As We May Study: Towards the Web as a Personalized Language Textbook

Mircea F. Lungu, Luc van den Brand, Dan Chirtoaca, Martin Avagyan
Johann Bernoulli Institute, University of Groningen
The Netherlands

ABSTRACT

We present a system designed to enable learners of a foreign language to read materials that are personally interesting to them from the web and practice vocabulary with interactive exercises based on their past readings. We report on the results of deploying the system for one month with three classes of Dutch highschool students learning French. The students and their teacher were positive about the system and in particular about the personalization aspects that the system enables.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces

Author Keywords

language learning; personalization; education

INTRODUCTION

At any given moment, numerous people are learning new languages. English, estimates the British Council, will be learned by two billion people by 2020. Although a plethora of tools and techniques exist for beginners, few exist to support the intermediates and advanced learners. for whom one of the best possible activities is extensive reading [8, 3, 24].

However, when reading in a foreign language, most of the intermediate learners still require language textbooks. Such textbooks are designed by experts who make sure that the texts are simple enough for the desired language level of a broad audience and that exercises that allow the readers to practice newly learned concepts acompany the texts. In spite of becoming more colorful, being sold with complementary audio or video lessons, their main limitation remains unchanged since the last century: by being designed for any average learner they are not exciting for any individual learner [18]. A student interested in sports might not be motivated to read about *Maria who is a babysitter in Spain*. The lack of motivation, a well known problem, especially with young learners [18, 33], might be solved if learners could read materials that are personally interesting. If they would spend more time reading, their capabilities would increase, and enjoyment would result in further reading, in a virtuous circle [4, 16].

Given the vast amounts of multi-language content available and added daily on the Internet (e.g. blogs, news articles, eBooks) it is likely that every student can find materials that are *personally interesting* for them in almost any language they are learning. This would fit a general trend where old systems designed for the average user are being replaced with personalized attention across domains: in medicine[1], computer security, web design [32], or mathematical education [30].

Although individual components have been proposed before for free reading (e.g. browser extensions) and interactive exercises (e.g. Duolingo), a system that integrates free reading with personalized exercises into a *personalized textbook* has not been evaluated. The first contribution of this paper is describing the architecture and user experience of a system which combines:

1. **Reading comprehension support on both desktop and mobile devices for texts that are interesting to the user.** If ideal comprehension support should work *"without requiring even a single click"* [31], the next best thing is one-click (or one-touch) support. This is combined with a system for estimating text difficulty that allows the learner to choose articles that are within their capabilities[2].

2. **Integration between vocabulary practice and reading history.** Instead of manually adding words to an external vocabulary practice system, the most relevant unknown words encountered in the readings are automatically scheduled for practice, when possible in the context of the original text, since learning a word in context is more effective [26].

The second contribution is the evaluation of the usefulness and usability of such a system by deploying it with three classes of French learning highschool students for one month. We analyze their usage of the system, their feedback, and the feedback of their teacher to understand when and how such a system fits the modern language classroom.

[1]The nascent discipline of *personalized medicine* suggests that analysis of the genetic makeup of an individual may guide health care decisions far more precisely than big group studies do

[2]This to address the problem of a learner chosing an article randomly from a website and discovering that it is too difficult and giving up.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3173912

RELATED WORK

The domain of computer assisted language learning has a rich history of applied research that aims to improve the effectiveness and efficiency of language learning through helping both teachers and students [22]. In this discussion we focus on the aspects that differentiate our work from prior art.

The Web as A Source of Content

Multiple authors have observed before that the World Wide Web represents an enormous language database at the fingertips of the students [14, 19, 36, 38, 37].

Wible et al. introduce SRP – a tool that provides teachers and students with search capabilities for supplementary readings online [38]. The tool discovers similar texts to one given by the teacher to provide supplementary readings that offer repeated exposure to new vocabulary. Unlike ours, the system is presented without any user evaluation.

Streiter et al. [36] argued for a a system that would support browsing the Internet and a local document repository by dynamically annotating HTML and PDF documents with open dictionaries resources. A word is annotated with translations and pictures based on web search. However, the idea is not evaluated with users.

Trusty and Truong augmented the web in a learners native language with translations of a fixed set of words in the language that they are learning [37]. They show that in a two month deployment, 18 participants, learned in average 50 new words.

Besides these research efforts, many users use browser extensions to help them understand foreign texts. Díaz [12] did a study of how the users augment web browsers with extensions in order to "personalize on demand" their browsing experience. Based on millions of web users they saw that Google Translate was the 16th most used browser extension. Translate allows exporting words and translations but unlike our system, does not provide an API that would allow other applications to build on top of a learner's history. Moreover, this learner history does not capture the context in which a word was translated, context which makes learning more effective [26].

Kindle provides translations for individual words in a text. Just like Google Translate, these translations are not available to other applications (not even devices!), and they can not be made available to the teacher of a class, or to researchers.

There are also plenty of websites that provide texts for beginners together with translations (e.g. Veintemundos, DeutscheWelle, etc.). However, all these websites require the human editors to provide texts and also annotate words with translations, activities not needed in a system like ours.

Interacting With Foreign Language Texts

Augmenting foreign texts with annotations in the form of pop-ups and overlays has been found to benefit several aspects of language learning [10] and reading comprehension [35].

In one of the earliest such works, Nerbonne proposed Glosser – a system which would provide dictionary information about a given word including translation, part of speech, declinations, etc. [27]. In a study with 22 people they observed learners using the system for twenty minutes [13]. In their work, they focus on individual words and a limited number of predefined and pre-processed texts. In our work we observed a larger number of learners for a longer period of time.

Azab et al.proposed a system entitled SmartReader which provides interactive annotations of English words for the advanced foreign students who learn English [2]. Pop-ups are displayed above the selected word with information about it. The study introduces and describes the system, however it does not report anything about the way the system is used.

DeRidder [10] studied the behavior of students reading with hyperlinks. The results indicate that when reading a text with highlighted hyperlinks, readers are significantly more willing to consult the gloss. Sanko [35] showed that hypertextual input enhancement favourably affects vocabulary learning. In our case, the interactive reader component we developed, considers every word to be practically a hyperlink since every word in the text can be interacted with.

Vocabulary Practice

Dasgupta argues that in the context of interactive books, self-contained exercises to be included [7]. However, most of the vocabulary practice systems are disconnected from the readings of the learners. Most popular (usually commercial) systems such as Babbel, DuoLingo, RosettaStone, and Memrise are mainly focused on vocabulary drilling for beginners.

