Auto-Generating Multimedia Language Learning Material for Children with Off-the-Shelf AI

Fiona Draxler
fiona.draxler@ifi.lmu.de
LMU Munich
Munich, Germany

Laura Haller
haller.la@campus.lmu.de
LMU Munich
Munich, Germany

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

Lewis L. Chuang
lewis.chuang@phil.tu-chemnitz.de
Chemnitz University of Technology
Chemnitz, Germany

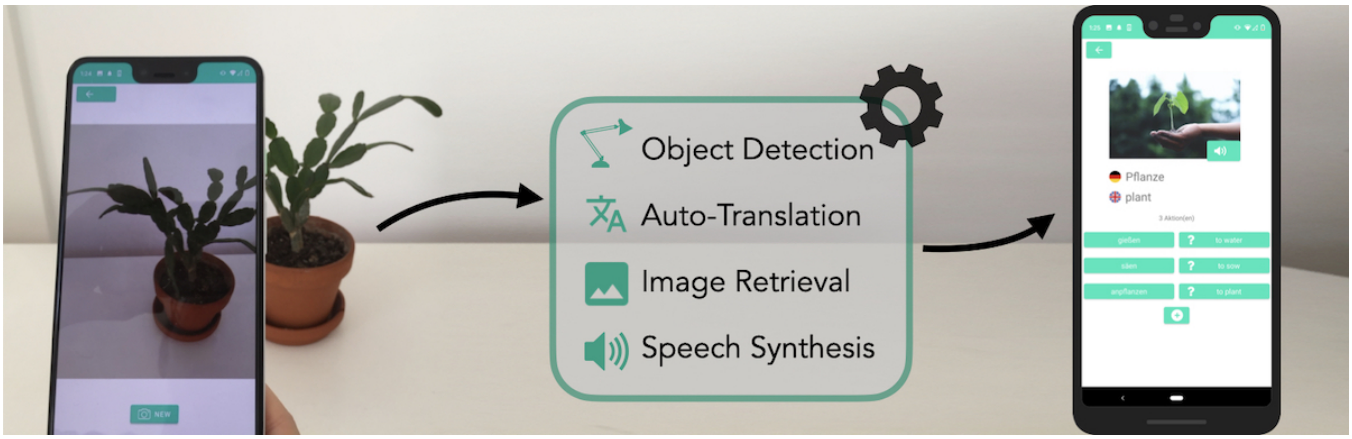


Figure 1: Auto-generating multimedia language learning material with object detection, machine translation, image retrieval, and speech synthesis

ABSTRACT

The unique affordances of mobile devices enable the design of novel language learning experiences with auto-generated learning materials. Thus, they can support independent learning without increasing the burden on teachers. In this paper, we investigate the potential and the design requirements of such learning experiences for children. We implement a novel mobile app that auto-generates context-based multimedia material for learning English. It automatically labels photos children take with the app and uses them as a trigger for generating content using machine translation, image retrieval, and text-to-speech. An exploratory study with 25 children showed that children were ready to engage to an equal extent with this app and a non-personal version using random instead of personal photos. Overall, the children appreciated the independence gained compared to learning at school but missed the teachers' support. From a technological perspective, we found that auto-generation works in many cases. However, handling erroneous input, such as blurry images and spelling mistakes, is crucial for children as a target group. We conclude with design recommendations for future projects, including scaffolds for the photo-taking process and information redundancy for identifying inaccurate auto-generation results.

MuC '22, September 4–7, 2022, Darmstadt, Germany

ACM ISBN 978-1-4503-9690-5/22/09…$15.00
https://doi.org/10.1145/3543758.3543777

CCS CONCEPTS

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

KEYWORDS

Mobile language learning, Content generation, Object detection, Applied machine learning.

ACM Reference Format:
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 (MuC '22), September 4–7, 2022, Darmstadt, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543758.3543777

1 INTRODUCTION

According to a 2020 study, more than 53% of 7-year-old children in the UK own a mobile phone [3]. Besides recreational and social usage, this also opens up promising new opportunities for educational purposes. In particular, combining smartphone capabilities such as cameras, microphones, and gyroscopes with internet services such as machine translation makes it possible to re-think the design of learning experiences. Learning material can now be auto-generated, contextualised, and personalised. For example, mobile devices have already been used for learning grammar in the learners' environment with automated exercise generation [7] and location-based vocabulary learning [13]. However, research on the user experience and design implications of automated generation of learning materials with smartphones and AI systems has rarely focused on children as users. Moreover, it is unclear whether auto-generated material should be personalised for each user or if shared material is equally suitable.

This is the perspective we provide in the present work. Specifically, we design and evaluate a mobile app for young children learning English. The non-personalised version of the app applies off-the-shelf algorithms—image retrieval, machine translation, and speech synthesis—to auto-generate multimedia learning material. Specifically, it creates learning objects comprising an English noun, an image tagged with this noun, its German translation, its pronunciation, and associated verbs. In the personalised version of the app, it additionally integrates computer vision: labels for photos that learners take with the app serve as a trigger for the generation of the multimedia learning materials (cf. Figure 1). We evaluated the technical feasibility and learning experience in an exploratory between-groups study with 25 children. One group used the personalised version of the app that included automatic image labels (Personal group). The second group used the non-personalised version (Non-Personal group); the words that could be added were determined from the word lists generated in the Personalised group. *Research Question 1* addresses the technology perspective: *Can we leverage the capabilities of mobile devices and off-the-shelf AI to auto-generate (personalised) multimedia language learning content for children?* Specifically, we evaluate the characteristics and technical quality of the learning material generated from the children's input and analyse the requirements of this target group. As the auto-generated content is of no use if children are not comfortable with learning from it, we then proceed to *Research Question 2: How do children experience learning with auto-generated learning material and what specific challenges and needs have to be addressed?* Here, we analyse patterns in the observed interaction and report findings from the interviews conducted with the children. Finally, we investigate the role of personalised input which the auto-generation makes possible in *Research Question 3: What difference is there between the interaction with non-personal, i.e. randomly selected, objects and the additional auto-generation layer where children individually add objects from photos?*

