Exploring Visualizations for Digital Reading Augmentation to Support Grammar Learning

Fiona Draxler, Christina Schneegass, Nicole Lippner, Albrecht Schmidt
LMU Munich
Munich, Germany
firstname.lastname@ifi.lmu.de

ABSTRACT

Reading foreign language texts is a frequently used strategy for language learning. Visual text augmentation methods further support the learning experience, e.g., by annotating vocabulary or grammar. Common approaches are integrated dictionaries or static grammar highlights. This work investigates how we can further support grammar learning with the dynamic visualization and interaction opportunities offered by digital reading devices. In collaboration with teachers and potential learners, we identify difficulties learners experience with English grammar and gather ideas for suitable interactive text augmentations. Based on this, we design four different concepts that augment adjectives and adverbs in English-language texts using typographic cues and interactive information displays. The concepts are evaluated in a within-subject study ($N = 16$). Results show that participants preferred concepts that presented case-specific support, did not distract too much from the text, and gave details on demand. We conclude with design recommendations for designing text augmentation for language learning.

CCS CONCEPTS

• **Human-centered computing → Information visualization**; *Graphical user interfaces*; *Empirical studies in interaction design.*

KEYWORDS

textual enhancement, language learning, grammar, reading

ACM Reference Format:
Fiona Draxler, Christina Schneegass, Nicole Lippner, Albrecht Schmidt. 2019. Exploring Visualizations for Digital Reading Augmentation to Support Grammar Learning. In *18th International Conference on Mobile and Ubiquitous Multimedia (MUM 2019), November 26–29, 2019, Pisa, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3365610.3365623

1 INTRODUCTION

Despite the availability of instant translation technology, learning a foreign language is still important and can, for example, be beneficial for the development of communication skills [6, 15]. One common way to improve in a foreign language is reading, which, in particular, facilitates vocabulary and spelling knowledge [4, 13].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MUM 2019, November 26–29, 2019, Pisa, Italy
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7624-2/19/11…$15.00
https://doi.org/10.1145/3365610.3365623

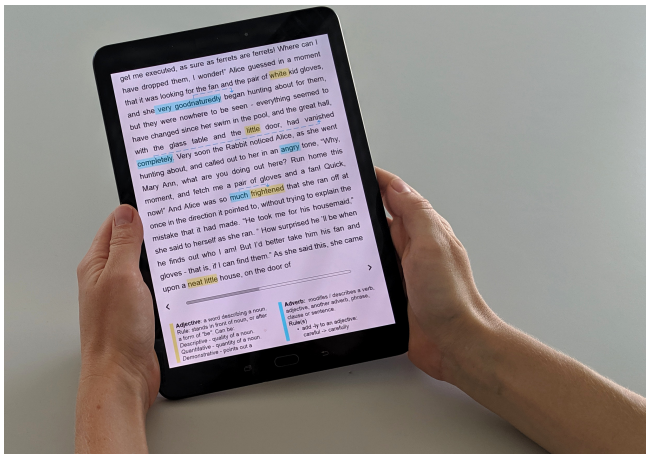


Figure 1: One of our four concepts for interactive grammar highlights in a digital text

Digital reading technologies, such as tablets or e-Readers, provide support for users to improve their language skills while reading. For example, the Amazon Kindle features explanations of complex vocabulary[1]. Moreover, focus-on-form approaches have dealt with various static textual enhancements for learning grammar [8, 14, 18]. However, only a few systems interactively teach grammatical structures, such as SMILLE [24] or WERTi [16]. The focus of these research prototypes was to correctly identify grammar concepts as accurately as possible. The user experience, on the other hand, remains understudied.

Within this work, we explore which (English-language) grammar concepts can be taught using interactive visual reading augmentation. Building on this, we design and evaluate different concepts for visual augmentation with the aim to maximize user experience and consequently support grammar learning while reading.

For the purpose of this work, we conducted two focus groups. In the 1st focus group ($N = 3$), we asked high school English teachers to generate a set of grammar topics which students commonly struggle with and which would benefit from visual augmentation. In a 2nd focus group ($N = 7$), we asked HCI students and potential users to come up with a set of different visualizations for the grammar topics prioritized in the teacher focus group. As a result of both sessions, we designed four different visualizations for reading augmentation on a tablet-sized device to provide explanations on the use of Adjectives and Adverbs, using (1) color *Highlights*, (2) a

[1]For more information on Amazon Word Wise, see https://www.amazon.com/gp/feature.html?ie=UTF8&docId=1002989731, last accessed July 8th, 2019

pop-up *Window*, (3) *Footnotes*, and (4) annotations *Above* the text (cf. Figures 4a-4c and 5). In a follow-up within-subject lab study ($N = 16$), we evaluated these four concepts regarding user interaction and experience. We investigated the perceived usability of the augmentations, as how suitable they are considered for different types of texts, their ability to encourage interactions with grammar explanations, and overall preferences.

In conclusion, the color highlights were perceived as most attractive visualization and scored highest in user experience. However, participants would also like to combine the highlights with case-specific information presented in pop-up windows (details on demand). We discuss the advantages and potential disadvantages of these concepts as well as possible improvements in future work. In a conclusive section, we derive six design recommendations to help researchers and practitioners in designing reading interfaces to support grammar learning.

2 RELATED WORK

In this section, we present existing methods for text augmentation with a focus on grammar, their benefits for learning, and interactive, digital solutions. We also address reading flow as a key issue of reading experience.

2.1 Text Augmentation

The implementation of annotation, or textual enhancement, for language learning has a long history (for an overview see [8, 14, 18]). The focus-on-form approaches commonly described mostly utilize static typographical cues to draw the readers' attention towards the concepts to be studied, e.g., by underlining words, setting them in bold font, or highlighting them in a different color. Moreover, they show that textual enhancement can increase awareness of the enhanced forms and contribute to users' understanding. It has to be noted, however, that there has not yet been a definite conclusion on long-term learning effects.

In the more recent past, reading on electronic devices, e.g., on e-Book readers, has made it possible to automatically adapt and enhance texts for language learning. Furthermore, digital devices have opened up new ways of interacting with the augmented content, e.g., by displaying additional information on demand [23]. Existing texts on web pages are often used as a basis, which has the advantage that quasi-infinite amounts of learning content can be generated. For instance, in one study, replacing words on web pages with the corresponding translations in a foreign language helped users learn an average of 50 new words within a month [22]. Similarly, songs were used as authentic material for teaching French [19].

Grammatical structures can now also be recognized with high accuracy. This leads to a wide range of further possibilities for digital learning enhancement. In WERTi, grammatical aspects such as determiners or prepositions on web pages are analyzed [16]. The system then creates fill-in-the-blank or multiple-choice quizzes for learners directly on the web page. Another system highlights various grammatical structures to make users aware of when and how they are used [24].

