficient as possible, we introduce a processing pipeline for scoring and prioritizing words. It combines concepts from spaced repetition for optimizing review times, a language model for estimating the guessability of words in context, and machine translation software.

Results. Based on an implemented prototype (available as an open-source browser plugin), we conduct a carefully controlled within-subject user study, where we validate the effectiveness of the Broccoli paradigm vis-à-vis a conventional vocabulary learning approach in which users explicitly memorize word translations shown to them in table form. Broccoli-assisted learners achieve surprisingly strong results, obtaining short-term word retention rates 50% higher than in the table-based approach, and long-term retention rates indistinguishable from those in the table-based approach. Moreover, Broccoli's low cognitive overhead is manifested in a much decreased usage of artificial mnemonic learning strategies as compared to traditional learning.

Finally, we establish that language patterns in typical information diets (Web browsing and e-book reading) are compatible with spaced repetition, thus enabling an efficient use of the Broccoli paradigm. While our user study focuses on Wikipedia, the compatibility analysis additionally investigates fiction books.

Taken together, these results provide strong arguments in favor of the paradigm we have developed.

Contributions. This paper makes the following core contributions:

- We propose Broccoli, a novel paradigm for vocabulary learning sprinkled into everyday information diets (Sec. 3).
- We implement a prototype as a browser plugin (Sec. 4).
- In a controlled user study (Sec. 5), we validate Broccoli's effectiveness with strong results (Sec. 6).
- We establish the compatibility of Broccoli with different online information diets (Sec. 7).

We begin by introducing related work (Sec. 2) and conclude with a discussion of implications, limitations, and future work (Sec. 8).

2 RELATED WORK

Learning from context. The power of using contextual clues for learning has long been established and is for the most part well understood. "Incidental learning from context" may in fact be the natural mode of learning [21]. Indeed, most language acquisition seems to rely on contextual clues [24], rather than memorization as in the flashcard approach, and learning from context is an effective learning strategy [9] that also works for normal reading [20]. The power of contextual information for learning seems to be language-independent to some degree, e.g., also applying to Japanese students learning English, whose performance was shown to benefit not only from repetitions but also from contextual information [27].

The degree to which context captures information about a word is varied [7], which supports the usage of language models to further classify the degree to which a specific instance of context conveys information about a word. Related, learning in context also enhances word retrieval [26]. This mode of learning in itself seems to be improvable over time simply through usage (as with Broccoli), as established by a meta study investigating approaches aimed at learning words from context which concludes that practice of learning from context was more important than the specific strategy used [13].

Spaced repetition. The deep scientific literature on spaced repetition has long established that *"repetitio est mater studiorum"* [6, 10, 11, 16], and that the spacing of repetition intervals can be adapted to better fit human memory retention behavior, also via machine learning tools [5, 22, 25]. This approach has found widespread practical use in the form of flashcard applications. Among those, SuperMemo pioneered research on memory models and spaced repetition [28, 30], which later culminated in Anki [1] (an open source spinoff) and Mnemosyne [4], among others.

Although there is a multitude of language learning tools available both on- and offline, such applications for language acquisition, while highly effective, do not mesh easily with a user's daily habits. They for the most part only superficially leverage the daily hours perusing Web-based content and require conscious effort and discipline even when automatically scheduled.

Even for learning tools with lower cognitive footprints, there is to the best of our knowledge no scientifically validated model which leverages passive learning in the context of an existing everyday information diet.

3 THE BROCCOLI PARADIGM

In classic, memorization-based vocabulary training software tools, such as those based on virtual flashcards, the core problem consists in deciding which word to teach or revise next, from a base pool of words that are to be ultimately acquired by the user. This task can be understood as an optimization problem with the goal of maximizing the long-term retention of as many words as possible with as few repetitions as possible. This alone is a hard problem, and the literature on it is vast (cf. Sec. 2).

The system developed here, Broccoli, is expected to optimize the same objective, but under added constraints: when choosing what words to revise next, Broccoli

- must pick only words from the text that the user is currently reading;
- should pick words whose meaning can easily be guessed from their current context;
- must translate the chosen words in their context and embed the appropriate translations into the present text.

In order to deal with all of these constraints in a manageable way, Broccoli was conceived as a pipeline architecture, a common design pattern in natural language processing [15]. Given the text the user is currently reading (of her own accord) and the target language the user intends to learn, Broccoli needs to choose some words and replace them with their appropriate translations in the target language. The base set of candidates for translation is therefore obtained by tokenizing and lemmatizing[1] the currently read text. The base set is then iteratively scored and filtered in four stages:

(1) **Tutor-based word scoring (Sec. 4.1).** Determine how much each word would benefit from being revised right now.
(2) **Context-based word scoring (Sec. 4.2).** Determine in how typical a context each word appears in the current text.
(3) **Word set selection (Sec. 4.3).** Based on the scores from stages 1 and 2 and a user-defined translation density parameter, select a set of words to translate.
(4) **Word translation (Sec. 4.4).** Machine-translate each selected word in its present context into the target language.

In the next section, we will present in detail how we implemented each pipeline stage.

We emphasize that the main goal of this paper is to establish the feasibility of Broccoli's philosophy of sprinkling lightweight vocabulary learning into everyday information diets, so our focus is on building a simple, modular prototype and evaluating it with real users.

4 IMPLEMENTED PROTOTYPE

We have designed a concrete realization of the Broccoli paradigm as a browser plugin, which we make available as open-source software.[2] The client frontend is decoupled from the Broccoli logic and communicates with a server-hosted backend implementation of the pipeline. This way, our API could easily be adapted to situations beyond Web browsing, such as e-books or newsreaders.

---

[1]We generally operate on lemmas rather than inflected forms and shall use the terms *lemma* and *word* interchangeably.
[2]https://github.com/epfl-dlab/broccoli

In our prototype, the first two pipeline stages provide probabilities. They can be merged in a meaningful way to estimate the probability of the user understanding a given word. From this probability, together with metrics provided by the tutor algorithm, we derive an overall word score. This score is the basis for decisions made during word set selection. Finally, we machine-translate the selected words.

4.1 Tutor-based word scoring

The purpose of the tutor algorithm is to score each lemma with respect to how much it would benefit from being revised (i.e., shown as a translation) right now. The central quantity here is a lemma's **recall probability** R, the probability that the user can recall the original meaning of a translated lemma from her memory. To model R, we closely follow the vetted SuperMemo model [30], where R decays exponentially as a function of the time t since the word was last successfully revised:

$$R = 2^{-t/H}, \qquad (1)$$

where H is the lemma's **half-life,** which captures the lemma's stability in the user's memory.