Our findings show that off-the-shelf algorithms can successfully be employed for auto-generating learning material from user input, such as personal photos. However, the more established algorithms for machine translation, image retrieval, and speech synthesis yielded better outcomes than the image labelling component. In addition, the quality of the learning material strongly depends on the given input. Notably, some children in our study tended to take blurry photos, which led to unsatisfactory recognition results. Similarly, several children made spelling mistakes when adding actions—although the generation system was robust with respect to some of the mistakes. The interviews with children and their parents suggest a good user experience overall and show particular appreciation for the integration of pictures and pronunciation support through speech synthesis. However, it also pointed to issues with the specificity and saliency of the image labels. Regarding the two variants, we observed no significant difference in engagement or usability between the Personalised and the Non-Personalised group, as manifested in the overall usage time and the number of objects and actions that were added. The relative share of study time was significantly higher for the Non-Personalised group.

Based on our results, we give recommendations for designing systems that auto-generate learning content for children with off-the-shelf algorithms. In particular, we address *error handling*, *activity guidance*, *possible content sources*, and *integration with school*.

In sum, we contribute a novel system for auto-generating multimedia language learning material based on children's input. We study the practical application of this system from a technical and a user perspective and derive recommendations for the design of future systems that use off-the-shelf algorithms and affordances of mobile devices to achieve diverse, ubiquitous, and personalised education. We believe that our findings are also interesting for other domains where users need to interact with imperfect AI systems.

2 RELATED WORK

The growing number of mobile devices available within and outside of classrooms enables novel applications in education. In this section, we summarise the affordances of mobile technology in children's education, the role of context in learning, and the design of appropriate learning material.

2.1 Mobile Devices as Learning Tools for Children

Mobile devices are promising learning tools because of their ubiquity and sensing capabilities. In particular, mobile devices are a preferred choice for problem-based and inquiry-based learning where students determine their own pace of learning—alone or in groups [9, 28]. For example, several projects have used the camera of mobile devices as a tool in learning, e.g. for documenting acted-out scenes in language learning [37]. The camera is also used in handheld augmented reality (AR) for education [9], where a camera view is augmented with additional virtual information on learning targets such as the current in electrical circuits [1], the life cycle of trees [41], or as an augmentation of contents in a book [11]. AR can benefit children's engagement but may also lead to cognitive overload [9]. In a science context, accelerometers in smartphones as experimental tools for learning about pendulum motion can promote children's interest in comparison to traditional experimental tools [16]. In our project, we utilise the device camera to generate learning material from a learner's environment and augment the experience with machine translation, image retrieval, and speech synthesis.

2.2 Context-Based and Automatically Generated Learning Material

A match between learning material and a learner's context can promote interest, engagement, and learning outcomes [10, 15, 27]. This match can be achieved through adaptive strategies, where the learning material is personalised to foster associations with a learner's context [34]. As contexts can be very diverse, context adaptation is often supported by automated content generation that simplifies manual preparation or even renders it unnecessary. In the domain of language learning, past research has proposed methods for automatically extracting vocabulary linked to context characteristics such as location [10, 13], based on objects detected in their environment [7, 8, 25], or virtual contexts such as texts on visited websites [31] and an image uploaded as a search prompt [30]. However, to the best of our knowledge, context-based content generation so far has focused on adults as users. In our work, we apply similar principles to content generation for children, but we investigate what specific needs arise from the younger target group.

2.3 Designing Multimedia Language Learning Material for Children

Besides the contextual relevance, additional points that influence (children's) engagement and learning outcomes in learning experiences include the modalities of presentation [4, 21] and interesting [29] and realistic learning tasks [14]. According to the multimedia learning theory, simultaneously encoding learning material for different channels fosters learning, while redundant information that addresses the same channel can increase cognitive load [21]. For example, an image with a verbal (narrated or textual) description is likely to be helpful, while presenting text with a narrative voice-over impacts the focus on the content. Furthermore, motivation is essential for successful learning. Intrinsic motivation, in particular, fosters conceptual understanding as opposed to rote memorisation, and learners are more likely to be intrinsically motivated when they have the autonomy to choose what they find interesting [29]. Other factors that influence learning motivation are summarised in [35]. In addition, authentic settings contribute to learning by fostering conceptual understanding [14]. Authentic learning activities simulate real-world tasks. Important characteristics include a realistic context, multiple perspectives, collaborative elements, and encouragement for reflection. This paper focuses on concrete words and actions that relate to children's everyday lives. We provide multiple representations: text, image, and audio, and we keep the presentation clear and simple to reduce extraneous load. We do not design a fully authentic learning experience but provide a tool that can easily be embedded in a larger social and cultural context.

3 LEARNING BY EXPLORING

We developed an Android app that enables children to independently learn a language with the objects in their environment. Specifically, children can add English nouns by taking photos of objects and augment the objects with corresponding actions to additionally practice verbs. The app comprises multiple layers of auto-generation: (1) noun extraction through image labels, (2) multimedia content creation through machine translation of nouns, image retrieval, and speech synthesis, and (3) machine translation of verbs.

3.1 Adding Objects and Activities

Tapping the "New" button on the main screen (cf. Figure 2a) opens a camera view. Once a child has taken a photo, we retrieve image labels for the image and, if successful, add the recognised object and its German translation to the child's object list. By making children take photos of their surroundings, we aim to spark situational interest [15] and support learning words in context [10]. The app also provides a list of all added objects and a detail view. In the detail view, children can trigger speech output of the word to learn its pronunciation [2] and see an automatically added image for better visualisation [4]. Through the combination of text, audio, and images, we aim to foster multimedia learning [18, 21]. In addition, the objects are used as anchors for verbs: the trigger question "What can you do with the object?" incentivises children to add actions via an input field (cf. Figure 2c). This is inspired by language learning games that also ask for actions connected to nouns [38, p. 42]. Actions are entered in German and automatically translated to English.

