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

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

UPDATED—December 27, 2017.

This paper describes the design, implementation, usage analysis, and evaluation of a "personalized language textbook" – a system that supports learners of a foreign language in reading materials that are personally interesting. The system does this by allowing them to read news and blogs, sourced from the Internet, in an interactive reader. The system provides translations for the unknown words at the touch of the screen while at the same time saving them in order to be able to monitor the current state of the knowledge of the learner and to later generate personalized exercises which are derived from past readings.

This paper reports 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 overall very positive about the system, and in particular about the study personalization aspects that it enables.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

Authors' choice; of terms; separated; by semicolons; include commas, within terms only; required.

INTRODUCTION

It is known that when learning a new language, free reading is one of the best ways of improving one's vocabulary. Reading something that is interesting to the learner will increase their desire to read and in turn it will increase the time spent reading. Additionally, the importance of reading is emphasized by the fact that it acts as a microcosm of all the other skills. [20]

However, to enjoy the benefits of free reading, the learner must already be sufficiently fluent in the target language: even knowing 95% of the vocabulary in a text, a learner still has to look up in average a word on every line. [16] This means that not any text is good for reading. Randomly choosing a text might actually end up being frustrating because the learner has to look up a lot of words in the dictionary.

Before fluency, learners use language textbooks as reading material. Textbooks are an artefact of the last century which still proves to be useful today. They are designed by experts who make sure that the texts that the readers are reading are simple enough for the desired language level and interesting enough for a broad audience. One of their advantages is that they have exercises which are based on the texts. Over the years, the textbooks have become more colorful, they even come with complementary audio lessons, but their static nature has not changed.

We believe that one of the main limitations of the textbook approach stems from the fact that textbooks are designed for the average learner and they are not exciting for any individual learner. As the US Air Force learned the hard way sixty years ago, "there is no average pilot": when cockpits, jumpsuits, and instructions were designed for the *average pilot*, the actual pilots had a hard time maneuvering the planes; performance improved only when the cockpit was designed in such a way as to be adjustable to the individual.

There are many other domains where the averages are being replaced with individualized attention: medicine[1], computer security , web design[27] , mathematical education[25]. In this paper we are focusing our attention on a subset of education - language learning - which in itself is a very broad issue with the potential to impact the lives of a very large number of people: the British Council estimates that by 2020 there will be 2 billion people learning English as a foreign language.

The fact that textbooks do not work that well is illustrated also by the fact that some of the teachers of foreign languages whom we have spoken to, do not even use a textbook anymore, but instead find articles that they deem interesting online and share them with their students. Note that this could be a step in the right direction, since the teacher has a better understanding of the interests of the class. However, in the end, the student is still not reading what they are passionate about, but rather, what the teacher deems relevant.

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

[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

This is sub-optimal, especially since given the vast amounts of multi-language content available on the Internet, it is very likely that every student could find materials that would be interesting for them. Blogs, News, eBooks exist for all the major languages. A student passionate about sports, might read with pleasure 10 articles on sports rather than one about "Maria, who is a babysitter in Spain"... So it would seem that, the best way towards allowing the learners to study what they are excited about, is to allow them to use the Internet as their textbook. There are nevertheless, three problems that prevent readers to read materials on the Internet:

1. The materials might be too difficult for them. The articles in the daily *Neue Züricher Zeitung* have a very high degree of variability in their textual difficulty. A learner that picks an article randomly might choose an article that is too difficult, and would eventually give up.

2. The existing reading tools might not be appropriate. The optimal comprehension support infrastructure would ideally work "without requiring even a single click from the user" argues Proszeky [26]. A single click / tap on a phone screen would still be simple enough, but any more complicated interaction (e.g. using an external translator) is likely to interrupt the flow of reading, and dramatically reduce the enjoyment of reading.

3. The texts available on the Internet do not come with exercises that would help a learner practice and retain newly learned words, and improve their own vocabulary. If a learner finds a word in a text, they have to manually add it to a vocabulary practice platform.

In order to address these three problems, we propose an infrastructure which allows learners to:

Exert agency over the materials they study by selecting sources and types of written content that are interesting for them

Access conveniently translations for unknown words in those cases when they are encountered, as it is unlikely that these words can be completely avoided.

Practice using personalized & contextual exercises that are generated automatically based on their past readings.

In the remainder of this paper we describe the design of such a system (Section 3) and we present our results from deploying the system for one month with a group of sixty Dutch high school French learning students (Sections 4 – 9). We then talk about the limitations of this study (Section 10) and then we list some of the challenges that we think our infrastructure and similar ones must face in order to increase the chance of their acceptance in practice (Section 11).

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 [18]. In this discussion we focus on several aspects that we combine in this work, which we argue, have not been combined together before.

Using the Web as A Source of Language Materials

Multiple authors have observed before that the World Wide Web represents an enormous language database at the fingertips of the students [12, 15]. Thus, the idea of augmenting texts with translations has been proposed before in various forms:

- One of the first occasions was in the work of Nerbonne [23] who proposed Glosser – system, which would provide dictionary information about a given word, including translation, part of speech, declinations, etc. In a follow up study with 22 people they observed users using the system for twenty minutes [11]. In their work, they focus on individual words. In our work we observed a larger number of learners for a longer period of time. In their study the words were previously annotated; our tools allow learners to translate sequences of words and not only individual words.

