Foreign language learning using a gamificated APP to support peer-assessment

Manuel Palomo-Duarte  
Escuela Superior de Ingeniería  
University of Cadiz, Spain  
C/ Chile 1, CP 11002, Cádiz  
+34 956 015 483  
manuel.palomou@uca.es

Anke Berns  
Facultad de Filosofía y Letras  
University of Cadiz, Spain  
Av. Doctor Gómez Ulla, s/n  
CP 1103, Cádiz  
+34 956 015 549  
anke.berns@uca.es

Juan Manuel Dodero  
Escuela Superior de Ingeniería  
University of Cadiz, Spain  
C/ Chile 1, CP 11002, Cádiz  
+34 956 015 784  
juanma.dodero@uca.es

ABSTRACT

Whilst in recent years there has been an increasing trend to design mobile APPs for foreign language learning, most of the available APPs support mainly one-way teacher-to-learner interaction and use mobile devices to deliver content rather than encouraging learners to interact amongst each other. To address this lack and to provide learners with more versatile opportunities to communicate amongst each other, by sharing, assessing and co-constructing their knowledge, we designed an APP based on a highly interactive, ubiquitous and constructive learning approach. This means that learning contents are not just being delivered to our learners but, instead of this, they are being integrated into versatile tasks that although individually performed, affect the community of learners. The current paper presents our first experience using the Guess it! Language Trainer APP. Firstly, we detail how the APP was used to support students’ language learning outside the classroom, secondly, how it helped learners to get actively involved in their own learning process by adding themselves content to the knowledge base, and thirdly, we show how the data on students’ usage of the APP and participation in the learning process is used to automate the assessment of different competences by the language instructor. A pre-posttest comparison showed very good results on students’ language acquisition. This is complemented with the analysis of the system logs that highlight firstly, that students dedicated to playing the APP much longer time than we estimate they usually employ for independent learning and secondly, that they significantly improved their learning outcomes whilst using the APP. Additionally, other objective indicators on proficiency of other skills such as the ability to explain terms in a foreign language or the competence to assess definitions of mates were obtained automatically.
The software we have developed specifically for the current project is available as GPL free software in its forge\(^1\). The system follows a client-server architecture as we can see in figure 1. Students install client software in their Android smartphones or tablets from Google Play\(^2\). The APP identifies the students and interacts with the server through the Internet to implement the game as well as to show the different rankings.

The typical application workflow begins by asking the user to select a language. At the moment we support English, German and Russian, but the system can easily include new languages by just translating a couple of text files. The language is used both, in the interface (buttons, announcements, etc.) as well as for the definitions to play. After this the student is asked to identify himself in the APP by using an email address and password. This way, he can play on different devices, such as for instance on a smartphone while waiting for the bus or on a more comfortable tablet at home.

\(^1\) https://github.com/AlbertoCejas/GermanLearningUCA

\(^2\) https://play.google.com/store/apps/details?id=es.uca.tabu

---

2. THEORETICAL BACKGROUND

The game we have designed for the current project, is based on the pedagogical framework of mobile learning. As outlined by Sharples and other researchers [21] mobile learning theories recognize not only “(...) the essential role of mobility and communication in the process of learning (...)” but also the importance of the context for meaning construction. Sharples has also described mobile learning as a process in which knowledge is gained through multiple contexts amongst people and personal interactive technologies”. In line with the aforementioned aspects Sharples et al. propose a framework that is based on Engeström’s expansive activity model [14] and which aims to analyze the interdependencies between learning and technology. The framework he proposes is mainly based on five factors (subject, object, context, tools and communication), which are themselves analyzed under two different layers: a technological and a semiotic layer. The technological layer describes learning as “an engagement with technology, in which devices such as mobile phones (...) function as interactive agents (...)” that help its users to acquire knowledge, to communicate amongst each other, to share and negotiate contents and meanings and, finally, to reflect on them. In contrast, the semiotic layer describes learning as “a cognitive system in which learners’ objective-oriented actions are mediated by cultural tools and signs”.