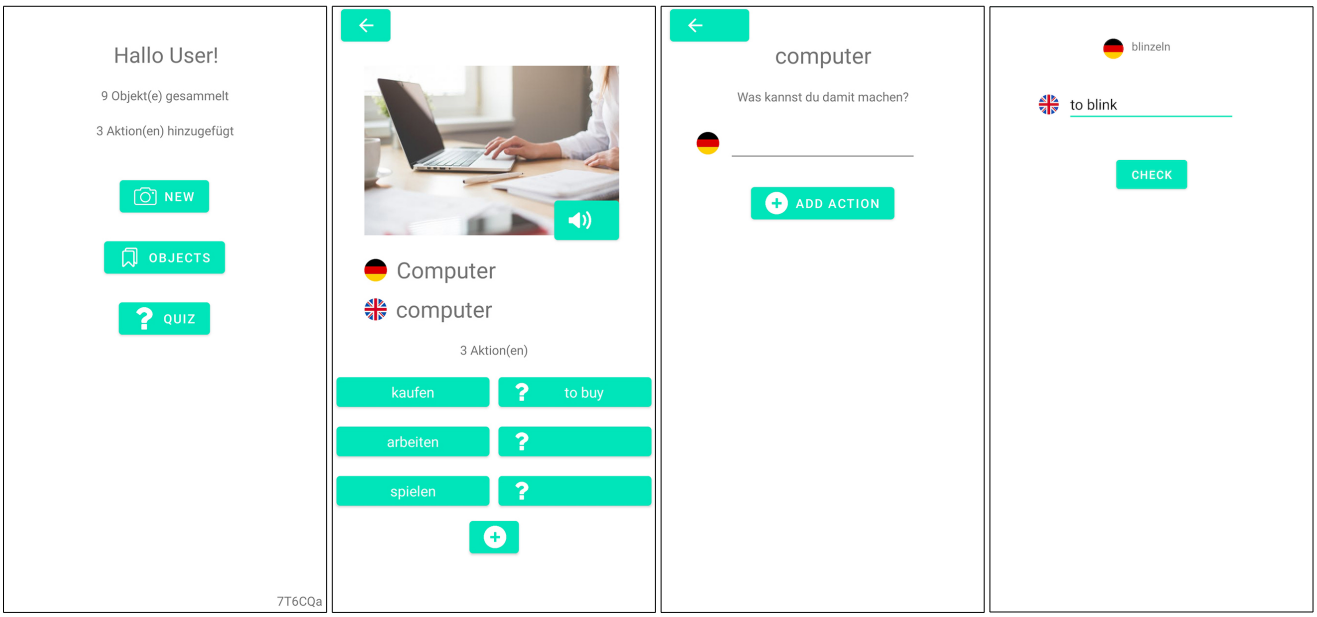
3.2 Testing Knowledge

The app also includes a quiz feature where children can study all the objects and actions they add (cf. Figure 2d). For each quiz round, five items are randomly selected from the child's object and action lists. They are each presented as one of three different types of questions: (1) translation tasks from English to German and vice versa, (2) listening to the English sound and transcribing it, and (3) a multiple-choice question where all actions associated with a noun need to be selected.

3.3 Implementation Details

The app was implemented for Android 5 and newer. We use the Google Vision API[1] to retrieve labels associated with the image. The top result is then translated to German with the PONS API[2] and the Google Translate API[3] as a fallback option. We use photos rather than the live camera preview to avoid bandwidth issues while retrieving image labels. For the detail view of an object (cf. Figure 2b), we request an image from the Pexels API[4], using the object as the search key. Retrieving an image via Pexels rather than showing the photo taken by a user makes the content generation more robust to detection errors, as the tags in Pexels are manually assigned and, therefore, the image is very likely to actually match the tags. Moreover, using a service with strict content guidelines means that we can preserve the child's privacy and avoid showing explicit or other non-suitable content that may be present in their personal photos. For the speech synthesis, we used Android's text-to-speech feature. The speed was set to 50% of the regular speed to make it easier to understand for novice learners. Finally, when children add actions, the input is cleaned and translated to English using Google Translate. As the same English action can sometimes

[1] https://cloud.google.com/vision, last accessed 2022-03-16
[2] https://en.pons.com/p/online-dictionary/developers/api, last accessed 2022-03-16. We chose PONS because their dictionaries are also used in schools.
[3] https://translate.google.com, last accessed 2022-03-16
[4] https://www.pexels.com/api/, last accessed 2022-03-16



(a) Home Screen **(b) Object Details** **(c) Adding an Action** **(d) Answering a Test Question**

Figure 2: Screenshots of *Learning by Exploring*

have various meanings in German, each translation pair is stored with a reference to the associated object it.

4 USER STUDY

We conducted a one-week between-groups study with 25 children to address the research questions stated in the introduction: technological feasibility (RQ1), the user experience from a children's perspective (RQ2), and the effect of exploring one's environment to add vocabulary items (RQ3). For answering RQ3, we implemented a second version of *Learning by Exploring* and used this as a control for the factor *personal context as a source for learning material*. The second version only differed in how new objects could be added: Instead of taking pictures, users with the non-personalised app could add a randomly selected word from a list containing all the objects the participants in the Personalised group had taken. We added a short wait time of 3 seconds after tapping the "New" button to make the process of adding objects comparably long in both groups.