In summary, textual enhancement approaches have been widely studied and appear to have potential for language learning. Yet, they are typically not interactive and do not provide further support such as explicit rule explanations. More recent interactive approaches lack a thorough usability concept and evaluation in user studies, even though good usability is a key issue for minimizing frustration caused by text augmentations and for motivating people to continue using a system.

2.2 Annotations and Reading Flow

Consulting annotations while reading a text may improve the understanding of words or concepts. However, at the same time, shifting attention towards other content means the reading flow is interrupted. Especially when synthesis of pieces of information across several paragraphs is necessary, interruptions decrease text comprehension when readers do not have time to prepare for them [7].

Reading on electronic devices has the advantage, though, that supplementary information can be flexibly located close to the associated text. Translations can be shown next to a specific word without the need to open an external dictionary. This is further aided by the possibility to hide annotations that are not currently necessary. Thus, it is likely that the disruptive effect is decreased [20].

In summary, since interruptions can impact the reading experience, possible causes of distractions should be considered when designing text augmentation concepts.

3 REQUIREMENT ANALYSIS

Before starting to develop concrete concepts for visually enhancing foreign language texts for grammar learning, we further extended our perspective from related work with insights gained in two focus groups. In a 1st focus group ($N = 3$) with high school English teachers, the focus was on problematic grammar concepts for English second language learners. A 2nd focus group ($N = 7$) with HCI students was then conducted to discuss feasible designs to implement the grammar concepts suggested by the teachers.

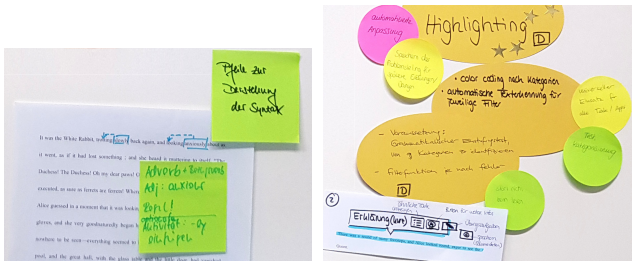
Both focus groups followed a similar procedure: we started with an explanation of the goals and a general outline. The participants then each signed a consent form and agreed to our data protection policy in line with the GDPR. We audio-recorded the sessions for post-processing with the participants' consent. Therefore, participants were encouraged to voice out their thoughts and concerns. Ideation was performed in an iterative and narrowing process aligned with previous studies (cf. [11]). To assist the creative process and foster discussion, we provided the participants with printouts of an excerpt of Carroll's "Alice's Adventures in Wonderland" [2] with increased spacing for them to draw on and discuss. The focus groups were conducted in German (as the participants' preferred language of communication) and took approximately 90 minutes. For their time and help, the participants were compensated with 15€ Amazon vouchers.

This section outlines the further procedure of our requirement analysis and the results of the focus groups, in particular, the ideas we used as a basis for the designs in our application prototype.

3.1 Focus Group 1 – Teachers

3.1.1 Procedure & Sample. To ensure the pedagogical validity of our concepts, the main goal of our 1st focus group was to develop a list of suitable and feasible grammar constructs to teach with our approach. The participants were teachers at a German high

Figure 2: Relevant grammar constructs and comments collected by teachers



(a) (b)

Figure 3: Selected sticky notes from focus groups: (a) arrows for showing relationships, (b) highlights and pop-ups

school (one male, two female), had 9, 14, and 20 years of teaching experience, and a mean age of 41.3 years (SD = 6.1). The main aim of the teachers was to focus on the grammar concepts and their pedagogical knowledge and less on the implementation and/or realism of their idea. Therefore, building upon their extensive experience in teaching, we asked the participants to start by listing all the grammar topics they teach in high school. In a second iteration, the teachers marked the concepts which students commonly struggle to understand. For those troublesome concepts, they wrote down which information students need to achieve greater comprehension. In a successive step, we asked teachers to imagine how the information could be visualized, using the text excerpts. Within the group, the three teachers debated the potentials and limitations of their ideas and finally decided on their favorite ideas.

3.1.2 Results. The grammar concepts the teachers considered most troublesome for high school students included tenses, differences between adjectives and adverbs, and irregular verbs (cf. Figure 2). Teachers emphasized the importance of providing learner with grammar rules, if possible comparing them to rules and terms in their native language. Furthermore, repetition was mentioned as a good tool to facilitate learning grammar, e.g., using example sentences. In terms of potential visualizations and interaction concepts, the participants came up with the idea of a filter option to choose from a list of grammar constructs to learn during the current reading session. By clicking on a certain word, the user could have the ability to highlight or color-coded elements of a sentence. Moreover, the text should show grammatical distinctions, for example, by highlighting irregular verbs in a different color. Displaying arrows between lines could highlight syntactical connections between words (e.g., a noun described by an adjective). A window next to the text or a footnote could display further details on demand.

3.2 Focus Group 2 – Users

3.2.1 Procedure & Sample. For exploring potential visualizations based on the teachers' theoretical and pedagogical insights, we recruited seven HCI students (six female, one male) with a mean age of 25.1 years (SD = 2.1). They also represented a typical target group for mobile learning applications and could thus contribute with their perspective as potential learners.

In contrast to the teacher group, we asked the students to view the potential application from an ego-perspective. Thus, we first asked them about English grammar constructs they would personally like to improve and which information they think they would need to strengthen their understanding. In a subsequent step, the students came up with several visualizations on how the information could be presented to a reader. We divided the group into smaller teams (two teams of two and one team of three) to foster creativity and enable quiet participants to engage more easily in the discussion. Within the smaller groups, we asked the participants to take 15 minutes to discuss and sketch visualizations of at least two distinct ideas. Each team presented their results and the others gave feedback on strengths, weaknesses, and optional revisions.

3.2.2 Results. The participants of the user focus group stated to be most interested in augmenting English texts with grammar definitions, rules, examples, exceptions, and translations in their mother tongue. Moreover, they reported that highlights and tooltips could be helpful ways to teach the grammar. Feedback, on the other hand, could be presented at different levels and should be personalized to the reader. For example, a personalized dashboard including a progress analysis could motivate the learner. Finally, the participants presented four concrete interaction methods:

(1) After clicking or tapping on a word or sentence, information about it would appear in an additional window. For the example of tenses, it would show an explanation when this tense is used, how it is formed and where it stands on a timeline.
(2) Eye-tracking could be used to identify weak points (assuming one reads those parts more slowly or repeatedly) to then provide rules, explanations, exercises, or further examples for the learner.
(3) Through selection of a text area, grammatical constructs could be color-coded with additional details displayed in between lines (bigger line spacing) or accessible as a pop-up.
(4) Selected categories based on skills could be highlighted through personalized color-coding. A pop-up window would provide an explanation and the option for additional information, exercises, documentation, and showing similar texts.