3. DESIGN AND ARCHITECTURE

This, coupled with the fact that, nowadays, devices such as smartphones are amongst our learners’ most frequently used gadgets [1] and that there is an increasing number of research studies stressing its potential for foreign language learning, led us to explore the possibility of integrating gamificated APPs in our course syllabus [9, 21, 22, 5].

The rest of the paper is organized as follows: Section 2 describes the teaching background in our experience. Then, section 3 introduces the learning design we implemented. Section 4 describes the experiment settings, and discusses results. Finally, in section 5 we provide conclusions and an outline of our future work.

The game we have designed for the current project, is based on the pedagogical framework of mobile learning. As outlined by Sharples and other researchers [21] mobile learning theories recognize not only “(...) the essential role of mobility and communication in the process of learning (...)” but also the importance of the context for meaning construction. Sharples has also described mobile learning as a process in which knowledge is gained through multiple contexts amongst people and personal interactive technologies”. In line with the aforementioned aspects Sharples et al. propose a framework that is based on Engeström’s expansive activity model [14] and which aims to analyze the interdependencies between learning and technology. The framework he proposes is mainly based on five factors (subject, object, context, tools and communication), which are themselves analyzed under two different layers: a technological and a semiotic layer. The technological layer describes learning as “an engagement with technology, in which devices such as mobile phones (...) function as interactive agents (...)” that help its users to acquire knowledge, to communicate amongst each other, to share and negotiate contents and meanings and, finally, to reflect on them. In contrast, the semiotic layer describes learning as “a cognitive system in which learners’ objective-oriented actions are mediated by cultural tools and signs”.

Guess it! has been designed on the aforementioned framework combined with a social constructivist approach. The latter considers learning as “an active process of building knowledge and skills through practice within a supportive community” [21]. Hence some of the key-issues, when starting with the design, were that in order to guarantee an effective learning environment, the APP should be learner-, knowledge-, assessment- as well as community centered [20, 2]. Hereby we mean firstly, that content builds on students’ skills, background knowledge and interests [21, 18], secondly, that the curriculum is built from “sound foundation of validated knowledge” (knowledge centered) [21], thirdly, that assessment is matched to our students’ ability offering them individualized feedback and formative guidance (assessment centered). And fourth, that the APP should engage learners in sharing knowledge and in supporting less able students (community centered) [21, 20].

Guess it! aims to offer learners a highly interactive and flexible learning environment that allows them to ubiquitously get and produce new language input/output. This is done by providing several tasks that focus on students’ reading and writing skills. All tasks are based on the idea of guessing and/ or explaining different words with the help of definitions in the target language. The initial base of definitions is gradually created, assessed and reported by the players themselves, being formative [25, 16, 24]. This way, our learners are encouraged to participate actively in their learning process by reviewing and reflecting on the target language and hence gradually co-constructing the system knowledge [19, 4]. In addition to this and in order to make the different learning tasks more challenging we gamificated the APP by adding features such as a score system, rankings, a time-limit as well as a competitive game model [13, 3, 17].

---

\(^1\) https://github.com/AlbertoCejas/GermanLearningUCA

\(^2\) https://play.google.com/store/apps/details?id=es.uca.tabu
Furthermore, before each game-session the players have to choose the level, categories and number of definitions they want to play. In this way, the learning process is highly customizable according to the educational needs: for instance, students with a higher level compared to the rest of the class can play more advanced levels, whilst those with a lower language proficiency can play already played terms, as many times as they need. Once within the game screen (figure 2), the APP will provide the players with a randomly selected definition of a term which they then have to guess with the optional help of a clue (e.g. Definition of cold: In Alaska it is always ____. Clue: In the summer it is usually hot, but in the winter it is often ____).