These systems employ various types of personalized scheduling for the vocabulary exercises but when it comes to the content, they either have predefined material or they require the learner to upload the vocabulary for study (e.g. Anki, Memrise).[3] The solution we propose adopts both personalized scheduling for the exercises and the automatic personalization of the content that is the result of retrieving the content from the readings of the learner.

One promissing vein of research in vocabulary practice has recently focused on discoveirng innovative opportunities for study in order to support the busy learners. In particular, micro-learning has been used in very creative ways: Dearman and Truong introduce a 'live wallpaper' interface always visible when a user uses the phone [11]; Cai introduces WaitChatter providing vocabulary exercises while the user awaits instant messaging responses [5]. The relationship between these micro-learning systems and our work is complementary: micro-learning exercises can be generated based on past readings of the learner as we showed with a smartwatch [29].

A MINIMUM VIABLE LANGUAGE TEXTBOOK

Our long term vision, of an ecosystem where various educational applications, created by different authors, interacting and sharing information in order to maximize the efficiency and enjoyment of the vocabulary improvement process is described in more detail elsewhere [23].

Figure 1 highlights two types of applications that are relevant for implementing a language textbook: the **interactive**

[3]The systems that have predefined content usually have a limited number of words: Babbel and DuoLingo offer 2000 to 3000 words per language

reader apps allow the learners to interact with texts in their preferred context (e.g. eBooks, News, Blogs), and the **vocabulary trainer apps** allow the readers to practice vocabulary exercises. The figure also presents several critical components of the ecosystem with which the applications interact: the learner model, the translation service, the content recommender, and vocabulary recommender. Before we convince other system creators to join such an ecosystem, we have decided to build a *minimum viable ecosystem* which includes basic implementations of the core components.

In this section we briefly describe the various back-end components, and in the next we describe the user interface of a unified, web-based reader and trainer app. The back-end services are implemented using Python. The front-end uses Javascript and HTML5. The source code for both is available online as open-source (the repo url is at the end of this paper).

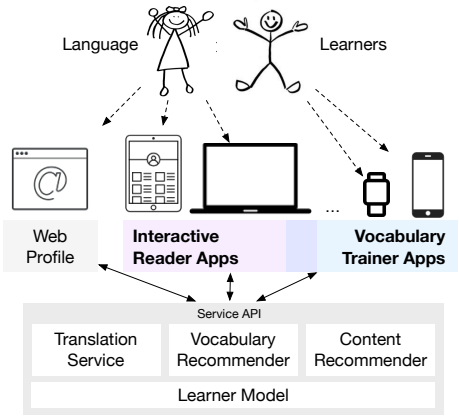

Figure 1. The architecture of the envisioned software ecosystem

Learner Model

At the core of the ecosystem a *learner model* tracks the evolving knowledge. Based on this model, algorithms can make recommendations for the individual applications regarding interesting content to read and appropriate vocabulary to study. The individual applications, in turn, report back to the learner model events from which the learner progress can be infered.

Currently, the model tracks the probabilities of a user knowing words based on interaction events in the apps: asking for a translation, repeatedly encountering a given word without asking for a translation, or answering an vocabulary exercise.

The Translation Service

The Translation Service is a subsystem implemented using Python which provides translations to all the applications in the ecosystem. Instead of implementing our own contextual translation engine, we rely on existing industrial grade translation APIs. To avoid depending on a single service and to also increase the likelihood that at least one of the alternative translations is the correct one, the translation service collects in parallel results from three third party translation APIs: Google Translate, Microsoft Translate, and Glosbe[4]. [9] In the next section we explain how the best guess is inserted in the text while the alternatives are available for the readers to consult.

The dependency of the translation service on multiple third party APIs allows for a higher reliability and a chance to guarantee a low response time: when a service is down or too slow to respond, the results from it are ignored. We detail elsewhere the strategies we use to keep response times low[9].

It is critical that the translation service be used by all the applications in the ecosystem since this allows the server to track the words and the context in which they are being looked up. This information is then used for estimating learner knowledge, for generating personalized recommendations, but also for allowing a teacher to gain insight into student activity.

Vocabulary Recommender

The goal of the vocabulary recommender is to program optimally-timed words to practice. To schedule the words to practice the system uses an adaptive, response-time-based scheduling algorithm aimed introduced by Mettler et al. [25].

The words scheduled for practice come from those that are translated by the readers. However, only a subset of words are actually scheduled for practice: those that are deemed *fit for study*. Not fit for study are words that can not be found in frequency lists of the language, expressions which are longer than three words, words whose translation is the same as the origin (e.g. digital(en) = digital(de)), and words whose context is too long.

Content Recommender

The content recommender aims to present the reader with texts that are both interesting and accessible at the same time. The current implementation requires the reader to select online sources (e.g. news, blogs) to be followed. The sources are constantly scanned for the latest articles and cached by using a custom-made library[5]. To add a new source, the teacher of the class (or the admin of the system) only has to add the url of the source.

The difficulty of a text is computed by aggregating the individual difficulty of its words. Individual word difficulty can vary from 0 to 10 and is computed in the following way:

- When the word is estimated to be known, its difficulty is considered to be zero. A word can be estimated to be known either based on past readings (i.e. encountered multiple times, but never looked up) or based on past vocabulary exercises (i.e. correctly identified in the most recent exercises)
- When the word is in the top 50K most frequent words in the target language, its difficulty is considered to increase with 0.1 for every 500 words; if the word is not in top 50K, its difficulty is ten.

With these strategies for computing word difficulty, the text difficulty is computed as the median of the words in the text. One limitation of this measure of difficulty is that it does not take into account the phrase length, as other measures do. [20]

[4]Google and Microsoft provide context-aware translations and multi-word translations. Glosbe is a simple dictionary

[5]Open sourced at: https://github.com/zeeguu-ecosystem/watchmen

A WEB-BASED READER AND TRAINER PLATFORM

In this section we present the user interface of the prototype *personalized language textbook* that we have built. It combines in a single responsive web application a reader applications and a vocabulary trainer with multiple exercise types, and thus, can be used from a variety of devices. In the user evaluation reported in this paper, it was used from Windows, Android, and iOS devices. Although not presented here, since it was not used in the user evaluation, a smartwatch application also exists as another vocabulary trainer [29].

Finding Personally Interesting and Accessible Texts

The current system allows the learners to subscribe to various online sources (i.e. news, blogs) and then monitors those sources for new texts. Figure 2 presents the source subscription dialog listing multiple text sources for French.