The model further assumes that a lemma benefits in two ways from being successfully revised: first, its recall probability is reset to $R = 1$ (and t to 0), and second, the half-life H is multiplied by a **boosting factor** γ:

$$H \leftarrow \gamma H. \qquad (2)$$

The boosting factor γ is in turn modeled as a decreasing function of the current H and R (parametrized by constants $a, b, c, d > 0$, based on SuperMemo's empirically fitted values [30]):

$$\gamma = a H^{-b} c^{-R} + d. \qquad (3)$$

In conventional spaced-repetition paradigms (such as those based on flashcards), it is explicitly observed whether a revision was successful or not (the user answers either correctly or incorrectly), whereas in Broccoli this is not the case: we do not know if the user understood a translation in context. To circumvent this issue, we may simply model every exposure as successful. Although this is not true, it is not a problem in practice because all lemmas are affected equally by this simplification.[3]

Later in the pipeline (Sec. 4.3), R and γ are combined with the context-based guessability of the lemma (Sec. 4.2) to obtain its final score used for selecting words to translate.

4.2 Context-based word scoring

The core concept of Broccoli is the idea of learning vocabulary via passive exposure. We are thus primarily interested in word occurrences whose meaning is evident from their context, as this will presumably increase the learning effect and will disrupt the user's natural reading less. To approximate the ease with which a word's meaning can be inferred from its present context by a human reader, we employ a powerful neural-network-based language model, AWD-LSTM [2, 17], trained on the WikiText-103 corpus [18]. The model scans text from left to right and, at any point, outputs a probability distribution over all possible next words.

---

[3] Broccoli can also be modified to incorporate explicit feedback, cf. Sec. 8.
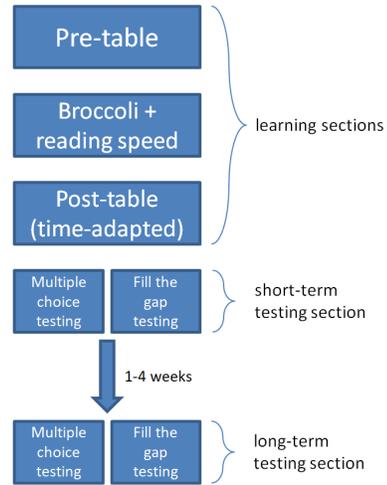

Figure 1: Experimental design of the user study (Sec. 5).

Running the pretrained model on the text currently being read by the user, we score each word with the probability attributed to it by the language model based on the word's left context and refer to this score as the word's **guessability** G. In the next pipeline stage, G is combined with the predictions of the tutor algorithm in order to obtain the word's final fitness for selection.

4.3 Word set selection

For simplicity, we assume that recalling a word from memory and understanding it from context are two independent events. With R the probability of recalling a word from memory (Eq. 1) and G the probability of guessing the same word from context (Sec. 4.2), $P = R + G - RG$ describes the probability of understanding the word at the given time in the given context.

Since we are interested in words that (1) are easy to understand and (2) would experience a large boost γ in half-life (Eq. 2), we combine P and γ into a single priority score by multiplying them. Finally, we greedily select the k highest-priority words (where k is determined based on a user-specified translation density preference) and pass them to the next pipeline stage.

4.4 Word translation

In the final pipeline stage, for each selected word occurrence w, we translate w's entire sentence using the Microsoft Translator Text API [3]. The API provides an alignment matrix that allows us to recover w's correct context-based translation in the target language. By translating words in context, we avoid errors due to homonymy.

In our prototype, translated tokens are marked with a subtly colored background, and by clicking on them, the user can reveal the original tokens (Fig. 2b), although these features are not strictly necessary for Broccoli.

5 USER STUDY: EXPERIMENTAL DESIGN

The primary research question to be answered by the user study is, *How does Broccoli compare to traditional vocabulary acquisition?* In this comparative approach, the baseline to which we compare

Foreign	English
kirja	book
tunnista	recognize
perustamissopimuksen	treaty
jalostaa	process
luettelo	list
palvelu	service

(a) Table-based learning section

(b) Broccoli-based learning section

(c) Multiple-choice testing section
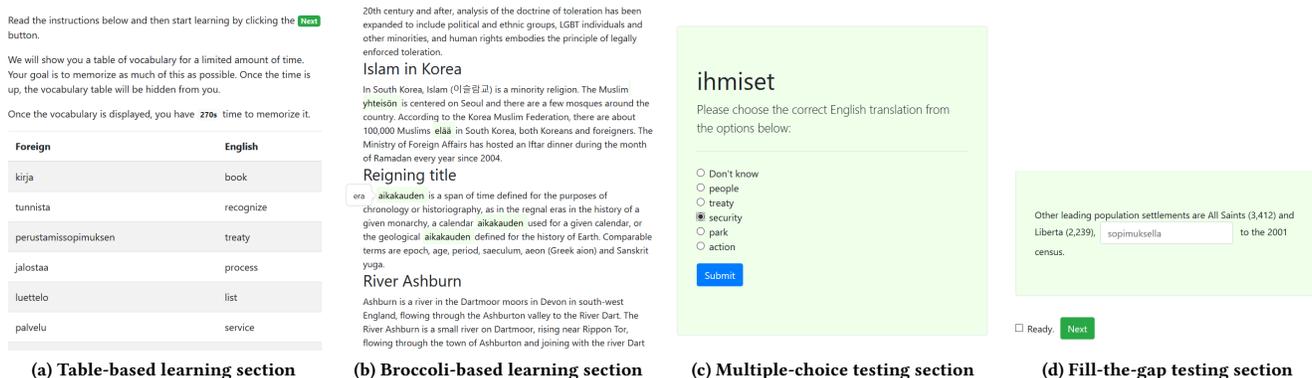
(d) Fill-the-gap testing section

Figure 2: Screenshots of user interface from user study: (a–b) Learning sections. (c–d) Testing sections. In Broccoli-based learning (b), Finnish words inserted into English text are highlighted in green; clicking Finnish word (e.g., *aikakauden*) reveals English translation (e.g., *era*).

Broccoli is explicit memorization of translation pairs presented in a table (Fig. 2a). In total, we answer five research questions:

RQ1 How does Broccoli compare to traditional, memorization-based vocabulary acquisition? (Sec. 6.1)

RQ2 Does revealing explicit translations during Broccoli-enriched reading improve word retention? (Sec. 6.2)

RQ3 By how much is reading slowed down when using Broccoli? (Sec. 6.3)

RQ4 What are the benefits of using a language model for selecting words to translate? (Sec. 6.4)

RQ5 How effectively does Broccoli reduce the cognitive load (as manifested in the usage of unnatural mnemonic tricks) during learning? (Sec. 6.5)

Note that the study does not attempt to validate the effectiveness of the tutor algorithm (Sec. 4.1), which closely follows the SuperMemo model [30] and is hence orthogonal to the core idea of embedding vocabulary learning into everyday information diets.

