Proceedings of the 9th Games and Natural Language Processing Workshop @LREC2022, pages 7–16
Marseille, 25 June 2022

Applying gamification incentives in the Revita language-learning system

Jue Hou[1], Ilmari Kylliäinen[2], Anisia Katinskaia[1], Giacomo Furlan[1] and Roman Yangarber[2]
[1]Department of Computer Science
[2]Department of Digital Humanities
University of Helsinki
{firstname.lastname}@helsinki.fi

Abstract

We explore the importance of gamification features in a language-learning platform designed for intermediate-to-advanced learners. Our main thesis is: learning toward advanced levels requires a massive investment of time. If the learner engages in more practice sessions, and if the practice sessions are longer, we can expect the results to be better. This principle appears to be tautologically self-evident. Yet, keeping the learner engaged in general—and building gamification features in particular—requires substantial efforts on the part of developers. Our goal is to keep the learner engaged in long practice sessions over many months—rather than for the short-term. In academic *research* on language learning, resources are typically scarce, and gamification usually is not considered an essential priority for allocating resources. We argue in favor of giving serious consideration to gamification in the language-learning setting—as a means of enabling in-depth research. In this paper, we introduce several gamification incentives in the Revita language-learning platform. We discuss the problems in obtaining quantitative measures of the effectiveness of gamification features.

Keywords: Language learning, Gamification, Natural language Processing, Intelligent Tutoring Systems

1. Introduction

Learning a language toward intermediate or advanced levels requires a massive investment of time on the part of the learner. Some statistics from the Foreign Service Institute, USA,[1], in Table 1, show the number of *contact* hours required for an English speaker, on average, to reach upper-intermediate level of proficiency, typically needed for diplomatic service.

In principle, a language learning platform can serve as a powerful research tool. On one hand, it can provide real value to learners. On the other hand, it can provide invaluable data to researchers—about possible learning paths, common patterns of mistakes, etc.,—which can drive research in educational data science (EDS), learning analytics, and computational didactics. We believe this kind of data is essential for real progress in EDS—we need to collect data on a massive scale, tracking learner progress over time.

This kind of longitudinal data cannot be collected without engaging the learner over extended periods of time. If the platform offers limited learning content, a "toy" learning environment, or repetitive, monotonous means of engagement, then it will allow us to collect sufficient data to serve as a foundation for in-depth research.

Gamification is the strategic attempt to enhance systems, services, and activities to create a user experience akin to playing a game—in order to engage and motivate users. Game-design elements and principles are implemented in several non-gaming contexts, including education, data collection, and data labeling (Chamberlain et al., 2013; von Ahn et al., 2006).

In this paper, we discuss several gamification strategies applied in an Intelligent Tutoring system (ITS) for language learning, Revita[2], and discuss the impacts that gamification has achieved so far in this experimental setting. Revita—a project for supporting intermediate-to-advanced language learners—is an international collaboration between several European universities. The collaborators include specialists in language teaching and didactics, currently with hundreds of university students using the platform on a regular basis. The experimental setting we describe involves applying Revita in the context of several universities.

In this paper, we evaluate the effectiveness of gamification incentives in Revita and discuss the preliminary results and problems highlighted by the evaluation. We believe that research in gamification can facilitate personalized tutoring and enhance the learning experience—which in turn will improve learner engagement, and lead to a positive feedback loop: more

Language	*Hours*
French, German, Italian, Portuguese, Romanian, Spanish, Swedish, Dutch, Norwegian, Afrikaans	600
Indonesian, Malaysian, Swahili	850
Albanian, Amharic, Azerbaijani, Bulgarian, Finnish, Greek, Hebrew, Hindi, Hungarian, Icelandic, Khmer, Latvian, Nepali, Polish, Russian, Serbian, Tagalog, Thai, Turkish, Urdu, Vietnamese, Zulu	1,100
Georgian, Mongolian	1,600
Arabic, Chinese, Japanese, Korean	2,200
Compare:	
4 years of college (8 semesters × 50 hr)	400
Child reaching fluency (2–4 years × 10 hr/day)	7.5–15K

Table 1: Estimates of *contact hours* required for native English speakers to reach fluency in various languages, on average. (Statistics: Foreign Service Institute)

[1]www.state.gov/foreign-language-training/

[2]https://revita.cs.helsinki.fi

learner data enables the development of better models, which provides a better service to the learners.