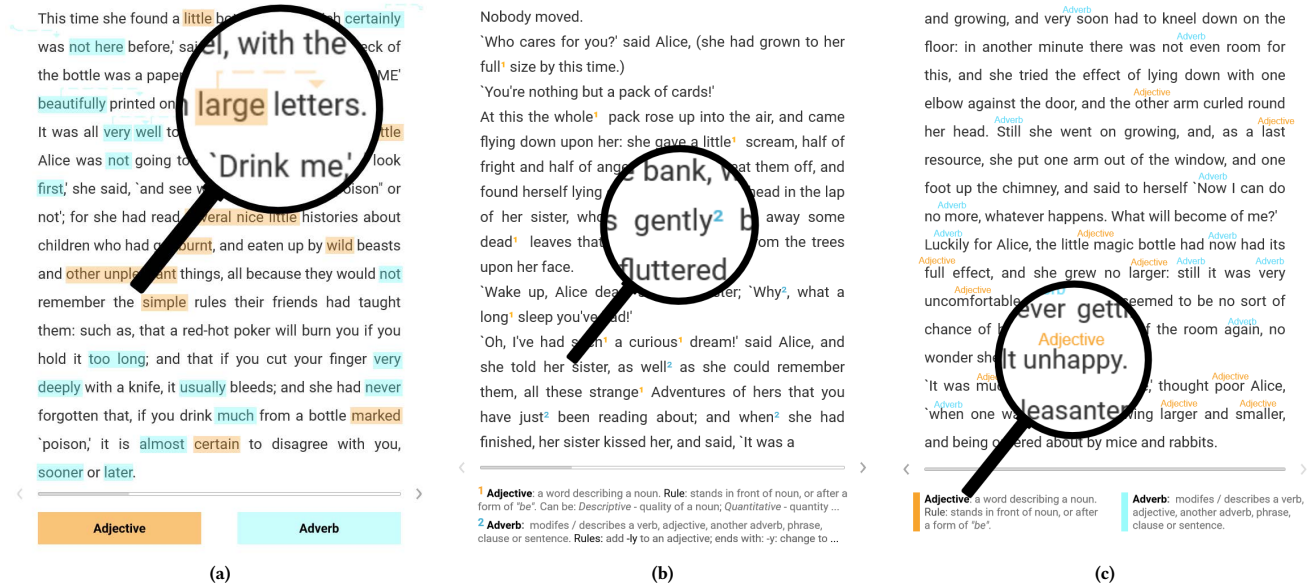


Figure 4: Visual augmentations in the concepts (a): *"Highlights"* - color marking and arrows to show references; (b): *"Footnotes"* - presenting more information in footnotes; and (c): *"Above"* - explanations above words in between lines. The magnifying glass shows details for illustration purpose.

4 CONCEPTS

The participants of the two focus groups came up with a variety of grammar constructs that could be supported by a reading augmentation tool.

We selected the use of adjectives and adverbs in English texts as the grammar concept we would enhance in our visualizations, as this is frequent pattern, but still poses challenges to learners. Difficulties are (1) the construction of the corresponding adverbs for an adjective, and (2) whether the descriptor refers to a noun or another type of word, and in the latter case, if it is a verb that requires adverbial use. While in principle, the use of adverbs and adjectives is very similar in German and English, adverb and adjective forms of a German word are usually identical and only differ in their function. This makes it harder for Germans to understand the conceptual differences [5].

For this grammar topic, we developed four visualization and interaction methods, summarized in Figures 4a to 4c and 5, all further explained below. The concepts were based on the ideas of the focus groups, but we organized them to cover a range of interactions. Furthermore, we included patterns from prior literature, such as text highlights and pop-ups with additional information.

Highlights. Highlighting different grammar constructs in individual colors is the first concept we chose (cf. Figure 4a). This form of visualization was frequently mentioned in the literature and was a preferred solution in the focus groups. We used orange and blue as representations for adjectives and adverbs respectively, so the constructs are easily distinguishable even for people with color-vision problems. The rectangles at the bottom work as the key for the color-coding and function as a button for opening up a description of the grammar concept. In addition to the color highlights, arrows link the adjectives and adverbs to their reference word within the sentence. Since focus group participants were worried that the arrows might be too disturbing, they can be toggled by clicking on the highlighted words.

Window. In the second visualization, the grammar constructs are highlighted in a less obvious fashion than in the color highlighting mentioned above. The participants of both focus groups emphasized the importance of reading without hindrance. Adjectives and adverbs are displayed in grey text color. By clicking on a grey word, a pop-up window opens (cf. Figure 5) to display further information on the respective grammar constructs.

In addition to the information displayed in the *Highlights* visualization, the grammar information in *Window* is specific to the individual word the user clicks on. The window is positioned either above or underneath the specific word depending on the word's position in the text. A little spike points at the selected word.

Footnotes. Furthermore, participants considered footnotes (see Figure 4b) a good option to mark the position of grammatical constructs with more explanation below the text. The footnote numbers, *1* for adjectives and *2* for adverbs, are colored blue and orange so that they can be distinguished from actual footnotes in the text. The footnote key and a short explanation of both constructs within the footnotes are placed below the progress bar for visual separation. By clicking on the respective footnote, a pop-up window with further information will appear. It is identical to the one applied in concept *Highlights* (cf. Figure 4a).

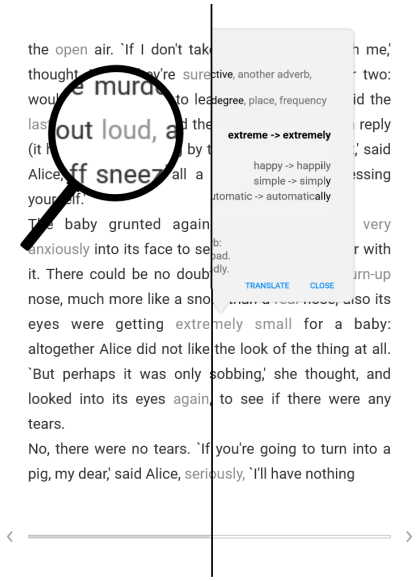


Figure 5: Concept *"Window"* - Details on demand. By clicking on a grey word in the text (left half), a pop-up window with additional details opens (right half).

For the participants of the focus group, the aim of footnotes as a subtle annotation was to recognize the grammar constructs without interrupting the reading flow. This concept was also included to broaden the range of abstraction within our choice of visualizations. Footnotes are a common method for referring to further information on something, but they are not expressive on their own.