- Azab et al. [3] proposed a system entitled SmartReader which provides interactive annotations of English words for the advanced foreign students who learn English. Popups 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.

- [32] introduce SRP – a stand-alone tool that provides teachers and students with search capabilities for supplementary readings online. The tool exploits text retrieval techniques and is based on the hypothesis that there is a parallel between text similarity measurement on the one hand and the pedagogical task of providing supplementary readings which offer repeated exposure to new vocabulary.

- [30] argued fora 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. It argues for the creation of personal word lists and exdrcises. But there is no study of the system being reported.

Interactive Texts

Augmenting foreign texts with annotations in the form of popups, and overlays, has been found to be beneficial to several aspects of language learning [8] and improvements in reading comprehension [29].

- [8] 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

- [33] showed that multimedia gloss groups noticed and recognized significantly more of the target words than the control group

- [29] showed that hypertextual input enhancement favourably affects vocabulary learning

- [10] did a study of how the users augment the web browsers with extensions in order to "personalize on demant" their browsing experience. Based on millions of web users they saw that Google Translate was the 16th most used browser extension. Limitations of Translate...

- [17] - have already conducted small supervised experiment to evaluate effect of text augmentation of reading speed. The results show that augmented webpage slows reading speed down on average by approximately 7

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[31]. They show that in a two month deployment, 18 participants, learned in average 50 new words.

Vocabulary Practice Exercises

Dasgupta argues that in the context of interactive books, self-contained exercises to be included [6]. Also studies show that learning a word in context is more effective [22].

The number of systems that can provide vocabulary exercises to the learners is very large with several very popular commercial systems such as Babbel, DuoLingo, RosettaStone, Memrise, etc. Babbel and DuoLingo offer 2000 to 3000 words per language. Rosetta Stone claims that one can reach past B2 with their advanced course. Almost all these previous systems are targeted at the beginners and intermediates. The system we present here can be used by any learner, no matter how advanced.

Most of the systems personalize the exercises based on the learning progress of the users but not the content. Contentwise, these systems have predefined content, or no content at all (e.g. Anki, Memrise) and the user is supposed to manually punch in the content.

Recently, creative applications of vocabulary exercises:

- [9] Dearman and Truong propose a 'live wallpaper' interface that is always visible to the user when he is using his phone. They also present words in context.

- [4] WaitChatter presents vocabulary exercises while the user awaits IM responses

Personalization

The idea of a Personal eLearning Environment has been proposed by Attwell in 2007 [1] who assumes that it will take place in different contexts and situations and will not be provided by a single learning provider.

In web design Reinecke et al. propose culturally adaptive interfaces which are able to adapt their look and feel to suit visual preferences of a given population [27].

In mathematical education, Polozov et al. propose a technique for automatic generation of personalized word problems[25].

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 learning process is described in more detail elsewhere [19].

Figure 1 highlights two types of applications that are relevant for implementing a language textbook: **reader applications** allow the learners to interact with texts in their preferred context (e.g. eBooks, News, Blogs) and **vocabulary trainers** allow the readers to practice vocabulary exercises. It also presents three critical components of the ecosystem with which the applications interact: the learner model, translation service, and the study 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.

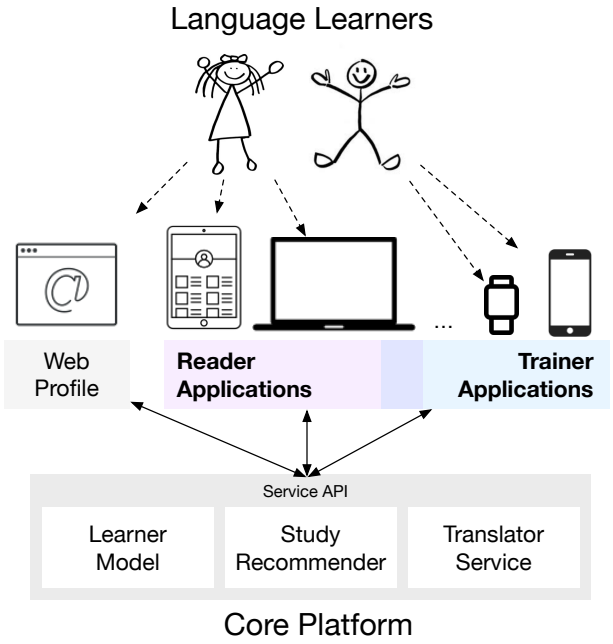

Figure 1. The architecture of the envisioned software ecosystem

Learner Model

The basic idea, sketched in Figure 1, is that at the core of the ecosystem a "learner model" tracks the evolving knowledge of the learner. Based on this it can make recommendations to the individual applications. The individual applications, in turn, report back to the learner model events from which the learner progress can be infered.

Translator Service

The core API of the ecosystem provides translations. It is critical that the translation service is 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 and for generating personalized recommendations.

The Translation Service is an API implemented using Python. Instead of implementing our own contextual translation engine, we decided to 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 dispatches in parallel requests to at least three third party translation APIs: Google Translate, Microsoft Translate, and Glosbe. [7] The first two provide contextual translations and multi-word translations, while the third is a simple dictionary.