The paper is organized as follows. Section 2 reviews relevant prior work. Section 3, reviews the Revita platform for language learning toward advanced levels. Section 4 describes "hard-value"—or *competency*-related—incentives in the learning system. In Section 5, we discuss "soft-value"—or *enjoyment*-related—incentives supported or planned in system. In Section 6, we discuss a preliminary evaluation of the gamification elements in our experimental environment. In Section 7, we summarize the contributions and the future work.

2. Prior Work

2.1. GWAP

GWAP—games with a purpose, introduced in (von Ahn, 2006)—is using games to leverage human brain power to solve open problems. As a side effect of the game, annotated data is collected. von Ahn and Dabbish (2008) propose three general gaming mechanisms:

- Output agreement games: Players are randomly paired, and given a shared visible input. They attempt to achieve agreement with each other on output (not shared).
- Inversion problem games: Players are randomly paired. One plays as the describer, while the other plays as the guesser.
- Input agreement game: Two randomly paired players are given an input object. They need to describe the inputs to each other, to decide whether their inputs are the same.

Research on games and psychology shows that 8 major elements make games entertaining and enjoyable (Koster, 2004; Sweetser and Wyeth, 2005; Csikszentmihalyi, 1991):

- Concentration
- Challenge
- Immersion
- Supporting player skills
- Clear goals
- Feedback
- Social interaction
- Player's sense of being in control

These gamification principles are taken into consideration in several GWAP applications, some of which have proven to be effective for collecting data from users. For example, von Ahn et al. (2006) and Ho et al. (2009) work on image recognition. Chamberlain et al. (2013), Madge et al. (2019b), Madge et al. (2019a) and Fort et al. (2014), work on text annotation. Several papers focus on collecting data for recommendation systems (Walsh and Golbeck, 2010; Banks et al., 2015) and knowledge repositories (Herdagdelen and Baroni, 2010; Herdağdelen and Baroni, 2012) via GWAP.

2.2. ITS

Computer-assisted language learning (CALL) is a research area introduced over 50 years ago. CALL is broadly defined as "the search for and study of applications of the computer in language teaching and learning" (Levy, 1997). It is not intended to be a replacement for the teacher. As CALL developed, ITS emerged with the goal of "computer as a tutor." ITSs have been adopted in various knowledge domains, including mathematics, sciences and language learning (Slavuj et al., 2015). One popular language-learning ITS is Duolingo.[3]

A key goal of ITS is to model the learners' knowledge and skill levels. Several approaches have been proposed, including Bayesian Knowledge Tracing (BKT) (Corbett and Anderson, 1994), Learning Factor Analysis (Cen et al., 2006), and its more advanced variant, Performance Factor Analysis (Pavlik Jr et al., 2009). In this paper, we discuss the application of the Elo rating system designed for zero-sum games.

3. Language Learning Platform

Revita is a freely available online platform, for supporting language learning/tutoring *beyond the beginner level*, (Katinskaia et al., 2017; Katinskaia et al., 2018). Many free and commercial resources and applications exist on the Web, which support beginners, some with millions of users. However, once the learner has passed the beginner level, and reached low-intermediate to advanced (LI-A) level—i.e., above A1/A2 on the CEFR scale—resources available to her become drastically limited. As surveys show, very few systems today provide substantial support for LI-A learners in multiple languages.

The Revita language-learning system primarily targets "high-stakes" learners—users who are invested in the learning for the long run, and have an internal motivation for learning, such as the need to pass university courses, for work, citizenship, etc.

Revita currently supports several languages—in various stages of development, ranging from initial, "beta" versions to fairly well-developed ones. The languages include "big" languages—Finnish, Russian, Italian$^{\beta}$, German$^{\beta}$, Kazakh$^{\beta}$, Swedish$^{\beta}$, Mandarin$^{\beta}$—and several endangered minority languages, including many Finno-Ugric languages in Russia.