Above. Participants also liked the idea of displaying the grammatical concepts in the blank space between lines, above the individual word (see Figure 4c). Thanks to the textual description, the reader does not need to remember a color-coding and does not have to move the gaze far away from the text. Thus, participants argued that the reading flow may be less interrupted than in the *Footnotes* or *Highlights* concept. As in the *Footnotes* concept, a short explanation of the grammar construct is shown at the bottom of the page, which, again, opens a pop-up with additional information on click.

In all four concepts, the pop-up windows display rules, examples, and common errors, accessible by clicking on the footer buttons in *Highlights*, the footer summaries in *Footnotes* and *Above*, and the words themselves in *Window*. Additionally, *Window* and *Highlights* explicitly show the word the adjective or adverb describes, *Window* also marked the rule applied for constructing the selected adverb.

The visualizations were first drafted as low-fidelity wireframes on paper and afterward implemented as high-fidelity clickable prototypes in the web-based tool *Axure*[2].

5 USER STUDY

We evaluated the four enhancement concepts in a lab-based user study ($N = 16$), using quantitative and qualitative measures with a focus on user experience and perceived helpfulness of the visualizations to learn about the underlying grammar constructs. For now, we applied the augmentations to a literary text, but in our qualitative analyses, we also explore their potential for other genres. Specifically, we wanted to answer the following questions:

- Q1: Which visualization is ranked highest in terms of usability and why?
- Q2: Which visualization is preferred for what type of text and why?
- Q3: Which visualization encourages interaction with the contents (grammar explanations) and why?
- Q4: Which visualization do the participants prefer overall and why?

5.1 Sample

We recruited participants using our university's Facebook page and Slack channel with the only requirement of them not being English native speakers. A total of 16 people (eight male, eight female) took part in the study, whereof 12 stated to be German native speakers, two Chinese, one Polish, and one bilingual (Afghan and German). Their mean age was 25.06 years ($SD = 2.33$) and they all had at least a high school diploma. Seven participants were university students, four computer scientists, one a medical doctor, and the remaining four worked in different industry domains. Self-assessed English skills on the Common European Frame of Reference were reported as fluent / near-native (C2) by two participants. Five participants considered to have an excellent command / highly proficient in spoken and written English (C1), three a very good command (B2) and one good command / good working knowledge (B1). Finally, five participants stated to have basic communication skills / working knowledge in the English language (A). Participants rated how often they used e-Readers and read digital texts on a 5-point Likert scale (1=*not at all*, 5=*a lot*). Experience with e-Readers was diverse ($M = 2.81$, $SD = 1.42$). However, with one exception, all participants frequently read digital texts ($M = 4.31$, $SD = 1.10$). We could therefore assume a certain level of familiarity with digital texts.

5.2 Study Design

In our within-subject user study, the participants interacted with all four visualizations in balanced order to avoid sequence effects. Since we were able to recruit 16 participants, all possible sequence combinations for the visualizations could be tested. Besides our independent variable *design*, with its four facets (*Highlights*, *Window*, *Above*, and *Footnotes*), we further assessed five dependent variables: (1) AttrakDiff scores, (2) SUS scores, (3) perceived suitability of visualizations for different text types, (4) interaction behavior during usage, (5) ranking of concepts according to individual preferences. We gathered additional qualitative data in post-hoc semi-structured interviews. All measurements used in this study are further outlined in the following subsections.

5.3 Procedure

The study took approximately one hour and the users were compensated with either a 10€ Amazon voucher or course credit. Initially, all participants were informed about the topic and the structure of

[2]Axure: https://www.axure.com/, last accessed August 7th, 2019

the user study. They gave consent to audio recording for the evaluation of think-aloud comments, and the data protection regulation. We offered participants to pick their own user study ID randomly by crossing out one of the numbers on a sheet showing numbers from 1 to 100. With this technique, we aimed to avoid sequence associations and increase anonymization even in our small sample.

At the beginning of the study, participants filled in a demographic questionnaire asking for age, gender, education, native language, English skills, and experience with digital reading. Furthermore, we assessed the participants' knowledge on the use of adverbs and adjectives with four task-specific fill-in-the-blanks questions (e.g., "She drives ___. (good/well)"). We explicitly decided against using a post-hoc knowledge test, since our within-subject design would not allow any conclusions about the origin of the measured learning effect. Exposure to the contents, independent from any visualization, will increase or refresh participants' knowledge on the topic and, thus, bias the results on any form of learning. Instead, we assess perceived change in the understanding of the grammar constructs. By doing so, we aim to receive feedback on the concept of grammar annotations in general.

Afterward, participants started with the first condition: Reading four pages of text, interacting with the visualization while thinking aloud, rating the concept's usability, and stating advantages and disadvantages. There was no explicit introduction to the respective features as we wanted to observe intuitive usage. This process was repeated for all four conditions. At the end of the study, participants ranked the visualization concepts in their order of preference and we asked them about their reasoning.

5.4 Material & Apparatus

As text materials for this study, we again used an excerpt from the book *Alice' Adventures in Wonderland* [2] since it is easy to understand and offers several Adjective and Adverb constructs. The study was conducted on an Android tablet with a 9.68-inch display, comparable to an e-Reader. Additionally, a Windows Notebook was set up for the participants to fill in the questionnaires.

After interacting with each visualization, the participants filled in two standardized usability questionnaires: (1) AttrakDiff [9], assessing a system's qualities and attractiveness, and (2) the System Usability Scale (SUS) [1] to measure usability. The AttrakDiff is a standardized questionnaire addressing a system's or product's perceived pragmatic quality (usefulness or practicality), hedonic quality (excitement, novelty), and attractiveness (overall appeal) [9]. We applied the short version containing ten item pairs, which are rated on 7-point Likert scales [10]. The highest and lowest values are represented by the facets' extreme values (e.g., rating the system's ease of use on a scale from 1=*complicated* to 7=*simple*). This scale was transformed to the range -3 to 3 for later analysis. Complementarily, the SUS comes with ten questions addressing the system's perceived usability on 5-point Likert scales ranging from 1=*strongly disagree* to 5=*strongly agree* [1].

The conclusive semi-structured interview revolved around the questions:

(1) What are advantages / disadvantages of the given concepts?
(2) What do you consider the best / worst concept to learn the grammar constructs and why?
(3) How helpful were the visual augmentations to understand the grammar?
(4) Did your understanding of adjectives and adverbs improve?
(5) For what type of text would you recommend what type of visualization?