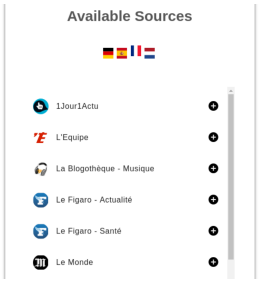

Figure 2. Different users subscribe to different sources

Once a reader is subscribed to a source, that source is constantly monitored for new articles, which are recommended to the learner in an article browser like the one in Figure 3. The browser displays for each article an icon representing its source, a title, a summary, and an estimated difficulty level. To visualize the reading difficulty of an article, there are three levels of information displayed: 1) a flag representing the language of the article since a learner could be actually registered to feeds in multiple languages; 2) a color coded difficulty from green to yellow to red, to allow the user to rapidly judge difficulty on an intuitive level; 3) a numerical difficulty score to allow a more quantitative judgment of the estimated difficulty.

Figure 3. Article browser presents estimated difficulty levels

Interacting With Unknown Words While Reading

To make reading as facile as possible, the reader is optimized for the most frequent action that a reader is likely to want to perform: translating a word. Thus, when a user clicks on a word, a translation is inserted right after the word, as Figure 4 illustrates[6]:

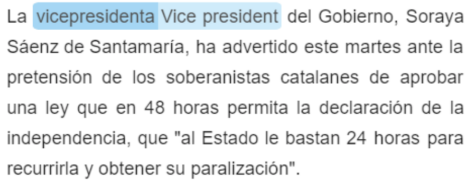

Figure 4. A translated word is inserted after the tapped word.

Two other alternatives that we explored and eventually dropped (for each had disadvantages) were:

1. Temporarily showing a popup of the translation and then hiding it again. This had a disadvantage for difficult sentences, where multiple words must be translated. The reader can forget translated words by the time they arrive at the end of an article, requiring them to re-translate.

2. Using the native selection mechanism to select text as opposed to click / touch. This had the disadvantage that native selection is not designed as a priority action and thus is slow to respond (e.g. on Android a user must hold their fingertip down for almost a second before the contextual menu is displayed).

Translating Multi-Word Expressions

The user can chain a few consecutive words into a single translation by simply tapping adjacent words which are then automatically merged in a translation bubble (Figure 5). This is useful for collocations and in cases where by expanding the translated set of words the precision of the translation increases.

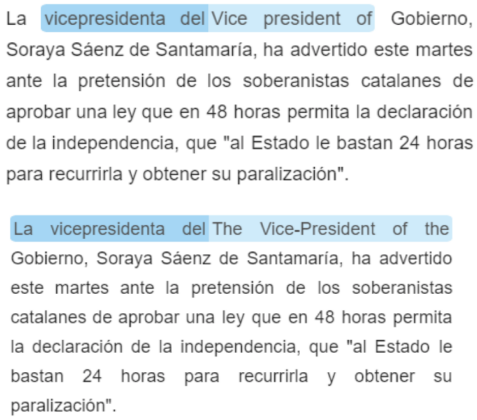

Figure 5. When adjacent words are tapped the translation bubble is extended accordingly

[6] A screencast is at https://vimeo.com/250152073

This minimalistic interaction model serves a double purpose - it enables and eases the translation of several chained words but it discourages users from translating entire sentences or phrases. This is in line with the recommendations of the literature (e.g. Renandya argues that extensive reading should discourage intensive use of translations[33]) but also because it reduces the amount of characters which are being translated by the learner (and thus the costs of the system, since some of the translation services have a per-character fee).

One of the limitations of this interaction is that it is not clear (at least at the moment) how to expand it for the situations in which expressions are present that are composed of words which are not adjacent (e.g. particle verbs in German).

Compensating for the Limits of Machine Translation

Due to the limitations of machine translation multiple translations might be possible in a given context. In such a case the system will insert the most likely alternative as described earlier right after the selected text, but it will allow the reader to discover alternatives. With a click on the translation, a drop-down menu appears in which alternatives are presented. Figure 6 shows that besides the predefined alternatives the learner can provide their own translation via an input box (the third line, "took place" is typed in by the learner in the figure).

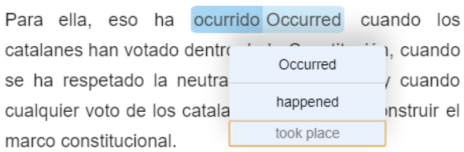

Figure 6. A translated word is inserted after the tapped word.

Discovering the Pronounciation of a Word

The process we followed while developing the reader was an iterative process, with short release cycles (one or two weeks), and frequent testing with members of the research team, and the occasional external user.

One of the features that we added following a suggestion of an early beta-tester – a teacher of Dutch as a foreign language – was the pronounciation of a translated word. After exploring several trade-offs between flexibility, ease of use, and a clean user interface, we settled on triggering the pronounciation of a word (or group of words) with a tap.

Practicing Personalized Vocabulary in Context

Given that the translation API captures the context together with every translation, exercises can be personalized for every user based on their past reading by using the original context in which the words have been encountered.

Figure 7 shows such a generated exercise which asks the reader to translate a given word in the context in which it was encountered in a past reading. The main interactive elements (IEs) that are specific to this exercise are an input box that allows the user to enter a solution (IE5); a button for checking the correctness of the input answer (IE2); a hint button which presents the correct answer (IE1). Two types of control that span exercise types are: a word pronunciation option (IE3) and a feedback option (IE4) which allows the user to provide feedback about the exercise.

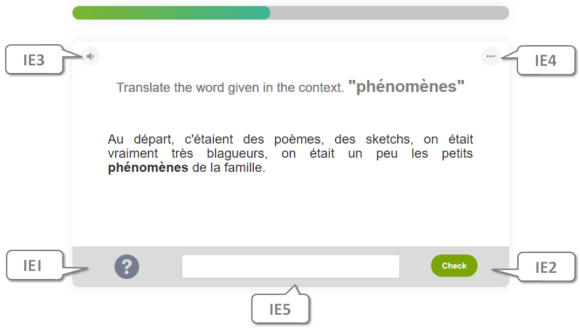

Figure 7. Translate exercises ask the user to translate a word in a given context (retrieved from the user's past readings)

The system currently implements three other types of vocabulary practice exercises[7], which can be split into two categories:

1. Free text input – where the text must be typed in the learned language (exercise type: Find)
2. Multiple choice – where the user is presented with a set of alternatives (exercise types: Choose, and Match).