Revita builds on educational data collected through a collaborative effort with language teachers at several universities. In this paper, we focus on the evaluation in our experimental setting, at several major universities, where hundreds of students enroll in Russian language courses at various levels. The teachers suggest to their students to use Revita to solidify their knowledge through practice sessions, and to prepare for exams. Currently, we collect data about the students' progress in three practice contexts:

Story exercises: Students practice by doing exercises based on texts. One set of exercises is given for each

[3] https://www.duolingo.com

snippet of the text—about one paragraph. Each exercise is linked to one or more linguistic "concepts"—technically known as *constructs*. Each concept is a "skill" that the learner must master, for example: the usage of genitive plural nouns belonging to a certain paradigm, or verb government, etc. The inventory of concepts for a well-developed language is many hundreds, up to about 1.5 thousand. The user response data for each exercise contains: the correct answer, student answer (if incorrect), concepts linked to the exercise, timestamp. The system offers various types of exercises: multiple-choice questions, "cloze" (fill-in-the-blank) questions, listening comprehension, etc. These exercises are generated automatically based on the text chosen by the learner.

Flashcards: While working with texts, the students can request translations for any unfamiliar words. All requested translations are stored in the student's deck of flashcards. Students practice their vocabulary by playing with flashcards, in batches with *timed repetition*. Two types of flashcards are currently available: translation, and gender selection—important for German, French, Swedish, etc., languages where the gender of most nouns is not obvious from the noun's form. The response data consists of: student's answers to a flashcard, timestamp. Learners can upload and edit their own flashcards. We assume that the reason a learner clicked a word in text for translation is because it is unfamiliar. Also, the sentence/context where the word was encountered is attached to each flashcard as a hint.

Tests: Students can take online tests through the platform (for some of the languages). Teachers can configure the topics of the test items and their number. Items are sampled from a database of about 2000+ multiple-choice questions. The test can also be *adaptive*, where the system picks the items depending in the learner's previous questions. Tests are timed—each question has a time limit, typically 30 seconds. Like the story exercises, each test item is linked to one of the concepts implemented for the language. The questions are prepared by language teachers and linguistic experts, e.g., (Kopotev, 2012; Kopotev, 2010). At the time of this writing, the response data consists of 875000 test answers, by over 5000 learners. For each question, the system records to which concept the question belongs, whether the answer was correct, and a timestamp.

4. Improving Competency as Incentive

A crucial aspect of gamification is providing value—or incentives—to motivate users to practice longer. We can informally distinguish two kinds of value: "hard" value relates to improving competency and growing skills; "soft" value relates to spending time in an entertaining and enjoyable fashion. In the context of high-stakes language learners, the primary motivation is obtaining *hard value* from the learning system by increasing competency. However, that does not mean no other motivators are in play. In fact, we believe that "soft value" or *enjoyment incentives*—discussed in the next section—affect the user's involvement in the learning process in equal measure with hard value incentives.

We next briefly discuss what we consider to be the primary hard-value incentives that Revita offers to learners: *interesting content, assessment, and feedback*. As a learner interacts with a human teacher, she expects to receive all of these, in order to stimulate and guide her progress toward linguistic mastery. Thus it is reasonable for an automated tutoring system to aim to provide similar value.

Assessment of user performance is considered to be an important incentive. Assessment brings incentives not only on the personal level, but also on the social level—since students can compare their performance with classmates, or other learners in the platform.

4.1. User-selected Content

A key motivator in Revita's approach is encouraging the learner to select arbitrary authentic texts—which correspond with her own, personal interest *outside* the language learning context—and using this arbitrary chosen material as content for learning. This is done by automatically generating a wide variety of exercises based on the text content chosen by the user, using language technology and AI. This is a key principle in the Revita approach to tutoring.

The principle is based on the assumption that if the learner can work with topics that pose an inherent interest to her—independently of the language learning objectives—then she will spend more time engaging with the content, and hence more time practicing. Recall, our overall goal is to maximize the time which the learner invests in practicing with the language.

4.2. Elo Ratings for Language Learning

Revita adopts the Elo rating system to rate learners. The Elo rating system was originally developed for chess, and has received wide acceptance in many of the currently popular online and e-sport games. Earlier attempts have been made to apply Elo in the context of ITS, (Klinkenberg et al., 2011; Pelánek, 2016).