The interviews were held in German. We transcribed them from the audio-recording and translated statements to English for reporting. For questions 4 and 5, we collected additional quantitative ratings to supplement the subjective statements of the participants. To estimate the change of comprehension of the grammar concepts (cf. Question 4), we asked the participants if they felt their understanding of adverbs or adjectives had changed. They rated accurateness of the statements "I now have a better understanding of ['adverbs' / 'adjectives']." each on a 5-point Likert-scale (1=*totally disagree* to 5=*totally agree*). To assess the suitability of the visualizations for different types of text (cf. Question 5), we defined three characteristics for text: (a) text length, short or long; (b) text difficulty, easy or difficult; and (c) text content, factual or literary. We provided examples for all possible combinations, eight in total. For instance, a short story for children would be characterized as short-easy-literary, whereas a scientific paper would qualify as long-difficult-factual. Participants could then rate the helpfulness of each of the four visualizations to be used for grammar learning in all of the eight text combinations. For this purpose, we used a 5-point scale from -2=*Not helpful* to 2=*Very helpful*.

Additionally to the subjective measures, we logged the time users spent reading the individual pages and texts, and the number and type of interactions (e.g., click on word, click on window) for each visualization.

6 RESULTS

The reporting of our results in this section is aligned along the four research questions stated in the beginning of Section 5, *User Study*. We complement the descriptive analyses we performed with statements and comments collected from participants during the interviews to give further insights into our findings. Please note that whenever we report on significance tests, we used a repeated-measures ANOVA and Bonferroni-corrected post-hoc tests at a 5% significance level. Mauchly's test for sphericity was not significant unless mentioned otherwise.

6.1 Q1: Usability and Attractiveness

To evaluate the perceived usability of our visualizations, we compare the concepts' AttrakDiff and SUS scores. Aligned with Hassenzahl, Burmester and Koller [9], we performed significance tests on the AttrakDiff and found that both the overall AttrakDiff score and the score for the pragmatic quality were significantly affected by the concept ($F(3, 45) = 3.66$, $p < 0.05$ and $F(3, 45) = 3.08$, $p < 0.05$, respectively). In the first case, post-hoc tests showed a significant difference between the conditions *Above* and *Highlights* ($p < 0.05$). In the latter case, no comparison was significant. This finding is confirmed by the visualization based on the AttrakDiff evaluation tool[3] shown in Figure 6. In both pragmatic and hedonic quality, *Highlights* achieved the highest scores ($M = 1.1$, $SD = 1.22$), followed by *Window* ($M = 0.86$, $SD = 0.90$), and *Footnotes* ($M = 0.45$,

[3] www.attrakdiff.de, last accessed August 27th, 2019

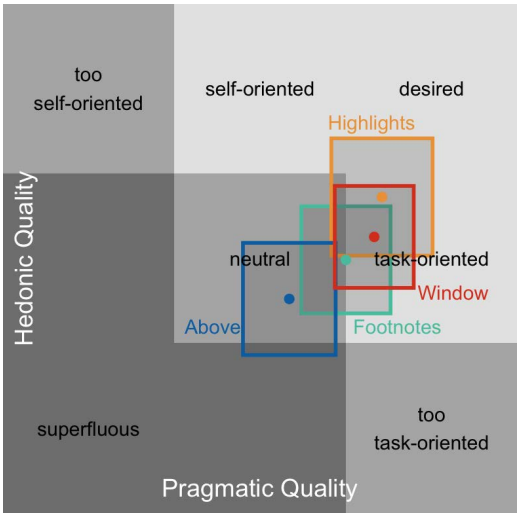


Figure 6: AttrakDiff – mean pragmatic and hedonic quality scores (dots) and confidence intervals (boxes) for the four concepts.

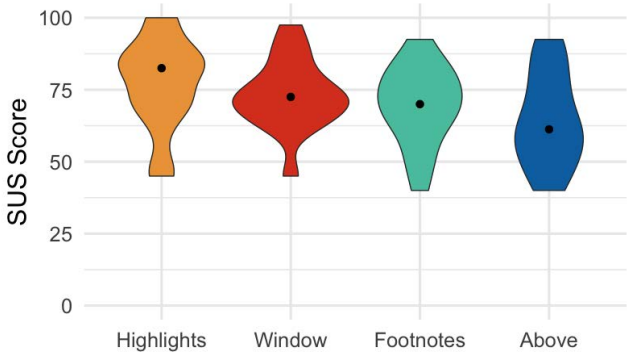


Figure 7: Violin plot of SUS scores for the four concepts, showing the median (dot) and the probability distribution.

$SD = 1.10$). Only the evaluation of the concept *Above* resulted in a slightly negative mean AttrakDiff score ($M = -0.14$, $SD = 1.17$).

The mean SUS score was also highest for the *Highlights* condition ($M = 77.19$, $SD = 15.41$, 100 being the highest possible score), followed by *Window*, *Footnotes*, and *Above* (cf. Figure 7). No significant differences between the four conditions could be found (all p-values $> .08$).

The interviews gave us further insights into aspects of the concepts that participants liked or found useful. For *Highlights* and *Window*, the participants appreciated that the visualizations/explanations specifically referred to the current words. Twelve participants positively noted the arrows connecting an adjective or adverb and the word it describes (e.g., P24: "direct reference with the arrow is very nice"). Similarly, eight participants appreciated the case-specific rules in *Window*. Eight participants explicitly missed the specific information in concepts *Footnotes* and *Above*. Regarding the design and layout, it was mentioned that in *Footnotes* and *Above*, too much space was used up for the descriptions at the bottom (P24, P25, P51) and the text-in-text layout in *Above* made the text appear chaotic (P55). P39 and P55 added that opening the full window in *Window* on each click was too much. Colors were considered useful for distinction between adjectives and adverbs (e.g., P47: "Overview thanks to colors"), but some people found them overwhelming (e.g., P55: "[*Highlights* was] too overloaded with colors").

6.2 Q2: Suitability for Different Text Types

Within the semi-structured interview we performed at the end of each session, participants rated the suitability of our visualizations to be used for the augmentation of different text types (cf. Figure 8). Texts were categorized along the dimensions *text difficulty*, *text length*, and *text type*, and examples for all possible combinations were given. For the visualization concept *Highlights*, participants saw the greatest helpfulness for literary-short-easy texts ($M = 0.56$). These ratings are similar to the ones of the concept *Above*, receiving a slightly lesser average score for the perceived helpfulness ($M = 0.25$). Still, for *Above*, the application in literary-short-easy texts was perceived as most suitable. For the visualization in *Footnotes*, participants gave the highest ranking for the application in short and easy factual texts ($M = 0.25$).

The only concept that was considered most helpful for difficult texts is *Window*, which received equal scores ($M = 0.81$) for literary-long-difficult texts as well as factual-short-difficult texts. Over all concepts, *Window* received the highest out of the four suitability scores for six out of the eight text combinations (except literary, easy texts, both short and long).