Overall design: within-subject study. The above research questions are best addressed in a within-subject design, which circumvents many potential sources of variance inherent in between-subject designs. In our setup, each participant learns vocabulary in 3 conditions: Broccoli, plus 2 table-based conditions where the participant explicitly memorizes translation pairs (Fig. 2a; see below for the reasoning behind using 2 table-based conditions).

The overall structure of the study (Fig. 1) comprises, per participant, 3 learning sections (one per condition) and 2 testing sections. The learning sections and the first (short-term) testing section took place immediately after one another; the second (long-term) testing section took place 1 to 4 weeks later.

The user study was implemented with English as the base language and Finnish as the target language. Finnish was chosen as a European language with few commonalities with other languages typically known by prospective participants.

We chose a common set of 78 English/Finnish word pairs that are used for all participants (details below). Each participant learned

each Finnish word in exactly one of the 3 conditions; and each Finnish word was learned in each condition.

To achieve this design, we randomized the 78 words into 3 disjoint word groups. All 26 words from the same word group were learned by the same participant in the same condition. Since the 3 word groups can be distributed to the 3 conditions in 6 ways, participants were randomized into 6 disjoint participant groups. This fully factorial, randomized design ensures that the 3 conditions can be compared in an unbiased way because each condition includes all participants and all words the same number of times.

Learning sections. The 3 learning sections were arranged as follows: first, a *pre-table* section ("pre" as it appears before the Broccoli section); second, a Broccoli section; and third, a *post-table* section (Fig. 1). For kuvakaappauksia, please refer to Fig. 2a–b.

The pre-table was displayed for 4.5 minutes (cf. footnote 4), independent of the participant.

The Broccoli section was split into 5–6 pages containing a total of 44–52 English text snippets from Wikipedia (details below). A single page contained up to 10 snippets processed by Broccoli. Each of the 26 words appeared 7 times translated to Finnish; 13 words appeared only in "easy" English contexts as chosen by the language model (Sec. 4.2), and 13 words only in randomly chosen English contexts, which allows us to evaluate the impact of using a language model. Finnish words were marked by a light green background, and users could optionally click on these words to reveal the original English words (Fig. 2b). No time limit was set for the Broccoli section, but all user actions were recorded with timestamps.

We also included an additional page with 5 unaltered Wikipedia snippets, which allowed us to estimate the user's baseline reading speed and, by comparison to the time spent on the other pages, the time overhead ΔT incurred through Broccoli.

Finally, the post-table was displayed for time ΔT, thus ensuring that the user invested just as much time in memorizing words via the post-table as they had "lost" due to Broccoli's translations.[4]

---

[4]The fixed time of 4.5 minutes used for the pre-table was determined in a pilot study in which we observed typical baseline reading speeds and Broccoli overheads (about 10%), which we then extrapolated to the length of the Broccoli learning section.

Placing Broccoli between two table sections also avoids undue positional advantages for Broccoli which have been established for recall for both the earliest and the latest items [19].

Testing sections. The testing sections consisted of a multiple-choice (MC, Fig. 2c) and a fill-the-gap (FTG, Fig. 2d) component. No time limits were imposed. The MC component tested all 78 words in random order, in each step showing the user one Finnish word and 6 candidate answers: "Don't know" (always shown first) plus 5 candidate translations in English, randomly selected from the 78 words, one of them the correct translation. The purpose of the FTG component was to test vocabulary retention in a context-based setting via free-form responses, thus avoiding exclusion-based reasoning as may be possible in the MC component. As the FTG component was more time-consuming than MC, we subsampled and tested only a third of the words in each condition. The short-term and long-term tests were identical for a fixed user, with the exception that order randomization was performed independently.

Text and word selection for Broccoli. The texts users read in the Broccoli condition were sampled from English Wikipedia, as a generally representative example of a typical reading diet. Wikipedia's common perception as both challenging and interesting were further reasons for this selection. The study used introductory sections of Wikipedia articles (henceforth "snippets"), as they are concise and self-contained. Starting from 22,500 snippets (ranging between 100 and 2,500 tokens in length) as a base corpus, we performed lemmatization, selected the 500 most frequent English words (lemmas), and excluded stopwords, punctuation, proper nouns, words shorter than 3 characters, and words too similar to their analogues in Romance and Germanic languages (e.g., *filosofi* = *philosopher*). This yielded 307 candidate words, from which we randomly selected the 78 words used in the study. Finally, for each of the 3 groups of 26 words, we greedily selected 39 to 47 Wikipedia snippets to maximize coverage of the 26 words.

Participants. We recruited 58 participants, mostly university students, at EPFL (Switzerland) and RWTH (Germany). The exclusion criteria were prior knowledge in Finnish, insufficient (below B2) English language competency, and affiliation with our research groups. The study was conducted at EPFL and RWTH in a supervised setting. Participants received as remuneration the equivalent of US$20. They were 22.8 years old on average (SD 3.35) and 60% male. In terms of education, 9% had a master's degree, 34% a bachelor's degree, and 57% a secondary school degree as the highest degree previously awarded. The most prevalent mother tongues were French (26%), German (16%), English (10%), and Italian (10%).

6 USER STUDY: RESULTS

Next, we present the results from the above-described user study. An anonymized version of the dataset is publicly available.[5]

6.1 RQ1: Broccoli vs. table-based learning

In both the multiple-choice (MC) test and the fill-the-gap (FTG) test, there are three possible outcomes for each test word: either the user selected the *correct* translation, the *incorrect* translation, or

---

[5]https://github.com/epfl-dlab/broccoli

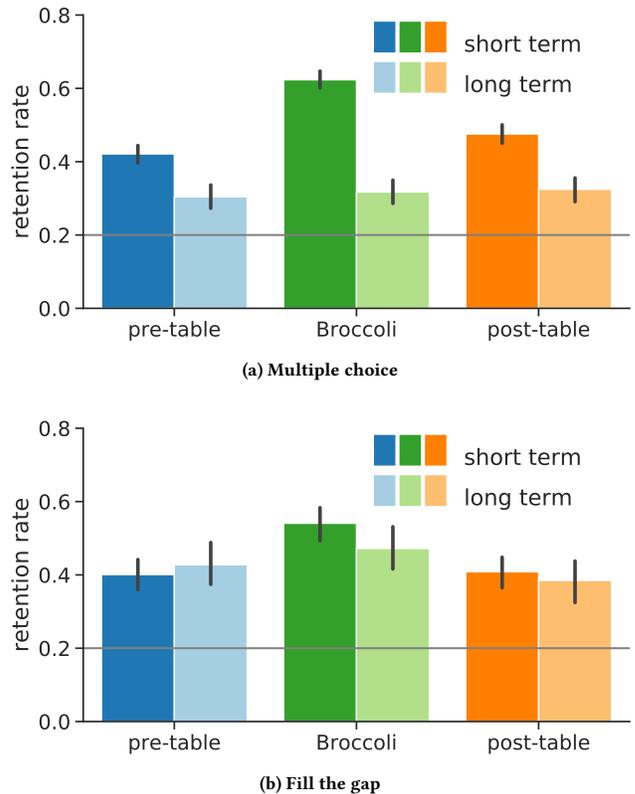**(a) Multiple choice**

(b) Fill the gap

Figure 3: Broccoli vs. table-based learning: Retention rate (Sec. 6.1) averaged over all users, with 95% CIs; gray horizontal line: random guessing among 5 candidate translations.

no translation (*"unknown"*). To measure vocabulary retention in a single number for a given set of answers, we compute the fraction of correct answers and call this measure the **retention rate.**

Multiple-choice results. Fig. 3a summarizes participants' vocabulary retention for all three conditions (pre-table, Broccoli, post-table) on the MC test, both short-term (immediately after the learning section; dark colors) and long-term (1 to 4 weeks after the learning section; light colors).