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[7].

Study Recommender

There are two types of recommendations that an intelligent language textbook should make:

- Interesting and accessible texts to read. The texts should be both interesting and of an appropriate difficulty level to maintain the motivation. The current implementation of the text recommender allows the reader to select given online sources (e.g. news, blogs) which are interesting for them. The difficulty estimation of everytext is done by taking into account the frequency in the target language of all the words in the text.
- Optimally-timed words to practice. To schedule the words to practice the system uses an adaptive, response-time-based scheduling algorithm [was developed] to increase the efficiency of perceptual learning by Mettler et al. [21]. After evaluating several alternative scheduling strategies we settled on the Mettler one since it has been proven to have gains with both familiar, seen items as well as with new, unseen instances and the benefits of adaptive scheduling were present at an immediate test as well as at a delay [21].

A Web-Based Reader and Trainer Platform

The system that we present in this paper as a *personalized language textbook* is composed of instantiations of the components. We present them in turn in this section, focusing the discussion on three key activities that a user of such an interactive textbook is interested in:

1. finding texts to read
2. reading the found texts
3. practicing vocabularyÂăin context

The system presented is implemented as a responsive web application, and thus can be used from a variety of devices. We also have implemented a trainer for smartwatch devices. However, in the experiments reported in this paper, it was used from Windows, Android, and iOS devices.

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.

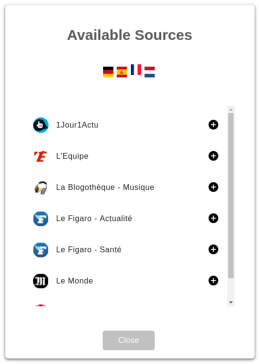

Figure 2. Different users subscribe to different sources

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 difficulty score from 0 to 5 to allow a more quantitative judgment of the estimated difficulty.

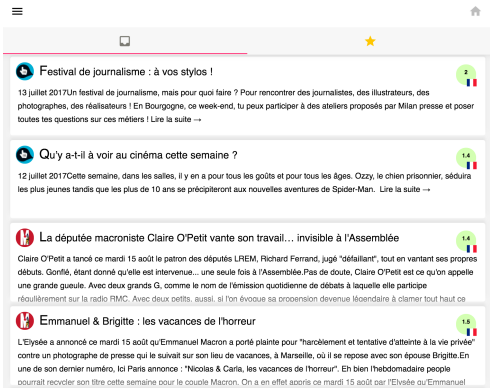

Figure 3. Article listing presents estimated difficulty levels

Difficulty estimation takes into account the frequency of the vocabulary items in the text. The system filters out articles that are too difficult given that it was deployed with non-advanced learners.[2]

[2]This is one of the limitations of the system that we discuss later.

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:

La vicepresidenta Vice president del Gobierno, Soraya Sáenz de Santamaría, ha advertido este martes ante la pretensión de los soberanistas catalanes de aprobar una ley que en 48 horas permita la declaración de la independencia, que "al Estado le bastan 24 horas para recurrirla y obtener su paralización".

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).

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.

La vicepresidenta del Vice president of Gobierno, Soraya Sáenz de Santamaría, ha advertido este martes ante la pretensión de los soberanistas catalanes de aprobar una ley que en 48 horas permita la declaración de la independencia, que "al Estado le bastan 24 horas para recurrirla y obtener su paralización".

La vicepresidenta del The Vice-President of the Gobierno, Soraya Sáenz de Santamaría, ha advertido este martes ante la pretensión de los soberanistas catalanes de aprobar una ley que en 48 horas permita la declaración de la independencia, que "al Estado le bastan 24 horas para recurrirla y obtener su paralización".

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

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[28]) 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 and Dutch).

Compensating for the Limits of Machine Translations

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).

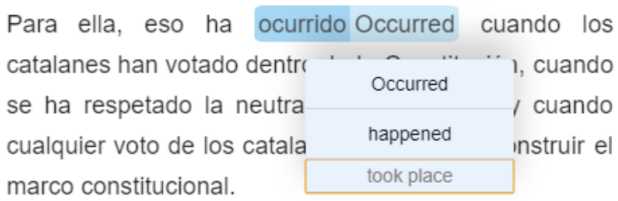

Figure 6. A translated word is inserted after the tapped word.

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 – 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 in a selection bubble) with a tap on them.

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 feebdack about the exercise.

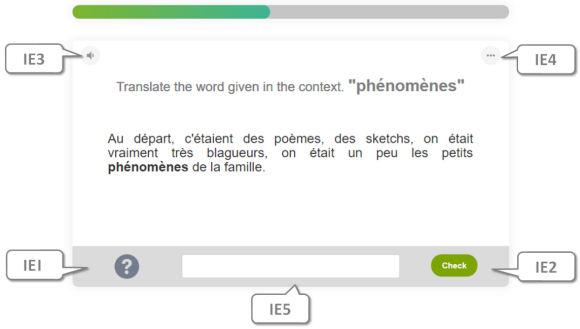

Figure 7. Translate exercises ask the user to translate a word in a given context (retrieved from the user's past readings)

The system currently implements three other types of vocabulary practice exercises[3], 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).

Selecting Words to Study

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.[4].
2. Context Quality. The system favors words that come with a context which is not too short but not too long.

