

An Instrumented Mobile Language Learning Application for the Analysis of Usability and Learning^{*}

Aigerim Aibsssova, Antonio Cerone, and Mukhtar Tashkenbayev

Department of Computer Science, Nazarbayev University, Nur-Sultan, Kazakhstan

`aigerim.aibassova@nu.edu.kz@nu.edu.kz`

`antonio.cerone@nu.edu.kz`

`mukhtar.tashkenbayev@nu.edu.kz`

Abstract. Mobile applications for language learning (MALL) is a field that is at large dominated by translation-based learning approaches. Moreover, MALL feature a number of common practices that may not effectively address learning or may even increase the number of user errors. In this tool paper, we introduce a language learning application equipped with instrumentation code to collect data about user behavior and use such data in different ways. The most obvious use is to provide statistics and patterns of learning of the users, which can be used by users to adjust their learning approaches and by researchers to study learning processes and attitudes. For the benefit of the user collected data can be also exploited to drive the synthesis of exercises that best suit the user’s language level and learning approach and are not likely to cause usability errors.

The main use of the application is, however, as a tool for research purposes. In fact, it is a tool for testing new forms of exercises and their combination on samples of users, thus providing valuable information for research in language learning as well as supporting the software development process of new MALL. Finally, an additional feature of the tool is the conversion of the collected data into a formal description of the user’s behaviour to be used for formal verification and validation purposes.

Keywords: Mobile Applications; Language Learning; Usability; Instrumentation Code.

1 Introduction

A large variety of mobile applications for language learning (MALL) has been developed during the last years. Such applications are very appealing to the large public and mostly inexpensive, at least in their freeware and shareware versions. Since they can be easily used anytime and anywhere in very short sessions of just

^{*} Work partly funded by Seed Funding Grant, Project SFG 1447 “Formal Analysis and Verification of Accidents”, University of Geneva, Switzerland.

a few minutes, they appear to many users as a panacea to learn new languages effortlessly.

However, although the effectiveness of such applications has not been studied systematically, the sparse studies on their usability and learning effectiveness seem to agree on the following issues:

- these applications can only be used as a secondary tool in learning a new language, i.e. the user cannot gain a reasonable fluency in a new language using only MALL tools [10];
- the opportunity to learn more new languages simultaneously using MALL creates its own unique set of challenges, that are not present in traditional learning, e.g. increased rate of confusion of grammar and/or vocabulary among the languages [3];
- the core issue is the translation-based learning approach, which provides little grammatical support [6, 9, 10].

However, since most research is based on self-reported user feedback, it could potentially be unreliable [5].

The main aim of our work is to empirically identify best linguistic and application design approaches that lead to learning effectiveness of language learning applications. For this purpose we have developed a MALL tool that includes instrumentation code to collect data on user behaviour and performance as well as analysis features for the visualisation and exploitation of such data. All examples in this paper refer to an English speaker learning the Kazakh language.

Data exploitation is carried out with two aims. First the way lessons are delivered is adapted to the characteristics and the progresses of the user. This is achieved by the data-driven synthesis and sequentialisation of exercises that best foster the user’s language level and learning approach. The aim is to avoid the usability errors that have been found to be a commonplace for the considered user. Second, by collecting data on the effectiveness of exercises and different exercise sequences, in terms of the number and type of errors made by the user, the application will be able to provide valuable information for research and software development purposes. This includes information on usability and learning-related errors. In terms of data exploitation, an important feature of our application is the automatic conversion of collected data into a formal description of the user’s behaviour to be used for verification and validation purposes.

In terms of data visualisation, the application can provide statistics and patterns of learning of the users. Some of these forms of visualisation can be observed by the users with the aim of adjusting their learning approaches. Other can be observed by researchers to study learning processes and attitudes.

The rest of the paper is structured as follows. Section 2 illustrates some of the review work carried out to compare various MALL as well as to analyse in depth some of the most popular MALL. Contextually, the section also provides a general background on MALL. Section 3 presents the architecture of our instrumented MALL. Sections 4, 5 and 6 illustrate the implementation components: database, web interface and mobile application, respectively. Section 7 evaluates mobile applications in terms of usability and their functionalities for analysing

learning effectiveness. Finally, Section 8 presents the current implementation status and concludes the paper, also proposing possible future work.

2 MALL Literature Review

Gangaianmaran and Pasupathi [6] review a wide range of different applications. They partition them into three main categories, depending on the addressed learners:

- *primary learners*, i.e. children 3-11 years old;
- *secondary learners*, i.e. teenagers 12-17 years old;
- *tertiary learners*, i.e. university students and adults.