4.1 Procedure

The study was organised in two phases: first, the Personalised group completed all steps, and then the Non-Personalised group. Thus, we could re-use the photos from the Personalised group for the participants in the Non-Personalised group. Besides the timing, the procedure was identical for both groups. After signing the data protection and consent form, the participants' parents received an e-mail with further information, customised for each group. The e-mail contained the link to download the respective version of the app and detailed instructions on how to use it. For the participants in the Personalised group, we additionally asked for permission to analyse the photos the children had taken. The children were then asked to use the app at their own pace for one week (outside of school). They could add objects and actions and take quizzes that included the items they already had on their list. During this time, we logged the interaction with the app. Specifically, we recorded start and stop events, the addition of new objects and actions, and all training questions the children answered. Parents were encouraged to ask for help at any time should any questions or problems with the app arise. After the active study period, we conducted online interviews with the children. We administered the UX Kids questionnaire [39], which includes 16 pairs of opposing items, e.g. "good for learning" and "bad for learning", where tendencies are expressed on a 5-point Likert scale. The questionnaire was initially evaluated in the learning domain, making it a good match for our purpose. We also asked children what they (dis)liked about the respective app, if they would like to continue using it, and how learning with it compared to learning at school. The parents additionally indicated if there were problems with the app and how they perceived their child's motivation and self-directed app use. Finally, we gave the children a translation task including 10 objects and 10 actions they had added (less than 10 when not enough objects or actions were available). Participation was compensated with a 10€ voucher for a bookstore. We obtained approval from our local ethics committee.

4.2 Participants

We recruited 25 participants (21 girls and 4 boys) via a sports club and advertising in schools. Their mean age was 9.92 years (SD = 1.15, min = 8, max = 12 years). Seven were in 3rd grade, five in 4th, six in 5th, and seven in 6th grade. The children were randomly distributed into the two groups while taking care to balance age. Finally, there were 12 children in the Personalised group and 13 in

the Non-Personalised group. All children were currently learning English at school and were non-native speakers. Several used their parents' mobile devices to participate in the study. One child was provided with a device because the family did not have a suitable one available.

5 RESULTS

In this section, we report the study results grouped by the three research questions. We include both quantitative and qualitative data gathered from the interaction logs and the post-study interviews. For the short open-ended interview questions, we clustered participant responses, counted occurrences of similar statements, and derived general themes by summarising the clusters (e.g. *independence of use* and *pronunciation support*). For the reporting, we use consecutive participant numbers with the prefix "P" for the Personalised group and "N" for the Non-Personalised group. For the quantitative analyses that compare the two groups, we apply Welch's t-tests at a significance level of $\alpha = 0.5$. In case of non-normality, we use Wilcoxon rank-sum tests instead. The technical assessment was done manually. We checked all available images, detection results, machine translation results, an exemplary set of retrieved images, and speech synthesis. We rated the matches either as a *good match*, a *partial matches*, or as a *mismatch*. To minimise subjectivity, we defined a rating scheme for each category.

5.1 RQ1: Evaluation of the Generated Learning Material

Based on the photos taken by the participants we were granted permission to analyse, we evaluated the content-generation process from a user and a technology perspective. We separately assessed all levels of the auto-generation process: (1) retrieving nouns from automatic image labels for the users' photos, (2) generating multimedia learning material from these nouns with image retrieval, text to speech, and machine translation, (3) associating verbs via machine translation of user input.

For Level 1, we assessed the labelling quality and popular motifs and characteristics of the images. In 62.4%, the image actually contained an object described by the retrieved label or could be associated with this label. However, of these, only 58.9% were salient aspects of the image, 11.0% of the top-ranked labels were monochrome colours ("brown", "grey"), and another 8.7% were labelled as "wood" (many of them contained wooden flooring or furniture). Our manually assigned tags revealed that 10.7% of the photos showed a person or body parts (e.g., a foot, a face), 25.4% showed decoration, art, or plants, 13.5% household items, 14% toys, and 3.5% pets. A small number of photos contained outdoor scenes or room interiors, and some photos were blank or too blurry to assign any category. In fact, 9.7% of all analysed images were somewhat or very blurry. The detected labels were often generic, e.g. a rabbit described as "vertebrate" and an apple or orange described as "food".

For Level 2, we assessed translation, image retrieval, and speech synthesis. The translation was generally correct, with only a small number of exceptions where the translation was too specific (e.g. "glasses" translated as "sunglasses" and "pen" as "fountain pen") or where less common terms were given as the preferred translation (e.g. "whiskers" in the sense of facial hair instead of the rabbit's whiskers that had been detected in the respective image). We then verified how well the images retrieved from Pexels matched the 194 search terms, i.e. the single words or composite words detected in photos. This was a momentary assessment, as the search queries do not always return the same results, but it is likely that the overall quality does not change substantially. The image well represented the English words in 177 cases (91.2%). 10 (5.2%) could be connected to the word but were not a prominent element, and 7 (3.6%) images did not match the word. The match for the words translated to German was slightly worse, as some ambiguities became apparent. 155 images (79.9%) were good matches, 22 (11.3%) were mediocre matches, and 17 images did not represent the words (8.8%). There was no instance of the most critical combination, where a German word is a match, but the English word does not fit: In this case, users have no means of recognising an error. Figure 3 shows exemplary retrieval results with varying success: the term "office supplies" is perfectly matched. The light bulbs for "electricity" work as a metaphor for the intangible phenomenon. (c) includes a "communication device", but this is not the focus of the scene, and "watch" is not matched because the label refers to the activity and not the noun. Finally, we reproduced the speech synthesis of all object terms. The pronunciation was correct in all cases, although this may differ for other devices that do not use the Google Text-to-Speech engine.

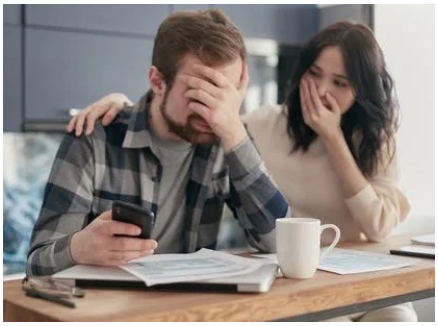
For Level 3, we first checked the validity of the children's input for actions (sense and spelling) and then verified if the translation was correct and matched the noun. 6.0% of the added actions were not a verb in the infinitive form, and another 3.5% were misspelt, e.g. "fereisen" instead of "**ver**eisen" and "gleten" instead of "gl**ät**ten". The translation was correct and matched the object in 83.4% of cases. Errors were caused by invalid input or by ambiguous translations of German activities, e.g. "anziehen" translated as "attract" instead of "put on" as it should have been in the context of clothing. In 3.7% of cases, the translation was correct despite invalid input, as the translation engine proved robust to typos such as "berüren" instead of "berü**h**ren" and returned the intended result. However, the first-language error was retained because our system did not correct the initial spelling.