[3]Detailed description of the exercise types are elsewhere[2]

[4]For word frequencies we use frequencies computed based on movie subtitles which have been shown to be highly representative to frequencies in human interacitons [24]

TESTING 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.

At the beginning of June 2017, we introduced the system and its usage in each of the three classes. With few exceptions the students created an account and started using the system the latest on June 9th. The system was to be used until the end of the month, which coincided with the end of the study year.

Usage Scenario

The teacher invited the students to use the system in order to perform **supplementary reading**, aside from the other activities that were done in the class. He encouraged them to read texts they found interesting and to build up their own *personalized portfolio of words* which would be complementary to the list of common mandatory words that each class had to study.

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. This was a requirement from the teacher and based on a strategy he used in the past with other software. This might have affected negatively the willingness of the students to spend time on the platform.

Deployment

Before creating accounts on our platform, the participants were directed to a survey form which asked them to provide information about their current level of knowledge, learning strategies, and interests. A handful of the participants did not fill the survey before using the system.

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"*.

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.

The choice of sources was done in collaboration with the teacher. 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.

Students used personal computers and Android/iOS phones.

We deployed the system with the translations from French to English instead of Dutch 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.[5]

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 also 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.

The Teacher View

The teacher can see the history of what his students have read and observe their chosen translations in context (as shown in Figure 8. This chronological and contextual view is available also for the student, but they evidently can see solely their own history.

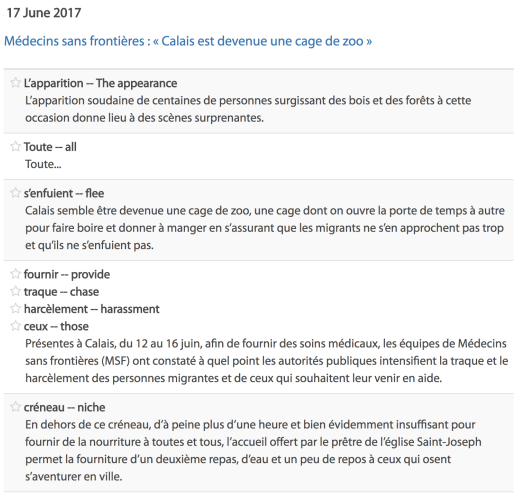

Figure 8. A teacher can see the log of the words that a student looked up, their chosen translations, and the corresponding contexts

Demographics

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

HOW DO STUDENTS USE THE POSSIBILITY OF READING PERSONALLY INTERESTING ARTICLES?

Figure 9 represents an incidence matrix collected at the end of the study interval: the columns represent students, and the rows represent article sources; if a student is registered to a given source, at the intersection of the corresponding row and column we place a ◇.

We would expect to see fully continuous horizontal rows of data-points if every user subscribed to the same feed, and fully continuous vertical rows if every user subscribed to all of the

[5]We made it clear to the students that they can ask us, and we will modify their personal account in such a way as to receive translations in Dutch. None of the students requested this.

feeds available. The fact that these patterns are largely absent in Figure 9 supports our assumption that different individuals prefer to subscribe to different reading sources.

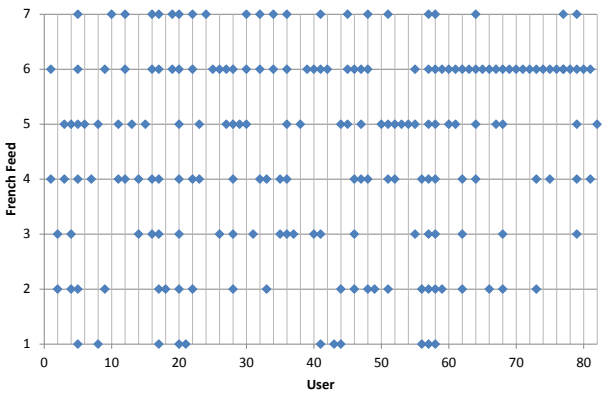

Figure 9. Different users subscribe to different sources

Of course some feeds are more popular than others. Projecting the data- points onto the horizontal axis and sorting the results leaves us with a histogram as can be seen in Figure 10.

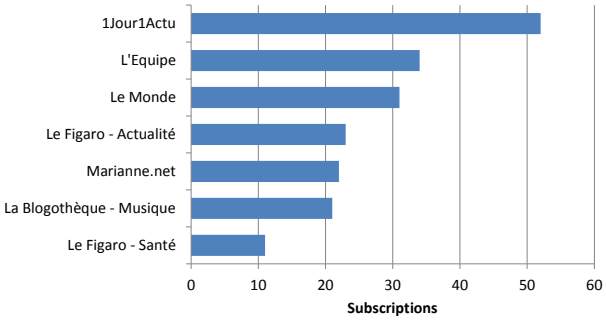

Figure 10. Some feeds are more popular than others

1Jour1Actu is the most popular article source and Le Figaro - Santé is the least. In order to see whether or not this might be related to how they are presented in the dialog window of our system (see Figure 2), in Figure 11 can compare the order of popularity with the order in which they are displayed. One can see how the second-to-last presented feed, Le Monde, is the second most popular feed by measure of subscriptions. Conversely, the feed listed above Le Monde is actually the least subscribed-to feed in our listing.