For each category, the authors select a number of applications that they have considered the best, and list them in a comparative table together with basic information, such as the name and a description.

Although Gangaianmaran and Pasupathi’s work is merely descriptive, with little or no exploratory attempts, some general information can be gathered from the tables. Devices running on iOS, such as iPhone and iPad have a larger number of high-quality apps in comparison with Android devices. This is especially noticeable in the primary learners category, which is quite diverse in terms of the study topics. Vocabulary and speaking skills are widely represented in the secondary learners category. However, only one of the considered applications focuses on grammar. The tertiary learners category adds focus on pronunciation, but still seems to lack applications that work on grammar intensively. The obvious conclusion of this work is that listening is the only skill that can be best developed via the use of mobile applications.

Lai and Zheng [8] provide the results of a survey and interview study held in Hong Kong on a sample size of 256 people. All participants were of Chinese descent, and 77% of them were females. The study focuses on the way students use their mobile devices for studying in their free time, and draw a number of important conclusions:

- Mobile devices can significantly improve the personalisation of the study, but are not necessarily as good in terms of *authenticity*, e.g increasing participation in target language communities, and *connectivity*, e.g connection with native speakers or peer learners.
- Social interaction appears to be a big obstacle for both authenticity and connectivity measures, with participants reporting uncertainty when using unknown languages both in personal or public spaces.
- Smart-phones are mostly considered to be leisure devices, unlike laptops or personal computers, but are used effectively to fill in “pockets of time” with activities such as watching short videos, studying during travel, or conversations on-the-go. Therefore, participants did tend to choose specific tools for different tasks.
- Time of device usage for language learning averaged between 1 and 3 hours per week.

Nushi and Eqbali [9] focus on the features of the most popular language application, Duolingo [1]. Duolingo is an application that actively uses the translation aspect of learning, with most exercises focusing on this aspect. Although such an approach can be effective, it actually has a major limitation. It addresses effectively only those people who are fluent in the supposedly known language for which the course is designed. For example, English users have a selection of 16 languages to learn, while Spanish ones have only 6. Duolingo provides five types of exercises for its users:

- *translation* exercises, either writing the translation or choosing words among a given set to compose the translation;
- *matching* exercises, selecting the appropriate figure and associated word to learn for a given word in the known language;
- *pairing* exercises, pairing words from a list of mixed words belonging to the two languages;
- *speaking* exercises, orally repeating a sentence presented in the language to learn;
- *listening* exercises, either writing or composing the sentence heard in the language to learn.

Only the last type of exercise does not involve translation.

One major aspect of the Duolingo is “gamifying” the experience through a system of rewards for completing daily goals, and inclusion of competition via XP points gained when finishing each lesson. Duolingo has social-media features, such as connecting with friends and competing with them, which can be encouraged by receiving notifications when some of the friends scores more points than the user. Notifications are a major part of the application, being sent also when users do not meet their daily goals. Such notification tend to use coercive language.

Nushi and Eqbali point out multiple problems with the application in terms of its teaching techniques, mostly in terms of limitations:

- users are not provided with grammar explanations, and have to figure them out by themselves;
- many words are introduced without information about their meanings or without pronunciation;
- a lot of sample sentences and the synthetic voice that is used to read them can be off-putting and unnatural.

They conclude that the application does not provide a complete learning experience by itself, although it can be helpful to a certain extent.

Nushi and Eqbali conducted a similar study for 50Languages, another popular application in the MALL category [10].

3 System Architecture

Our MALL tool supports the synthesis of exercises and their delivery modalities, as featured by various existing MALL, including the ones considered in Section

2. In this way approaches used by distinct MALL can be emulated and compared in terms of their usability and learning effectiveness.

The system we have developed consists of two parts:

- a *front-end* comprising
 - an *Android application*, and
 - a *web interface*.
- a *back-end* comprising a database and the APIs for data collection, storage, data analytics, data presentation, client control and data exploitation.

The interaction between such components is shown by the UML sequence diagram in Figure 1.