5.2 RQ2: User Experience of Children as Users

In the final interview, we asked children how they perceived the user experience with the auto-generated multimedia learning material and how they compared it to learning at school. The results of the UX Kids Questionnaire are shown in Figure 4a. The average UX rating in the Personalised group was 13.8 ($SD = 7.6$) and 16.2 ($SD = 9.9$) in the Non-Personalised group (lower is better), i.e. in the top 25% of possible scores. The difference was not significant. Eight children in the Personalised group (66.7%) and eight (61.5%) in the Non-Personalised group stated that they would like to continue using the app. The parents of P4 and P9 said that it was difficult to get started with the app, while the parent of P10 described the app as "very intuitive" besides some challenges while taking targeted photos. Suggested additional features were audio output for the action words (P2, P5), adding words from a list (N2, parent of N7), more varied activities (N3), a feature for correcting pronunciation (N12), and the possibility to delete items (P6, N6). In addition, several children and parents mentioned that they would have liked the app

(a) Office supplies - by Tim Gouw

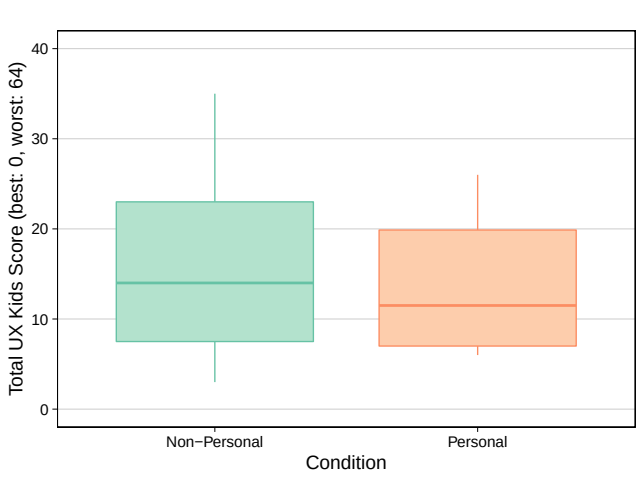
(b) Electricity - by Rodolfo Clix



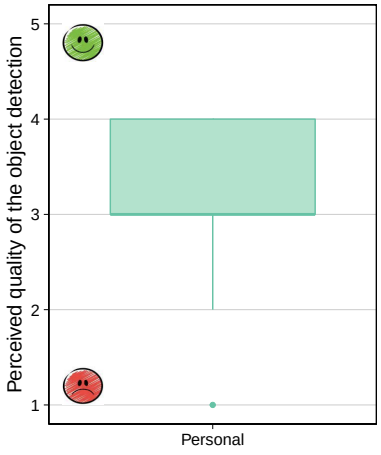
(c) Communication device - by Mikhail Nilov

(d) Watch - by Ketut Subiyanto

Figure 3: Example images retrieved via Pexels queries



(a) UX Kids Score [39]. Possible scores range from 0 (best) to 64 (worst)



(b) Perceived quality of the image labels

Figure 4: User experience and perceived image label quality

to be more colourful (parent of P12, N4, N10) and game-like (parent of P3, P4). When comparing learning with the app to learning in school, the children found the main benefits of the school to include the teacher's support (3×P, 7×N) and social aspects (3×P, 2×N), such as meeting friends. On the other hand, the app allowed for self-paced use (4×P, 2×N), and the pictures (3×P, 2×N) and speech synthesis (P5, P9, N12) were helpful. However, P7 reported that the synthesis was sometimes unclear or started in the middle of the word, and N10 described the sound as "sometimes a bit weird".

In the Personalised group, we additionally asked for a rating of the image labelling (cf. Figure 4b. Here, the average value was slightly better than neutral at $M = 3.1$ ($SD = 0.9$; 1 = bad quality and 5 = good quality). When asked about issues with the app in the interviews, 3 of the 12 children and 10 of the 12 parents mentioned the quality of the image labelling. For example, the background

("wood") was sometimes detected instead of the targeted object (mentioned by P10 and P12) and an object was described with a category (e.g. "food") instead of a concrete term (e.g. "apple", P9). We also found this in our analysis in Section 5.1.

5.3 RQ3: Engagement with Personal or Non-Personal Content

From the interaction logs, we counted the number of objects and actions each child added and computed usage statistics. As some children continued to use the app after the end of the study, we capped all interaction logs eight days after the first registered interaction. As shown in Table 1, the engagement was higher for the Non-Personalised group than in the Personal group for all measures of activity that we assessed. However, there is substantial individual variance, and the only significant difference is observed in the share of the total time dedicated to answering questions on the previously added items. Three children in the Personalised group and one in the Non-Personalised group did not answer any test questions. The number of added objects ranged from 8 to 86 in the Personalised group and 4 to 105 in the Non-Personalised group. Although not a primary focus of the current study, Table 1 also summarises the correctness rate achieved in the test questions in the app and in the final interviews.

6 DISCUSSION

Learning by Exploring showcases the generating authentic context-based material for language learning with off-the-shelf algorithms such as computer vision. The technical assessment and user evaluation show that this is a promising approach. However, they also highlight challenges that need to be addressed to further advance the field of auto-generation for learning. In this section, we discuss the current state from an algorithmic perspective and with a focus on optimising the learning experience.

