

Relevance, Effort, and Perceived Quality: Language Learners' Experiences with AI-Generated Contextually Personalized Learning Material

Fiona Draxler
LMU Munich
Munich, Germany
Munich Center for Machine Learning
(MCML)
Munich, Germany
fiona.draxler@ifi.lmu.de

Albrecht Schmidt
LMU Munich
Munich, Germany
albrecht.schmidt@ifi.lmu.de

Lewis L. Chuang
TU Chemnitz
Chemnitz, Germany
lewis.chuang@phil.tu-chemnitz.de

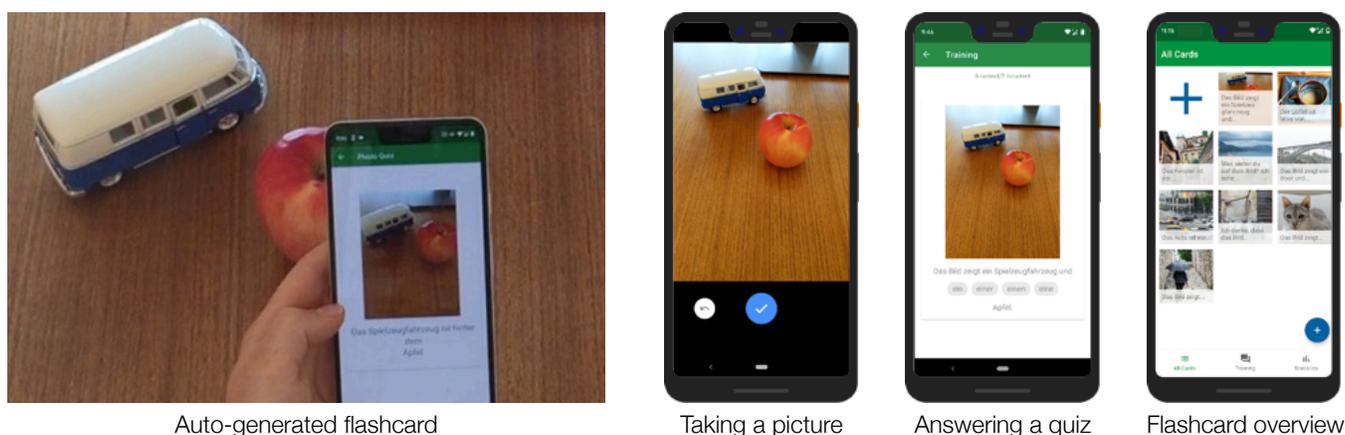


Figure 1: Example of a flashcard for German case grammar, created with the Photo Flashcards app.

ABSTRACT

Artificial intelligence has enabled scalable auto-creation of context-aware personalized learning materials. However, it remains unclear how content personalization shapes the learners' experience. We developed one personalized and two non-personalized, crowdsourced versions of a mobile language learning app: (1) with personalized auto-generated photo flashcards, (2) the same flashcards provided through crowdsourcing, and (3) manually generated flashcards based on the same photos. A two-week in-situ study ($n = 64$) showed that learners assessed the quality of the non-personalized auto-generated material to be on par with manually generated material, which means that auto-generation is viable. However, when the auto-generation was personalized, the learners' quality rating was significantly lower. Further analyses suggest that aspects such as prior expectations and required efforts must be addressed before

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 the author(s) 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.
DIS '23, July 10–14, 2023, Pittsburgh, PA, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9893-0/23/07...\$15.00
<https://doi.org/10.1145/3563657.3596112>

learners can actually benefit from context-aware personalization with auto-generated material. We discuss design implications and provide an outlook on the role of content personalization in AI-supported learning.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Computing methodologies** → *Object detection*; • **Human-centered computing** → *Interactive systems and tools*.

KEYWORDS

mobile language learning, object detection, content generation, personalization, expectations

ACM Reference Format:

Fiona Draxler, Albrecht Schmidt, and Lewis L. Chuang. 2023. Relevance, Effort, and Perceived Quality: Language Learners' Experiences with AI-Generated Contextually Personalized Learning Material. In *Designing Interactive Systems Conference (DIS '23)*, July 10–14, 2023, Pittsburgh, PA, USA. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3563657.3596112>

1 INTRODUCTION

Content personalization in learning can support motivation and increase learning success [47, 59]. In recent years, intelligent systems have made personalization more scalable and widely accessible, e.g.,

by automatically scheduling revision cycles based on previous performance [51]. Others have also started to investigate context-aware content generation, where learning materials are personalized to a learner’s context. Matching materials to the context can trigger situational interest [26] and increase perceived relevance [19]. Objects in our everyday environment, texts we read, and videos we watch are rich sources of learning material extracted from or embedded into personal contexts. Related work has proposed automated methods that enable scalable content personalization. They source content from texts (e.g., [43, 55]), a learner’s GPS location (e.g., [19, 22]), or life-logs [44]. In addition, recent projects in mobile language learning employ computer vision (CV) in the learner’s environment (e.g., [16, 18, 57]).

These works tend to focus on learning outcomes [19, 22, 55] or technological feasibility [18, 22, 43, 44, 57]. Meanwhile, little is known about what learners think of context-aware personalized auto-generated learning material. Nonetheless, understanding this is essential for the successful design and long-term application of learning systems that incorporate AI technologies for content personalization. Therefore, this work explores the following main research questions:

- RQ 1: *Can we achieve comparable usability, user engagement, and recall with auto-generated learning material compared to manually generated material?*
- RQ 2: *Can context-aware content personalization influence the learners’ quality assessment of auto-generated learning material?*
- RQ 3: *What are the opportunities and challenges of context-aware personalized auto-generation for learning?* Here, we report on our lessons learned and identify steps for future development.

To address these questions, we developed three versions of a mobile learning app as a study tool. We focused on designing a realistic scenario that is feasible (albeit not perfect) with current technology. Version 1 auto-generates learning material with state-of-the-art computer vision. Namely, it creates German case grammar exercises based on objects detected in photos that learners create from their everyday contexts (*auto-personalized*). Version 2 includes the same learning material but delivers it through learnersourcing¹, i.e., it is not personalized to the learner context (*auto-learnersourced*). With Version 3, we assess how our auto-generation method fares in comparison to manually generated learning material. As in Version 2, the learning material is based on the learnersourced photos taken in Version 1, but this time, the learning exercises are manually created (*manual-learnersourced*). Subsequently, we conducted a two-week between-subject user study with 64 participants where we evaluated how participants perceived the different types of learning material after repeated use. Additionally, we measured usability and user activity (added flashcards and app interaction) and administered a short pre-test and post-test for German. As a first step, we verified if the (personalized) auto-generation affected these latter measures in comparison to our manually sourced control and found no significant differences between the three versions. However, personalization actually had a negative impact on the perceived quality of learning materials: perceived correctness, quality, and relevance were comparable in both learnersourced conditions,

but significantly lower in the personalized condition. We followed up on this finding with a correctness analysis of the auto-generated material, which showed that some, but clearly not all, of the auto-generated material was as good as the human-generated material. In particular, there were several instances of flashcards where the auto-generated material was generally correct but too imprecise (e.g., referring to “packaged goods”) or where the detected objects were not salient elements of the depicted scene.

Taken together, our findings indicate that learners may have higher expectations towards the learning material when it is personalized and that the additional effort for personalization can impact the overall experience. The benefits of content personalization that are often observed in learning could not compensate for this. Accordingly, we summarize recommendations for researchers and designers of future learning experiences. Notably, because the performance of AI algorithms will not be perfect in the near future, imprecision and a lack of saliency will continue to negatively influence quality perception. Therefore, we recommend applying case-based mitigation strategies, namely feedback mechanisms and crowd strategies.

To summarize, we (1) developed a personalized learning app with state-of-the-art methods for auto-generating context-based learning materials. (2) With this tool, we studied perceived quality in a real-life setting and identified challenges such as efforts and expectations. (3) We summarize our lessons learned as design recommendations. Altogether, we contribute to research on context-based personalization effects in AI-supported learning and to the iterative development of auto-generated materials for learning.

2 RELATED WORK

This section summarizes prior work that motivates context-based content personalization in learning, including crowdsourced and automated approaches to content generation. Moreover, we discuss how people’s perception of AI-generated content could influence learning with automatically personalized content.

2.1 Context-Aware Content Personalization for Learning

Personalized and adaptive learning can be beneficial for learning performance [47, 66] and increase individual motivation to learn [26, 29]. In this work, we focus on content personalization based on *extrinsic* [54] context factors such as the learner’s location, elements in their surroundings, or media they consume (as opposed to *intrinsic* factors such as the learner’s cognitive state). Thus, we follow a context-aware learning approach [32], where learning is situated in the personal learner context by selecting, adapting, or generating learning content. Context-aware approaches are well-suited for language learning because learners can experience language in an authentic setting [39] and because contextual relevance fosters situational interest or even individual interests, which are important motivational factors for learning [26].