Article Interactions

Figure 12 shows that the articles that the users interact with present a similar pattern: each user explores their own interest, and there is no one article that is interesting for all. The vertical "line " in the figure represents an over-active reader.

Interacting with means that the user translated at least one word in that article. to check with Dan that this is the definition! we currently do not have information about whether a user read the article to the end.

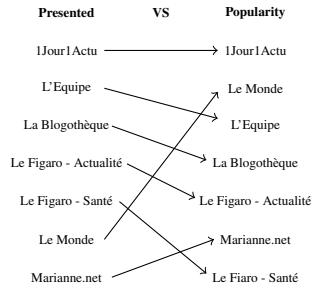

Figure 11. The popularity of the feeds vs. their ranking in the UI

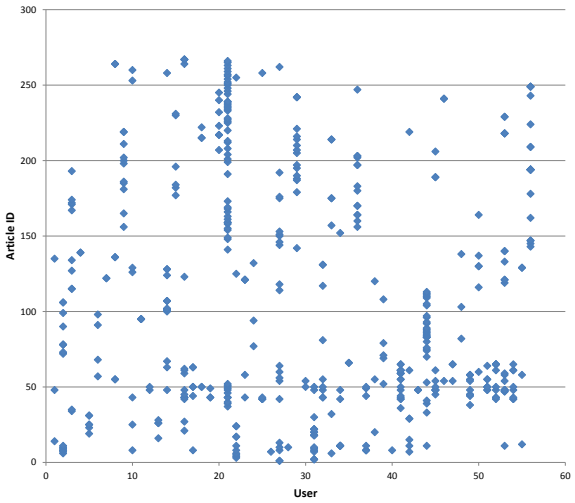

Figure 12. Every student has their own article reading preferences

After investigating some of the patterns of reading of students we observed that there are those who enjoy reding a variety of topics, but also those who like to read a single topic. From the latter category student #608 has read twelve articles exclusively on topics about sports in five different days and student #617 has read 6 articles exclysively about health topics over two distinct days.

HOW ARE THE READER FEATURES USED?

Given that the reader interaction more innovative and more complex than the one to be found in exercises, we decide to use telemetry to investigate how are the learners using these features. Telemetry has been successfully used for understanding user behavior in games [13] but also more generic contexts, such as automatically detecting personas from large scale interaction data [34]. In our study, we used telemetry to track the usage of various relevant features in the personalized textbook in order to better understand the usage of our system.

Based on logging every interaction of every user, Figure 13 shows the six most used features of the system.[6]

With 6.700 occurrences, *requesting a translation for a word* is the most used interactive feature of the system. The second

[6]An extended analysis that includes more features is presented elsewhere. [5]

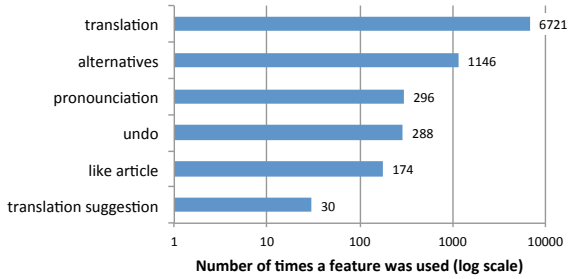

Figure 13. Popularity of features by their recorded usage-events

most used feature is *opening the translation alternatives* drop-down list. The six-to-one ration between the two features is an indicator of the limitations of the automatic translation – it seems that one in six translations are not satisfying to the learners.

The third most used feature is *pronunciation*. 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 to be useful.

A *like* button found at the bottom of an article was clicked by the readers 174 times. Although not used at the moment, this information can be used in the future to improve article recommendations.

The least used feature presented in the Figure 13, *translation suggestion*, allows users to contribute their own translations when they are not satisfied with the one automatically provided by the system.

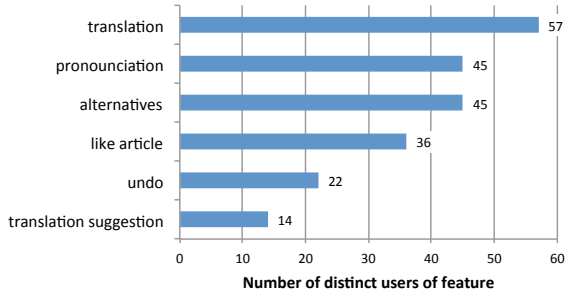

Figure 14. The usage of the various reader features by the various users

Figure 14 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. We see that:

- Not all the users of the system use translations
- *Translation suggestion* is used by very few users. It still is to be determined whether this is due to readers overwhelmingly being satisfied with the automatic translations and their alternatives, or due to a low involvement.

HOW DO STUDENTS USE THE PERSONALIZED VOCABULARY 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 exercises being presented to the students. Figure 15 presents the number of exercises which had a "correct" outcome (red) vs. exercises which had a "wrong" outcome (blue). The figure shows one student who did 3.000 (!) exercises during one month, and about six eager students who did more than 700 exercises.