![Figure 2. Screenshot of the APP displaying a definition to guess.](image)

For each term players get three opportunities to guess it. After guessing or failing the answer and wasting their attempts, the system will give players instant and explicit feedback by providing them the right answer. Feedback aims here to be constructive by supporting and guiding students during their autonomous learning process. Feedback information has therefore been designed to help learners noticing possible strengths as well as weaknesses, to either build on these or revise them and thus to guarantee that learners can succeed in their learning process [23]. Additionally, players get the opportunity to listen to the pronunciation of the displayed definition which, due to some linguists, enhances learners’ incidental vocabulary acquisition [7]. Furthermore learners can also copy the definition (or part of it) to their notebook and store it for a later review. Finally, before passing to the next definition, users are always asked to assess the quality of the previously played one. In case they consider the definition inappropriate, they can report it to their supervisor, but they must indicate the reason for their report: improper or offensive content, serious linguistic mistakes and not accurate definition.

At the end of each game, an overall summary is shown along with some stats to gamify the APP. However, the Statistics option from the main menu will preserve the last game data along with a bulk of additional useful information regarding all games and definitions played. And thus the Statistics provide data concerning the success ratio, the number of played definitions, the categories where the player guessed most or least definitions along with the played levels as well as with the most selected ones. Furthermore there is also provided some additional information concerning the definitions which have previously been introduced by the user him/herself, including the average rating, the average success ratio and the number of reports.

This is the usual game flow. But, for every twenty definitions a student works on he is asked to create a new definition for a term the system automatically provides him with. That term will be included in the database together with those introduced by other mates as well as those previously introduced by the teacher. As our target students were from the A1.1 level of the CEFR (Common European Framework of Reference for Languages) the terms, we have introduced in the current APP, have been selected from that level.

This makes the game highly dynamic and interactive, since learners can actively take part in the game development by tailoring it for their personal learning needs. As new definitions will be played and peer-assessed by the other mates, students try to do their best when providing a new definition. This way, knowledge is being increased and continuously reviewed by the users themselves. As a result of this process, the server will contain different graded definitions for each term.

The server program stores the knowledge base, which includes the following data: the definitions available for each term, the definitions each student did guess or fail during each game-session, the assessment of each player and the terms that each player considered interesting to be restudied. With that information the system provides its users with some useful feedback showing their learning progress. Additionally this knowledge is being increased and continuously assessed by the users themselves. As a result of this process, the server will contain different definitions and grades for each term.

In order to monitor and assess the process the course coordinator can have access to different learning analytics that include: the usage which has been made by each student of the APP, the grades he obtained (according to the terms that have been guessed and the grades her/his definitions have been given to), the grade he provided for each definition, a report of the low-graded definitions can provide insights of the difficulties a certain student (or a group of them) might have, another with those definitions that unfairly received high grades can highlight those students who are unable to detect mistakes in the target language. In addition to this, the system will provide the supervisor with different...
recommendations. For example, if the ratio of students guessing a certain definition is below a certain threshold, probably its linguistic level is more complex than it should be.

Finally, the server can automatically create personalized tests for each player according to the supervisor’s instructions. In particular, the tests can include a certain number of random definitions as well as those definitions that the student previously guessed several times. In this way, the test can detect if he guessed them by cheating, for instance, a dictionary or if he guessed them by really having the required knowledge. Additionally, the personalized tests might also include a certain number of definitions that a student either did not guess or did not focus on during the different game sessions.

4. EXPERIMENT

4.1 Settings

The experiment was done with more than 120 students from an A1.1 level German Foreign Language course (CEFR) at the University of Cadiz (Spain). All participating students were used to be exposed, from the very first moment of their German language course, solely to the target language and to use this as the only vehicle to interact and communicate with other classmates.

The experience lasted about 4 weeks. During that time students were asked to focus on levels 1 and 2 (first week), then on level 3 (second week) and finally on levels 1 to 4 (third and fourth week). Before giving students access to the APP Guess it! we aimed to test their previous vocabulary knowledge by asking them to take part in a pretest. This pretest focused on about 60 questions that were randomly selected by us from the knowledge base and comprised vocabulary from all four levels. Once students had taken part in the pretest, they were allowed to access the APP and to get familiar with its interface, toolbox and learning contents. To give students some initial guidance on how to use the different tools provided by the APP (notebook, statistics, game, etc.) we gave them a short training session, which took place within the German language seminar.