The Elo rating system is designed for zero-sum games, and is usually applied for Player vs. Player games (PvP). Its formula defines the **expected** result of actor A in a match against actor B according to the formula:

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{\sigma}}} \tag{1}$$

E_A is the expectation (probability) that actor A will succeed, or win. R_X refers to the current Elo rating of actor X, and σ is a scaling factor.

After a match with another actor is completed, the rating of actor A is updated according to the formula:

$$R_A^{i+1} = R_A^i + K(S_A - E_A) \tag{2}$$

where S_A refers to the actual score achieved by actor A in the match: loss, draw and win for A are counted as 0,

0.5 and 1 points, respectively. The factor K determines the maximal change in the rating at one time.

In Revita, the Elo equations are used so that, rather than playing against each other, users "play against" exercises in a text, language concepts, or vocabulary items. Revita scores users in the various practice modes: story exercises, flashcards, and tests. Experiments have shown that this approach to rating the user's competency gives consistent results between the exercise Elo rating and the test Elo rating, and correlates well with external competency judgements made independently by human teachers, (Hou et al., 2019).

4.2.1. Elo Ratings in Tests

In the test setting, one "match" refers to attempt by a student to answer a question related to a given concept from the concept inventory. The two rated "actors" are the student and the concept. The rating R_A of student A represents the ability of the student. The rating R_C of a question involving concept C models the *difficulty* of the concept.

One difference compared to the original Elo system, is that students have some chance of guessing correctly on *multiple-choice* problems. To compensate for this bias, Revita adopts the approach recommended by Pelánek (2016), penalizing the expected value by the probability that a random guess is correct:

$$E_A = \frac{1}{k} \cdot 1 + (1 - \frac{1}{k}) \cdot \frac{1}{1 + 10^{\frac{R_C - R_A}{\sigma}}}, \qquad (3)$$

where k is the number of choices in the multiple-choice question.

We expect that the Elo ratings for concepts will approach their "true" value after a large number of data points—"games," or test answers—have been collected from learners. To improve the quality of concept ratings, they are learned by re-adjusting all ratings by replaying all games in chronological order over several epochs. This corresponds to the Elo "*burn-in*" period, used to obtain stable ratings for all concepts currently implemented in the system for the given language.

4.2.2. Elo Ratings in Story Exercises

Revita generates exercises for each snippet of text (about one paragraph), one snippet at a time. Exercises are of different types. Each exercise is linked to one or more linguistic concept. An exercise can be rated by taking the maximum rating of the concepts linked to the exercise.

Alternatively, the system can make the simplifying assumption that the exercises in a given text will correspond *on average* to the *difficulty* of the entire text. Revita currently has models that estimate the difficulty of a text for several languages. When the learner selects a text and uploads it to the system, its difficulty is estimated by a model trained on a corpus of texts whose difficulty had been manually rated by experts.

Modeling the difficulty of a text—or its readability, complexity, etc.—is a well-studied problem, (Dubay, 2009). The model can use lexical and grammatical features, e.g., (Chen and Meurers, 2016; Heilman et al., 2008). Revita uses linear models and standard features, recommended, e.g., by Kincaid et al. (1975), Flesch (1979), and Chen and Meurers (2016), to estimate the difficulty of a text: including lexical frequency, mean token length, mean sentence length, etc.

When the exercise rating is defined in terms of average text difficulty, S_A can again denote the actual score that student A received when answering a given exercise. E_A for the exercise is assigned according to the difficulty of the text, from which the exercises are drawn.

The output of the model is scaled onto the Elo rating scale. This allows the system to estimate the performance of any rated learner on any rated text. The learner's Elo is updated after *each* answer. Further, the system updates the Elo rating of the entire text *relative to this learner* after a complete pass by the learner through the text. The rationale for updating the relative difficulty of the text is that every time the learner goes through the text, the text becomes more familiar, and therefore relatively "easier" for the given learner. Note, that since Revita selects the exercises presented to the user on each pass randomly, the actual exercises will, in general, be different on repeated passes through the text.