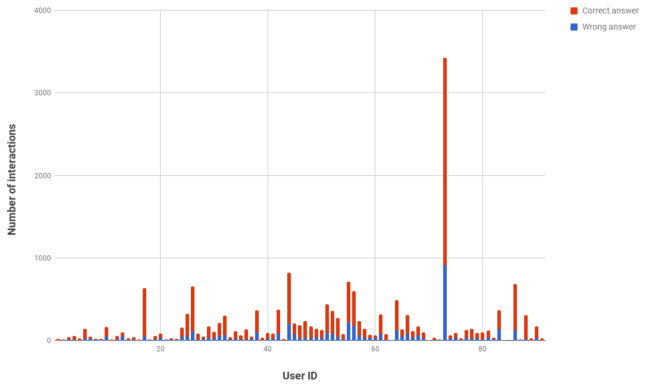

Figure 15. Correct (red) and wrong (blue) exercise outcomes per student

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 interactions	7180	6249	2643	2010
Hint requests	29	529	847	16

Figure 16 shows the days when learners practice exercises. It suggests constant activity over the entire period of the study with a more intensive period towards the end.

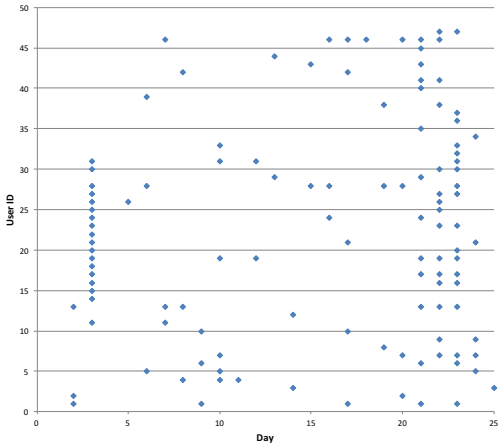

Figure 16. The students are doing exercises at their own pace throughout the one month interval

DO STUDENTS IMPROVE THEIR VOCABULARY?

The value of extensive reading and vocabulary practice can be found, besides the new words that are learned, also in the strengthening of the knowledge of the existing words. This is hard to quantify, but by analyzing the learner interaction with the exercises we can provide a glimpse into it. Looking at the outcomes of all the exercises done by the students we observe that:

- In total 5149 words were used in the exercises platform during the learning period. They are words for which the learners requested a translation before. Therefore the learners either did not know them or at least were unsure about them.
- For 80% (4110) of the words the learners were able to correctly identify the meaning in the last associated exercise. Out of these:
  - 14% were wrong during their first interaction in exercises but were correct in the final iteration of the exercises. These are **likely to be learned via the exercises**.
  - 66% were recognized already for the first time in the exercises. These are **likely to be words for which the knowledge was strengthened by using the system**: the students were unsure when encountering them initially in texts but eventually knew their meaning when encountering them later in the exercises.
- For 20% (1037) of the words the final outcomes showed incorrect answers, thus we can assume that they remained **unlearned at the end of the experimental period**.

WHAT IS THE PERCEPTION OF THE LEARNERS?

Post-Usage Survey

After the semester was over, we sent an email to the students, asking them to answer several questions about their experience. Our survey was answered by 20 students in total. Figure 17 shows that most of the respondents found the system easy to use and useful.

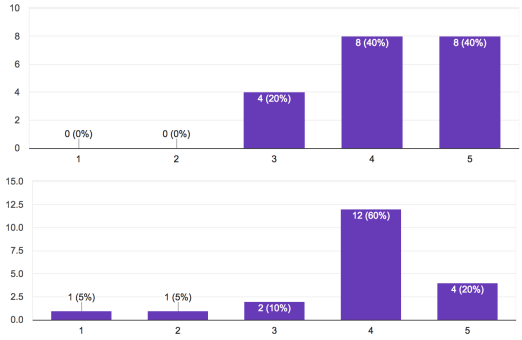

Figure 17. Students assess ease of use and the usefulness of the reader

We asked the participants what would make their experience better. Many thought that the system was good the way it was, and a few others had some more specific answers (e.g. night reading mode, etc.). [7]. The most requiring requests were:

- Choice of Topics: *"Order in different subjects like animals, politics, fashion..."*, *"Add a choice for different topics not only for the sources"*, *"Better display of the articles and tags such as 'gaming' or 'news"'*, etc.
- More freedom for chosing materials: *"Would be nice to be able to add website to the list"* and *"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."*

Figure 18 shows that when asked to provide their personal rating of the the quality of the exercises, the majority of the respondents expresses their appreciation.

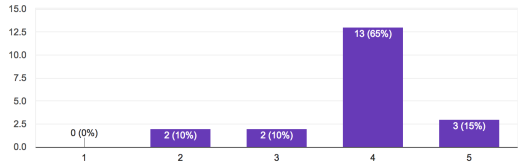

Figure 18. Students assessment of the generated exercises

[7]The complete feedback is available in the GitHub repository of the paper.

When asked about what they dislike about exercises, many said literally "nothing". However, several also had concrete feedback:

- During exercises, translations are disabled, so if one does not understand a word from the context, they will have a difficulty. *"There aren't translations"*, *"Doesn't give the translations"*
- Difficulty is not always appropriate: *"Some exercises are too easy"*
- Contexts are always the same: *"I would like to see the words I practice in a different context"*

Finally, we asked whether our learners would prefer our system to a textbook if offered the choice. We thus asked them what would they choose between the our system (Zeeguu Reader in the figure) and a textbook. We also gave them the possibility to answer something else with a free-form text field. Figure 19 shows that the majority of the learners who answered our post-usage survey would prefer our system. However, some still prefer a textbook.