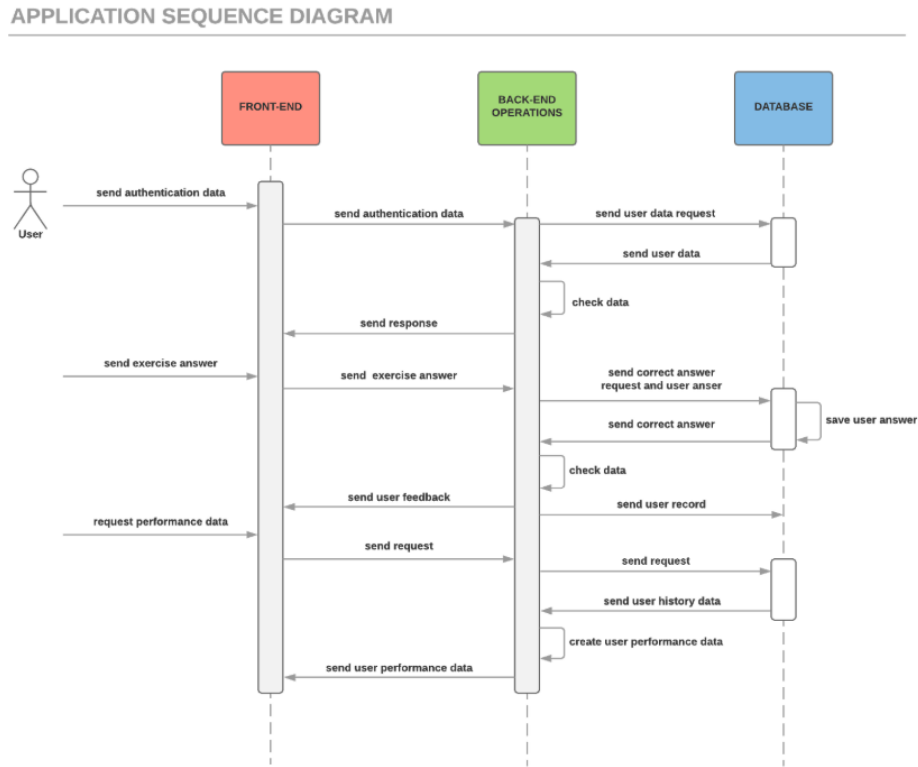


Fig. 1. Application sequence diagram.

The development of the front-end has been carried out using Android NDK for the Android app and HTML/CSS/JS web tools for the web interface. The back-end part of the system comprises a NoSQL Firebase realtime database to store all data necessary for the application.

The Android application serves the purpose of teaching the language as well as recording the data. The web interface is used for designing the delivery of exercises and the display of analytic data. The types of the exercises that are illustrated in this paper are chosen from the most popular language learning applications such as Duolingo and Rosetta Stone. The web interface supports the possibility of choosing from different variants in the exercise presentation. For example, Duolingo matching exercises consist always in the selection of a figure and the associated word to learn to be matched with a given word in the known language. Our application can provide several variants of this exercise. One variant is given in Figure 5(c), where the choice is between words in the language to learn to be matched with just one figure with no word in the known language. The design of such variants is carried out using the web interface.

Section 4 describes in details the kind of data that can be stored and collected. Sections 4–6 highlight the system functionalities, including exercise classification, user functionality and data analytics and exploitation.

4 Database and Analytics

The system uses Firebase NoSQL realtime database. This choice was motivated by several factors. First, it is easy to integrate Firebase with the Android application. Second, NoSQL provides a more flexible framework, which was crucial during the development process, when changes to the specification of the overall system and its functionality changed frequently. Third, Firebase offers unique tools like authentication and database manipulation through the usage of additional functions. This was essential in developing the research-oriented features of the tool.

The database stores:

- exercises, as shown in Figure 2(a);
- personal user data;
- user exercise history, as shown in Figure 2(b);
- analytical data, as shown in Figure 2(c).

Figure 2(a) contains the information of two exercises of the same type (**BS**, i.e. fill in the Blank Space as **type** field) in which the user has to choose the right, among four possible alternatives, to insert in the missing part of a sentence (**question**). The **subject** field defines the grammatical topics covered, i.e. the language skills this exercise is aimed at training. Values for this field in the example considered in this paper are **V** for ‘Vocabulary’, **P** for ‘Pronouns’ and **N** for ‘Number’ (i.e. singular vs. plural).

All data used for data analytics are stored in the database in the form of user exercise history. For example, the information in Figure 2(b), contains: correctness of the given exercise answer (value 0 in the example means that the answer was wrong), subject of the exercise (**P** for pronouns), type (**TS**, i.e. Translate Sentence), and time taken to complete the exercise (4848 ms).