Prior work has explored different context characteristics as triggers for content personalization. Below, we summarize empirical findings and key considerations for location-based, object-based, and media-based personalization that guided our implementation. In the domain of location-based personalization, Edge et al. [19]

¹Crowdsourcing by a community of learners

derived contextually relevant vocabulary via Foursquare. This increased the variance of locations where learners studied in comparison to frequency-based vocabulary suggestions. Hautasaari et al. [22] additionally found benefits for vocabulary recall when audio-based vocabulary was presented in situ.

Object-based approaches integrate concrete elements of the learner's environment into the learning process. For example, a common practice for novice language learners is to attach sticky notes to objects around the house². There are also automated approaches, such as an augmented-reality system proposed by Ibrahim et al. [33] that annotates real-life objects for vocabulary. In a user study, this improved performance on recall tests in comparison to flashcard-based learning. However, a similar system that generated case grammar exercises from detected objects showed no improvement over non-personalized content [17]. There are also multimodal systems such as SCROLL, where learners can create and annotate learning items from life-log photos in a semi-automated manner Ogata et al. [44].

Media-based personalization includes approaches that individually propose media learning material or adapt media for learning. For example, Coleman and Hine [9] used Twitter posts as a source of authentic target language sentences on a topic of the learner's choice. Labutov et al. [37] applied data-mining techniques to extend textbooks with content available online, enabling learners to acquire more information on topics they are interested in. Trusty and Truong [55] replaced words in texts on websites that users visited with target-language words. Thus, the learning material was personalized to the context of the media that learners selected. Their user study showed evidence of vocabulary learning. Similarly, an interactive browser extension by Meurers et al. [43] created grammar exercises in several languages within website texts; however, this system was not evaluated with users. Rüdian et al. [50] explored using auto-correction tools can provide feedback on errors in texts that learners write but found that it was not yet widely applicable for language learners.

In this work, we opted for object-based personalization because it is particularly suitable for interactive approaches where learners themselves select contents they are interested in. Notably, we use photographs that learners in our experimental group take, i.e., a visual representation of their context, including objects in their surroundings. Moreover, the photographs are suitable for generating authentic multimedia learning material that combines visual and verbal information for improved memory processing [42]. Finally, the sense of autonomy and ownership in asking learners to provide photographs as input is likely to benefit the connection to prior knowledge [60].

2.2 Creating Context-Aware Personalized Learning Material

The better the learning materials are matched to an individual learner, the more likely it is that personal goals and interests can be satisfied. On the other hand, an increasing degree of personalization demands scalable solutions that are computationally feasible without increasing extraneous load [59]. Below, we summarize related

work on the two approaches we follow in this paper: crowdsourced learning (in our case, "learnersourcing") and auto-generation.

In learnersourcing, individuals contribute to the overall repository of available learning material [24, 62]. Key advantages of learnersourcing are that the effort for each individual contributor is small and that learners can actually learn while creating content [11]. However, personalization is limited because the content is not explicitly matched to an individual learner. Examples of learnersourcing include student-generated collections of questions [3] or crowdsourced explanations [65] and are supported by sharing platforms such as the Anki flashcard repository³. Lee et al. [38] studied promoting situatedness in learnersourcing, e.g., through authentic roles and micro-tasks in a realistic setting such as website creation. Abou-Khalil et al. [1] recommend vocabulary items to migrants that other migrants in similar situations have logged. And Weir et al. [62] ask learners to label subgoals in video-based learning. As in any crowdsourcing scenario, quality control is essential to avoid that learners work with erroneous material [13]. Typical methods for quality assessment include self-assessment based on a set of criteria, peer rating, and algorithmic checks. When the quality assessment is well done, learning outcomes can match those achieved with material collected by instructors [61, 62, 65].

Automated content adaptation or generation further decreases the effort for personalized content authoring. They often apply rule-based or more advanced algorithms, e.g., natural language processing (NLP), computer vision, or location and activity recognition. NLP approaches include the text-based personalization methods mentioned above (e.g., [37, 43, 55]). Computer-vision projects used object detection [16, 31, 57], reverse image search [52], or caption generation [21] to generate learning content. Other projects created environment-based grammar exercises but relied on fixed object markers [17] or a limited set of predefined objects [18]. Some personalization mechanisms are automatically triggered (e.g., at timed intervals [22] or through RFID tags attached to objects [4]). Others require simple user input (e.g., a search term [9] or selection in a camera preview [17]) to trigger generation processes.

In this paper, we follow a user-triggered approach because it balances relevant results and required effort on the user side. In addition, a sense of learner autonomy can foster intrinsic motivation for learning [49, 56].

2.3 Perception of AI-Generated Content

Auto-generating (personalized) learning content also has implications for the learners' experience. Notably, whether an AI or a human executes an action influences how people assess the performance [8, 27]. In the context of personalized learning, this means that the learners' perception of auto-generated content may be different from that of manually curated materials. Consequently, whether or not learners are willing to work with auto-generated content depends not only on objective performance metrics (e.g., quantified grammatical correctness) but also on the perceived quality and quality expectations. Past studies have assessed the user perception of auto-generated content in related domains. For example, in a study by Chiarella et al. [8], artworks were often rated as less aesthetically pleasing when described as AI-generated than

²see, e.g., <https://www.babbel.com/en/magazine/language-learning-with-sticky-notes>, last accessed 2023/05/04

³<https://ankiweb.net/shared/decks/>, last accessed 2023/02/02

when attributed to a human artist. In addition, a study in the news-writing domain showed that AI-written articles were considered more credible but less readable [20]. The perception of AI-generated content also depends on prior expectations: Hong et al. [28] showed that the attitude towards creative AI was generally positively correlated with liking the auto-generated content. However, they could also observe *expectancy violation* effects [7]. Specifically, participants who were positively surprised by an AI's compositions gave better ratings (violation) than those who had already expected good results (confirmation) and vice-versa. We expect that similar effects play a role in this work.

3 PHOTO FLASHCARDS: CONCEPT AND IMPLEMENTATION

As a reference implementation and study tool for learners' perspectives on auto-generated personalized learning content, we developed the *Photo Flashcards*⁴ app. We implemented three variants for the conditions *auto-personalized*, *auto-learnersourced*, and *manual-learnersourced*, respectively. The *auto-personalized* version app creates authentic content for language learning by captioning photos that learners take, using a state-of-the-art object detection and image classification service. It then generates flashcards with multiple-choice options for learning German case grammar from image captions. Thus, it provides a means for learners to choose content that they personally find interesting. The *auto-learnersourced* and *manual-learnersourced* versions assemble digital flashcards created in the personal version and provide them to other learners as crowdsourced content. This means that the choice of content is limited, but the effort for adding it is lower, as users do not need to actively decide on a motif and take a picture.

Version 1: Personalized Content with Auto-Generation. Flashcards are generated from photos that a user takes from within the mobile app. We process the image and create multiple-choice questions using a variety of caption templates. The overall process for quiz generation is as follows:

- (1) Flashcard generation is triggered with a button in the flashcard overview. This opens a camera preview where learners can take a photo. This photo is sent to our server (without metadata such as camera model or location data).
- (2) The server forwards the photo to a computer vision service, which detects objects, and, in case no objects are found, also retrieves labels that characterize the overall scene or its elements.
- (3) The server ranks the detected objects or image tags by confidence levels. If at least two different objects are found, a heuristic algorithm also determines their relative position from the objects' bounding boxes to create more varied exercises.
- (4) The server selects a random caption template and fills it with information about the detected objects or image tags.
- (5) The caption is automatically translated from English to German to obtain a grammatically correct caption in the target

⁴The source code for the auto-personalized version is provided as supplementary material

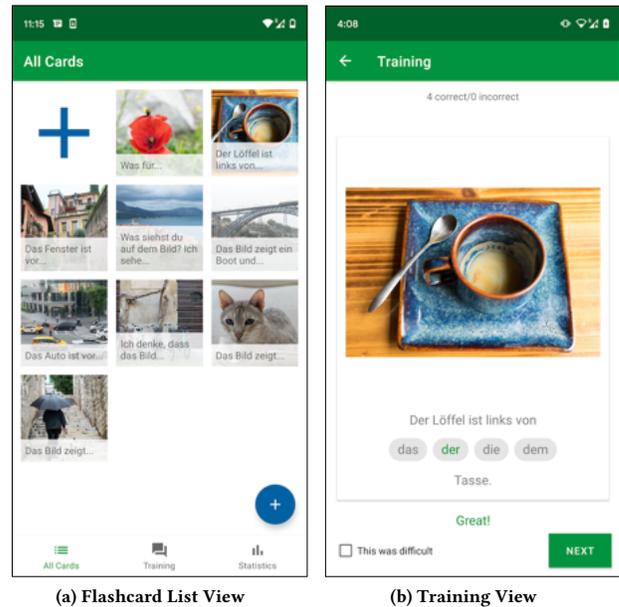


Figure 2: Screenshots of the Photo Flashcards app with flashcards for learning German case grammar.

language. In a few cases, the translations are changed to more commonly used phrases and less formal expressions.