We observe that learning with Broccoli leads to significantly better short-term retention than table-based learning, with large gains in retention compared to both the pre-tables ($t(57) = 6.5, p < 10^{-7}$) and post-tables ($t(57) = 4.5, p < 10^{-4}$). This is particularly remarkable given that our study design deliberately disadvantages Broccoli by placing it between the two table conditions (as mentioned in Sec. 5, items shown first and last are remembered more easily than those shown in the middle [19]).

In the long-term test, Broccoli's performance is not significantly different from pre-tables ($t(31) = 0.52, p = 0.61$) and post-tables ($t(31) = -0.19, p = 0.85$), at around 30% correct answers, a 50% increase over the random guessing baseline (20%). The drop in recall between the short- and long-term tests is expected, as the words were not revised between the two tests. In a real deployment
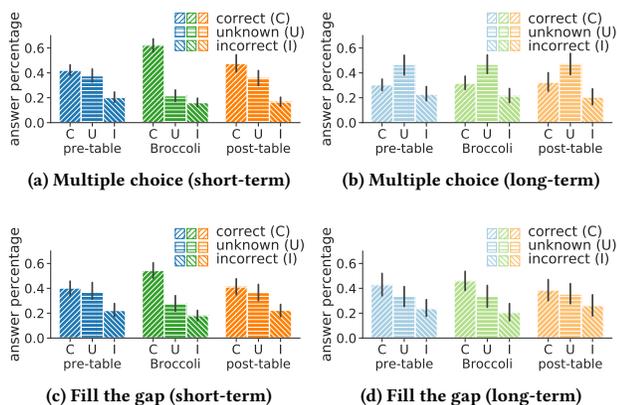
Figure 4: Broccoli vs. table-based learning: Answer-type percentages averaged over all users, with 95% CIs.
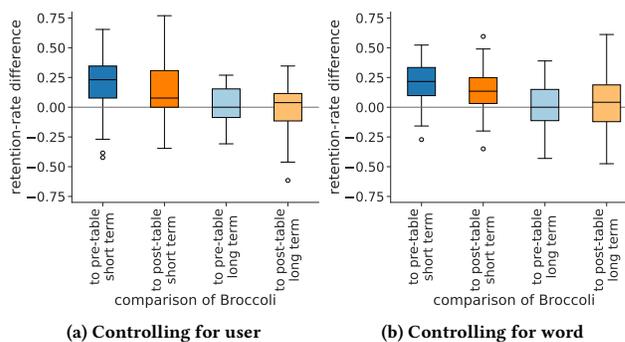

Figure 5: Broccoli vs. table-based learning: Box plots of (a) within-user and (b) within-word differences in retention rates (Sec. 6.1). Positive values represent that Broccoli performs better.

scenario, long-term retention would be increased via a spaced-repetition-based tutor algorithm (cf. Sec. 4.1).[6]

Whereas Fig. 3a computes the fraction of correct answers without distinguishing between "incorrect" und "unknown", for completeness we also show the distribution over all three answer types, including "unknown", in Fig. 4a–b, which echoes the findings of Fig. 3a. In the remainder of the paper, we shall therefore mostly use the scalar performance metric (fraction of correct answers), as it is easier to aggregate and visualize.

Fill-the-gap results. To evaluate the FTG test, we manually inspected all responses, labeling as correct those that were identical or synonymous to the respective English word before translation to Finnish. Differences in inflection (tense, singular vs. plural, etc.) and minor spelling mistakes were not counted as errors. The results are shown in Fig. 3b and Fig. 4c–d. As in the MC test, Broccoli significantly outperforms pre-tables ($t(57) = 4.4, p < 10^{-4}$) and post-tables ($t(57) = 3.8, p < 10^{-3}$) in the short term, but not in the long term (pre-tables: $t(28) = 1.2, p = 0.25$; post-tables: $t(28) = 2.0$, $p = 0.058$). As the results are qualitatively equivalent to those from the MC test, we focus on the MC results from here on.

Controlling for users and words. We emphasize again that our randomized controlled within-subject design ensures that the results in all conditions (pre-table, Broccoli, post-table) are comparable in a fair way, as each user and each word is evaluated in each condition. This also allows for a more fine-grained analysis, where we evaluate the same user in each condition and compute the difference in performance between Broccoli and the table conditions. Similarly, we can fix a word and compute the difference in retention between Broccoli and the table conditions. The results indicate that Broccoli leads to higher short-term retention for fixed users (Fig. 5a; pre-tables: $t(57) = 6.6, p < 10^{-7}$; post-tables: $t(57) = 4.5$, $p < 10^{-4}$) as well as for fixed words (Fig. 5b; pre-tables: $t(77) = 11$, $p < 10^{-15}$; post-tables: $t(77) = 7.4, p < 10^{-9}$). In the long term, the three conditions are again indistinguishable (all $p > 0.46$).

---

[6] Spaced repetition may be considered an orthogonal addition to Broccoli as evaluated in this user study. Sec. 7 provides evidence that everyday information diets are indeed amenable to spaced repetition.

Summary. In a nutshell, (1) all evaluated learning conditions (Broccoli and tables) outperform random guessing; (2) Broccoli outperforms both table conditions short-term; (3) we cannot reject the null hypothesis that Broccoli and the tables perform equally long-term.

6.2 RQ2: Impact of revealing translations

Recall that the Broccoli prototype tested in the user study allowed users to reveal the original word (before it was translated into the target language) by clicking on the translated word. We now investigate the impact of such clicks on word retention.

As explained, in the Broccoli condition each word appeared in 7 instances for each user, and each instance could be clicked zero or more times. Fig. 6a plots the distribution of clicks per word–user pair (the 7 instances of the same word are counted toward the same word–user pair). The probability of clicks per word is strictly decreasing, and in particular, one quarter of the words are never clicked by the average user, an indicator that even in a lab setting where users are aware that they will be tested on word retention, Broccoli's ideal usage mode of learning without explicit interaction comes naturally to users. We expect that in a real-world usage scenario, the fraction of words whose translation is never requested would be even higher.