Since a learner might encounter many words that are not understood, we need to prioritize those that are to be studied in exercises. We use two aspects to prioritize words:

1. Word Importance. The system prioritizes words based on the frequency with which they appear in the language.[8].
2. Context Quality. The system favors words that come with a context which is not too short but not too long.

TESTING THE SYSTEM WITH HIGH SCHOOL STUDENTS

We tested our system with sixty students from a public high-school in Netherlands, representing three classes that have the same French teacher and are bilingual in Dutch and English. All the students are below eighteen.

Usage Scenario. The teacher asked the students to use the system for **supplementary reading**. He encouraged them to read texts they found interesting and to build up their own *personalized portfolio of words*, complementary to the list of mandatory words that were common in every class.

For every half an hour of usage, the students had to write a brief report on how they spent their time and submit it to the teacher. The teacher could then decide to selectively test them on the basis of their reports. The teacher had used this strategy in the past with other software that he used in class.

Deployment. At the beginning of June 2017 we introduced the system and its usage to each of three classes. With few exceptions the students created an account and started using

[7]Detailed description of the exercise types are elsewhere[1]

[8]For word frequencies we use frequencies computed based on movie subtitles which have been shown to be highly representative to frequencies in human interacitons [28]

the system the latest on June 9th and until the end of the month, which coincided with the end of the study year. Students used personal computers and Android/iOS devices.

Before creating accounts on our platform, we asked the participants to answer a survey about their current level of knowledge, learning strategies, and reading interests. A handful of the participants, who were not in class when we presented the system, did not fill in the survey.

When asked whether they have favorite topics they would like to read about, half of the students mentioned various topics while the other half did not answer the question. From the topics that they mentioned as possible interests some of the more popular were: sports, music, travel, lifestyle, fashion, movies, and somebody mentioned as interest *"no politics"*.

In collaboration with the teacher we seeded the system with a variety of French news and blogs that cover the aforementioned aspects: 1Jour1Actu, L'Equipe, La Blogoteque, Le Figaro, Le Monde. Even if the source of readings was not actually the entire web, practically, having many dozens of news articles daily (only Le Figaro has usually more than forty in a day) offers sufficient opportunity for the free choice of individually interesting articles.

We deployed the system with the translations to English since, based on our experience translation APIs are of higher quality when one of the languages is English and because the students and their teacher were comfortable with the idea.[9]

We also invited the students to send us feedback at any time if they encounter problems or if they have ideas for improvement. Several of them did email. Towards the end of the month, we deployed several in-app focused pop-up questions using a customer opinion elicitation service called HotJar. After the month was over we sent out a follow-up questionnaire.

We also provided the teacher with a dashboard to see the activity of the students: the texts that were read and translations they requested. This chronological activity view is available also for the student who can solely see their own history.

Demographics. The participants that filled the survey were 54 female and 15 male with ages below 18. Based on their self characterization, 53 students are level B1 (i.e. can understand the main points of clear standard speech, can narrate an event or experience) and 16 are level A2 (i.e. can describe their surroundings and communicate immediate needs).

Usage. In average a student logged in on 2.83 different days[10](median 2.5). The most active student used the system in 8 different days while 19 students used it in a single day. The students interacted with 279 articles in total for an average of 5 and median of 3 articles each. The average number of words translated by the students is 71 and median 70. In total, during the entire duration of the study students solved 14,609 vocabulary exercises with a median of 85 exercises and a maximum of 2,865.

[9]We told the students to ask if they want their account switched to Dutch. None of the students requested this.

[10]We can not tally the actual logins, so if a user logs in multiple times in a day, we count that only once.

HOW IS READING PERSONALIZATION BEING USED?

Figure 8 represents an incidence matrix in which the columns represent students and the rows represent article sources. If a student is registered to a given source, the intersection of the respective row and column is a ◇. We would expect to see full horizontal rows of data-points if every user subscribed to a given feed, and full vertical rows if every user subscribed to all of the feeds available. The absence of such patterns proves that different individuals prefer to subscribe to different sources.

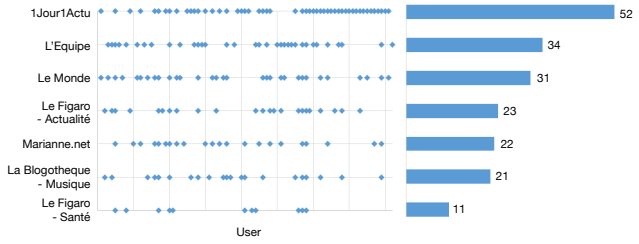

Figure 8. Different students subscribe to different sources

Projecting the data points onto the horizontal axis results in the histogram to the right of Figure 8 which shows that source popularity varies. To show that popularity is not related to the order in which they are presented in the subscription dialog, Figure 9 compares the popularity order with display order.

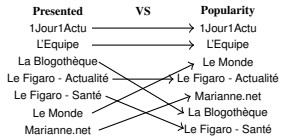

Figure 9. The popularity of the feeds vs. their ranking in the UI

Figure 10 shows an incidence matrix of users (columns) and articles that they interact with (rows). The distinct column patterns hint at the fact that each user explores their own interest: a few students read exclusively articles about sports (e.g. one student read twelve articles only about sports), one student reads exclusively articles about health; nobody reads on all the topics but the majority read on multiple ones.

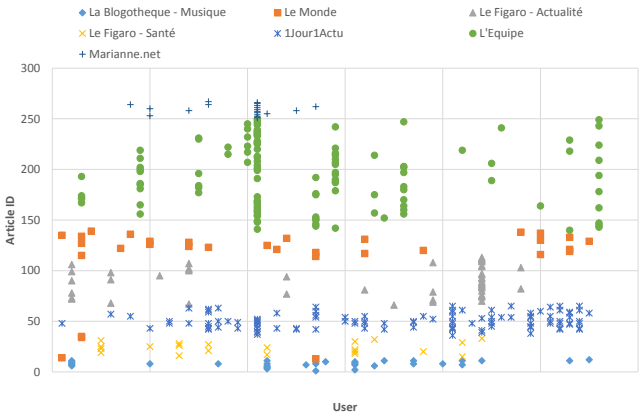

Figure 10. Each student (column) reads a different article mix

IS THERE AN IMPACT ON STUDENT VOCABULARY?

The value of extensive reading can be found, besides the new words that are learned, in the strengthening of the knowledge of the existing words, increased fluency, and increased grammar knowledge. Some of these benefits can be reliably measured only after a time longer than our deployment [33].