4.2.3. Elo Ratings in Flashcards

In the context of practicing with flashcards, the notion of a "game" is similar to the notion of a game in the context of story-based exercises, above. S_A is defined as the actual score that student A received when attempting a batch of flashcards, for example, 20 or 50. The expectation E_A for a batch of flashcards is the average Elo score of each flashcard (word). The Elo score of a flashcard/word is scaled from its Inverse Document Frequency (IDF), which is considered to be a good estimate of its difficulty level. The scaling is a mapping from the ranges of lexical frequencies to the corresponding ranges of Elo scores; this is done by experts in language pedagogy.

4.3. Feedback

In the story-based exercise mode, the learner can make *multiple attempts* to answer an exercise. After the learner answers the exercise, the system does not simply reply "correct" or "incorrect," and show the learner the correct answer in case the answer was incorrect. Rather, after each attempt, for each exercise that has not yet been answered correctly, the system returns to the learner personalized *feedback* based on her answers. Feedback comes in the form of additional hints, which *gradually* guide the learner toward the correct answer. The goal is to help the user to learn to arrive at the correct answer on her own, by developing the habit of searching the context of the exercise for clues, which indicate the correct answer.

This graduated feedback follows the foundational didactic principles of Dynamic Assessment in second

Figure 1: Examples of feedback for story exercises (in Russian). The green part of the tool-tip contains feedback to the learner: why her answer is incorrect, and hints about how to correct it. (The user can click on the blue part to request a translation for the given word, which is available for *all* words in the text).

language teaching, e.g., (Poehner, 2008). Revita's feedback module 1. analyzes the learner's answer, and 2. tries to establish which hints are most suitable, given how the learner has answered so far. Feedback is based on syntactic information found in the context of the exercise. For example, agreement—elements of a noun phrase must agree in number, case, gender, etc.—or syntactic government—a verb has certain *valence*, or its arguments are required to be in a certain case, etc.

Feedback is also based on a detailed *hierarchy* of linguistic features—which features of a word or phrase have higher priority than other features. For example, the priorities for language L might indicate that if a verb form is incorrect, then the learner should first try to get the correct mood and tense—before correcting the person and number. This hierarchy of priorities are defined in collaboration with experts in linguistics and didactics, for each language.

Figure 1 shows examples of feedback that a learner may receive after attempting to answer a story exercise. The circled border shows the phrase structure surrounding the cloze exercise, and hints at the *agreement* relationships that must not be violated within the phrase. The blue underline shows that there is a *government* relationship between the verb and a phrase that it governs. The green part of the tool-tip contains the feedback and hints that the user receives after the previous attempt.

The examples on the bottom show how the progressive feedback becomes more specific as the learner proceeds, until she finds the correct answer—or exceeds the maximum number of attempts. On the left, the hint says that the gender is incorrect; on the right, it gives the specific gender needed in this context.

5. Enjoyment as Incentive

As discussed in (von Ahn and Dabbish, 2008)—in the context of GWAP—users play not (only) because they are personally interested in solving an instance of a computational problem, but because they like to be entertained.

We next describe several features that Revita tries to provide as enjoyment incentives.

5.1. Crossword

The crossword stimulates further practicing with grammar and vocabulary problems based on the text that the user may have worked with earlier, but while working in a different setting, which is more akin to solving a puzzle. A crossword is based on any text chosen by the learner; words in the crossword are automatically and randomly selected from the text. To complete the crossword, the learner inserts each missing word into the story, in its correct inflected form. The clues are the translations of the missing words, rather than their lemmas, as in story-based exercises. Figure 2 shows an example of a crossword built from a news story.

5.2. Social Interaction

Friend and Sharing: As a social feature in Revita, it allows learners to share any content that they find interesting. Stories can be shared among friends, with a message attached. When a learner shares a story with another, an email notification is sent. The receiver can accept or reject the shared content, and accept the sender as a "friend", so future sharing will require no notification, or block the sender. User can also share arbitrary own *notes* that they can attach anywhere in the story.

Sharing with a group of learners is also possible. A teacher can create a group, and invite learners into the group. This feature supports the collaboration with teachers, since it allows the teacher to supervise a class of students. The teacher can invite them to join a group through the platform (which requires an email confirmation by the student), or send the invitees an encrypted pass-key to the group.

Competition Mode: The competition mode in Revita is related to story-based exercises. Regular exercises, described in section 4.2.2, allow the user unlimited time to answer. The purpose of the competition mode is to challenge the learners to make correct answers, but under time constraints.