Fig. 6b plots the easiness of words as a function of the number of clicks. Here we measure the easiness of a word as the fraction of users who learned the word in one of the table conditions and recalled it correctly in the short-term test (but clicks are counted in the Broccoli condition, of course). We hypothesized that easy-to-guess words would be clicked less frequently, and indeed this is confirmed by Fig. 6b, but the effect is subtle, with easiness decreasing only by 10 percentage points over the full range of clicks.

Ideally, Broccoli users would not need to see any "solutions" of correct translations in order to successfully retain vocabulary. We investigate if this ideal is feasible in practice by plotting the retention rate for words learned in the Broccoli condition as a function of the number of clicks the respective user made on them, for both the short-term (Fig. 6c) and long-term (Fig. 6d) test. We see that the retention rate does not systematically vary with the

(a) Distribution of clicks per word **(b) Word easiness vs. clicks** **(c) Results vs. clicks (short-term)** **(d) Results vs. clicks (long-term)**
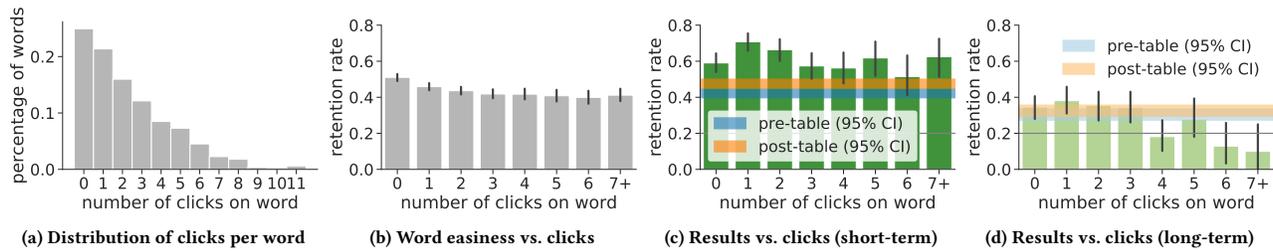
Figure 6: Impact of revealing translations by clicking on words in Broccoli condition: (a) Distribution of clicks per word, averaged over all users. (b) Word easiness (measured as word's retention rate in table conditions). (c) Retention rates in short-term multiple-choice test. (d) Retention rates in long-term multiple-choice test. Errors bars are 95% CIs; gray horizontal lines in (c) and (d): random guessing among 5 candidate translations.

number of clicks and that, in the short-term test (Fig. 6c), even words that were never clicked are retained by the respective users at a rate significantly higher than words learned in either table condition. In the long-term test (Fig. 6d), too, words that were seen in the Broccoli condition and clicked between 0 and 4 times are indistinguishable with respect to retention, and they are recalled at about the same rate as words learned in the table conditions. These are among the most important results of our study, as they establish that Broccoli can teach vocabulary at competitive rates purely via passive exposure in context, without any explicit user interaction.

6.3 RQ3: Broccoli-induced slowdown

We expect the fact that text in the Broccoli condition contains words in a foreign language to decrease users' reading speed. As described in Sec. 5, we designed the study to allow for quantifying this effect: Next to the article snippets augmented with translated words, the Broccoli section also included one page into which no translations were inserted. On this unprocessed page, we can estimate the user's natural reading speed. Fig. 7 summarizes the distribution over all users' reading speeds with and without Broccoli. Users' median reading speed is 199 words per minute without Broccoli, vs. 163 with Broccoli, implying a Broccoli-induced slowdown of 18%. Note that this does not necessarily imply a *conscious* slowdown, such as would be indicated by the usage of mnemonic strategies for memorizing words. Indeed, in Sec. 6.5 we will see that such strategies are applied extremely rarely in the Broccoli condition (we will also discuss the slowdown in Sec. 8).

6.4 RQ4: Impact of language model

Next, we investigate the effect of using a language model when selecting words for translation (cf. Sec. 4.2). The goal of the language model is to identify instances where the given word appears in a typical context, which we expected to lead to increased word retention. As seen in Fig. 8, this is not the case: words translated in contexts that scored highly according to the language model (green, left) have a retention rate indistinguishable from that of words translated in random contexts (gray, right). We further discuss potential causes of this effect, as well as the (positive) implications of the language model's weak role, in Sec. 8.
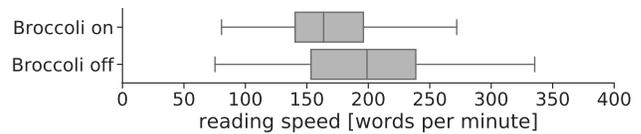

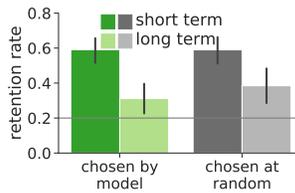**Figure 7: Broccoli-induced slowdown: Reading speed with vs. without Broccoli.**

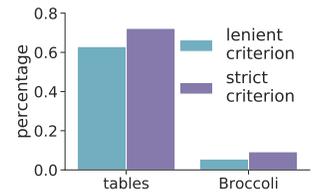**Figure 8: Impact of language model on retention rate.**

Figure 9: Mnemonic-strategy use in tables vs. Broccoli.

6.5 RQ5: Use of mnemonic strategies

The ideal followed by Broccoli is to avoid explicit cognitive efforts on the learner's behalf and enable vocabulary learning in a purely passive fashion. A common strategy in classical, memorization-based vocabulary learning (e.g., via flashcards) is to invent ad-hoc mnemonic devices, or "memory hooks", that facilitate memorization by establishing salient phonological, visual, semantic, etc., connections between the foreign word to be memorized and words in the learner's mother tongue.[7] While effective [8], such strategies require self-discipline and creativity, and come at a high cognitive cost. If foreign-language vocabulary learning could be achieved without them, it would become a more attractive activity for anyone—which is the very philosophy behind Broccoli.

We therefore conclude our study evaluation by analyzing the prevalence of mnemonic devices in the Broccoli and table conditions. The exit survey of the study included the two questions, "Did you use any tricks or strategies to remember translations (1) from the vocabulary tables? (2) when reading the Wikipedia articles?", asking

---

[7] One participant in our study reported this mnemonic device for Finnish *karhu* (*bear*): "*kar* sounds like *car, hu* is close to *hood.* I imagine a bear ripping off a car hood."

for a free-text response where participants could describe the tricks or strategies they had used in each condition. We manually coded the responses into two classes, according to whether they had or had not employed mnemonic strategies.