6.1 Off-the-Shelf Algorithms for Generating Learning Material for Children

By capturing elements of our physical surroundings, image labelling algorithms are a source of imageable items, i.e. items that easily produce a mental image. Words with high imageability are typically early elements in a child's language acquisition [22]. But while current algorithms achieve very high precision in benchmark tests [17], this does not mean that they are fully ready to deploy in real-life learning applications. For instance, a label such as "food" on an image of an apple is certainly correct but transports very little information for a language learning use case. Language learning demands specific and versatile results, and to date, this demand can only partially be satisfied with pre-trained models and APIs. Custom models may be trained to provide more suitable labels, but this is impractical because the training requires large data sets and substantial computational power. From the children's perspective, the results were often not as good as they would have wanted them to be. They certainly took some blurry pictures or did not centre the target objects, but even good pictures did not always produce the desired result. In a future iteration, we will directly apply object detection in the image and only use image labels as a secondary source. Exemplary tests suggest that this may already lead to more precise results. Overall, image labelling or object detection can be an interesting source of learning material for children, but they should not be the only ones. Instead, a hybrid approach could additionally include a simple search box or word lists from school—as suggested by some children and their parents. Alternatively, permanently placed objects in confined spaces like a classroom could be pre-labelled as in [7] or labelled with an interactive approach as proposed in [17]. This means that detection can be run on the device, and results can be shown in the live camera preview without bandwidth issues. However, pre-labelling would re-introduce an effort on the teacher's side and lessen the overall versatility.

The multimedia learning material created based on the object trigger proved very accurate overall: For the objects, the translation services we used (PONS and Google Translate) only performed non-optimally in a small number of cases where multiple translations were possible, or the first retrieved translation was too specific. The image retrieval provided an exact match for more than 90% of the search queries for the English words, thanks to the accurate labels the image creators generally provide[5]. After the translation to German, approximately 80% were still good matches, mostly due to the ambiguous translation options. Most importantly, in our test set, there was no instance of a combination where the German word matched the image, but the English word did not. Hence, there was always a means for the children to gauge the probability that the translation was correct. However, one problem was that for the added actions, the children did not have a second channel for verifying the translation. And indeed, the verb translation did not always match the context of the associated object. Future implementations should provide a second channel, e.g. through example sentences. Finally, the speech synthesis in the app correctly pronounced all objects. The truncated beginnings reported by one child were most likely caused by a heavy load on the UI thread, which could be avoided by moving the audio initialisation to a background thread.

Overall, we believe that if they are well-combined, off-the-shelf algorithms can be applied in real-world applications. Thanks to internet access, the device camera, and system-integrated speech synthesis, this can even be achieved on mobile devices. Combining multiple representations not only produces multimedia learning material [21] but also helps identify potential errors in one of the representations. This is essential because people sometimes overly trust AI systems [33] and remember incorrect information even when it has been corrected [20].

6.2 Children's Engagement with Auto-Generated Learning Material

Most parents reported that their child was generally motivated to use our app, and their engagement manifested in a variety of added objects and actions. The children who used the version with object detection took photos of their homes, family members, numerous toys, and pets—motifs that appeared interesting in the current situation. They spent less time practising than in the Non-Personalised group, which suggests that they were more curious to add new objects than to rote learn. Past research has shown that interest-driven learning is generally beneficial for attention and achievements [15],

[5]Image retrieval from Pexels is not an "AI" algorithm because people create the labels. However, it can still contribute to an intelligent system.

Table 1: Group averages and standard deviation of interaction measures and correctness rate in recall tests.

Measure	Personalised	Non-Personalised	Welch's t'/Wilcoxon rank-sum
Number of added objects	37.0 (22.9)	43.1 (32.8)	$t'(23.1) = 0.55, p = 0.59$
Number of added actions	11.8 (12.6)	19.9 (12.5)	$t'(18.6) = 1.43, p = 0.17$
Total questions answered	62.9 (72.1)	121.0 (156.0)	$W = 57.5, p = 0.43$
Total usage time	3958s (2670s)	5708s (4202s)	$W = 58, p = 0.29$
Time share spent on quizzes	18.1% (18.6%)	37.8% (25.8%)	$W = 40.5, p < 0.05*$
Correctness rate in questions	38.6% (29.1%)	51.1% (22.0%)	$t'(17.9) = 1.66, p = 0.12$
Correctness rate in final test: objects	57.5% (29.9%)	65.1% (22.3%)	$t'(20.3) = 0.71, p = 0.48$
Correctness rate in final test: actions	44.3% (47.2%)	53.1% (33.2%)	$W = 58, p = 0.92$

but we could not find a benefit of personal interest on overall engagement in our study. In fact, the average engagement correctness measures were slightly higher for the Non-Personalised group. This also suggests a stronger learning focus than in the Personalised group.

In both groups, the added actions showed that the children actively engaged with the learning content. They particularly liked being able to study at their own pace, for example, when they wanted to repeatedly listen to the pronunciation of a word. The UX Kids Questionnaire achieved good ratings in both groups, although the final interviews showed that the design of the app should be further adapted to children by adding more colours and playful elements. The children also felt that teachers are still better at explaining than a mobile app, and they enjoy the social part of being in school. Therefore, a mobile app is more suitable as a study companion than a standalone learning tool. Despite the possibility to use the app anywhere, almost all pictures were taken indoors. It remains to be seen if this was a seasonal effect or if a learning app is still associated with studying at home. However, even at home, the variety of perspectives in the photos showed that the flexible use of the device made a difference. Another advantage was that children did not need to sit at their desks.

6.3 Limitations

The current evaluation addresses the user experience and technical performance of the full and non-personal versions of our mobile learning app with different degrees of auto-generation. However, future investigation is required to assess the impact of adding auto-generated and personalised content compared to other means of self-studying as a baseline, e.g. flashcard systems [19]. The current study also did not measure learning outcomes, as the short duration of the study was not sufficient to reliably track recall over time and because the words to be learned were not known in the beginning and could not be pre-tested. A longer study would further help to rule out novelty effects. Another important point to note is the influence of the participants' parents. On the one hand, the app was sometimes installed on a parent's phone, which limited the child's access. On the other hand, a parent's encouragement to engage with the app may have increased the overall usage time. However, the parent-child interaction can contribute to academic achievements [36] and should not be cast aside. Instead, future designs could include parents as additional stakeholders.