However, since our system combines free reading with vocabulary exercises and tracks all word interactions, by analyzing the learner interaction with the reader and the exercises[11], we can provide a glimpse into two measures of progress visible after one month of usage: increasing confidence about words and learning new words.

Figure 11. Word encounters in the Reader and Exercises

Figure 11 summarizes the interactions of the students with words in the reader and the exercises. The figure is based on the analysis of the user interaction database and shows that:

- From the 6,721 words that were looked up in the reader, 4,759 were used in the exercises platform. Since the learners requested a translation, they were either not sure about these words or did not know their meaning.
- More than 1,900 words were not practiced in the exercises. Some did not get their turn to be scheduled by the algorithm and others were not presented because the system deemed them not fit for study (cf. Vocabulary Recommender, p.3).
- For 3,626 words the learners were able to correctly identify them the latest in the last associated exercise. Out of these:
  - 3,261 words were recognized already for the first time in the exercises. These are **words likely to be strengthened by translating while reading**: the students were unsure when encountering them initially in the reader but eventually recognized their meaning when encountering them later in the exercises[12]. They represent 48% of all the words that the students translated in the reader.
  - 365 words were wrong during their first exercise interaction but were correct in the final one. These **words are likely to be learned via the exercises** by the students. [13]
- For 1,133 words the outcome of the final exercise that involved was not a correct answer. Thus we assume that they were **still not learned at the end of the experimental period**. Some of them might have been learned after the last exposure via the testing effect [34], but we can not be sure.

[11] A more detailed analysis can be found elsewhere [1]

[12] It could also be that the students learned them after the first encounter in the text, but we keep the more conservative hypothesis

[13] This number is conservative. We consider an answer to be correct only if it was right from the first attempt, without the use of a hint, and without typos.

HOW DO LEARNERS INTERACT WITH THE READER?

The reader interaction is more innovative and complex than the exercises. This is why we use telemetry to investigate how do learners use the features of the reader.

Telemetry has been successfully used for understanding user behavior in games [15] but also more generic contexts, such as automatically detecting personas from large scale interaction data [40]. In our study, we used telemetry to track the usage of various relevant features in the reader of the personalized textbook in order to better understand the usage of our system.

Based on logging every interaction of every user, Figure 12 (left) shows the six most used features of the system.[14] Figure 12 (right) shows the number of distinct users for each category of events. A larger number of distinct users indicates a feature that is more important to the students.

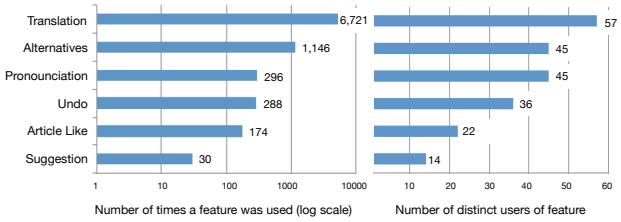

Figure 12. Popularity of features by their recorded usage-events (left) and number of users that use them at least once (right)

Requesting a translation is the most used interaction of the system and *showing translation alternatives* is the second most used one. The six-to-one ratio between the two features (6,721 to 1,146 as in Figure 12) is an indicator of the limitations of the automatic translation. The fact that they are both achievable with one click or touch proves to have been a good decision.

Pronuncing a word is he third most used interaction. On average, there are about 1.66 pronunciations for a given translation, suggesting that users are often asking for a second pronunciation after hearing it the first time.

Undo-ing a translation is used when the user wants to remove the last translation that was inserted in the text. For the proposed interaction mechanism this feature seems useful.

Liking an article that was just read by clicking the corresponding button at the bottom of an article happened 174 times. This information can be used in the future to improve article recommendations.

Suggestion of an alternative allows users to contribute their own translations when they are not satisfied with the one automatically provided by the system. This interaction is used seldom and by only a minority of users. It still is to be determined whether this is due to readers being satisfied with the automatic translations and their alternatives, or due to a low involvement. It might also be that more advanced readers would benefit more from this feature.

[14] An extended analysis that includes more features is elsewhere. [6]

HOW DO STUDENTS INTERACT WITH EXERCISES?

The system presented four types of vocabulary practice exercises to the students. In total, during the entire duration of the study we observed 18,082 attempts being submitted by the students in 14,609 exercises[15]. Figure 13 presents the number of answers which had a "correct" outcome (red) vs. exercises which had a "wrong" outcome (blue). The figure shows one student who submitted 2,865 answers during one month, and about six eager students who submitted about 700 answers each.

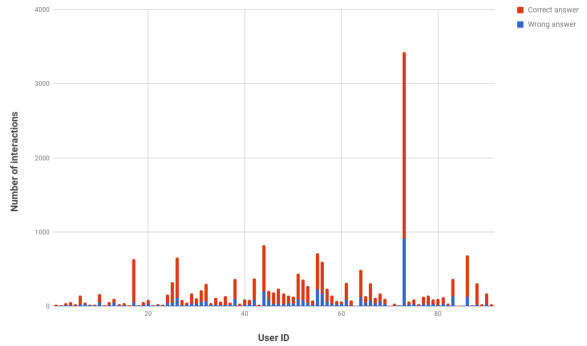

Figure 13. Correct (red) and wrong (blue) exercise outcomes per student

The figure does not include one other type of outcome, *requesting a hint*, which is presented in the table below grouped per exercise type. The corresponding number of hints suggests that the multiple-choice exercises (i.e. Match, Choose) are simpler than free text entry exercises (i.e. Find, Translate).

	Choose	Find	Translate	Match
Total attempts	7,180	6,249	2,643	2,010
Hint requests	29	529	847	16

Figure 14 shows the days when learners practice exercises. The x-axis has the days of June and the y-axis has the different user ids. The figure suggests that the students are doing exercises at their own pace over the observed period. The activity is rather sparse, with a more intensive period towards the end

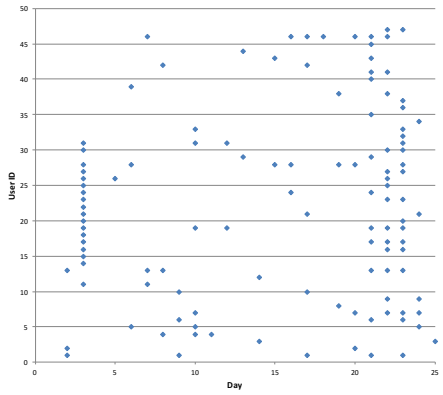

Figure 14. The students are doing exercises at their own pace throughout the one month interval

[15]Attempts include wrong submissions, and requests for hints

WHAT IS THE PERCEPTION OF THE LEARNERS?