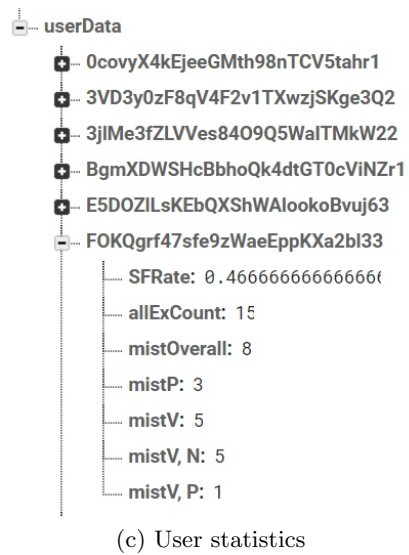
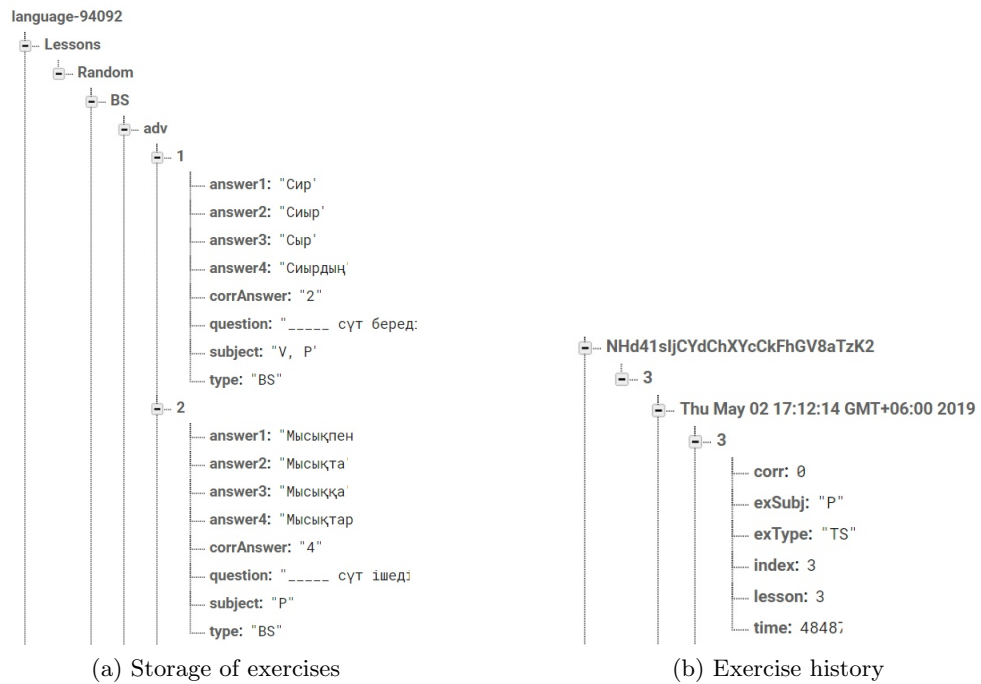


Fig. 2. Database contents

This data is then analysed using Firebase cloud functions, which allow developers to host any custom JavaScript functions on Firebase servers. These functions retrieve and transform data to create new statistical data then stored back in the database. Three category of *statistical data* are stored in the database:

Individual user data comprise success/failure rate of exercises (**SFRate**), number of total exercises completed (**allExCount**), number of exercises completed with mistakes (**mistOverall**) and number of mistakes for each subject. An example is given in Figure 2(c).

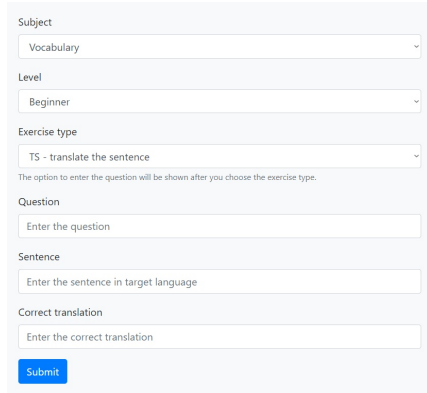
Population data comprise the same kinds of data as for individual user data, but for the entire population of users.

Exercise data comprise success/failure rate for every type of exercise and for every subject on which the exercises focuses.

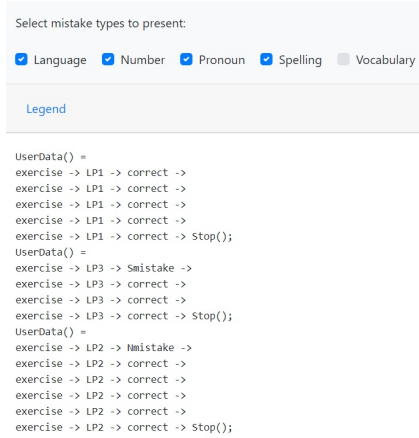
5 Web Application and Research-oriented Functionalities

The web application has a limited use for language learners (whom we call users). They may use it just to view their own individual statistics (individual user data).