We expected a lower prevalence of mnemonic strategies in Broccoli, compared to the table conditions. Fig. 9 shows how strikingly this expectation bears out in practice: the prevalence of mnemonic strategies in table-based learning is at least 7 times as high as in Broccoli (namely, when applying a strict criterion for counting mnemonic strategies to Broccoli, and a lenient criterion to tables), at 63–72% for tables and 6–9% for Broccoli.

These results, together with Broccoli's competitive retention rates (Sec. 6.1), present a strong argument for the paradigm advocated in this paper: in-context vocabulary learning is possible with very little unnatural memorization effort and no loss in retention.

7 COMPATIBILITY OF BROCCOLI WITH CONVENTIONAL INFORMATION DIETS

In the above user study, the texts shown to users were compiled into a predetermined sequence specifically such that each translated word was seen by the user multiple times (7 times). In a real setting, the sequence of texts would be determined not by us, but by the user's natural information diet, and it would be the role of the tutor algorithm (Sec. 4.1) to make sure that words are revised (translated) in a rhythm that allows them to be effectively retained.

Broccoli can be expected to work well if the words that the tutor algorithm wants to translate at a specific time occur in the text that the user reads at that time. If all words were frequent (which is of course not the case) any word could be scheduled for repetition at any time. Hence, when attempting to assess how realistically Broccoli can be applied to natural information diets outside of a lab setting, the crux is to determine how badly the tutoring algorithm's natural rhythm is hampered by infrequently occurring words.

In the following, we first introduce our methodology for answering this question, assuming we are given a corpus that the user reads top to bottom. Then we explain how we generate corpora and discuss the results obtained on them (Sec. 7.1 and 7.2).

Lemma revisitation time. In order to quantify how much the tutor algorithm is constrained by infrequent lemmas,[8] we proceed as follows. Given a corpus and a lemma x, (1) count the number of tokens between all subsequent occurrences of x, (2) translate each token distance into a time difference by assuming a reading speed of 200 words per minute (cf. Fig. 7) and a reading time of 3 hours per day, and (3) define the **revisitation time** of x as the 90th percentile of the time differences for x. Finally, we aggregate lemma-specific revisitation times into a corpus-wide revisitation time via the 90th percentile over the set $\mathcal{L}$ of all lemmas that are to be taught. Intuitively, a corpus-wide revisitation time of t days implies that nearly all subsequent occurrences of nearly all lemmas appear within t days of one another. Revisitation times depend linearly on reading speed and daily reading time: for a user who reads half as fast as our assumed reading speed, or spends half as much time reading, revisitation times would double.

---

[8]We operate on lemmas rather than raw tokens because we do not want to distinguish different inflected forms of the same word. Punctuation and numbers are discarded before the analysis.

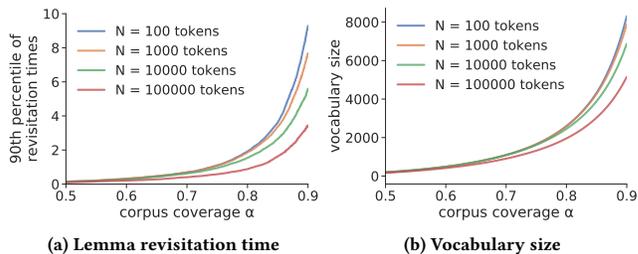(a) Lemma revisitation time (b) Vocabulary size

Figure 10: Compatibility of Broccoli with Wikipedia browsing: Growth of (a) lemma revisitation time and (b) vocabulary size as functions of corpus coverage α.

In our analysis, we work with various lemma sets $\mathcal{L}_\alpha$, where $\mathcal{L}_\alpha$ is defined as the smallest set of most frequent lemmas accounting for a fraction $\alpha \in [0, 1]$ of all tokens in the corpus. (We hence call α the **corpus coverage.**) The intuition is that after learning the translations of the lemmas $\mathcal{L}_\alpha$ in the target language, a user would understand a fraction α of a similar corpus in the target language.

Scenarios. In the following, we consider two common scenarios, Web browsing (Sec. 7.1) and fiction e-book reading (Sec. 7.2). We consider English as a source language, but given the universality of statistical laws of language [32], we anticipate the results to be similar for different source languages.

7.1 Web browsing

In the first scenario, the corpus consists of the concatenation of Web pages consumed by a user in temporal order. As real Web browsing histories are hard to obtain, and considering that Wikipedia is one of the world's most prominent websites, covers a wide variety of topics, and comes with publicly available statistics about its usage, we focus on the special case of (simulated) Wikipedia browsing.

Corpus generation. In order to generate Wikipedia browsing histories, we use the publicly available Clickstream data [31], which specifies the click frequency of every Wikipedia link (u, v), as well as the frequency of visits to u without an ensuing click. The Clickstream data allow us to generate synthetic yet realistic **browsing sessions** via biased random walks according to empirically estimated transition probabilities.[9] The starting page u_0 of a session is sampled with probability proportional to u_0's overall visit count. From there, we perform a biased random walk that, at any step, stops with the present page's no-click probability, in which case we jump back to u_0 and start a fresh walk from there. The goal of repeatedly restarting from u_0 is to generate topically coherent reading sessions containing an arbitrarily large number N of tokens. Once the concatenation of pages in a session surpasses N tokens, we terminate the session[10] and repeat the process (from a newly sampled u_0) until the concatenation of all sessions reaches 2 million tokens. The resulting sequence of 2 million tokens then serves as the corpus.

---

[9]It has been shown that first-order Markov chains, as defined by the Clickstream, are good models of Web browsing, including Wikipedia [23].
[10]When concatenating walks into sessions, we include u_0 only the first time.

By varying N, we can model users with varying breadth of interests (smaller N implies more frequent topic switching), which is useful as topical breadth might affect the applicability of Broccoli.

Results. The lemma revisitation time for simulated Wikipedia reading is shown in Fig. 10a for various combinations of corpus coverage α and session length N. As rarer words are included in the vocabulary to be learned (larger α), and as the reading history becomes more topically diverse (smaller N), revisitation times increase. Remarkably though, even in the most extreme setting ($\alpha = 0.9$, $N = 100$), nearly all lemmas (namely, 90% of them) are typically (namely, 90% of the time) encountered at least every 10 days. In less extreme settings (smaller α, larger N), revisitation times drop below 2 days. This regularity of occurrences implies that the vast majority of lemmas can be revised shortly after the tutor algorithm schedules them for revision. Additionally, the large vocabulary sizes (thousands of lemmas, Fig. 10b) imply that any text will presumably contain many lemmas that the tutoring algorithm currently considers worth revising. In practice, the tutor algorithm will be faced with an overabundance of word occurrences. In the extreme case of $\alpha = 0.9$ and $N = 100$, using the revisitation time of 10 days to cycle just once through the 8,000 unique words would require the user to review an average of 800 words each day. Together, these findings confirm the practical applicability of Broccoli for Wikipedia information diets.