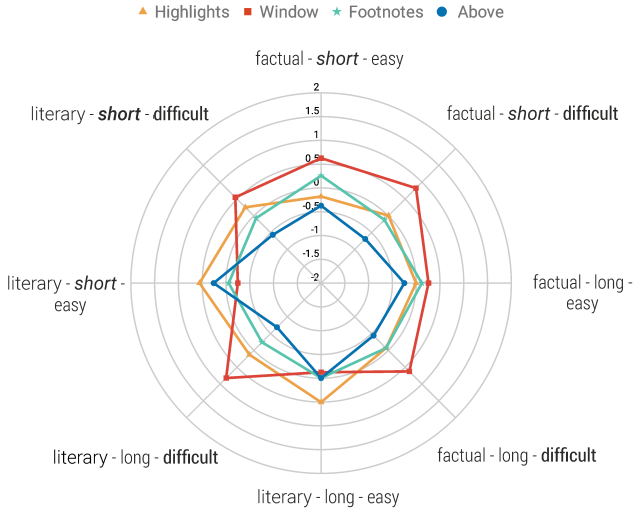


Figure 8: Perceived suitability of the visualizations for different text types characterized along the dimensions text length, difficulty, and content type.

6.3 Q3: Encouraging Interactions

When we look at the mean click counts for all visualizations, we notice an increase in interactions when the concept is presented on second, third, or forth position in the study design. Table 1

	1st		2nd		3rd		4th		Avg.
	T	E	T	E	T	E	T	E	
Highlights	79	19	147	18	139	15	192	25	158.5
Window	42	-	87	-	54	-	85	-	67
Footnotes	0	6	26	15	73	12	48	25	51.25
Above	6	3	49	19	20	11	44	2	38.5

Table 1: Total number of clicks on the text (T) and explanations (E) of the different visualizations when presented as first, second, third, or forth visualization in the study design. Every visualization appeared exactly four times at every position.

summarizes all clicks on the text visualizations and the explanations presented at the bottom of the page[4].

Across all concepts, the number of interactions with *Highlights* was higher ($M_H = 34.81$) than the interactions with the other concepts ($M_W = 17.63$; $M_F = 9.19$; $M_A = 7.43$). The concept had a significant impact on the click count ($F(2.1, 31.5) = 14.7, p < .001$ with Greenhouse-Geisser correction). Indeed, the post-hoc tests showed that *Highlights* had a significant difference compared to all other conditions: at $p < .001$ with *Above*, $p < 0.05$ with *Footnotes*, and $p < 0.05$ with *Window*. Since the presentation of the conditions was randomized, they appeared in every possible order. Comparing only the interactions on the first screen the participants saw, *Highlights* still had the highest click rate on the text on average ($M_H = 19.75$; $M_W = 10.5$; $M_F = 0$; $M_A = 1.5$) as well as the descriptions ($M_H = 4.75$; $M_F = 1.5$; $M_A = 0.75$).

Every participant interacted with the adverbs and adjectives in *Window* and all but one participant interacted with the *Highlights*. According to the interviews, it was not always obvious that interaction with the text was possible: this was mentioned three times for *Highlights* and twice for *Window*; for *Footnotes*, P25 assumed clicks on the (non-interactive) numbers to show more details.

Furthermore, it was seen as an issue that the general explanations and examples in the pop-up window did not change (e.g., P35: "did not open everything anymore because it said the same"). Consequently, P55 asked for more example sentences using a specific word or the current type of adjective/adverb (e.g., a demonstrative adjective). P54 suggested to show the full rules in the beginning and then minimize them.

As a further solution for adaptive content, the details on demand in *Highlights* and *Window* were well received (e.g., P30: "get additional information when necessary" and P32: "unpack only when necessary [...] good that it is not so obvious"). P54 and P55 added that it would be helpful if the arrows in *Highlights* faded out on their own. The details on demand also made it possible for participants to self-check (P35, P39, P51, P55): for example, P39 said that with the arrows, she was able to think first and then verify her solution.

Regarding individual interaction strategies, we could observe that some participants first read the whole page and then started interacting with the interface, whereas some participants interacted with it during reading. P33 stated that it was "annoying that you always had to click to get more information" in *Window*.

[4]Note: In the concept *Window*, the explanations were presented in the window itself and not separately as in the other visualizations.

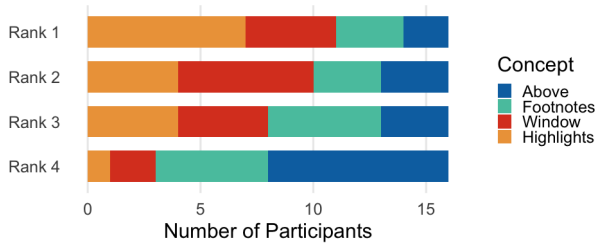


Figure 9: Preference ranking of the four concepts.

6.4 Q4: Ranking and Comparison of Concepts

After the interaction with all visualizations, participants had to bring the concepts in an order from most to least favorite. For this task, we did not specify any rating criteria but asked participants for an overall rating according to their own measures. Figure 9 depicts the overall ranking, showing that *Highlights* was picked as first favorite by seven participants and as second favorite by another four participants. While *Footnotes* and *Window* were named almost equally across all ranks, the concept *Above* was rated least favorite by half of the participants.

6.5 Feedback Overall Concept - Perceived Learning Effect

The four preceding fill-in-the-blank questions addressing the use of adverbs and adjectives showed that only some participants were able to correctly apply their knowledge. The four questions were answered correctly in 9/16, 11/16, 14/16, and 16/16 instances, respectively. We have to note here that participants did not receive any feedback on the correctness of their answers. We used the assessment only as a tool to control potential effects of prior knowledge on the participants' interaction with and opinion on the concepts in general. As mentioned in Section 5.3, we did not quantify the learning effect. At the end oft the user study we asked the participants if they felt their understanding of adverbs or adjectives had changed. They rated the change in understanding from 1=*totally disagree* to 5=*totally agree*. Neither the perceived understanding of adverbs nor adjectives changed notably. Participants rated the change in understanding as neutral (Adverbs: $M = 3.44$, $SD = 0.86$, Adjectives: $M = 3.06$, $SD = 1.30$). One participant noted to have refreshed her knowledge on these grammar constructs (P33), while others stated to have a better understanding of the rules now (P32, P54, P46), or noticed interesting facts about the grammar. For example, participant P57 realized that more words than she thought were actually adverbs. In addition, P47 highlighted that her ability to recognize adjectives in a text had improved, while P55 emphasized the system's quality to foster 'active reading', which she described as an improved way of exercising.

6.6 Reading Flow and Focus on Grammar

The concepts varied in their level of obtrusiveness; this also showed in the comments we received. Eight participants felt that *Highlights* supported their reading flow, while four stated that the color highlights were distracting. P39 said that through the highlights, there

was a stronger focus on grammar and words than on content and the reading was therefore like an exercise.