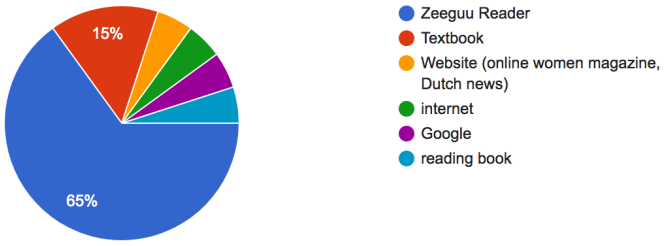

Figure 19. Answers to the question: *"If you wanted to read something in the language you study, what would you reach out for first?"*

In-App Feedback

Besides the analysis based on the observed telemetry data, we also asked the students a series of questions by popping up questions while they were using the system. To do this, we used an online tool called HotJar. Among the questions was whether they preferred the reading platform and why. Some of the answers can be seen in the screenshot below. It becomes clear that the students appreciate the possibility of reading what is interesting for them.

	OS	DATE	Do you prefer this reading platform (not considering the exercise platform) to y ...
		5th July	The textbook materials are repetitive and boring
		3rd July	I can choose the topic of the texts I read
		2nd July	Yes, it's easier to find translations of difficult words. Also, you can choose what article you want to read.

Figure 20. The students appreciate the freedom of choosing what to read

Reports to the Teacher

To triangulate the answers that they provided to us, we report here also what the students wrote in a separate evaluation of the tools they use in the class, which the teacher runs always at the end of the school year. Several of the answers are:

"It works well but it were better if the translation would have been in Dutch. It is good that you can choose what to read."

"Good for reading skills. Would have been best if it were in Dutch."

"Your vocabulary is really moving forward, but then you have to do it more often than a few times. Overall a nice website, easy and fun subjects"

The main message in the feedback is that learners appreciate the freedom of choosing materials to read that are personally interesting for them. The second implied observation is that they appreciate the translations, but they would want to have them in their native language.

The Evaluation of the Teacher

After the deployment with the school children was finished, we conducted a semi-structured interview with the teacher of the class to gain insight into his perception of the usefulness, benefits and limitations of the system. Overall, the conclusion is that the integration of the system in the three classes worked well. Some of the other ideas that stem from the interview are:

- Such a system is critical for language education in school, since the choice of topics is motivating for the students (Q_3)[8].
- The system should only be used for students who had already two or three years of prior foreign language experience (Q_5).
- The sources that were used were mostly general. There was only one source with sports and the number of students mostly boys found this to be very interesting. Maybe more sources for other specific subjects would be good (Q_7).
- There is no danger that every student will develop his little individual vocabulary bubble. The teacher believes that once the students have a solid basic vocabulary, it is perfectly acceptable that they study the words which interest them. Moreover, he believes that there might still be an overlap between the words studied by different learners due to the skewed distribution of word frequency (Q_5)
- The fact that the translations are not perfect, and every now and then a student must find an alternative translation is "more than acceptable". This might have the students be more actively engaged with the texts that they are reading (Q_9).

The teacher of the class appreciated the system, and decided to introduce it in the entire new academic year with a larger group of students.

[8]The Q_n annotations refer to the questions in the full text of the interview which is available online at: https://github.com/zeeguu-ecosystem/CHI18-Paper/blob/master/data/teacher-interview.txt

LIMITATIONS OF THIS STUDY

The feedback from the users was overall positive, with many of them showing appreciation for the personalization aspects of the system. However, more studies would be welcome 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 number of students who answered our survey was limited: only 20 students which represents less than 50% of the participants who actually used the system.

We showed that the users are using the system extensively. However, this might be because the students were encouraged to use the system as part of their assignment in the class. We showed that the majority of the students used the system constantly throughout the one month period. If they used it only for the final grade, we would have expected a more focused cramming at the end of the period (which we actually saw with few of the students...).

The students we worked with are not necessarily representative for the Dutch highschool student population since they are bilingual. Even in this case, during the feedback multiple of them remarked that they would prefer to use the system in their native Dutch as opposed to English.

The algorithms for scheduling vocabulary exercises are the state of the art in spaced repetition. However, we did not have a control group to see whether this approach works better than others. Moreover, note that other approaches for using spaced repetition already exist; what is unique in our approach is that the students learn based on personalized exercises generated based on the context of their past readings.

CHALLENGES

In this section we explore some of the challenges that we perceive need to be addressed by our system and similar "personal textbook" systems. We base our list of challenges on our observations and on the feedback that we received from our learners. The full list of recommendations from our users can be found in the GitHub repository online.

Registering for "topics" instead of "sources"

Multiple learners asked for the possibility of registering to article topics rather than "article sources". A future system should consider this.

Ensuring the appropriateness of articles

The advantage of a textbook is the fact that the quality control of the texts in it is guaranteed by the editors. How are we going to ensure the quality of the texts that re to be found online? What we did was to limit the possible sources from where the students can read. Nevertheless, one of the students, wrote in the feedback form *"I would like to avoid articles which have information about accidents with human casualties"*. One thing that we plan to investigate in the future is crowdsourcing and "teachersourcing" 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 of the upcoming adaptive learning [14].