- (6) The server generates multiple-choice options from the translated caption by omitting an article. The provided options are alternative articles (defined or undefined). In some cases, we also omit the word preceding the article to accommodate for contractions (e.g., “von dem” is usually shortened to “vom”).
- (7) The generated multiple-choice quiz is then returned to the user's smartphone and displayed on the device along with the selected photo. Table 2 shows exemplary results.

All created flashcards are presented in a list overview (see Figure 2a), and there is also a detail view for each card. Following the example in [17], the caption templates target German case grammar. The German language uses four cases, which indicate subjects, (in-)direct objects, and possessive constructs. With the case, both undefined and defined articles adjust (e.g., a word with neutral grammatical gender has the article “das” in the nominative case and “dem” in the dative case). Currently, we employ 19 different templates asking for articles in the correct grammar case to be inserted in different places. Exemplary templates are “The picture shows a(n) <object1> and a(n) <object2>” and “I see the <object1>”. By redefining the caption templates, the app can easily be adapted for other languages and topics.

Version 2: Learnersourced Content with Auto-Generated Captions. Instead of requesting a camera image, adding a flashcard in the crowdsourced variant of the Photo Flashcards app means that learners choose from a set of 200 already pre-processed *photo + caption* flashcards retrieved via the personal version. Selected items are added to the personal library. All other interaction remains the same, and the quality of the content is identical to Version 1.

Version 3: Learnersourced Content with Manual Captions. This version provides the images retrieved via the personal version with a human-generated caption and multiple-choice exercise. It served as a benchmark for the auto-generation. We used the same 200 images as for the *auto-learnersourced* version, and all user interaction was the same. Two researchers formulated captions in a similar style to the caption templates, and the first author translated them and transformed them into exercises.

Training and Additional Features. To measure recall and engagement with the app, we added a revision feature: Users can start a training round, where a random selection of multiple-choice quizzes [5] for their flashcards is presented one after the other (see Figure 2b). By setting the number of questions per round to 15, we follow a microlearning approach [30]: learning content is divided into small units that can be completed in short learning sessions and, therefore, allow flexible scheduling. At the end of a training round, the app presents a summary of the performance in the current round. Besides the training, additional features include a performance history view displaying the number of cards added, the average number of attempts per quiz, and the time elapsed since the last training round. Daily push notifications with the message “Add new cards to continue with the Photo Flashcard Study” remind learners to keep participating in the study. With these features, we aim to increase overall engagement.

Implementation Details. The client system was implemented as an Android app and runs on devices with Android 6 or newer. It is supported by an Express server that provides the quiz generation functionality, launches the requests to object detection and translation APIs, and anonymously logs requests as well as the interaction with the app. For object detection and tagging, we currently use the Google Vision API⁵, a top-rated system for image recognition⁶. The Translate API⁷ is used for translating sentences to our target language. We found these algorithms to be reliable overall, but the app is designed such that APIs can easily be exchanged. In order to ease the learners' familiarization process, we further included nine predefined flashcards.

4 USER STUDY

We evaluated the user perspective on personalized auto-generation of learning materials in an in-situ study with a mixed-methods between-groups design. The study was conducted using the online survey platform *Prolific*⁸. Group A used the personal version of the Photo Flashcards app, i.e., they created flashcards from photos taken while using the app (*auto-personalized condition*). Participants in Group B (*auto-learnersourced condition*) and Group C (*manual-learnersourced condition*) could choose flashcards based on the photos taken by participants of Group A. Thus, the type and quality of flashcards were identical in *auto-personalized condition* and *auto-learnersourced condition*, but only the participants in the *auto-personalized condition* were able to generate flashcards from their own environment. On the other hand, adding flashcards in the

auto-learnersourced condition and *manual-learnersourced condition* was a simpler and potentially faster process overall. We opted for an in-situ study because we were particularly interested in real-life usage scenarios and user experiences after repeated use [36], even though this can introduce confounds [48].

4.1 Measures

We collected data via a pre-study and a post-study questionnaires and the app usage logs. Demographic information was collected via Prolific. For RQ1—validity of the auto-generation—we measured the usability, user engagement, and performance in German tests with the three versions of the app. *Usability* was measured with the System Usability Scale [6] and with open-ended questions on what learners liked or disliked. *User engagement* was derived from app usage activity, notably the number of added flashcards and the number of training sessions. Prior knowledge of German was assessed with a translation task, and post-intervention knowledge with an image description task. For RQ2, we asked learners to rate the *perceived quality* (correctness, understandability, and relevance) of the learning material with 5-point Likert scales. A full list of questionnaire measures can be found in Appendix A.

4.2 Procedure

As the photos taken in the *auto-personalized condition* were utilized as input for the *auto-learnersourced condition* and *manual-learnersourced condition*, the study was organized in two phases, starting with the *auto-personalized condition*. In all conditions, the first step for participants was to read and consent to the study information and our data processing guidelines on Prolific before downloading the Photo Flashcards app from the Play Store. When they first opened the app, they were asked to enter their Prolific ID and to fill in a short pre-study questionnaire asking about their prior experience with German and mobile language learning. In addition, they completed a short translation exercise and viewed an introduction to the main features of the respective version of the app. On submitting the form, the remaining UI of the app was unlocked, and participants used the app at their own pace. They could add new flashcards, view flashcards, and complete training rounds based on the existing cards whenever and as often as they wanted. In the *auto-personalized condition*, participants took photos with the app to trigger flashcard generation (cf. Figure 3). Before processing a photo, we asked them for permission to store and use it for subsequent steps of the study. In the *auto-learnersourced condition* and *manual-learnersourced condition*, participants could add photo flashcards by choosing them from a predefined list compiled from the *auto-personalized condition* photos we were allowed to use⁹. After approximately two weeks of app usage, the participants received an invitation for the post-study questionnaire.

The study procedure adhered to ethics standards at our institution, and we obtained ethics approval. The questionnaires are provided as supplementary material.

⁵<https://cloud.google.com/vision>, last accessed 2023/02/03

⁶<https://www.perficient.com/insights/research-hub/image-recognition-accuracy-study>, last accessed 2023/02/03

⁷<https://cloud.google.com/translate>, last accessed 2023/02/03

⁸<https://prolific.co>, last accessed 2023/02/02

⁹Overall, there were 412 flashcards, and we were granted permission for using 403 of these. We manually removed all photos and that showed identifiable people, were completely blank, or looked identical to another photo. Finally, we randomly capped the list to 200 photos.

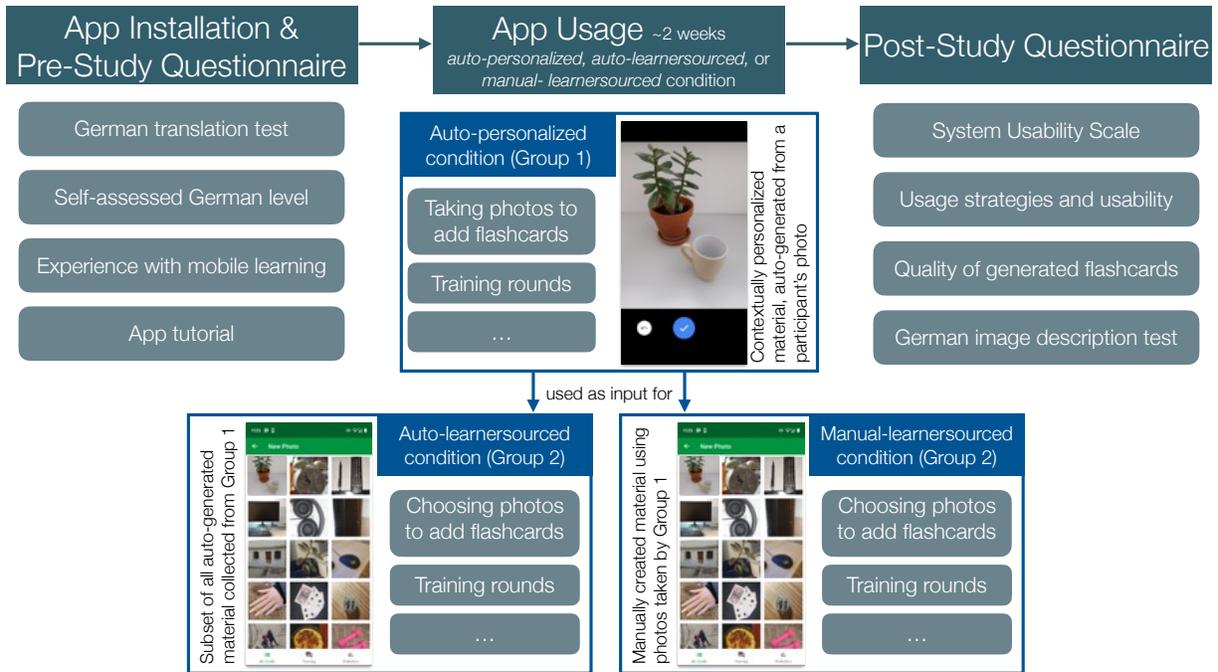


Figure 3: Overview of the between-groups study procedure. The photos taken by participants in the *auto-personalized condition* were used as input for the *auto-learnersourced condition* and the *manual-learnersourced condition*, hence the *auto-personalized condition* was run as the first condition.