About The Reader

After the semester was over, we sent an email asking the students to answer the survey about their experience and received 20 answers. Figure 15 summarizes the answers to the first question that asked the respondents to rate the ease of use (top) and usefulness (bottom) of the reader.

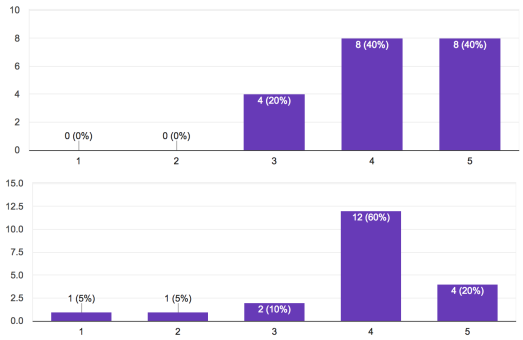

Figure 15. Feedback on Reader ease of use (top) and usefulness (bottom)

When asked what would make their experience with the Reader better many of the students thought that the system was good the way it was and few had some very specific features in mind (e.g. night reading mode[16]). There was however, one request which was expressed multiple times, even if in slightly diffferent ways by multiple respondents: the need for more specificity in the selection of materials to read. The students suggested: *"Order articles in different subjects like Animals, Politics, Fashion..."*,*"Better display of the articles and tags such as Gaming or News"*, *"Add a choice for different topics not only for the sources"*, *"Add a search engine"*.

When asked about what they dislike about the Reader, the majority of feedback was related to translations: two people complained about them being in English (*"The translations are always in English"*), five people complained about the translation quality (e.g. *"Some weak translations"*). The English translations are the reason for which one learner reported that they prefer the textbook: *"The translations are always in English. This is why I would grab a textbook first. I don't want to look up the (English to) Dutch translation."*

We also asked students how would they prefer reading texts in their foreign language. Figure 16 shows that the majority of the learners who answered our post-usage survey would prefer our system (Zeeguu Reader). However, some still prefer a textbook, probably for the reasons enumerated above.

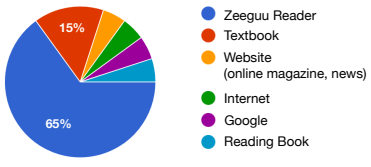

Figure 16. Answers to the question: "If you wanted to read something in the language you study, what would you reach out for first?"

[16]Complete feedback available in the GitHub repository of the paper.

About The Exercises

Figure 17 shows that when asked to provide their personal rating of the the quality of the exercises, the majority of the respondents are positive:

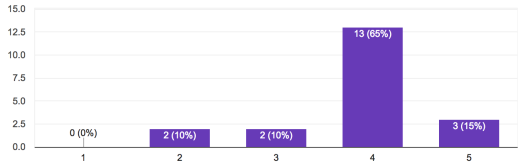

Figure 17. Students assessment of the generated exercises

When asked about what they dislike about exercises, many said literally "nothing". However, several also had concrete feedback that can be classified in two main directions:

1. Contexts are always the same: *"I would like to see the words I practice in a different context"*. This would indeed be an idea worthy of further investigation.
2. Exercises can be too easy. One learner wrote in their feedback that *"some exercises are too easy"*.
3. Exercises can be too difficult. Some learners encountered exercises in which they did not understand well the context, and would have wanted translations for it. However, translations are not enabled during exercises, so they reported that: *"There aren't translations"*, *"Doesn't give the translations"*.

About The Overall System

At the end of the semester, the students had to provide feedback to their teacher about all of the software tools they used in the class during the year. This is something that the teacher always does at the end of the school year. Since the students had to write also about our system, we asked the teacher if we could access the corresponding feedback and they were glad to oblige.

Six students wrote more detailed opinions. The main ideas in the feedback are illustrated by the three example quotes listed below[17]:

"It's good for improving reading skills. It would be even better if this tool was available in Dutch"

"Works well, but if it were possible to translate to Dutch it would have been better. Good that you can choose what you read."

"My vocabulary truly is improving, but you do have to use it more than a few times. A very nice website, easy to use and with nice topics"

Thus, learners appreciate the freedom of choosing materials to read that are personally interesting for them. They also appreciate the translations, but they would want to have them in their native language.

[17]Although the original feedback is in Dutch, we translated all the detailed answers and uploaded them online in the associated data repository of the paper as *feedback-to-the-teacher.txt*

THE PERCEPTION OF THE TEACHER

After the deployment at the school was finished we conducted a semi-structured interview with the language teacher of the three classes to gain insight into his perception of the benefits and limitations of the system. The teacher, who self-describes as *an experienced teacher and language scientist as well*, argues that:

- Such a system is critical for language education in schools, since the possibility of choosing their topics of interest is motivating for the students, and motivation is critical (Q_3)[18].
- The system should be used for students who had already two or three years of foreign language experience (Q_5).
- There is no danger that every student will develop his little individual vocabulary bubble. Once the students have a solid basic vocabulary, it is perfectly acceptable that they study the words which interest them. (Q_5)
- The used sources were maybe too general. One source which was very focused on sport news was found very interesting by several students, mostly boys. More highly specific sources for other topics could be good (Q_7).
- It is "more than acceptable" that the translations are not perfect and every now and then a student must look up an alternative translation. This might help students become more actively engaged with the texts (Q_9).
- The most important missing feature of the system is the possibility of comprehensively verifying that the students put quality and time in using the system at home (Q_2, Q_{11}).

The teacher decided to extend the use of the system during the academic year 2017–2018 with a larger group of students.

LESSONS LEARNED

Learners Appreciate Personalization But More Is Needed

Based on telemetry, we confirmed that the opportunity of reading personalized materials is used by the students. Also in their feedback, students appreciate highly the personalization of reading content. The teacher thinks that student motivation is increased due to the possibility of reading texts they like. However, one of the recurring themes of the student feedback is the need for a better way to find personally interesting articles. One possibility is suggested by the learners: the possibility of browsing articles by topics rather than sources. Another is a recommender system that takes as input the learner feedback on existing articles (e.g. the use of the Like button).

The vocabulary practice scheduler tries to optimize the times when the words are being repeated based on the state of the art in spaced repetition. However, we received multiple requests from learners who want to practice the words in a given text, once they are finished reading it, in the vein of traditional textbooks. The ideal system would allow the learners to personalize the vocabulary scheduling algorithms.