It is instead an essential interface for the researcher. It implements the fol-



(a) Create and upload exercises



(b) Output CSP code

Fig. 3. Database input and output

lowing functionalities:

- creation and upload of exercises, as shown in Figure 3(a);
- generation of a formal specification format from the user's exercise history, as shown in Figure 3(b);

- presentation of statistical data in form of charts and diagrams as shown in Figure 4;
- set the strategy for delivering the exercises.



Fig. 4. Visual data display

Setting appropriate strategies for delivering the exercises is an essential functionality for the researcher. The simplest setting is a *random* sequentialisation of the exercises. The researcher can also manually *control* the presentation and sequentialisation of the exercises

- by choosing the presentation form, for example whether or not to add the word to learn to the figure in a matching exercise;
- by force or preventing a pair of exercise types to occur consecutively.

This first control may be used to empirically evaluate which presentation form may lead to more mistakes in general or with specific users. The second control allows the researcher to emulate approaches used by a specific, existing MALL, in order to compare them, or to analyse the learning process in general terms, for example by identifying interferences between questions.

5.1 Formal Analysis

Our tool may also generate a formal specification from the user's exercise history to be used for formal verification and validation purposes. Following the approach by one of the authors [4], the tool generates *CSP (Communicating Sequential Processes)* [7] code to be used within the *PAT (Process Analysis Toolkit)* [2]. For example, the web page in Figure 3(b) produces a CSP representation of data on user interaction, in which we consider all errors apart from vocabulary mistakes ('Vocabulary' is the only unchecked type of mistake). The generated CSP code shown in Figure 3(b) concerns three users: the first does not make any mistake of the selected types, the second makes one 'Spelling' mistake (**Smistake**) and the third a 'Number' mistake (**Nmistake**).

Users perceive, focus and act in different ways while interacting with a MALL. For example, considering Duolingo, in translation exercises to the known language the sentence in the language to learn is presented simultaneously in written and audio modality. In this case, there are two alternative categories of users with two distinct cognitive profiles: focussing on the audio modality and focussing on the written modality. It would be interesting to understand whether, in general, such a user’s cognitive profile has a correlation with the level reached by the user in the learning language. Although the user’s level can be assessed by our MALL tool, assessing the cognitive profile is challenging. Interviewing user is not helping since focussing on a specific modality is a form of implicit attention, of which the user is normally unaware. Moreover, using special technologies, such as an eye-tracking system, does not provide a definite answer: the user may actually read the sentence while the attention mechanism selects the audio information and discharges the written information, so that only the read information progresses to mental processing.

Both the MALL system and user’s cognitive profiles may be formally specified using CSP [4] and composed in parallel to produce a constrained model of the system. The CSP process that formalises real data on the interaction of user at a specific level in the language to learn (i.e. beginners or advanced), which is automatically generated by our tool, can be then composed with the constrained system model. The formal verification of a temporal logic property that characterises the MALL system functional correctness against such a further constrained CSP model may then be used to validate a research hypothesis such as “A learner at an advanced level in the foreign language always focuses on the hearing modality.”

6 Mobile Application

The mobile application runs on Android NDK and implements different lesson functionalities, several types of exercises and data collection features. Although the full plan is to cover written sentence construction and comprehension, listening comprehension and spoken sentence construction, the current implementation does not include any audio functionalities. Therefore lessons and exercises are restricted to the written form.

Researchers are registered through the system setting and have special access rights, which allow them to use the full functionality of the web interface. Users, instead, have to register through the authentication page of the mobile application and they can only access their own individual statistical data using the web interface. Researchers may also register as users through the mobile interface.

When users first register in the system, using their email addresses, a unique token and entry in the database are created to be used in order to match the data that is being sent by the application from this specific account. Authentication through Firebase creates learner profiles that are used to track their progress, using encrypted email-password pairs. At registration, users are also asked to self-rate themselves in the language they intend to learn. This rating is recorded

in the system and used to assign a beginner, intermediate or advanced level to the user, in order to present the user with exercises appropriate for that level.

Depending on the setting, *random* or *controlled*, defined by the researcher as explained in Section 5, *lessons* can be

adaptive users are presented with exercises appropriate for their levels, which are chosen at random from the pool in the database, and their performance is tracked and contributes to change their ratings and, as a result, their levels; ***controlled*** the exercise sequences are controlled by the researcher, but the user's performance does not affect rating and level.