4.3 Participants

Participants were recruited via Prolific. We filtered potential participants to users with Android devices and conducted a pre-screening to identify people that had studied German for at least three months and had a general interest in learning the language. The 64 data sets used for analysis only include participants that added at least one photo so that we could be sure that they had actually used condition-specific features of the Photo Flashcards app. These comprise 24 participants in the *auto-personalized condition*, 17 in the *auto-learnersourced condition*, and 23 in the *manual-learnersourced condition*. For two participants, the pre-study questionnaire was not recorded correctly, and another two participants did not complete the post-study questionnaire. They were excluded from the respective analyses.

Eighteen participants identified as female, 45 as male, and data for one participant was missing (*auto-personalized condition*: 6f/18m; *auto-learnersourced condition*: 5f/11m/1na; *manual-learnersourced condition*: 7f/16m). Their ages ranged from 18 to 62 years ($M = 28.4$, $SD = 11.3$). The most frequent native languages were Polish (18) and English (17). The self-assessed level of German on the European reference scale¹⁰ (CEFR) was predominantly A1 and A2, i.e., at a beginner’s level (88.7%). Six participants selected B1 (4 in the *manual-learnersourced condition*, 2 in the *auto-learnersourced condition*), and one *manual-learnersourced* participant selected C1. The absolute German knowledge did not have an impact on our analyses, as we only measured relative changes and never compared

absolute German knowledge. Participants were compensated with £2.70 for participating in the surveys posted on Prolific and a bonus of £8 if they added a new photo (i.e., flashcard) on at least 12 days. Forty-two participants received the full payment of £10.70.

5 RESULTS OF THE USER STUDY

For quantitative results, we report Bayes Factors BF_M of Bayesian ANOVAs, showing the likelihood of the data under a model including the independent variable in comparison to the null model. In other words, this is the probability of including the independent variable when modeling the data. We add Bayesian post-hoc tests, where $BF_{10} > 1$ indicates that the alternative hypothesis H_1 is more likely than the null hypothesis H_0 , while $BF_{10} < 1$ indicates the opposite [58]. Using a Bayesian approach allows us to analyze the results in an exploratory fashion, comparing the likelihood of different models. The Bayes Factors also reflect uncertainty caused by factors such as small sample sizes [35]. As a reference, we additionally report frequentist ANOVAs. When a Levene’s test indicated a violation of homogeneity, we applied a Welch correction. For isolated Likert-scale items, we use Kruskal-Wallis tests with Games-Howell post-hoc tests. All tests were computed with JASP [34]. Moreover, we include participant quotes from the questionnaires that showcase opinions or possible explanations for individual ratings on aspects such as ease of use. Here, we use consecutive participant numbers with the prefix “P” for the *auto-personalized condition*, “L” for the *auto-learnersourced condition*, and “M” for the *manual-learnersourced condition*.

¹⁰<https://www.coe.int/en/web/language-policy/cefr>, last accessed 2023/02/02

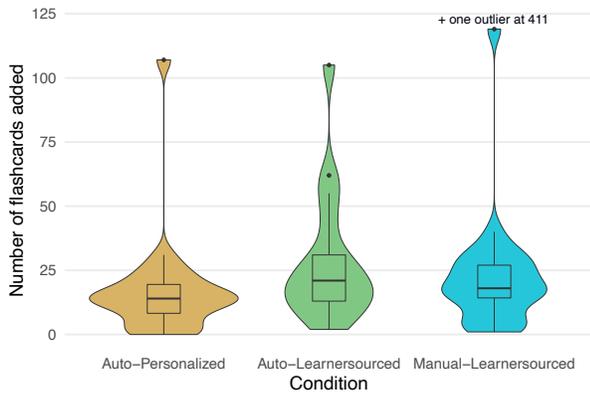


Figure 4: Distribution of the number of added flashcards.

5.1 RQ 1: Validating the Auto-Generation

First, we compare the two app variants with auto-generated material against the one with manually generated material, with a focus on usability, user engagement, and performance gains from the language pre-test to post-test.

Usability. Overall, the Photo Flashcards app was considered easy to use in all conditions. The personal version of the app obtained a slightly lower average SUS score than the *auto-learnersourced* ($BF_{10} = 1.058$) and *manual-learnersourced* versions ($BF_{10} = 2.376$; see Table 1). There were 21 explicit mentions of the ease of use in the post-study questionnaire (seven in the *auto-personalized condition*). One difference in the interaction that may have contributed to the different scores was the number of steps from opening the camera or picture chooser to the display of the final flashcard (interaction time + server processing). This means that the overall process for adding a flashcard in the *auto-personalized condition* took longer than adding a pre-processed flashcard. Moreover, for some pictures, it was not possible to create flashcards because a server bug prevented the correct parsing of captions already in use for another image.

User Engagement. We counted the number of flashcards and the number of training sessions that participants started as indicators of engagement. Quantitative results of these measures are summarized in Table 1. Overall, during the study period, the study participants added 1,816 flashcards. Figure 4 illustrates the distribution of added cards per participant. There were no clear differences in the number of flashcards added in each condition. On the other hand, there was strong evidence indicating that the number of training sessions per day in the *auto-personalized condition* was lower than in the *auto-learnersourced condition* ($BF_{10} = 6.820$) and the *manual-learnersourced condition* ($BF_{10} = 2.918$). Because of a logging error in the first two conditions, we were not able to capture the overall usage time.

Performance in German Tests. To assess learning effects, we compared normalized scores in the German tests before and after app usage (between 0 and 1, where 1 was perfect performance). The tests after app usage were slightly more difficult as they required free recall instead of translation; this explains the negative mean

value in the *manual-learnersourced condition*. Overall, changes were small, and we could observe no difference between the conditions.

5.2 RQ 2: Personalization and Perceived Flashcard Quality

The app variant was an important predictor of perceived correctness at $BF_M = 33.380$ and understandability at $BF_M = 6.340$, and lower for relevance at $BF_M = 1.302$. Post-hoc tests showed that the correctness and relevance of flashcards were rated highest in the *auto-learnersourced condition* and lowest in the *auto-personalized condition*, while understandability was highest for the *manual-learnersourced condition* (see Figure 5). For example, P22, who had rated the correctness as *bad*, commented that “the worst thing about the app was the A.I. Image Recognition, I often had to check whether the word the A.I. generated was correct”. Another participant pointed out that the “quizzes could be more about the object presented in the photo” (P14). Several participants felt that the automatic generation process was not good enough (e.g., P2, P10, P22) or that the variety of questions was too limited (e.g., P22). In the *auto-learnersourced condition*, L11 recommended “using better quality photos and mak[ing] the text more evident and understandable.” Similarly, L17 remarked that the quality of some of the proposed images was rather low.

5.3 RQ 3: Opportunities and Challenges

The participants’ responses to the open-ended questions illustrate what learners generally think about (personalized) auto-generation and how they engage with their variant of the Photo Flashcards app, but also what needs to be improved before adopting such an app in everyday life.

Comments on the Concept. Eleven of the 24 participants in the *auto-personalized condition* explicitly mentioned that they liked the overall idea of creating their own flashcards. For example, P13 explained that “the experience [is] really customized and personal with the addition of your own photos” and P9 that it is “like a game, you can learn [while] you are having fun” (P9). P21 found the “idea of making pictures [...] amazing”, as it “makes studying better” and “makes you actively learn about your surroundings.” P3 described the app as “one of the best systems [they] used to learn. It combine[s] learning with fun and doing photos.” P22 considered the app “a nice addition” but felt that it “cannot replace other learning methods.” P9 suggested “topics [that] can be filled [...] with relevant photos.” In the *auto-learnersourced condition* and *manual-learnersourced condition*, six participants said that they liked the visual focus of the app and the associations of words with objects.