After this students were invited to use the APP for their independent learning beyond the classroom focusing each week on different levels (see above). The idea was to familiarize students first with the APP itself and the opportunities it provides learners with to get actively involved in their own learning process (e.g. by guessing, reporting and even defining themselves words). For this purpose we recommended students first to focus solely on the levels 1 and 2, which comprise mainly the already known vocabulary from the previous language classes. Once students had practiced the levels 1 and 2 they were asked to pass to the following levels. Unlike the first two levels, levels 3 and 4 were only in part known to our students and aimed to face learners with new challenges to widen their knowledge.

After one week using the APP, students were asked to fill in a first posttest, followed by a second and third posttest, each of which took place in an interval of one week. All tests were done off-line and focused as the pretest did, on about 60 questions that were randomly selected from the knowledge base. Nonetheless, unlike the pretest, the three posttests focused all on different levels. Whilst the first posttest focused on levels 1 and 2 (223 definitions), the second posttest focused on level 3 (65 definitions) whereas the third posttest focused again on levels 1 to 4. However, this time its content had been selected on the basis of a much larger corpus (826 definitions) compared to when we took the pretest (282 definitions), since students started introducing, from the second week of the experiment on, new definitions and hence made the knowledge base growing each week.

4.2 Analysis and discussion

If we look at the tests we get the following insights. Firstly, all of the 100 students who took part in all 4 tests (1 pretest plus 3 posttests) improved very much their test scores, when comparing their results from the pretest with those obtained in the last posttest. The latter was completed by all students at the end of our experience. The positive results are especially valuable if we take into account that the knowledge base increased every day. When preparing the tests, we were always concerned about including at least 50% of the definitions proposed by the students themselves blended with another 50% definitions by the supervisor. The definitions were selected following linguistic as well as content-based criteria. With regard to the linguistic criteria we tried to select solely those definitions, which were linguistically speaking good definitions, whereas with regard to the content-based criteria we tried to include mostly very clever and outstanding definitions.

Additionally, in the last posttest we also included a number of extra terms to be defined. This way, we aimed to reward especially those students, with additional scores, who enriched the knowledge base by introducing a number of new definitions, instead of playing solely those provided by their mates or instructor. The difference between the pre and the last post is noteworthy: an average of 5,35 points of difference in a range from 0 to 10. We might assume that this difference could be due to the extra scores, but nonetheless the difference remains still significant: 4,8 points of gain average from pretest to first posttest and 5,7 from pretest to the second one. In fact, we can see that students made their best in the second posttest, as the average grade evolved from 1,2 in the pretest to 6 in the first posttest, getting to its maximum in the second posttest with 7,1 and ending up in 6,5 points in the last posttest.

An analysis of the server logs provide interesting automated indicators of different skills. For instance, if we look at the time dedicated to play, we could easily check constant study skill checking if each student played a certain minimum number of definitions every period of time (day, week) or total working effort (if we check the whole experiment length). In fact, 150 of the students registered in the system, playing a total amount of 165178 definitions. We estimate that every definition needs a minimum of 20 seconds to be read and answered, and another 10 seconds in average to be assessed and if considered relevant, to be copied down by the users in their APP notebook. According to
this estimation, each student had dedicated at least more than 9 hours to study using the APP. We consider this is quite a lot of time bearing in mind that our pilot study lasted only 4 weeks.

To complement previous indicators, we can retrieve from the system if students did actually guess the definitions or not. In fact, some students acknowledged, once the experiment was finished, that they only played the APP to note down the vocabulary in a physical notebook and then, to study it following the traditional way, rather than using the APP as their main learning device. So we can check their success ratio to get objective evidences of the level of definitions included in the APP and also if students did prefer to use the APP compared to the traditional way of studying vocabulary (flashcards, etc.) In figure 3 we can see the general statistics of guessing: each column indicates the ratio of definitions played and which of them were guessed in the interval from the column label to the new one. For instance, the first column on the left indicates that 4% of the definitions played were scarcely guessed, ranging from 0% to 10%. The second column indicates that around 2% of the definitions played by the students were guessed between 10% and 20%, etc. If we look at the definitions guessed more than 50% appeared in a game (adding the results of the five columns on the right, we can see that 65% of the definitions played were actually guessed. This means that students usually succeeded in guessing the definitions (note that this includes definitions that the server randomly assigned several times to each student).