In competition mode, the learner and the opponent work on identical exercises (based on the story chosen by the learner). The objective is to complete the exercises faster than the opponent, while making fewer mistakes than the opponent. The competition ends when one of the players—the learner or the opponent—reaches the end of the story. Whoever answered more exercises correctly is the winner. This effectively combines the drive for A. answering exercises correctly, and B. doing so within shorter time.

Revita creates an opponent—a "bot"—with which the learner will compete. The bot's parameters are tuned to match the human learner's previous performance: the learner's own reading speed, the learner's answering speed, and the learner's answer accuracy—these are all calculated based on the learner's past history.

Figure 2: Example of crossword for a story (in Finnish). Left to right: the crossword board, the text, the clue and translation box.

Figure 3: Leaderboard for *time spent* practicing on the platform. The board shows the top 3 learners from last week, and the leaders for the running week. Previous leaderboard achievements are denoted by numbers inside gold, silver and bronze medals. (The users' names have been blurred to protect their privacy.)

Thus, the bot aims to imitate a learner's performance as closely as possible. In this way, the learner is assured that the opponent is optimally matched to her skills—not much weaker and not much stronger. Since the opponent is optimally matched to the learner, the competition is optimally challenging, and the learner is essentially trying to surpass *her own prior performance*—to reach above her current skill level.

In the future, we plan to collect more detailed information about the learner's performance, e.g.: key-stroke frequency, expected response time per concept, etc.

Leaderboards and Achievements: Learners pass milestones on several metrics; currently the system awards "achievements" to the user based on A. the amount of time spent practicing, B. the number of stories the learner has uploaded to the system, and C. the number of stories the learner has practiced through to completion. Each of these metrics has five milestones. Once the learner reaches a milestone, a permanent badge will appear in the learner's achievement collection.

In addition, to encourage a wider-scale competition, Revita maintains a weekly *leaderboard*, tracking the time that the learners spend practicing across all types of exercises.[4] The three top performers each week also receive an achievement—a medal. Figure 3 shows a example leaderboard from a recent week.

6. Evaluation

Our experimental setting involves analyzing data from students at several European universities who are studying Russian and using Revita in conjunction with their coursework. The experimental period spans 10 months—41 weeks—from beginning of July 2021 through beginning of April 2022 (the time of this publication). We chose to begin compiling statistics in July, because that was the time when several major improvements to the support for Russian were released, which spurred the language teachers toward heavier utilization of the system in their teaching.

The activity of the students is recorded in Revita's database. At the time of this writing, the learning activities for which timing information is available in Revita include: story-based exercises, flashcard exercises, creation of new flashcards (which means that the user requested a translation for some unfamiliar word, thereby adding new flashcards to her card deck), and reading a story (without doing exercises).

Other activities—crosswords, competitions, etc.—at present do not have timing information recorded in the database. Therefore, these activities are not included in the present study; they will be the subject of more in-depth investigations on the impacts of gamification on learning in the near future.

[4]To ensure privacy, learners will appear in the leaderboard only if they agree to show their record.

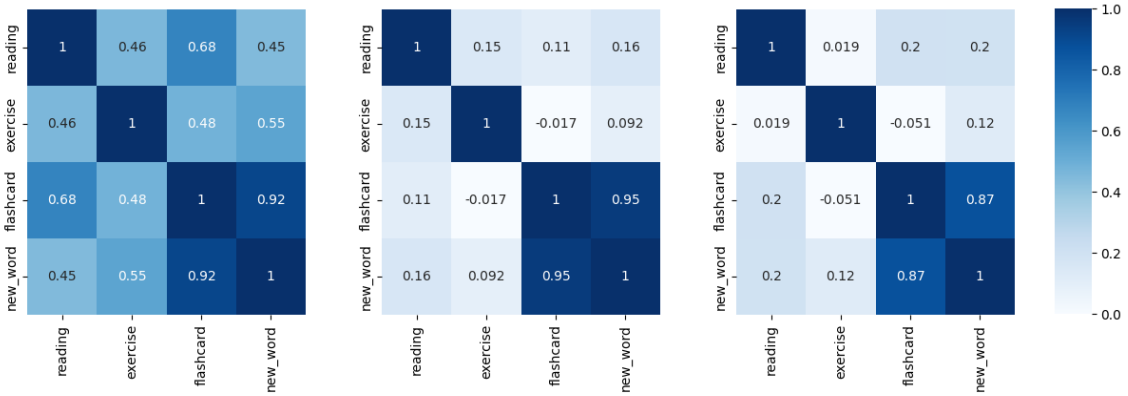

Figure 4: Correlation matrix between four types of user activities, for three populations: top 200 most active students (left), 200–400 (middle), and 400–600 (right).