The current implementation of our tool features one kind of *learning practice*

word learning in which the user is presented with a number of pictures of objects or a concept representations together the words that express them in the language to learn (grouped in lessons characterised by topics of increasing difficulties), as shown in Figure 5(b) where the Kazakh word means 'mother'.

and the following kinds of *exercises*:

word matching whose purpose is to recall the words learned previously, reinforce them and verify whether the user knows them, as shown in Figure 5(c); ***filling in the blank space*** whose purpose is to recall words in context. ***sentence translation*** whose purpose is to test the user's ability to build full sentences and which may be skipped if the user does not feel confident.

Learning practice and exercises are normally combined together in lessons but, for research purposes, may also be used as stand-alone.

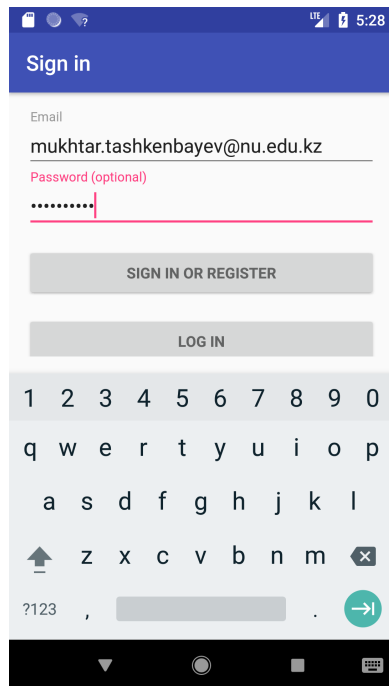
All lesson practices and exercises are stored in the database as the user progresses. Information on user performance is recorded by the application, and is stored in the database following the structure shown in Figure 2(b). The application requires internet connection to function properly.

7 Evaluation

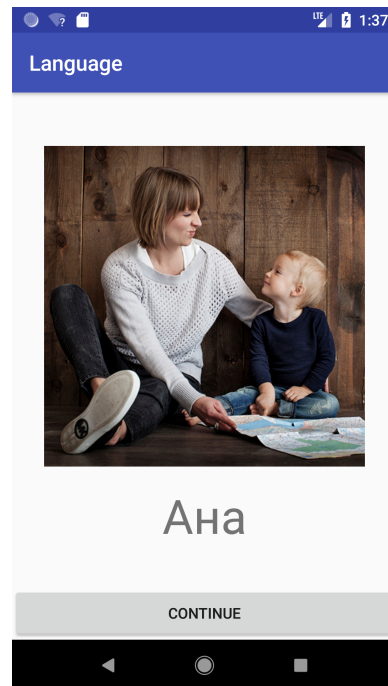
The tool was evaluated in terms of usability and in terms of its functionalities for analysing learning effectiveness. We have evaluated our application using convenience samples of university students as users.

7.1 Usability

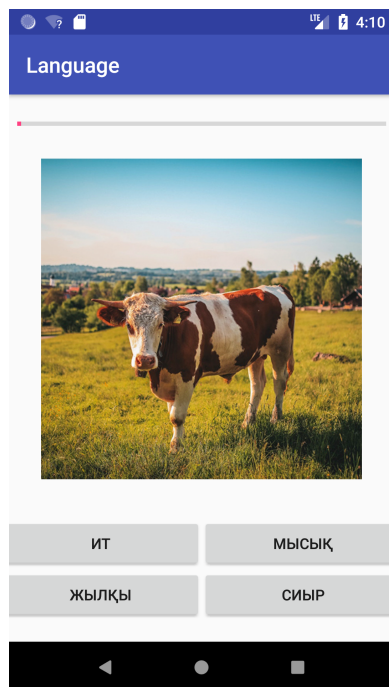
The sample consisted of 15 subjects with the following levels in Kazakh language: 6 native/advanced speakers, 3 intermediate level speakers, and 6 beginner level speakers. The users were asked to test the application for around 10 minutes, going through most of the functionalities, starting from the registration process and ending with language lessons. The whole process was monitored and the interview process was carried out in an informal manner with the addition



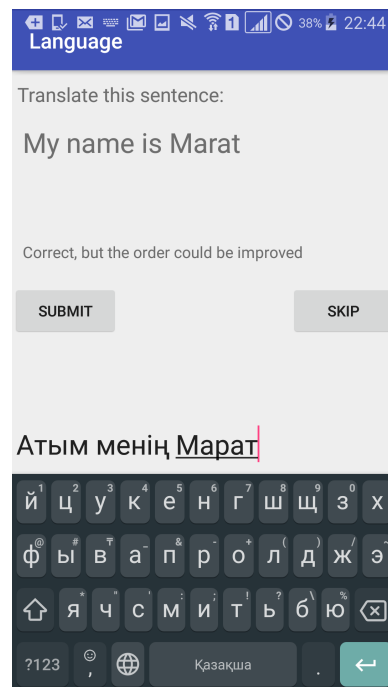
(a) Login



(b) Word learning



(c) Matching exercise



(d) Translation exercise

Fig. 5. Mobile application screenshots

of some guiding questions. Moreover, the respondents were given a chance to request clarifications about questions.