Window was considered least obvious: twelve participants stated that in this condition, their reading was barely interrupted (e.g., P24: "does not impact reading flow, but explains" and P30: "does not distract when functionality is not needed"). On the other hand, three participants remarked that the grey color was sometimes hard to notice. Furthermore, the interaction had a disruptive effect (e.g., P33: "reading flow interrupted by clicks") and the pop-up window was considered too large for short checks (e.g., P39: "too annoying, always a big window, even if you only want to have a quick check").

The perception of *Footnotes* was more diverse: five participants explicitly said that the footnotes compelled them to jump back and forth from text to explanation, while seven said that they were able to ignore the footnotes, especially once they had become used to them (e.g., P51: "You get used to the numbers, you do not look down each time anymore"). P24 remarked that *Footnotes* were more suited for learning than the highlights in *Window* since they were more obvious.

Thirteen participants found that in *Above*, the text-on-text had a negative effect on readability or was distracting (see also Section 6.1). Four participants liked the idea that they could immediately identify the word type almost subconsciously (P33, P35, P47) and without looking down at the footnote legend (P55).

Overall, the interview statements strongly suggest that *Window* was best for maintaining the reading flow, followed by *Highlights*. However, there were also comments on a potential negative correlation of reading flow and learning effect, e.g., P51: "[*Above*] could be more efficient for learning because you cannot avoid it".

7 DISCUSSION

In this section, we summarize our findings specific to the questions guiding our study. In our guidelines in Section 9, we discuss further details and consequences and related work in a broader sense.

SUS, AttrakDiff, and the ranking show a preference for *Highlights*, followed by *Window*. Important advantages of these concepts were probably the information related to the current word and the details on demand. Color was generally helpful – when it was not too bold and distracting. The preferred subtlety of annotations was also an important factor when it comes to suitability for texts with different purposes and characteristics. As a general trend, we see that for easier texts, more subtle annotations were preferred, whereas for more difficult texts, more support would be necessary. This probably also had an influence on why *Windows* was considered best for six out of the eight text type variations. In this concept, thanks to the details on demand, it is possible for readers to decide how much support they need at a given point.

The analysis of interactions showed that participants engaged a lot with the text and the additional information. However, they did not always realize which opportunities for interaction were available and could, thus, not fully benefit from each concept's potential. The visualizations differed slightly with respect to the specificity of the information they provided and how those could be accessed. This may have influenced the users' interaction strategies, but it also meant that we were able to assess the potential of different levels of granularity (such as the helpfulness of showing the reference word of an adjective or adverb as in *Highlights*).

Moreover, we observed varying reading strategies: some people checked annotations while reading, some after reading a page. The choice of strategies was also influenced by the way participants studied, e.g., by guessing word types or connections and then checking if their assumption was correct. Such self-checks are actually quite promising from the point of view of cognitive theory, as retrieval tests have been shown to improve retention [17].

The overall perceived learning effect was not very pronounced. With our specific task of teaching the use of adjectives and adverbs, we expect the learning effect to change in regards to the proficiency of the target user group. However, we did get some specific comments on details that participants had found out while reading the augmented texts. Learning was often mentioned in connection with the reading flow: Participants noted that the learning effect might be smaller when the main focus is on the text, so a certain level of interruptions might have to be accepted. The reading flow itself was considered best for the concept *Window*, where the only typographical cue used was grey text color and further annotations were explicitly triggered. A number of participants regarded bright colors, text-on-text design in *Above*, and frequent changes from text to the bottom of the page in *Footnotes* as distracting elements.

8 LIMITATIONS AND FUTURE WORK

The concepts we designed showcased interactive text augmentation for adjectives and adverbs and one pre-processed text only. In the future, we plan to extend this work to cover other grammar topics relevant to learners. For example, literary texts could be a useful resource for practicing tenses in a more enhanced way than it is currently done in text books for language learning. In combination with natural language processing mechanisms as in WERTi [16] or SMILLE [24], it would also be possible to automatically prepare texts for studying specific grammar aspects with our system.

In addition, the concepts we built were based on outcomes of two focus groups. We made sure to get insights from teachers as people with pedagogical experience, as well as potential users that could focus on the usability, but further study is needed to identify ideal augmentation methods. For instance, we assume that personalization with respect to the learner's knowledge and current context would be helpful. Furthermore, the current study already suggests that learners could benefit from integrating exercises such as clozes or multiple-choice tests, as several participants used the concepts for self-checking.

Regarding the evaluation, we have to note that we did not consider the actual learning effect we were able to achieve with our reading concepts. Because of the within-subject design, we would not have been able to identify which concept actually contributed to the learning. Now that we have obtained feedback on promising characteristics for augmentation concepts, evaluating the learning potential in a more focused study is logical next step. For now, participants also read four pages per concept only. A longer usage study is necessary to show long-term acceptance and if there are risks that text augmentation becomes too distracting and annoying when reading an entire story or book. Moreover, using an eye tracker

during a study could give further insights into the reading/study behavior and their reading flow.

9 DESIGN RECOMMENDATIONS

Based on our focus groups, the evaluation of our concepts, and related work, we can give the following recommendations for designing interactive grammar augmentations in text:

Emphasize case-specific rules. Information was particularly valuable to readers when it concerned the current construct. In our concepts, this was the case for the arrows in *Highlights* and the connections and rules explained in *Window*. Giving specific and varied information might also help readers learn to differentiate cases, as it follows the principle of interleaved practice: varying practice of different skills rather than practicing them one by one improves the capability to discriminate [21].

Make elements distinguishable from commonly used elements with other functions. Some of our participants associated annotations with other meanings: the numbers used for footnotes were confusing in combination with text footnote and highlights were seen as a tool to personally mark important parts, and not grammar. Yet, highlights are also commonly used in focus-on-form approaches and in the digital tools VIEW and SMILLE (see Section 2.1). Therefore, if possible, we suggest to give preference to annotations that do not override typical usage, for instance by using symbols instead of numbers for footnotes. In some cases, however, it might already be sufficient to use colors that are not typically used by readers or to give them the option to change them.

Avoid overly repetitive information. In our case, several participants mentioned that they found it annoying that the rule descriptions and examples did not change (except for the small changes in *Window*, where the current rule was highlighted). Instead, the displayed information could be made more adaptive, e.g., by introducing rules one after the other, by reducing the amount of information as learners progress, or by fading out points they have already seen. A certain degree of redundancy, however, may be helpful, as spaced repetition has been shown to improve recall [12].

Adapt topics and content to language level and text. Information should not only be adjusted over time, but also depend on the learner's knowledge and needs. Within one topic, a simple approach is to give details on demand. In our concepts, this was realized with the short and long (pop-up) descriptions in *Footnotes*, and *Above* as well as the pop-up on demand in *Highlights* and *Window*.