Usage Patterns. An essential part of understanding engagement with the app is the type of content they add and in what situations they do so. For example, participants mentioned situations where they wanted to know the names of an object (P20, P22, M19), everyday objects (P13, L17, M7), or random objects (P38, L4, M10, M11). These choices are also reflected in the motifs: in the *auto-personalized condition*, 76.5% of photos we could analyze showed everyday objects, followed by outdoor scenes (10.9%), food (7.4%), and pets (3.0%).

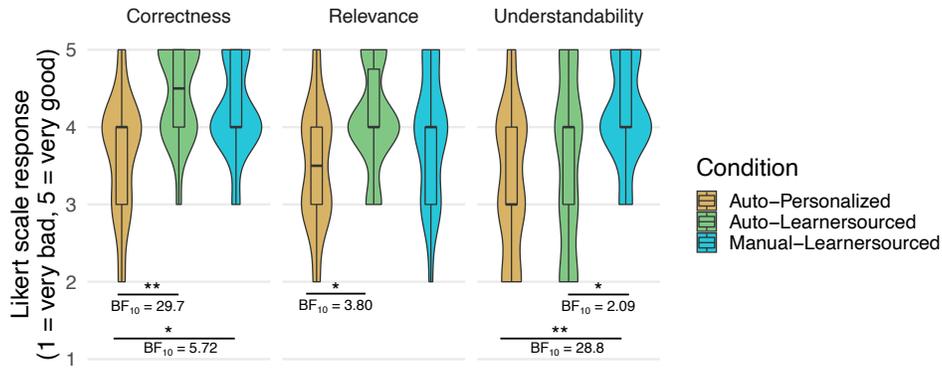


Figure 5: “How would you rate the quality of the generated quizzes regarding their correctness, understandability, and relevance?” – Perceived quality of the auto-generated and human-generated flashcards, including post-hoc Bayes Factors and significance markers for Games-Howell tests.

Table 1: Mean average, Bayes Factors indicating the likelihood of the data under a model including the factor *condition*, and ANOVA results of engagement, learning, and usability values obtained in the *auto-personalized condition*, the *auto-learnersourced condition*, and the *manual-learnersourced condition*.

	<i>auto-personalized</i>		<i>auto-learnersourced</i>		<i>manual-learnersourced</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Added flashcards	17.2	20.9	28.6	25.7	39.9	84.2
	$F(61) = 1.050, p = 0.356, \eta^2 = 0.033$				$BF_M = 0.289$	
Training sessions per day	0.28	0.27	0.63	0.51	0.58	0.56
	$F(32.265) = 5.205, p = 0.011, \eta^2 = 0.113$				$BF_M = 2.318$	
Change in normalized German test score	0.17	0.35	0.04	0.26	-0.01	0.25
	$F(57) = 1.97, p = 0.149, \eta^2 = 0.065$				$BF_M = 0.594$	
System Usability Scale	71.8	12.3	79.3	14.4	80.2	11.9
	$F(59) = 2.792, p = 0.069, \eta^2 = 0.086$				$BF_M = 1.047$	

Feature Suggestions. For future developments, eight participants requested English translations as an additional feature, as this would improve their understanding of the captions. P19, L4, and L7 would add pronunciation support, and P18 and L13 missed grammar explanations. Some participants wanted to see more interactivity, such as the ability to create their own cards (L4, M7, M12, M13) and the ability to provide feedback (M2). L4 would like to have flashcards sorted by category. When adding cards, photos already in use should be filtered out (L3).

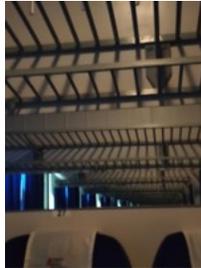
6 TECHNICAL EVALUATION OF THE FLASHCARD GENERATION

To gain additional insights into how well the personalized auto-generation worked from a technical perspective and to follow up on the points raised in Section 5.3, we performed an in-depth analysis of the flashcards collected during the user study. These flashcards provided us with authentic examples of photo flashcards that learners generate from their environment.

We assessed the 403 flashcards that were successfully created from the photos that participants in the *auto-personalized condition* took and that we were granted permission to store. Specifically, we analyzed the object detection results, the match of captions and images, and the caption translations. Thus, we analyzed the performance of the object detection in a real-world use case rather than on existing data sets commonly used for evaluation. We manually checked the final flashcards for all available photos based on a fixed list of criteria for the correct detection of objects, their salience in the picture, and the computed relative position (if applicable). In the second step, we checked if the translation of the caption still matched the photo.

Table 2 gives an overview of the final captions shown on the flashcards. *Correct and salient* comprises all flashcard captions with a correct caption that matches the image well. To fulfill this criterion, all objects detected in a picture must be salient objects in the scene. If several objects are included, their relative position must be correct. Similarly, image labels included in the caption

Table 2: Quality classification and example images of auto-generated captions (translated to English)

Correct & Salient	Correct & Not Salient	Correct & Imprecise	Incorrect
189 (46.9%)	23 (5.7%)	21 (5.2%)	170 (42.2%)
 <p>What do you see in the picture? I see a houseplant.</p>	 <p>Here you can see a toy and a computer keyboard.</p>	 <p>What is behind the headphones? There are packaged goods.</p>	 <p>What a beautiful grooming trimmer!</p>
 <p>The candle is in front of the stuffed toy.</p>	 <p>The picture shows a car.</p>	 <p>I would describe this image as an image of a concrete.</p>	 <p>Do you see the shoe in this picture?</p>

must match the scene (e.g., “landscape”). *Correct, but not salient* refers to captions where one or more objects were correctly detected, but a human would most likely not consider them the most relevant items of the scene. *Correct, but imprecise* captions contain generic descriptions, e.g., “packaged goods” or “luggage & bags”. In addition, some sentences contained countable versions of words that are often used as mass nouns (e.g., “fruit” better matched the context than “a fruit”). Whenever an error occurred in the process, we labeled the caption as *Incorrect*. This includes incorrect object detection, wrong translations (8 instances), grammar mistakes (6 instances), or capitalization (2 instances).

7 DISCUSSION

Research in psychology and pedagogical sciences readily attest to the importance of interactive learning material that is personally and contextually relevant [26, 42, 47]. Auto-generation is a promising approach to reducing the immense effort required to personalize such material. Our study shows that artificial intelligence methods such as computer vision and automatic translation can generate material for language learning and that it comes close to manually prepared material in terms of usability, engagement, and short-term learning performance (RQ 1). However, it also reveals that personalization may actually harm perceived quality if the generation does not live up to users’ expectations (RQ 2). Participant statements and an in-depth analysis of the generated material provide insights into the opportunities and challenges of personalized auto-generation

(RQ 3). Below, we discuss these findings in more detail and summarize our lessons learned as design recommendations for future systems.

7.1 Context-Aware Personalization in Auto-Generation for Learning

Contrary to our assumptions, learners did not benefit from personalization in our scenario. We discuss the learners’ expectations and effort as two possible reasons.

Personal relevance affected the perceived quality of auto-generated learning material. Learning material derived from personal photos was rated lower for correctness, relevance, and understandability compared to non-personal learnersourced photos, even though the actual material was identical. There are at least two explanations for this. The personal significance of one’s own photos could bias how one describes the image, resulting in greater dissonance with an auto-generated description. In addition, there may have been an expectancy violation [7, 28]: learners may have higher prior expectations of object detection performance for images they are personally familiar with and were subsequently negatively surprised by the detection results. Learners who only chose a picture may have had lower expectations, and therefore, their evaluation of the result was less negative. In our scenario, we further identified several instances of picture series showing the same motif, which suggest that participants were unhappy with the initial results. So

while participant liked using their own photographs, the generated materials were not deemed sufficient for improved learning or engagement.

Creating personal flashcards requires effort. Another reason may be the required effort. Adding flashcards required more active engagement in the *auto-personalized condition*, as selecting a motif and taking a picture requires more steps and more decisions than scrolling through a list and tapping on an existing item. Future projects should minimize the perceived effort required to create personal learning content to avoid compromising interest [2]. In Section 8, we list two concrete recommendations, namely batch processing and themed challenges.

7.2 Lessons Learned from the Deployment

Our app extends the space of auto-generating material for learning. And while it was sufficient for studying personalization challenges (RQ 1), it is by no means a perfect solution. Through the *in-situ* deployment of the Photo Flashcards app and the technical analysis, we also gained insights into challenges and opportunities (RQ 3).