7 DESIGN RECOMMENDATIONS

The lessons learned during the implementation and findings obtained from the exploratory user evaluation can serve as a guide for future mobile learning experiences using off-the-shelf AI. Below, we discuss our main recommendations, focusing on the user experience design.

Handling Erroneous Input with Pre-Processing and Interactive Approaches. Especially when working with children, non-optimal input must be expected. For example, spell checking makes sense for textual input. In case of doubt, an option to choose from several suggested alternatives could be added to the user interface. For photo input, past work suggested a training phase so learners can use what images are suitable [30]. Alternatively, users could be provided with scaffolds: an overlay in the camera preview could indicate what object is currently in focus, and a hint could be displayed when an image was blurry and should be retaken. Moreover, once a photo has been analysed, all detected objects and bounding boxes could be shown so that children can interactively select what they are most interested in. Generally speaking, error handling and scaffolding can and should be applied when creating input and pre-processing it. However, not all input errors are equally problematic. For example, the final multimedia learning material generated with our app was still correct even when target objects could not be detected. The material only had limited personal relevance because the original photo was not included. Depending on a system's goals, such "errors" may, therefore, be safely ignored or addressed with lower priority. In potentially critical cases, learners should be provided with a means to verify the validity of automatically generated content, e.g. through redundant representations (cf. Section 6.1), and an option to flag content.

Activity Guidance. Metacognitive skills are still in development for children around the age of 10, and this affects their chosen learning strategies [24]. In our study, the lack of regulation showed in the time spent revising the existing learning material in the Personalised group, which was below 20% of the total time. However, repeated practice is important for long-term recall [19]. Hence, children should also be encouraged to practise previously learned items without making them feel too strongly controlled or pressured, as this could, in turn, impact conceptual understanding [12]. One way to guide activities and add incentives to complete tasks is to gamify the experience. However, this primarily increases extrinsic, not intrinsic, motivation [23].

Possible Content Sources. Our apps utilised object detection, image retrieval, speech synthesis, and machine translation for nouns and verbs. Thus, it only explores a subset of what could potentially serve as a content source. On the one hand, future systems could integrate existing texts, videos, or social media content to enrich diversity while preserving authenticity. For example, the video-based dictionary proposed in [40] could be adapted for children by focusing on educational videos. On the other hand, research has only started to tap the potential of generative systems such as GPT-3[6]. For instance, a few keywords could serve as cues for generating a compelling and personalised short story. Moreover, automatic pronunciation assessment [26] could be integrated to check the children's utterances against the text-to-speech audio for pronunciation practice. Materials could be shared with other learners (just as the children in the Non-Personalised group were provided with the Personalised group's objects) [5]. However, this requires a means to assure quality [6]. With children as users, it is also particularly important to preserve privacy and to filter out explicit content.

Integration with School. To better align learning material with school curricula, teachers could provide vocabulary lists that a system automatically augments with multimedia content. Similarly to [37], they could also provide inspiration for objects or scenes that children capture in their photos. This would enable children to study learning items in their curriculum from a new perspective. In addition, prior work has shown that collaborative learning with peers can be beneficial for vocabulary retention [32]. For example, children could collaboratively assemble auto-generated media items associated with a target word. Finally, school could serve as a channel for questions and support when children encounter something they are unsure about while using a mobile learning app, e.g. to clarify potential issues with or errors in auto-generated learning content.

8 CONCLUSION

With their mobile devices, children have access to a powerful learning tool. Through the combination of integrated tools (such as the device camera) and off-the-shelf AI, learning experiences can tap a potentially infinite resource of learning materials. Our work has given an example of how device capabilities and external tools can be combined to provide rich multimedia language learning material for children. The implementation work and an exploratory evaluation showed that the combination of machine translation, image retrieval, and speech synthesis produced high-quality results. Object detection did not live up to the children's expectations and showed no benefit to their engagement. However, the integration of object detection did not impact the learning material's correctness because it merely used a search trigger. For future systems, we recommend methods for compensating input and output errors, scaffolds, balanced activity guidance, and the interplay with school curricula. Moreover, we discuss potential content sources provided through off-the-shelf algorithms.

[6]https://openai.com/api/, last accessed 2022-06-20

ACKNOWLEDGMENTS

This work was partly conducted within the Amplify project which received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 683008). It was also funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. LLC was funded 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] Elham Beheshti, David Kim, Gabrielle Ecanow, and Michael S. Horn. 2017. Looking Inside the Wires: Understanding Museum Visitor Learning with an Augmented Circuit Exhibit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, Denver, Colorado, USA, 1583–1594. https://doi.org/10.1145/3025453.3025479

[2] Walcir Cardoso. 2018. Learning L2 pronunciation with a text-to-speech synthesizer. *Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL* (2018), 16–21.

[3] Childwise. 2020. Childhood 2020: new independent report. http://www.childwise.co.uk/uploads/3/1/6/5/31656353/childwise_press_release_-_monitor_2020_-_immediate_release.pdf

[4] Dorothy M. Chun and Jan L. Plass. 1996. Effects of Multimedia Annotations on Vocabulary Acquisition. *The Modern Language Journal* 80, 2 (June 1996), 183–198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

[5] 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

[6] 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

[7] 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

[8] 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.

[9] Matt Dunleavy and Chris Dede. 2014. Augmented Reality Teaching and Learning. In *Handbook of Research on Educational Communications and Technology*, J. Michael Spector, M. David Merrill, Jan Elen, and M. J. Bishop (Eds.). Springer New York, New York, NY, 735–745. https://doi.org/10.1007/978-1-4614-3185-5_59

[10] 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

[11] Raphaël Grasset, Andreas Dünser, and Mark Billinghurst. 2008. Edutainment with a mixed reality book: a visually augmented illustrative childrens' book. In *Proceedings of the 2008 International Conference in Advances on Computer Entertainment Technology - ACE '08*. ACM Press, Yokohama, Japan, 292. https://doi.org/10.1145/1501750.1501819

[12] Wendy S Grolnick and Richard M Ryan. 1987. Autonomy in children's learning: an experimental and individual difference investigation. *Journal of personality and social psychology* 52, 5 (1987), 890. Publisher: American Psychological Association.