We said before that adjectives and adverbs are an important topic for German native speakers, but for other (native) languages, other aspects will have higher priority. If topics are not defined algorithmically or with a machine-learning model, we recommend giving users the option to filter topics and possibly also to define the level of detail in explanations. SMILLE and VIEW both give the option to select grammar topics. Tools such as FLAIR are useful to identify texts that are suited for instruction of specific topics [3].

Balance the trade-off between focus on text and focus on form. Reading flow and distraction were a recurrent theme in our interviews. However, opinions on what concept was most disrupting diverged. This disruptiveness of a concept might furthermore depend on the readers' goals. If the focus is on reading and grammar should only be perceived on the side, highlights can be simple (as the grey text color in *Window*); avoid markers that make the eyes jump (as in *Footnotes*). If the focus is on learning, then more obvious elements and colors can be used. Ideally, the annotations provide scaffolds to continue reading at the correct position once additional information has been consulted.

Clearly indicate what information is available and how to obtain it. In our case, the participants did sometimes not even try to click on items because they did not know that interaction was possible. As a solution, clickable items could be designed to look like buttons and small info buttons could be added.

10 CONCLUSION

Reading foreign texts is a common way to improve language skills and the augmentation of such texts can support the user for effective language learning. Prior work has already explored several possibilities for reading augmentation, including vocabulary translations and highlighting grammar constructs. However, the evaluation of such concepts in terms of interactivity and user experience was previously neglected. In this work, we presented four different visualization concepts to support English grammar learning based on two focus groups, namely color *Highlights*, a pop-up *Window*, details in *Footnotes*, and annotations *Above* the text. We furthermore evaluated these concepts in a within-subject user study ($N = 16$) in terms of their usability, suitability for text types, users' interaction behaviour, and general feedback. The results show that participants appreciated simple, non-distracting designs with details on demand. *Highlights* was perceived as most attractive and encouraged the interaction with the text. However, the feasibility of the concepts differed when participants were asked for application in different types of texts.

Based on the rich feedback we received in the evaluation, we derive six design recommendations to help researchers and practitioners in designing reading interfaces to support grammar learning. In the future, we plan to develop an improved prototype to be tested in a long-term in-the-wild user study. We aim to investigate the possibility for user adaptation and personalization as well as the applicability of our concepts for different types of grammar learning and texts. Overall, we are confident that our evaluation poses interesting insights into the design of grammar augmentations for reading to increase user experience and facilitate effective language learning.

REFERENCES

[1] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[2] Lewis Carroll. 1983. *Alice's adventures in wonderland.* Collins.
[3] Maria Chinkina and Detmar Meurers. 2016. Linguistically aware information retrieval: providing input enrichment for second language learners. In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications.* 188–198.
[4] Richard R Day, Carole Omura, and Motoo Hiramatsu. 1992. Incidental EFL vocabulary learning and reading. *Reading in a foreign language* 7 (1992), 541–541.
[5] Burkhard Dretzke and Margaret Dretzke. 2015. *Englisch üben fürs Lehramt: Fehler erkennen, korrigieren, vermeiden.* UTB.
[6] Samantha P Fan, Zoe Liberman, Boaz Keysar, and Katherine D Kinzler. 2015. The exposure advantage: Early exposure to a multilingual environment promotes

effective communication. *Psychological Science* 26, 7 (2015), 1090–1097. https://doi.org/10.1177%2F0956797615574699

[7] Cyrus K. Foroughi, Nicole E. Werner, Daniela Barragán, and Deborah A. Boehm-Davis. 2015. Interruptions disrupt reading comprehension. 144, 3 (2015), 704–709. https://doi.org/10.1037/xge0000074

[8] ZhaoHong Han, Eun Sung Park, and Charles Combs. 2008. Textual Enhancement of Input: Issues and Possibilities. *Applied Linguistics* 29, 4 (05 2008), 597–618. https://doi.org/10.1093/applin/amn010

[9] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. (2003), 187–196.

[10] Marc Hassenzahl and Andrew Monk. 2010. The inference of perceived usability from beauty. *Human–Computer Interaction* 25, 3 (2010), 235–260.

[11] Marion Koelle, Katrin Wolf, and Susanne Boll. 2018. Beyond LED Status Lights-Design Requirements of Privacy Notices for Body-worn Cameras. In *Proc. of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 177–187.

[12] Nate Kornell. 2009. Optimising learning using flashcards: Spacing is more effective than cramming. 23, 9 (2009), 1297–1317. https://doi.org/10.1002/acp.1537

[13] Stephen Krashen. 1989. We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The modern language journal* 73, 4 (1989), 440–464.

[14] Ryan M LaBrozzi. 2016. The effects of textual enhancement type on L2 form recognition and reading comprehension in Spanish. 20, 1 (2016), 75–91. https://doi.org/10.1177/1362168814561903

[15] Zoe Liberman, Amanda L Woodward, Boaz Keysar, and Katherine D Kinzler. 2017. Exposure to multiple languages enhances communication skills in infancy. *Developmental science* 20, 1 (2017), e12420. https://doi.org/10.1111/desc.12420

[16] Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 10–18.

[17] Henry L Roediger III and Jeffrey D Karpicke. 2006. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science* 1, 3 (2006), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

[18] Daphnée Simard. 2009. Differential effects of textual enhancement formats on intake. 37, 1 (2009), 124–135. https://doi.org/10.1016/j.system.2008.06.005

[19] Ross Sundberg and Walcir Cardoso. 2019. Learning French through music: the development of the Bande à Part app. *Computer Assisted Language Learning* 32, 1-2 (Jan. 2019), 49–70. https://doi.org/10.1080/09588221.2018.1472616

[20] Alan M Taylor. 2014. Glossing frequency and L2 reading comprehension: the influence of CALL glossing. 31, 3 (2014), 374–389.

[21] Kelli Taylor and Doug Rohrer. 2010. The effects of interleaved practice. *Applied Cognitive Psychology* 24, 6 (2010), 837–848.

[22] Andrew Trusty and Khai N. Truong. 2011. Augmenting the web for second language vocabulary learning. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, 3179. https://doi.org/10.1145/1978942.1979414

[23] Rafael Veras, Erik Paluka, Meng-Wei Chang, Vivian Tsang, Fraser Shein, and Christopher Collins. 2014. Interaction for reading comprehension in mobile devices. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services - MobileHCI '14*. ACM Press, Toronto, ON, Canada, 157–161. https://doi.org/10.1145/2628363.2628387

[24] Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. 2017. Enhancing grammatical structures in web-based texts. *CALL in a climate of change: adapting to turbulent global conditions* (2017), 345.