7.2 Fiction e-book reading

Fiction books constitute a relevant application domain for Broccoli because here texts are topically coherent, which presumably facilitates learning in context. They are also well suited for evaluating the real-world feasibility of Broccoli, since the exact sequence in which words will be encountered is predetermined, such that real reading histories are readily available and need not be simulated.

Corpus generation. We use a collection of freely available English-language books from Project Gutenberg [14]. After filtering out books with fewer than 25k tokens (about 100 pages), we are left with 2,221 books from 134 authors (median length: 77k tokens; IQR [52k, 113k]). In our analysis, we fix the text coverage to $\alpha = 0.9$ and treat each book as a corpus. Fig. 11b contains a distribution over the books' vocabulary sizes at $\alpha = 0.9$.

Results. Fig. 11a summarizes the distribution over corpus-wide lemma revisitation times across the 2,221 books (median: 0.55 days; IQR [0.42, 0.70]), showing that in nearly all books, nearly all lemmas tend to reappear within less than a day. This implies that, whenever the tutor algorithm considers a word worthy of revision, it is likely to become a translation candidate for Broccoli that very day. This further corroborates the compatibility of Broccoli with everyday information diets.

8 DISCUSSION

Summary of results. While we considered Broccoli to be a potentially effective learning paradigm from the onset, our expectation prior to the user study was that classic, memorization-based vocabulary learning would vastly outperform Broccoli, due to the conscious and focused effort inherent in dedicated learning.

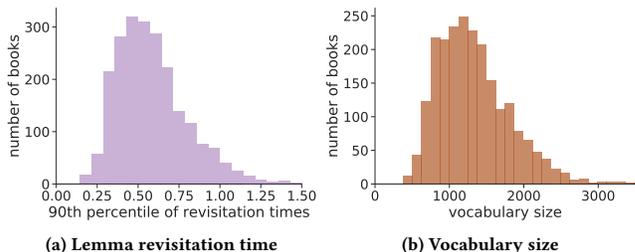**(a) Lemma revisitation time**   **(b) Vocabulary size**

Figure 11: Compatibility of Broccoli with fiction e-book reading: Distribution of (a) lemma revisitation time and (b) vocabulary size for 2,221 books. Each book treated as an individual corpus; analysis done for corpus coverage $\alpha = 0.9$.

Surprisingly, though, Broccoli led to retention rates competitive with memorization-based vocabulary learning via tables (Sec. 6.1), both in the short term (where Broccoli surpassed the tables) and in the long term (where the two conditions were indistinguishable). While Broccoli still incurs a slowdown in reading speed (Fig. 7), the cognitive work causing this slowdown—indicative of a learning process—seems to require substantially less conscious cognitive effort than table-based learning, as indicated by the pronounced lack of consciously used mnemonic strategies (Fig. 9), which were, on the contrary, heavily employed in table-based learning.

The surprising effectiveness of Broccoli might be due to the fact that incidental learning from context is a very natural mode of learning [21]. While contextual vocabulary learning is natural and effective, until now it has lacked optimized repetition intervals. Conversely, while flashcard-based vocabulary learning is able to freely use sophisticated models to schedule words, it lacks the natural mode of learning from context. Combining context and repetition has been shown to help learning in general [12], and we believe that Broccoli's effectiveness stems from applying this approach to vocabulary learning. Also, Broccoli minimizes the conscious cognitive load by piggybacking on the user's natural Web browsing or e-reading, activities in which they would engage regardless.

Role of language modeling. We observed the somewhat surprising result that replacing words in contexts for which the language model (Sec. 4.2) predicted a high guessability were *not* retained more easily than words translated in random contexts (Sec. 6.4). In our attempts to understand this observation, we hypothesized that the effect of easy contexts might be overpowered by the effect of explicit teaching. To test this hypothesis, we repeated the same analysis as in Sec. 6.4 for the subset of words whose explicit translation was never seen by the respective user (i.e., words on which the user never clicked), but the results remained qualitatively identical, thus refuting our first tentative explanation.

In the initial phase of the project, when exploring how to implement context-based word scoring (Sec. 4.2), we conducted a regression analysis in which the context-based human guessability of hidden words (for which we had collected a small set of ground-truth data) was modeled as a function of features of the hidden word itself and its context. Inspection of the fitted coefficients revealed that by far the most powerful predictors were the hidden word's

frequency in Wikipedia and in the present text, followed by its context-based probability according to the language model. Since all words used for learning in the user study were by construction frequent in Wikipedia and the input text (cf. Sec. 5), all words selected for translation in the user study could be expected to have a high guessability by default, such that the marginal contribution of the third most powerful feature—the language model's output—was only minor, which may in turn explain the ineffectiveness of the language model in our particular user study.

If the impact of the language model were to remain low even in real application settings, this would in fact have positive consequences: it would mean that even a computationally inexpensive, bare-bones implementation of Broccoli, along with a rudimentary replacement strategy, already confers a significant portion of the benefits. This could justify a deployment without a neural-network-based language model, thus allowing Broccoli to work on low-power devices such as cellphones and e-readers with little or no connectivity to cloud computing resources.

In long-form reading such as e-books, one may also leverage strong signals in addition to language models when picking words to translate: the predictability of the reading trajectory enables look-ahead strategies that prepare the reader in a targeted way for the words that will become important in her reading down the line.

Limitations of the user study. The evaluation of vocabulary retention only tested the direction of Finnish to English, corresponding to a passive knowledge of vocabulary, e.g., being able to understand the foreign language without necessarily being able to speak it. While this is well aligned with Broccoli's philosophy of learning vocabulary via passive exposure, future work should also evaluate retention in the opposite direction.

The user study was aimed at measuring the efficiency of Broccoli and making it comparable to traditional methods of vocabulary learning. This has led to the design of our survey with an identical number of words per condition (pre-table, Broccoli, post-table; cf. Sec. 5). A fixed number of words, together with constraints on the duration of the survey (we did not want participants to lose motivation before reaching the post-table) have forced us to condense the interactions with Broccoli. The text presented during the Broccoli section of the study was selected to allow for a high annotation density, with passive encounters of all words, in a limited time frame. Normally, when operating on a real-world information diet, we would likely annotate fewer words.

In this light, our study is not able to capture a realistic slowdown induced by Broccoli. Moreover, participants in our study knew that they would be tested on the vocabulary, which presumably gave them a motivation to focus on translated words and contributed to a higher slowdown recorded in the survey. At the same time, the results from our study emphasize that learning purely from context, without mnemonic devices and even without ever clicking on a word to reveal its translation, is an effective strategy when using Broccoli. Over time, users might learn to rely more on this passive strategy, which, together with the novelty aspect of Broccoli wearing off, might lead to a reduced slowdown.