[13] 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

[14] Jan Herrington and Ron Oliver. 1995. Critical Characteristics of Situated Learning: Implications for the Instructional Design of Multimedia. In *ASCILITE 1995 Conference*. Melbourne, 253–262.

[15] 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

[16] Katrin Hochberg, Jochen Kuhn, and Andreas Müller. 2018. Using Smartphones as Experimental Tools—Effects on Interest, Curiosity, and Learning in Physics Education. *Journal of Science Education and Technology* 27, 5 (Oct. 2018). https://doi.org/10.1007/s10956-018-9731-7

[17] 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

[18] Daesang Kim and David A Gilman. 2008. Effects of text, audio, and graphic aids in multimedia instruction for vocabulary learning. *Journal of educational technology & society* 11, 3 (2008), 114–126. Publisher: JSTOR.

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

[20] 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

[21] R.E. Mayer. 2017. Using multimedia for e-learning. *Journal of Computer Assisted Learning* 33, 5 (Oct. 2017), 403–423. https://doi.org/10.1111/jcal.12197

[22] Colleen McDonough, Lulu Song, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Robert Lannon. 2011. An image is worth a thousand words: why nouns tend to dominate verbs in early word learning: Imageability and early word learning. *Developmental Science* 14, 2 (March 2011), 181–189. https://doi.org/10.1111/j.1467-7687.2010.00968.x

[23] Elisa D. Mekler, Florian Brühlmann, Alexandre N. Tuch, and Klaus Opwis. 2017. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior* 71 (June 2017), 525–534. https://doi.org/10.1016/j.chb.2015.08.048

[24] Janet Metcalfe and Bridgid Finn. 2013. Metacognition and control of study choice in children. *Metacognition and Learning* 8, 1 (April 2013), 19–46. https://doi.org/10.1007/s11409-013-9094-7

[25] Dan Motzenbecker. 2017. Thing Translator. https://experiments.withgoogle.com/thing-translator

[26] Mauro Nicolao, Amy V. Beeston, and Thomas Hain. 2015. Automatic assessment of English learner pronunciation using discriminative classifiers. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, South Brisbane, Queensland, Australia, 5351–5355. https://doi.org/10.1109/ICASSP.2015.7178993

[27] 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

[28] Y. Rogers, D. Stanton, M. Thompson, M. Weal, S. Price, G. Fitzpatrick, R. Fleck, E. Harris, H. Smith, C. Randell, H. Muller, and C. O'Malley. 2004. Ambient wood: designing new forms of digital augmentation for learning outdoors. In *Proceeding of the 2004 conference on Interaction design and children building a community - IDC '04*. ACM Press, Maryland, 3–10. https://doi.org/10.1145/1017833.1017834

[29] Richard M Ryan and Edward L Deci. 2009. Promoting self-determined school engagement: Motivation, learning, and well-being. (2009).

[30] 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

[31] 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

[32] Wen-Ta Tseng, Hao-Jyuan Liou, and Hsi-Chin Chu. 2020. Vocabulary learning in virtual environments: Learner autonomy and collaboration. *System* 88 (Feb. 2020), 102190. https://doi.org/10.1016/j.system.2019.102190

[33] Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. 2021. The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction. *Frontiers in Robotics and AI* 8 (April 2021), 554578. https://doi.org/10.3389/frobt.2021.554578

[34] Candace Walkington and Matthew L. Bernacki. 2014. Motivating Students by "Personalizing" Learning around Individual Interests: A Consideration of Theory, Design, and Implementation Issues. In *Advances in Motivation and Achievement*, Stuart A. Karabenick and Timothy C. Urdan (Eds.). Vol. 18. Emerald Group Publishing Limited, 139–176. https://doi.org/10.1108/S0749-742320140000018004

[35] Kathryn R. Wentzel and Allan Wigfield (Eds.). 2009. *Handbook of motivation at school*. Routledge, New York ; London. OCLC: ocn268957399.

[36] S. Wilder. 2014. Effects of parental involvement on academic achievement: a meta-synthesis. *Educational Review* 66, 3 (July 2014), 377–397. https://doi.org/10.1080/00131911.2013.780009

[37] L-H. Wong and C-K. Looi. 2010. Vocabulary learning by mobile-assisted authentic content creation and social meaning-making: two case studies. *Journal of Computer Assisted Learning* 26, 5 (Oct. 2010), 421–433. https://doi.org/10.1111/j.1365-2729.2010.00357.x

[38] Andrew Wright, David Betteridge, and Michael Buckby. 2006. *Games for Language Learning* (3rd ed ed.). Cambridge University Press, Cambridge ; New York.

[39] Lea Wöbbekind, Thomas Mandl, and Christa Womser-Hacker. 2021. Construction and First Testing of the UX Kids Questionnaire (UXKQ): A Tool for Measuring Pupil's User Experience in Interactive Learning Apps using Semantic Differentials. In *Mensch und Computer 2021*. ACM, Ingolstadt Germany, 444–455. https://doi.org/10.1145/3473856.3473875

[40] Yeshuang Zhu, Yuntao Wang, Chun Yu, Shaoyun Shi, Yankai Zhang, Shuang He, Peijun Zhao, Xiaojuan Ma, and Yuanchun Shi. 2017. ViVo: Video-Augmented Dictionary for Vocabulary Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 5568–5579. https://doi.org/10.1145/3025453.3025779

[41] Heather Toomey Zimmerman, Susan M. Land, and Yong Ju Jung. 2016. Using Augmented Reality to Support Children's Situational Interest and Science Learning During Context-Sensitive Informal Mobile Learning. In *Mobile, Ubiquitous, and Pervasive Learning*, Alejandro Peña-Ayala (Ed.). Vol. 406. Springer International Publishing, Cham, 101–119. https://doi.org/10.1007/978-3-319-26518-6_4