Scheduling vocabulary exercises

We have implemented the vocabulary practice scheduler in such a way that it 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 the users who are asking for the possibility of rehearsing the words in a given text, once they are finished with its study, more in the vein of traditional textbooks. It might be that in the future it would be useful to allow the learners to influence the scheduling algorithms.

Evaluating the quality of examples

It is indeed desirable to find good examples of practice exercises from past readings. Sometimes, the context in which the learner looks up a word is too long and sometimes it is too short. How to estimate the quality of an exercise? One measure that we are considering is: ensuring that all the words in the context are simpler than the tested word.

Estimating article difficulty in a personal way

The way we did difficulty ranking was sub-optimal. Most of the difficulties were very close to each other in value, between 1 and 4 and they were generic instead of being personalized. We think that as a result, the difficulty estimations that the system presented were too abstract for the readers. And as a result, one of the readers reported that he disliked about the Reader the fact that *"My level of the language is quite low for now, so I clicked to get a translation very often. Too often."*

A more advanced strategy is needed, one that is less abstract, and personalized for the individual learner. One approach would be to estimate the number of words that are likely to

be unknown in an article for a particular learner. A complementary information about the article could be the number of words that we know are being learned at the moment which are to be found in that article. In this way, a learner can choose a text that also gives them the chance to reëncounter words being learned.

The teacher perspective

The system we presented here has a (limited) teacher dashboard for those users who have a teacher. The dashboard currently shows a chronological log of the words that the student has looked up in the context. However, the system could present more advanced analytics that could enhance the teacher's understanding of the class. This is something that is a clear opportunity when moving to a digital textbook.

Investigating more possible classroom workflows

The system was initially designed for self study. However, when invited to test it in a formal classroom we were happy to oblige. We plan to work more with teachers to better understand how to combine the individuality of the system with the shared experience of the learners in a classroom. Indeed, new workflows and classroom activities must be discovered.

CONCLUSION AND FUTURE WORK

We have presented here a system that we aimed to be a minimal viable product for a personalized language textbook. We report on using the system with high school students for about one month. We observe that overall the students make use of the personalization features, and when asked about it, they appreciated them. The teacher of the class also appreciated the system, and decided to introduce it in the new academic year with a larger group of students.

AVAILABILITY OF 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 this paper with another population, one can deploy their own version.

The anonymized telemetry data, representing the interactions of more than sixty learners with the system for one month, is available as a MySQL database dump on GitHub at the following link: https://github.com/zeeguu-ecosystem/CHI17-Paper. The same link holds the full questionnaire data used in this paper. That data is also anonymized.

We hope that the availability of the system, code, and the open data that we publish here will make it easy for other researchers to investigate problems related to personalization in foreign language reading.

REFERENCES

1. Graham Atwell. 2007. Personal Learning Environments - the future of eLearning? *eLearning Papers* 2, 1 (2007), 1–9.
2. Martin Avagyan. 2017. Building Blocks for Online Language Practice Platforms. (July 2017). Bachelor Thesis, University of Groningen.
3. 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.
4. 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
5. Dan Chirtoaca. 2017. Apollo: Simplicity and Intuitiveness in a Personalized Multilingual Reading Tool. (July 2017). Bachelor Thesis, University of Groningen.
6. 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*.
7. Johan De Jager. 2017. A Self-Adaptive API Multiplexer. (Aug. 2017). Bachelor Thesis, University of Groningen.
8. Isabelle De Ridder. 2002. Visible or invisible links: Does the highlighting of hyperlinks affect incidental vocabulary learning, text comprehension, and the reading process? (2002).
9. 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.
10. 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.
11. Duco Dokter, John Nerbonne, Lily Schurcks-Grozeva, and Petra Smit. 1998. Glosser-RuG: A User Study. In *Language Teaching and Language Technology*. 169–178.
12. Gregory L Friedman. 2008. Learner-created lexical databases using web-based source material. *ELT journal* 63, 2 (2008), 126–136.
13. 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

14. 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

15. 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.

16. David Hirsh and Paul Nation. 1992. What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure? *Reading in a Foreign Language* 8, 2 (1992), 689 – 696.

17. Róbert Horváth and Marián Šimko. 2013. Enriching the Web for Vocabulary Learning. In *European Conference on Technology Enhanced Learning*. Springer, 609–610.

18. Mike Levy and Glenn Stockwell. 2013. *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.

19. 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

20. C McCarthy. 1999. Reading theory as a microcosm of the four skills. *The Internet TESL Journal* 5, 5 (1999), 1–6.

21. 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.

22. 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.

23. 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

24. 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.

25. 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.

26. Gábor Prószéky. 2002. Comprehension Assistance Meets Machine Translation. *Tomaš Erjavec* (2002), 1–5.

27. 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.

28. Willy A. Renandya. 2007. The Power of Extensive Reading. *RELC Journal* 38, 2 (2007), 133–149. DOI: http://dx.doi.org/10.1177/0033688207079578

29. 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.

30. 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.

31. 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

32. 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.

33. Iñigo Yanguas. 2009. Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. (2009).

34. 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