Quality varies, and learners cannot always tell. The flashcard generation varied from very low to very high quality. However, learners often lack the language proficiency to detect mistakes in incorrectly generated captions. Compared to human-generated materials, learners perceived auto-generated non-personalized learning materials to be equally relevant and correct, albeit less understandable. So while the auto-generated material may be sufficiently relevant and correct from a learner’s perspective, verifying material quality is essential, especially in light of the fact that incorrectly remembered information can persist even when it has been revoked (*continued influence bias*; [40]). This raises two challenges: identifying and dealing with low-quality results. We discuss possible strategies for this in Section 8. More broadly speaking, this is related to current discussions in human-AI interaction and human-centered AI that also highlight iterative and collaborative approaches as a solution to work with imperfect AI algorithms (e.g., [53, 63]). Furthermore, the interaction patterns we observed where learners took multiple similar pictures resembles prompt refinement in working with large-language models [12] or exploratory search patterns [64] to gradually approach targeted results. Similarly, the image generation model DALL-E¹¹ automatically proposes several results that users can pick from.

Translation engines can iron out inconsistencies. The fact that popular CV algorithms are almost exclusively trained with English-language labels means that translations are required. However, in our case, the translations were actually useful because the Translate API proved robust with respect to inconsistencies (e.g., “a”, “an”) in our sentence construction, as the Translation API was robust to such errors. Thus, the translations worked exceedingly well—with very few exceptions. Of course, this may be different for languages that are less similar than English and German. On the learner side, adding translations to the quizzes—which some participants requested—could also serve as a means to identify incorrect captions.

¹¹<https://openai.com/blog/dall-e/>, last accessed 2023/02/02

7.3 Limitations

The current study explored the possible contribution of automated content generation towards mobile-assisted personalized language learning. In other words, it is intended to innovate a procedural step of mobile-assisted language learning. We assumed that this automated procedure could reduce the effort of personalization and mitigate moments of low participant motivation, which has been reported even with the use of popular language learning apps, such as Duolingo [41], as well as make the content more personally relevant. While personalized content has been shown to facilitate individual interest and motivation [25, 59], we found effort and expectations to be limiting factors. The current work is also limited by the constraints and challenges of the *in-situ* deployment. While it enabled us to study learner experiences in a realistic usage scenario, a longitudinal study will be required to investigate the impact of automatically generated flashcards with context-aware personalization on learning and to validate the user experience and engagement in comparison to traditional methods. In addition, we compared the automated procedure to a manual caption approach. However, we did not provide a comparison to personalized manually generated material: this would have introduced time delays required for the individual processing and, consequently, would have additionally influenced the user experience.

8 RECOMMENDATIONS FOR USING CONTEXT-AWARE AUTO-GENERATED LEARNING MATERIAL IN PRACTICE

To address the challenges mentioned in the discussion, this section summarizes recommendations based on the study results and lessons learned during the development and deployment.

Balancing correctness and personalization. To balance the correctness of learning material and the load on learners and instructors, a possible solution could be crowdsourced support. We suggest aligning material with the learners’ level of proficiency through a network of peer and instructor support. In detail, we propose a multi-step assessment and correction based on the quality categories derived in Section 6. The first step is the categorization. This should be performed by an instructor or native speaker but can be completed within seconds. Following the initial assessment, the learning material can then be treated as follows:

- **Correct and salient:** No action necessary. The material can be used as is.
- **Correct, but not salient:** No action necessary for high-proficiency learners. We expect that they can integrate the information with their prior knowledge. For beginners, the material could be annotated. For example, the bounding boxes of detected objects could be highlighted to clarify references. If not already provided by the CV algorithm, these can be automatically proposed [23] and manually confirmed.
- **Correct, but imprecise:** No action necessary for high-proficiency learners. For beginners, we see two main opportunities: (1) leveraging the crowd to propose alternatives (e.g., replacing “animal” with “cat”), (2) providing a feedback channel by annotating the material (e.g., with a translation to the learner’s native language).

- **Incorrect:** This is the most urgent category that demands correction by an instructor or native speaker. Once a critical mass of users has been reached, this could be approached with concepts as employed by *Be My Eyes*¹²: native speakers could caption images for learners.

Furthermore, we recommend primarily using already verified learnersourced items for beginners and personalization mostly for high-proficiency learners. Ultimately, with such a pipeline in place, we hope that this can re-establish the added value typically found for personalization in learning [47].

Interaction design for lower perceived load. The perceived effort for authoring flashcards should be as low as possible [45]. In particular, suggestions and challenges could be an incentive. For example, users could first take a series of photos in a spare moment, which could then be batch-processed in the background. This would speed up the perceived creation time. If learners lack inspiration or are not sure what kind of photos to take, themed challenges (e.g., “find your favorite items in this room” or “take a picture outside”) could serve as an incentive to keep adding material. Another possible approach is content generation without explicit user input, e.g., by guessing interesting items based on eye gaze. Follow-up studies could also investigate how more effortful (yet personalized) approaches could be intermixed with less effortful (yet more generic) approaches in sync with the learners' low and high periods of participant motivation. This could be an integrative approach that would provide learners with the benefit of personalization as well as the flexibility to cope with variable motivation. This could be achieved by adapting the level of personalization to the learner's ongoing agenda, context, or even their circadian rhythm [14, 15].

Mitigating biases and preserving privacy. Current debates on stereotypes reproduced by generative models stress the importance of ethical considerations. In our case, exercises were primarily created from objects or abstract nouns, and those that referred to a person did not imply a gender (e.g., “person”). However, when content generation is extended to other domains, it is essential to assess potential biases. Crowdsourcing has also been found to confirm stereotypes [46], which means that it cannot be relied on for resolving biases. Instead, active counter-strategies need to be integrated.

Moreover, individuals must be able to control which data is shared and with whom. For example, users should be asked before images taken in private surroundings are shared with others. In our app, we provided a check box to give permission before analyzing an image. Future systems should also give the option to retrospectively delete shared data. Data sharing could also be limited to a network of trusted peers for small-scale crowd feedback.

Redundant information. To avoid continued influence biases, interfaces should include a means for learners to check the quality of generated language material (besides the mitigation strategies listed above). In addition to translations, image search results for detected objects would provide a simple means to do so: If the retrieved image matches the original image, the detection is very likely to be correct.

¹²<https://www.bemyeyes.com>, last accessed 2023/02/02

Algorithm improvements for better and more varied results. We recommend adding salience detection to identify the most likely regions of interest. This could be done with heuristic methods, e.g., using the size and position of bounding boxes or with special-purpose algorithms [10]. In addition, provisioning templates for vocabulary and different grammar concepts would enable a broader application of captions for language learning purposes. For example, it would be possible to generate multiple-choice quizzes from detected objects to study vocabulary or verb forms to study conjugation.

9 CONCLUSION

The auto-generation of learning material with AI technologies, such as image captioning and object detection, enables personalized learning experiences. At the current state of technology, varied and sufficiently accurate material can be created. Using our mobile app Photo Flashcards, we explored this approach to content generation with a novel computer vision system. However, our *in-situ* user study with 64 participants showed that learners perceived the quality of auto-generated learning material to be lower when it was personalized than when it was chosen from a crowdsourced library, even though the material in the two auto-generation conditions was actually identical. So while the Photo Flashcards system facilitates the personalization of learning material and follows the trend towards self-directed and seamless ubiquitous learning, the perceived quality compromised the benefits of personalization. Based on the study and quality assessments of the generated learning material, we summarize opportunities and challenges in content generation for learning and propose design recommendations.

In future work, we plan to deploy a revised version of the Photo Flashcards app in a real-life setting. In particular, we will focus on adapting the user experience such that personalization actually fosters learning. Based on our insights, we will further explore usage for other languages and linguistic concepts as well as potential additional application domains, such as science education.

ACKNOWLEDGMENTS

We thank Eleanor Colligan for her support with the data collection and evaluation. We also thank the reviewers of this paper for their helpful comments. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. LLC was supported by the research initiative “Instant Teaming between Humans and Production Systems” co-financed by tax funds of the Saxony State Ministry of Science and Art (SMWK3-7304/35/3-2021/4819) on the basis of the budget passed by the deputies of the Saxony state parliament.