[18]The Q_n annotations refer to the questions in the full text of the interview, found in the data repository as: *teacher-interview.txt*

Improving Text Difficulty Reporting

One of the learners reported: *"My level of the language is quite low for now, so I clicked to get a translation very often. Too often."*. Since the feedback was anonymous, it is not clear whether this situation came about due to the limitations of the difficulty computation, the limitations of the user interface, or because the student was simply not as advanced as their colleagues. In any case, a more personal approach to difficulty computation and reporting seems to be needed in order to steer students away from articles which are too difficult for them.

Ensuring the Quality of Content

As opposed to a traditional textbook, a personalized textbook like the one we present has no editors and no quality control. In our study we limited the possible sources of articles together with the teacher. Even so, one of the students, wrote in their feedback: *"I would like to avoid articles which have information about accidents with human casualties"*. Ideally, this kind of personal preferences can be specified by the readers. One possibility would be integrating foreign language search, as Lappas and Vlachos propose [21]. Another is crowdsourcing where learners (and teachers, or more generally, trusted advanced learners) can provide feedback on existing materials. Crowdsourcing has been identified by Heffernan et al. as one of the driving technologies in learning [17].

Limitations of Automatically Generated Exercises

Although it is practical and effective to reuse the original context of a word in exercises[26], sometimes the context in which the learner looks up a word is too long, sometimes too short, sometimes too difficult, and sometimes too easy. Estimating the quality of the automatically extracted context of an exercise, and ensuring that it is in the *zone of proximal development* – where they are not too easy, and not too hard [39] – is a challenge for future builders of similar systems.

Insight Into Student Activity Is Important

One of the advantages of our (eco)system architecture in comparison to other alternatives for online reading (e.g. browser extensions for translations) is that it allows the teacher to gain insight into the reading activity of students. The deployed system has a teacher dashboard showing a chronological list of the words that a student looked up in context. In the final interview, the teacher observed that the biggest missing aspect of the system is a more complete insight into student activity, in particular the time students spend and the quality of their work. What other kinds of information are critical for teachers and how to collect and present them is an open question.

LIMITATIONS OF THE STUDY

Although results are promising, further studies are needed since there are multiple reasons for which these results might not extend to the broader population. The students might have been influenced by our enthusiastic presentation of the system at the beginning of the testing month. Also, the students we worked with are not necessarily representative for the Dutch highschool student population since they are bilingual. Also, the number of students who answered our survey was limited: only 20 students which represents only about 30% of the participants who actually used the system.

Student interaction with texts and exercises indicates that at the end of the month they have learned words that they did not know at the begining and strengthened words they were not sure of. However, it is not clear whether this knowledge will remain for the long term. Currently only once a student has correctly handled a word three times in a row in exercises the system considers it learned and removes it from the exercises (even if it should probably be verified once more much later).

We encountered an enthusiastic teacher who thinks that such a system is critical to his classroom. Although we think he is right, he might not be representative of the general teacher population so more studies with language teachers are needed.

AVAILABILITY OF THE SYSTEM, CODE AND DATA

The system described in this paper is deployed and available online. If the readers of this article want to test it they can use the *CHI2018* invite code while following the "Become a Betatester" link at https://zeeguu.unibe.ch/.

The source code is open under a MIT license and available online at https://github.com/zeeguu-ecosystem. The code is covered by tests and documentation. To replicate a study like the one presented in this paper with another population, a researcher can deploy their own version of the system.

Telemetry data representing all the interactions of the learenrs with the system, all student feedback, and a full transcript of the teacher interview are available on GitHub at: https://github.com/zeeguu-ecosystem/CHI18-Paper.

CONCLUSIONS AND FUTURE WORK

We presented a system aimed to be a minimal viable product for a personalized language textbook that uses the web as its content source. We deployed the system with sixty high school students for one month. Based on telemetry we see that students take advantage of the possibility of personalization by reading articles that are interesting and by practicing words in exercises generated from their past readings. Based on their interactions we can see that their vocabulary is enriched with new words and the knowledge of other words is strengthened.

In their feedback, the students appreciate the possibility of reading personally relevant texts and the ease of interaction with the texts that is provided by the Reader component of our system. However, they also want better ways to find personally relevant content. The teacher of the three classes thinks that such a system is critical for the modern classroom, but wants more detailed data about student activity within the system.

As future work we see two salient directions. First, more work with teachers is needed to better understand how to combine the individual focus of personalized textbook with the collective experience of the learners in a classroom. Second, improved content recommenders and difficulty estimators are needed in order to provide an even better personalization of content and thus increase learner interest and motivation.

Acknowledgements. We thank the anonymous reviewers for very valuable feedback. We thank Sara Vanini and Mark Langheinrich for advice, Wim Gombert and Jeroen van Engen for collaboration, and Oscar Nierstrasz for hosting the system on the SCG servers.