All subjects in the sample underwent *adaptive lessons*, finished multiple lessons, and were further interviewed by one of the researchers. The interview process consisted of ranking questions and a free feedback part. Ranking questions aimed at evaluating the application's usability, intuitiveness, and overall design on a scale from 1 to 10. Questions were as follows:

1. How easy it is for you to use the language learning application?
2. Does the visual presentation of the application have any distracting details that, while not confusing you, might create a distraction?
3. How do you rate the visual appeal of the application?

Users found the application to be both usable and useful. Average ratings for the three questions were 8.27, 8.67 and 7.00, respectively. Standard deviations were 0.80, 1.29 and 1.07, respectively.

Common critiques from the feedback part included:

- loading time issues and minor bugs;
- repeated and easy questions;
- poor exercise input checking, e.g. accepting only one of the possible correct answers in translation exercises.

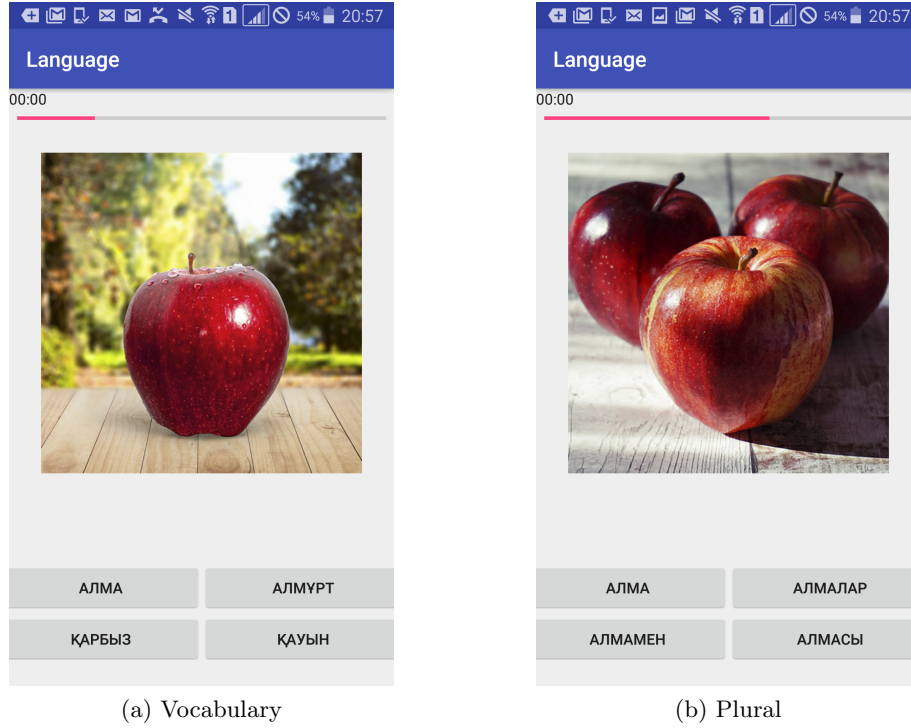
7.2 Learning Effectiveness

The sample consisted of 20 subjects. Purpose of this small scale evaluation was to test the approach of using the tool for evaluating learning effectiveness. This evaluation made use of *controlled lessons*. The test was successful in identifying some common patterns in the user's behaviour:

1. immediately after matching the correct singular word with the given picture (vocabulary-type exercise, as in Figure 6(a)), users were very prone to make a mistake in the matching exercise that followed immediately, if this required them to find a plural form of the same word (number-type exercise, as in Figure 6(b)).
2. translation exercises have gathered significantly more mistakes than any other type of exercise;
3. after learning a new word through a learning practice, some users still made mistakes while testing the same word.

We can interpret such common patterns respectively as follows:

1. users tended to click on the singular form as soon as they saw it, without actually trying to discern the difference between exercises.
2. translation exercises should be presented only when the learner has acquired sufficient confidence, with the granted option of skipping the exercise, without this affecting the performance score;
3. users did not focus enough during learning practice and this shows that learning practice needs to be more engaging.