REFERENCES

- [1] Victoria Abou-Khalil, Samar Helou, Mei-Rong Alice Chen, Brendan Flanagan, Louis Lecaille, Niels Pinkwart, and Hiroaki Ogata. 2021. Vocabulary recommendation approach for forced migrants using informal language learning tools. *Universal Access in the Information Society* (May 2021). <https://doi.org/10.1007/s10209-021-00813-3>
- [2] Marilyn P. Arnone, Ruth V. Small, Sarah A. Chauncey, and H. Patricia McKenna. 2011. Curiosity, interest and engagement in technology-pervasive learning environments: a new research agenda. *Educational Technology Research and Development* 59, 2 (April 2011), 181–198. <https://doi.org/10.1007/s11423-011-9190-9>

- [3] Simon P. Bates, Ross K. Galloway, Jonathan Riise, and Danny Homer. 2014. Assessing the quality of a student-generated question repository. *Physical Review Special Topics - Physics Education Research* 10, 2 (July 2014), 020105. <https://doi.org/10.1103/PhysRevSTPER.10.020105>
- [4] Jennifer S. Beaudin, Stephen S. Intille, Emmanuel Munguia Tapia, Randy Rockinson, and Margaret E. Morris. 2007. Context-Sensitive Microlearning of Foreign Language Vocabulary on a Mobile Device. In *Ambient Intelligence*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Bernd Schiele, Anind K. Dey, Hans Gellersen, Boris De Ruyter, Manfred Tscheligi, Reiner Wichert, Emile Aerts, and Alejandro Buchmann (Eds.), Vol. 4794. Springer Berlin Heidelberg, Berlin, Heidelberg, 55–72. https://doi.org/10.1007/978-3-540-76652-0_4 Series Title: Lecture Notes in Computer Science.
- [5] Elizabeth Ligon Bjork, Jeri L. Little, and Benjamin C. Storm. 2014. Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition* 3, 3 (Sept. 2014), 165–170. <https://doi.org/10.1016/j.jarmac.2014.03.002>
- [6] John Brooke and others. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
- [7] Judee K. Burgoon. 1993. Interpersonal Expectations, Expectancy Violations, and Emotional Communication. *Journal of Language and Social Psychology* 12, 1-2 (March 1993), 30–48. <https://doi.org/10.1177/0261927X93121003>
- [8] Salvatore G. Chiarella, Giulia Torromino, Dionigi M. Gagliardi, Dario Rossi, Fabio Babiloni, and Giulia Cartocci. 2022. Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Computers in Human Behavior* 137 (Dec. 2022), 107406. <https://doi.org/10.1016/j.chb.2022.107406>
- [9] Graeme W. Coleman and Nick A. Hine. 2012. Twasebook: a "crowdsourced phrasebook" for language learners using Twitter. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction Making Sense Through Design - NordiCHI '12*. ACM Press, Copenhagen, Denmark, 805. <https://doi.org/10.1145/2399016.2399157>
- [10] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. 2019. Review of Visual Saliency Detection With Comprehensive Information. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (Oct. 2019), 2941–2959. <https://doi.org/10.1109/TCSVT.2018.2870832>
- [11] Gabriel Culbertson, Solace Shen, Erik Andersen, and Malte Jung. 2017. Have your cake and eat it too: Foreign Language Learning with a Crowdsourced Video Captioning System. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 286–296. <https://doi.org/10.1145/2998181.2998268>
- [12] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. (2022). <https://doi.org/10.48550/ARXIV.2209.01390> Publisher: arXiv Version Number: 1.
- [13] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *Comput. Surveys* 51, 1 (Jan. 2018), 1–40. <https://doi.org/10.1145/3148148>
- [14] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building Cognition-Aware Systems: A Mobile Toolkit for Extracting Time-of-Day Fluctuations of Cognitive Performance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 1–15. <https://doi.org/10.1145/3132025>
- [15] Fiona Draxler, Julia Maria Brenner, Manuela Eska, Albrecht Schmidt, and Lewis L Chuang. 2022. Agenda- and Activity-Based Triggers for Microlearning. In *27th International Conference on Intelligent User Interfaces*. ACM, Helsinki Finland, 620–632. <https://doi.org/10.1145/3490099.3511133>
- [16] Fiona Draxler, Laura Haller, Albrecht Schmidt, and Lewis L. Chuang. 2022. Auto-Generating Multimedia Language Learning Material for Children with Off-the-Shelf AI. In *Mensch und Computer 2022 - Tagungsband*, Bastian Pfefling, Kathrin Gerling, and Sven Mayer (Eds.). ACM, New York, 96–105. <https://doi.org/10.1145/3543758.3543777>
- [17] Fiona Draxler, Audrey Labrie, Albrecht Schmidt, and Lewis L. Chuang. 2020. Augmented Reality to Enable Users in Learning Case Grammar from Their Real-World Interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. <https://doi.org/10.1145/3313831.3376537>
- [18] Fiona Draxler, Elena Wallwitz, Albrecht Schmidt, and Lewis L. Chuang. 2020. An Environment-Triggered Augmented-Reality Application for Learning Case Grammar. In *DELFI 2020 - Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V.*, Raphael Zender, Dirk Ifenthaler, Thiemo Leonhardt, and Clara Schumacher (Eds.). Gesellschaft für Informatik e.V., Bonn, 389–390.
- [19] Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A. Landay. 2011. MicroMandarin: mobile language learning in context. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, Vancouver, BC, Canada, 3169. <https://doi.org/10.1145/1978942.1979413>
- [20] Andreas Graefe, Mario Haim, Bastian Haarmann, and Hans-Bernd Brosius. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism* 19, 5 (May 2018), 595–610. <https://doi.org/10.1177/1464884916641269>
- [21] Mohammad Nehal Hasnine, Brendan Flanagan, Gokhan Akcapinar, Hiroaki Ogata, Kousuke Mouri, and Noriko Uosaki. 2019. Vocabulary Learning Support System Based on Automatic Image Captioning Technology. In *Distributed, Ambient and Pervasive Interactions*, Norbert Streitz and Shin'ichi Konomi (Eds.), Vol. 11587. Springer International Publishing, Cham, 346–358. https://doi.org/10.1007/978-3-030-21935-2_26 Series Title: Lecture Notes in Computer Science.
- [22] Ari Hautasaari, Takeo Hamada, Kuntaro Ishiyama, and Shogo Fukushima. 2019. VocaBura: A Method for Supporting Second Language Vocabulary Learning While Walking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (Dec. 2019), 1–23. <https://doi.org/10.1145/3369824>
- [23] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. 2019. Bounding Box Regression With Uncertainty for Accurate Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 2883–2892. <https://doi.org/10.1109/CVPR.2019.00300>
- [24] Neil T. Heffernan, Korinn S. Ostrow, Kim Kelly, Douglas Selent, Eric G. Van Inwegen, Xiaolu Xiong, and Joseph Jay Williams. 2016. The Future of Adaptive Learning: Does the Crowd Hold the Key? *International Journal of Artificial Intelligence in Education* 26, 2 (June 2016), 615–644. <https://doi.org/10.1007/s40593-016-0094-z>
- [25] Catherine Regina Heil, Jason S. Wu, Joey J. Lee, and Torben Schmidt. 2016. A Review of Mobile Language Learning Applications: Trends, Challenges, and Opportunities. *The EuroCALL Review* 24, 2 (Sept. 2016), 32–50. <https://doi.org/10.4995/eurocall.2016.6402>
- [26] Suzanne Hidi and K. Ann Renninger. 2006. The Four-Phase Model of Interest Development. *Educational Psychologist* 41, 2 (June 2006), 111–127. https://doi.org/10.1207/s15326985ep4102_4
- [27] Joo-Wha Hong, Ignacio Cruz, and Dmitri Williams. 2021. AI, you can drive my car: How we evaluate human drivers vs. self-driving cars. *Computers in Human Behavior* 125 (Dec. 2021), 106944. <https://doi.org/10.1016/j.chb.2021.106944>
- [28] Joo Wha Hong, Qiyao Peng, and Dmitri Williams. 2021. Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society* 23, 7 (July 2021), 1920–1935. <https://doi.org/10.1177/1461444820925798>
- [29] Ching-Kun Hsu. 2015. Learning motivation and adaptive video caption filtering for EFL learners using handheld devices. *ReCALL* 27, 01 (Jan. 2015), 84–103. <https://doi.org/10.1017/S0958344014000214>
- [30] Theo Hug. 2007. *Didactics of microlearning*. Waxmann Verlag. <https://books.google.de/books?hl=de&lr=&id=j0-KAwAAQBAJ&oi=fnd>
- [31] Brandon Huynh, Jason Orlosky, and Tobias Hollerer. 2019. In-Situ Labeling for Augmented Reality Language Learning. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Osaka, Japan, 1606–1611. <https://doi.org/10.1109/VR.2019.8798358>
- [32] Gwo-Jen Hwang, Chin-Chung Tsai, and Stephen JH Yang. 2008. Criteria, strategies and research issues of context-aware ubiquitous learning. *Journal of Educational Technology & Society* 11, 2 (2008).
- [33] Adam Ibrahim, Brandon Huynh, Jonathan Downey, Tobias Hollerer, Dorothy Chun, and John O'donovan. 2018. ARbis Pictus: A Study of Vocabulary Learning with Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 11 (Nov. 2018), 2867–2874. <https://doi.org/10.1109/TVCG.2018.2868568>
- [34] JASP Team. 2022. JASP (Version 0.16.3)[Computer software]. <https://jasp-stats.org/>
- [35] John K. Kruschke and Torrin M. Liddell. 2018. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review* 25, 1 (Feb. 2018), 178–206. <https://doi.org/10.3758/s13423-016-1221-4>
- [36] Sari Kujala, Ruth Mugge, and Talya Miron-Shatz. 2017. The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service. *International Journal of Human-Computer Studies* 98 (Feb. 2017), 51–61. <https://doi.org/10.1016/j.ijhcs.2016.09.011>
- [37] Igor Labutov, Yun Huang, Peter Brusilovsky, and Daqing He. 2017. Semi-Supervised Techniques for Mining Learning Outcomes and Prerequisites. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Halifax NS Canada, 907–915. <https://doi.org/10.1145/3097983.3098187>
- [38] David T. Lee, Emily S. Hamedian, Greg Wolff, and Amy Liu. 2019. Causeway: Scaling Situated Learning with Micro-Role Hierarchies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland UK, 1–12. <https://doi.org/10.1145/3290605.3300304>
- [39] Sangmin-Michelle Lee. 2022. A systematic review of context-aware technology use in foreign language learning. *Computer Assisted Language Learning* 35, 3 (March 2022), 294–318. <https://doi.org/10.1080/09588221.2019.1688836>
- [40] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (Dec.