REFERENCES

1. Martin Avagyan. 2017. Building Blocks for Online Language Practice Platforms. (July 2017). Bachelor Thesis, University of Groningen.
2. Mahmoud Azab, Ahmed Salama, Kemal Oflazer, Hideki Shima, Jun Araki, and Teruko Mitamura. 2013. An NLP-based Reading Tool for Aiding Non-native English Readers. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. 41–48.
3. Timothy Bell. 1999. Extensive Reading: Why? and How? *The Internet TESL Journal* 4, 12 (1999), 1–6. http://iteslj.org/Articles/Bell-Reading.html
4. William G Brozo, Gerry Shiel, and Keith Topping. 2007. Engagement in reading: Lessons learned from three PISA countries. *Journal of Adolescent & Adult Literacy* 51, 4 (2007), 304–315.
5. Carrie J. Cai, Philip J. Guo, James R. Glass, and Robert C. Miller. 2015. Wait-Learning: Leveraging Wait Time for Second Language Education. In *Proceedings of the 33rd ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3701–3710. DOI: http://dx.doi.org/10.1145/2702123.2702267
6. Dan Chirtoaca. 2017. Apollo: Simplicity and Intuitiveness in a Personalized Multilingual Reading Tool. (July 2017). Bachelor Thesis, University of Groningen.
7. Sayamindu Dasgupta. 2010. Interactive Ebooks: Experiments on the OLPC XO-1 Book-reading System. In *International Conference on Designing for Children - With focus on Play + Learn*.
8. Richard R Day, Julian Bamford, Willy A Renandya, George M Jacobs, and Vivienne Wai-Sze Yu. 1998. Extensive reading in the second language classroom. *RELC Journal* 29, 2 (1998), 187–191.
9. Johan De Jager. 2017. A Self-Adaptive API Multiplexer. (Aug. 2017). Bachelor Thesis, University of Groningen.
10. Isabelle De Ridder. 2002. Visible or invisible links: Does the highlighting of hyperlinks affect incidental vocabulary learning, text comprehension, and the reading process? (2002).
11. David Dearman and Khai Truong. 2012. Evaluating the implicit acquisition of second language vocabulary using a live wallpaper. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1391–1400.
12. Oscar Díaz and Cristóbal Arellano. 2015. The augmented web: rationales, opportunities, and challenges on browser-side transcoding. *ACM Transactions on the Web (TWEB)* 9, 2 (2015), 8.
13. Duco Dokter, John Nerbonne, Lily Schurcks-Grozeva, and Petra Smit. 1998. Glosser-RuG: A User Study. In *Language Teaching and Language Technology*. 169–178.
14. Gregory L Friedman. 2008. Learner-created lexical databases using web-based source material. *ELT journal* 63, 2 (2008), 126–136.
15. André R. Gagné, Magy Seif El-Nasr, and Chris D. Shaw. 2011. A Deeper Look at the Use of Telemetry for Analysis of Player Behavior in RTS Games. In *Proceedings of the 10th International Conference on Entertainment Computing (ICEC'11)*. Springer-Verlag, Berlin, Heidelberg, 247–257. DOI: http://dx.doi.org/10.1007/978-3-642-24500-8_26
16. John T. Guthrie, Allan Wigfield, Jamie L. Metsala, and Kathleen E. Cox. 1999. Motivational and Cognitive Predictors of Text Comprehension and Reading Amount. *Scientific Studies of Reading* 3, 3 (1999), 231–256. DOI: http://dx.doi.org/10.1207/s1532799xssr0303_3
17. Neil T. Heffernan, Korinn S. Ostrow, Kim M. Kelly, Douglas Selent, Eric Van Inwegen, Xiaolu Xiong, and Joseph Jay Williams. 2016. The Future of Adaptive Learning: Does the Crowd Hold the Key? *I. J. Artificial Intelligence in Education* 26, 2 (2016), 615–644. DOI: http://dx.doi.org/10.1007/s40593-016-0094-z
18. Suzanne Hidi and Judith M Harackiewicz. 2000. Motivating the academically unmotivated: A critical issue for the 21st century. *Review of educational research* 70, 2 (2000), 151–179.
19. Yoko Hirata and Yoshihiro Hirata. 2007. Independent research project with web-derived corpora for language learning. *The jALT CALL Journal* 3, 3 (2007), 33–48.
20. J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report. Naval Technical Training Command Millington TN Research Branch.
21. Theodoros Lappas and Michail Vlachos. 2012. Customizing search results for non-native speakers. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1829–1833.
22. Mike Levy and Glenn Stockwell. 2013. *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.
23. Mircea F. Lungu. 2016. Bootstrapping an Ubiquitous Monitoring Ecosystem for Accelerating Vocabulary Acquisition. In *Proccedings of the 10th European Conference on Software Architecture Workshops (ECSAW '16)*. ACM, New York, NY, USA, Article 28, 4 pages. DOI:http://dx.doi.org/10.1145/2993412.3003389
24. C McCarthy. 1999. Reading theory as a microcosm of the four skills. *The Internet TESL Journal* 5, 5 (1999), 1–6.
25. Everett Mettler and Philip J. Kellman. 2014. Adaptive response-time-based category sequencing in perceptual learning. *Vision Research* 99 (2014), 111 – 123. DOI: http://dx.doi.org/10.1016/j.visres.2013.12.009 Perceptual Learning – Recent advances.

26. William E Nagy. 1995. *On the role of context in first-and second-language vocabulary learning*. Technical Report. University of Illinois at Urbana-Champaign, Center for the Study of Reading.
27. John Nerbonne and Duco Dokter. 1999. An intelligent word-based language learning assistant. *Traitement Automatique des Languages* 40, 1 (1999), 125–142. http://urd.let.rug.nl/nerbonne/papers/tal.pdf
28. Boris New, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied psycholinguistics* 28, 4 (2007), 661–677.
29. Rick Nienhuis and Niels Haan. 2016. Time to Learn – Learning With the Use of a Smartwatch. (Aug. 2016). Bachelor's Thesis, University of Groningen.
30. Oleksandr Polozov, Eleanor O'Rourke, Adam M. Smith, Luke Zettlemoyer, Sumit Gulwani, and Zoran Popović. 2015. *Personalized mathematical word problem generation*. Vol. 2015-January. International Joint Conferences on Artificial Intelligence, 381–388.
31. Gábor Prószéky. 2002. Comprehension Assistance Meets Machine Translation. *Tomaš Erjavec* (2002), 1–5.
32. Katharina Reinecke and Abraham Bernstein. 2013. Knowing what a user likes: A design science approach to interfaces that automatically adapt to culture. *MIS Quarterly* 37, 2 (2013), 427–453.
33. Willy A. Renandya. 2007. The Power of Extensive Reading. *RELC Journal* 38, 2 (2007), 133–149. DOI: http://dx.doi.org/10.1177/0033688207079578
34. Henry L Roediger and Andrew C Butler. 2011. The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences* 15, 1 (2011), 20–27.
35. Gyula Sankó. 2006. The effects of hypertextual input modification on L2 vocabulary acquisition and retention. *University of Pécs Roundtable 2006: Empirical Studies in English Applied Linguistics* (2006), 157.
36. Oliver Streiter, Judith Knapp, Leonhard Voltmer, and Daniel Zielinski. 2005. Browsers for autonomous and contextualized language learning: tools and theories. In *Information Technology: Research and Education, 2005. ITRE 2005. 3rd International Conference on*. IEEE, 343–347.
37. Andrew Trusty and Khai N. Truong. 2011. Augmenting the Web for Second Language Vocabulary Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 3179–3188. DOI: http://dx.doi.org/10.1145/1978942.1979414
38. David Wible, Chin-Hwa Kuo, Feng-yi Chien, and Nai Lung Taso. 2001. Automating repeated exposure to target vocabulary for second language learners. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*. IEEE, 127–128.
39. VK Zaretskii. 2009. The zone of proximal development: What Vygotsky did not have time to write. *Journal of Russian & East European Psychology* 47, 6 (2009), 70–93.
40. Xiang Zhang, Hans-Frederick Brown, and Anil Shankar. 2016. Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5350–5359. DOI: http://dx.doi.org/10.1145/2858036.2858523