**Fig. 6.** Matching exercises

8 Conclusion and Future Work

We have introduced a tool capable to emulate a variety of MALL approaches, collect data on the user interaction and performance, present the collected data in a visual format, convert such data into a formal representation to be used in formal analysis, and exploit the data to drive the synthesis and sequentialisation of exercises. The tool is not just another MALL. In fact, it is not intended for teaching languages but, instead, as a research tool. In this respect, it can emulate approaches used by distinct MALL in order to compare their learning effectiveness.

More generally, the tool may contribute to learning theory through the analysis of learning processes. In Section 5.1 we have seen how the tool may be used to validate a research hypothesis on cognitive approaches to learning. Finally, in Section 7.2 we have illustrated the use of the tool to identify interferences between questions.

In the current implementation of the system

- audio functionalities are not included and lessons and exercises are restricted to the written form;
- adaptive lessons are limited to changing the user’s rating and levels.

Concerning the implementation, our proposals for future work in the *short term* include:

1. add audio functionalities in order to analyse interference between audio and written presentation either when they are merged through a multimodal presentation or when they occur in sequence;
2. perform the analysis proposed at Item 1 using a formal methods approach as the one described in the previous work of one of the authors [4];
3. exploit the collected data on the user's performance not only to adapt the delivered exercises to the user's rating and level but also to automatically adjust the presentation of the exercises and their sequentialisation in such a way to prevent the errors that are a commonplace for that user and maximise learning effectiveness.

The future work proposed at Item 3 involves the definition of appropriate measures to characterise learning effectiveness. This is a non-trivial task due to the difficulty in detecting the underlying causes of user errors. For instance, if a user is prone to do mistakes with a specific exercise, this may be due either to the fact that the user is weak in the subject of that exercise or that the exercise is inappropriate for the user's attitude or learning approach. In the former case learning effectiveness may improve by intensifying the use of that exercise. In the latter case, instead, removing that exercise would be the best strategy.

If we consider again the sequence of the two exercises in Figure 6(a) and 6(b), the error might be actually due to the fact that the user has not masterised the rule to form the plural of nouns. In our testing, the recurrence of such a mistake was frequent only when the two exercises were in a sequence. This suggests that the error was caused by an interference between the two exercises. Although this is likely to be true in most cases, such a sequence of exercises would actually be beneficial during the stage when the user has not masterised the formation of plural yet. In such a situation, the automatic control of the sequentialisation could force the sequentialisation during the learning phase of plural formation and could prevent it during the reinforcement phase. In this context, the use of our MALL tool would be twofold: first to realise the automatic control described above, then to collect and analyse data on the effectiveness of such a strategy.

Currently the database is populated only with lessons and exercises for teaching the Kazakh language with the Cyrillic writing systems to English speakers. Considering other language is just a matter of further populating the database. This is obviously a time-consuming task, which requires a lot of human resources and high expertise in linguistic, languages and language learning. This future work is therefore part of our *long term* plans.

Finally, a more extensive evaluation of the system is needed, both in terms of usability and learning effectiveness. In particular, it is essential to test our MALL tool on a large number of real users of various demographic groups.

References

1. Duolingo. <https://www.duolingo.com>.

2. PAT: Process Analysis Toolkit. pat.comp.nus.edu.sg.
3. H. B. E. Ahmed. Duolingo as a bilingual learning app: A case study. *Arab World English Journal*, 7(2):256–267, 2016.
4. A. Cerone and A. Zhexenbayeva. Using formal methods to validate research hypotheses: The duolingo case study. In *STAF 2018 Collocated Workshops (DataMod)*, volume 11176 of *Lecture Notes in Computer Science*, pages 163–170. Springer, 2018.
5. R. Gafni, D. B. Achituv, and G. Rahmani. Learning foreign languages using mobile applications. *Journal of Information Technology Education: Research*, 16:301–317, 2017.
6. R. Gangaianmaran and M. Pasupathi. Review on use of mobile apps for language learning. *International Journal of Applied Engineering Research*, 12(21):11242–11251, 2017.
7. C. A. R. Hoare. *Communication Sequential Processes*. Prentice All Int., 2004.
8. C. Lai and D. Zheng. Self-directed use of mobile devices for language learning beyond the classroom. *ReCALL*, 30(3):299–318, 2017.
9. M. Nushi and M. H. Egbali. Duolingo: A mobile application to assist second language learning. *Teaching English with Technology*, 17(1):89–98, 2017.
10. M. Nushi and M. H. Egbali. 50languages: a mobile language learning application. *Teaching English with Technology*, 18(1):93–104, 2018.