- 2012), 106–131. <https://doi.org/10.1177/1529100612451018>
- [41] Shawn Loewen, Dustin Crowther, Daniel R. Isbell, Kathy Minhye Kim, Jeffrey Maloney, Zachary F. Miller, and Hima Rawal. 2019. Mobile-assisted language learning: A Duolingo case study. *ReCALL* 31, 3 (Sept. 2019), 293–311. <https://doi.org/10.1017/S0958344019000065>
- [42] Richard E. Mayer (Ed.). 2014. *The Cambridge handbook of multimedia learning* (second edition ed.). Cambridge University Press, New York.
- [43] 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. <https://doi.org/10.5555/1866795.1866797>
- [44] Hiroaki Ogata, Bin Hou, Noriko Uosaki, Kousuke Mouri, and Songran Liu. 2014. Ubiquitous Learning Project Using Life-logging Technology in Japan. *Educational Technology & Society* 17 (2014), 85–100.
- [45] Thomas Olsson and Markus Salo. 2012. Narratives of satisfying and unsatisfying experiences of current mobile augmented reality applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 2779–2788. <https://doi.org/10.1145/2207676.2208677>
- [46] Jahna Otterbacher. 2015. Crowdsourcing Stereotypes: Linguistic Bias in Metadata Generated via GWAP. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 1955–1964. <https://doi.org/10.1145/2702123.2702151>
- [47] Jan L. Plass and Shashank Pawar. 2020. Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education* 52, 3 (July 2020), 275–300. <https://doi.org/10.1080/15391523.2020.1719943>
- [48] Yvonne Rogers, Kay Connelly, Lenore Tedesco, William Hazlewood, Andrew Kurtz, Robert E. Hall, Josh Hursey, and Tammy Toscos. 2007. Why It's Worth the Hassle: The Value of in-Situ Studies When Designing UbiComp. In *Proceedings of the 9th International Conference on Ubiquitous Computing (UbiComp '07)*. Springer-Verlag, Berlin, Heidelberg, 336–353. https://doi.org/10.1007/978-3-540-74853-3_20
- [49] Richard M. Ryan and Edward L. Deci. 2009. Promoting self-determined school engagement: Motivation, learning, and well-being. In *Handbook of motivation at school*. Routledge/Taylor & Francis Group, New York, NY, US, 171–195.
- [50] Sylvio Rüdian, Moritz Dittmeyer, and Niels Pinkwart. 2022. Challenges of using auto-correction tools for language learning. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, Online USA, 426–431. <https://doi.org/10.1145/3506860.3506867>
- [51] Burr Settles and Brendan Meeder. 2016. A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1848–1858. <https://doi.org/10.18653/v1/P16-1174>
- [52] Rustam Shadiev, Ting-Ting Wu, and Yueh-Min Huang. 2020. Using image-to-text recognition technology to facilitate vocabulary acquisition in authentic contexts. *ReCALL* 32, 2 (May 2020), 195–212. <https://doi.org/10.1017/S0958344020000038>
- [53] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction* 36, 6 (April 2020), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- [54] Hendrik Thijs, Mohamed Amine Chatti, Esra Yalcin, Christoph Pallasch, Bogdan Kyryliuk, Togrul Mageramov, and Ulrik Schroeder. 2012. Mobile learning in context. *International Journal of Technology Enhanced Learning* 4, 5/6 (2012), 332. <https://doi.org/10.1504/IJTEL.2012.051818>
- [55] 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, Vancouver, BC, Canada, 3179. <https://doi.org/10.1145/1978942.1979414>
- [56] Carmen Tschofen and Jenny Mackness. 2012. Connectivism and dimensions of individual experience. *The International Review of Research in Open and Distributed Learning* 13, 1 (Jan. 2012), 124. <https://doi.org/10.19173/irrodl.v13i1.1143>
- [57] Christian David Vazquez, Afika Ayanda Nyati, Alexander Luh, Megan Fu, Takako Aikawa, and Pattie Maes. 2017. Serendipitous Language Learning in Mixed Reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*. ACM Press, Denver, Colorado, USA, 2172–2179. <https://doi.org/10.1145/3027063.3053098>
- [58] Eric-Jan Wagenmakers, Jonathon Love, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F. Gronau, Damian Dropmann, Bruno Boutin, Frans Meerhoff, Patrick Knight, Akash Raj, Erik-Jan van Kesteren, Johnny van Doorn, Martin Šmíra, Sacha Epskamp, Alexander Etz, Dora Matzke, Tim de Jong, Don van den Bergh, Alexandra Sarafoglou, Helen Steingroever, Koen Derks, Jeffrey N. Rouder, and Richard D. Morey. 2018. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review* 25, 1 (Feb. 2018), 58–76. <https://doi.org/10.3758/s13423-017-1323-7>
- [59] Candace Walkington and Matthew L. Bernacki. 2018. Personalization of Instruction: Design Dimensions and Implications for Cognition. *The Journal of Experimental Education* 86, 1 (Jan. 2018), 50–68. <https://doi.org/10.1080/00220973.2017.1380590>
- [60] Candace Walkington and Matthew L. Bernacki. 2020. Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education* 52, 3 (July 2020), 235–252. <https://doi.org/10.1080/15391523.2020.1747757>
- [61] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing Student Open-Ended Solutions to Create Scalable Learning Opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*. ACM, Chicago IL USA, 1–10. <https://doi.org/10.1145/3330430.3333614>
- [62] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. 2015. Learnersourcing Subgoal Labels for How-to Videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, Vancouver BC Canada, 405–416. <https://doi.org/10.1145/2675133.2675219>
- [63] Justin D. Weisz, Michael Müller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection Not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 402–412. <https://doi.org/10.1145/3397481.3450656>
- [64] Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query—Response Paradigm*. Springer International Publishing, Cham. <https://doi.org/10.1007/978-3-031-02260-9>
- [65] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*. ACM, Edinburgh Scotland UK, 379–388. <https://doi.org/10.1145/2876034.2876042>
- [66] Tzu-Chi Yang, Gwo-Jen Hwang, and Stephen Jen-Hwa Yang. 2013. Development of an adaptive learning system with multiple perspectives based on students' learning styles and cognitive styles. *Journal of Educational Technology & Society* 16, 4 (2013), 185.

A USER STUDY MEASURES

Table 3: Pre-study questionnaire: list of measures

Prior Knowledge of German	
Text entry	Please translate the following sentences without using external aids. If you don't know the answer, just leave the field blank. <ul style="list-style-type: none"> The person is in front of the building. There is a cup on the table. I have many flowers in my garden.
Radio	How would you rate your level of German on the European reference scale?
Motivation and Experience with Language Learning	
Multiple choice	Which of these reasons best fit your motivation for learning a language?
Radio	Have you already used mobile apps for language learning (e.g., Duolingo, Busuu)?

Table 4: Post-study questionnaire: list of measures

Perceived Quality of the Auto-Generated/Manually Generated Content	
Likert matrix	How would you rate the quality of the generated quizzes regarding their correctness, understandability, and relevance?
Text entry	How could the quizzes on the flashcard be improved?
Usability	
Likert matrix	System Usability Scale
Text entry	How could the app be improved?
Text entry	What did you like about the app?
Motivation	
Likert scale	Adding new flashcards from photos increased my motivation to study.
Usage Situations and Usage Patterns	
Text entry	In what situations would you use an app like the Photo Flashcards app?
Text entry	When you added photos, what kind of photos did you select and why?
Text entry	What learning strategy would you say you used during the study? For example, when did you start training rounds?
Text entry	How would you compare your learning experience with this system to other methods you have used?
Post-Test German	
Text entry	What can you see in the images below? Add a one-sentence description in German without using external aids.