In the user study, the tutor algorithm was not used when selecting word occurrences to translate, since spaced repetition works on much longer time scales than our study, which lasted only about

an hour. The effectiveness of spaced repetition has been well established [6, 10, 11, 16, 29, 30], and we are confident that Broccoli would profit from the additional influence of a tutor algorithm. Nevertheless this is an effect that cannot be quantified by the present study. Future work should therefore conduct long-term, longitudinal studies of the behavior of real-world Broccoli users (beyond students in a lab setting) and of the word retention rates they achieve.

A further minor limitation stems from the fact that the current prototype and the study presented here operate on the level of individual words. It would, however, be straightforward to extend the method to multi-word collocations and common expressions (e.g., *for instance*, *table tennis*, *birds of a feather*).

Extending the Broccoli paradigm. Broccoli's design could be evolved in two opposite directions, both toward less and toward more user interaction. In the direction of **less interaction,** one could remove the ability to click a word to reveal its translation altogether. A lack of such a "temptation", while occasionally inconvenient, could over time condition the user to ignore the Broccoli aspect of her reading diet to an ever higher degree. This would be even more aligned with Broccoli's core philosophy, without in general detracting from the learning efficiency, as indicated by the result that learning is effective even without clicking (Fig. 6c–d).

The only user interaction possible in the current prototype and study is clicking on a word to reveal its translation. In the direction of **more interaction,** the user interface could be extended to allow the user to provide additional feedback, such as whether a translated word was understandable or not. Alternatively, one might embed small quizzes into the interface, e.g., asking the user to guess the meaning of a translated word. Such feedback and gamification elements could keep the user engaged and provide data that could lead to better models of word guessability (language model, Sec. 4.2) and memory (tutor algorithm, Sec. 4.1). By keeping such interactions entirely optional, the core idea of Broccoli—learning vocabulary through passive exposure—would remain untainted.

A long-term, more aspirational goal of a Broccoli-derived approach would be an adjustable annotation density that would very gradually transform the text until culminating in a total conversion from the original language to the target language. At some point during this process, the translation direction would be flipped: e.g., rather than an originally English corpus being sprinkled with an increasing number of Finnish words, the corpus would at some point become a preeminently Finnish text with a decreasing number of English words. This would also convey an increasing amount of grammatical information, as sequences of adjacent translated words start accumulating.

Whether Broccoli can really be pushed to such a radical level is of course far from certain, but we hope that this paper opens the door to further explorations of creative solutions for sprinkling learning—of vocabulary and beyond—into everyday activities, guiding us systematically to "learn without intending to learn".

REFERENCES

[1] 2019. *Anki website.* Retrieved 2019-10-08 from https://apps.ankiweb.net/
[2] 2019. *Fastai text models website.* Retrieved 2019-10-08 from https://docs.fast.ai/text.models.html#AWD_LSTM
[3] 2019. *Microsoft Translator Text API.* Retrieved 2019-01-20 from https://www.microsoft.com/en-us/translator/business/translator-api/
[4] 2019. *Mnemosyne Project website.* Retrieved 2019-01-21 from https://mnemosyne-proj.org/
[5] Richard C Atkinson. 1972. Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology* 96, 1 (1972), 124–129.
[6] David P Ausubel and Mohamed Youssef. 1965. The effect of spaced repetition on meaningful retention. *The Journal of General Psychology* 73, 1 (1965), 147–150.
[7] Isabel L Beck, Margaret G McKeown, and Ellen S McCaslin. 1983. Vocabulary development: All contexts are not created equal. *The Elementary School Journal* 83, 3 (1983), 177–181.
[8] Francis S Bellezza. 1981. Mnemonic devices: Classification, characteristics, and criteria. *Review of Educational Research* 51, 2 (1981), 247–275.
[9] Douglas Carnine, Edward J Kameenui, and Gayle Coyle. 1984. Utilization of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly* 19, 2 (1984), 188–204.
[10] Frank N Dempster. 1989. Spacing effects and their implications for theory and practice. *Educational Psychology Review* 1, 4 (1989), 309–330.
[11] Robert Godwin-Jones. 2010. Emerging technologies from memory palaces to spacing algorithms: Approaches to second-language vocabulary learning. *Language, Learning & Technology* 14, 2 (2010), 4.
[12] Jessica S Horst. 2013. Context and repetition in word learning. *Frontiers in Psychology* 4, 149 (2013), 1–11.
[13] Melanie R Kuhn and Steven A Stahl. 1998. Teaching children to learn word meanings from context: A synthesis and some questions. *Journal of Literacy Research* 30, 1 (1998), 119–138.
[14] Shibamouli Lahiri. 2014. Complexity of word collocation networks: A Preliminary structural analysis. In *Proc. Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics.*
[15] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.*
[16] Arthur W Melton. 1970. The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior* 9, 5 (1970), 596–606.
[17] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182* (2017).
[18] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843* (2016).
[19] Bennet B Murdock Jr. 1962. The serial position effect of free recall. *Journal of Experimental Psychology* 64, 5 (1962), 482–488.
[20] William E Nagy, Richard C Anderson, and Patricia A Herman. 1987. Learning word meanings from context during normal reading. *American Educational Research Journal* 24, 2 (1987), 237–270.
[21] William E Nagy, Patricia A Herman, and Richard C Anderson. 1985. Learning words from context. *Reading Research Quarterly* 20, 2 (1985), 233–253.
[22] Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics.*
[23] Philipp Singer, Denis Helić, Behnam Taraghi, and Markus Strohmaier. 2014. Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *PLOS ONE* 9, 7 (2014), e102070.
[24] Robert J Sternberg. 1987. Most vocabulary is learned from context. *The Nature of Vocabulary Acquisition* (1987).
[25] Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gómez Rodríguez. 2019. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 3988–3993.
[26] Gesa S van den Broek, Atsuko Takashima, Eliane Segers, and Ludo Verhoeven. 2018. Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning* 68, 2 (2018), 546–585.
[27] Stuart Webb. 2008. The effects of context on incidental vocabulary learning. *Reading in a Foreign Language* 20, 2 (2008), 232–245.
[28] Piotr A Woźniak. 2007. SuperMemo 2004. *The Electronic Journal for English as a Second Language* 10, 4 (2007), 1–12.
[29] Piotr A Woźniak, Edward J Gorzelańczyk, and Janusz A Murakowski. 1995. Two components of long-term memory. *Acta Neurobiologiae Experimentalis* 55, 4 (1995), 301–305.
[30] Piotr A Woźniak, Edward J Gorzelańczyk, and Janusz A Murakowski. 2005. The two-component model of long-term memory: Symbolic formula describing the increase in memory stability for varying levels of recall. In *Proc. Cybernetic Modelling of Biological Systems.*
[31] Ellery Wulczyn and Dario Taraborelli. 2017. Wikipedia Clickstream. (February 2017). https://doi.org/10.6084/m9.figshare.1305770.v22
[32] George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press.