![Figure 3. Ratio of guessed definitions chart](image)

To measure their contributions to the system (what could be participation skill) we can check the number of contributions from each student. In our case, the APP started up with 282 definitions provided by the teacher, but ended up with 850, so students contributed 568 definitions. This is a mean of almost 4 definitions per student. Although this figure is not quite high, looking at the previous commented high guessing rate, we think that they we good definitions, as mates could usually guess them.

As for peer-assessment skills, the mean difference between supervisor grades and peer-grades is 1.28 in a scale from 0 to 5 points. The difference between supervisor grades and peer assessment range from 0 points (more than 32,000 assessments were exactly the same grades given by the supervisor) to 5 (more than 4,800 times). In this second case, it was usually the supervisor reporting and the students giving the maximum grade (only 22 times was the opposite). As for individual assessment only 11 students got a mean difference below 1 point when comparing their assessment to that of the supervisor, 115 got between 1 and 2 points, and only 9 got between 2 and 3 (none of the students got more than 2.5 points of difference).

Students assessed 85% of the definitions with more than 4 points out of 5, and 12% with between 3 and 4 points. So 97% of the definitions were assessed as good/very good for the students. Those definitions assessed with 0 points were automatically reported to the supervisor. There were more than 11,000 reports. One hundred definitions were never reported. Some definitions were reported more than one hundred times, so there is little doubt that they must be wrong. But other 395 definitions were reported occasionally (less than 10 reports), so probably the reporting students were wrong. Surprisingly, several students reported even definitions provided by the teacher. And thus 19 reports considered definitions as improper or offensive content, whereas 80 considered that they contained serious linguistic mistakes and 7 of them were due to not accurate definition.

5. CONCLUSIONS

In this paper we have shown how our gamificated multilingual App Guess it! works and what learning objectives it can be used for. It is a client-server computer software which has been developed specifically for the current experiment. Whilst most of the APPs only offer static and supervisor created content, Guess it! is dynamically feed with new content from each user of the App. This way, learners get actively involved in their own learning process. In line with this, each student has to write definitions for a specific term which has previously been delivered to him by the system. Once the student has done this, the definition enters in the game and other mates have to guess the term defined. Since all definitions are peer-assessed by other mates, whilst playing the game, the author of each definition gets the opportunity to receive constant meaningful feedback on his performance. This feedback aims at providing him with the opportunity to critically review and to develop his language skills.

Furthermore this process facilitates monitoring and automated assessment by the language instructor.

In our first experiment using Guess It! we employed the APP in a Higher Education German Foreign Language course (A1.1 level of the CEFR). Students played the APP for a period of four weeks, focusing on different levels of vocabulary. Results show a significant increase when comparing post and pretests. Probably this is related to the time students spent on playing the APP, which was a much longer time than we estimate they usually employ for their independent learning beyond the classroom. These results are complemented with objective indicators on the acquisition of other skills that are automatically extracted from system logs. Those refer, for instance, to students’ ability to explain the terms in a foreign language or the competence to assess definitions of mates. Of course, these results need further study to get a stronger conclusion on the validity of the specific usage of the information retrieved.
Our next future work will focus on making a more detailed analysis of the results, trying to find relations between the different independent indicators obtained and the final course grades. Additionally we aim to identify students’ profiles to early detect those students who could fail the course.

6. ACKNOWLEDGMENTS

This work has been funded by the Andalusian Government under the University of Cadiz programme for Researching and Innovation in Education. And by European Union under the OpenDiscoverySpace (CIP-ICT-PSP-2011-5) project. We would also like to thank the OSLUCA, AUHR, as well as Elhí Salcedo, Pilar Romero, Ignacio Calleja, Andrea Calderón and Sebastián Pitula for their support.

7. REFERENCES