6.1. Correlation between activities

The matrices in Figure 4 show the pairwise correlations between the various learning activities, for several "populations" of students. We examine the 600 most active students during this period, and split them into three groups according to their activity rank: 0–200, 200–400, and 400–600. Activities labeled *flashcard*, story *exercise* and *new word* indicate the total number of items that a user has answered while practicing with flashcards and story exercises, and the number of translation lookups for unfamiliar words, respectively. We can make several observations from the Figure. The matrices show a high correlation between the *flashcards* and the *new word*.

This is very encouraging, since it shows that those learners who frequently request translations for unfamiliar words, also come back at a later time to practice with the vocabulary flashcards that they have collected over time—rather than looking up translations and never taking the trouble to review them and practice with them.

The lighter squares in the correlation matrix for the top-200 students also provide an interesting insight: they indicate a lower correlation between reading and the creation of new cards (*new word*). That means that people tend to look up unfamiliar words more during exercising than during reading. At the same time, the correlation between reading and card-based exercise is higher than the correlation between reading and story-based exercise. This may suggest that some people prefer to practice with the vocabulary flashcards after reading a story. This confirms that there is *added value* in offering multiple kinds of activities in the system, since different people prefer different activities.

Lastly, we can see that when we move from the top-200 population to the others, all correlations drop substantially (except the correlation between flashcard practice and new words, mentioned above). This may mean that the activities in which the "less-motivated" students engage are less varied and less spread out, more concentrated on one (or very few) types of activities. These observations are further explored in Section 6.4.

6.2. Weekly time spent on practice

The learners in our experimental setting are mainly university students: they are high-stakes users, since working with Revita is part of their curricular activity. The metrics presented in this section show the amount of activity during the given time period.

We measure the time that the students invest in working with Revita. Figure 5 shows the total activity time of the top 200 most active learners across the 41-week experimental period. The patterns that emerge from the Figure reflect the real-world situation:

- Reduced activity between semesters, and at the start of a new semester when students are being introduced to system: Dec 2021–Feb 2022,
- More activity in the middle of semester: Oct 2021–Nov 2021, and Feb 2022–Apr 2022,
- A spike of activity toward the end of semester and near exams: Aug 2021–Sept 2021.

6.3. Correlation between practice and leaderboards

Since the students invested considerably more time from September 2021, during these weeks we calculated the correlation between the user's *leaderboard position* (rank) on a given week N, and extra time spent on during the *following* week $N+1$ compared to week N. The correlations were computed only for students who reached a top-10 position during any of the 41 weeks of activity. The result is a positive correlation of 0.50, which suggests that a high rank on the leaderboard tends to measurably stimulate also more activity during the following week!

This suggests that being closer to the top is a strong motivator for students to work harder: that the leaderboard is an effective incentive to motivate our learners. The leaderboard may have a limited influence on students who do not achieve a relatively high rank. The leaderboard currently indicates only the student's absolute rank, rather than a relative position. We plan to

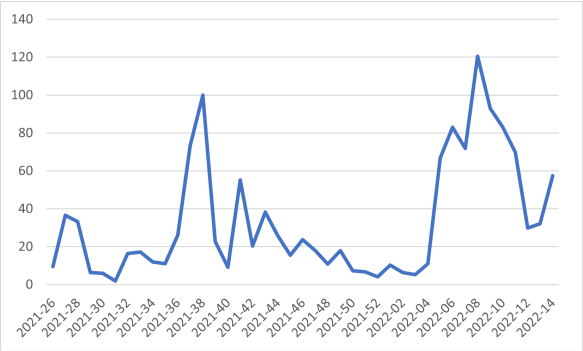

Figure 5: Total weekly hours spent, for 600 most active learners over the last 10 months.

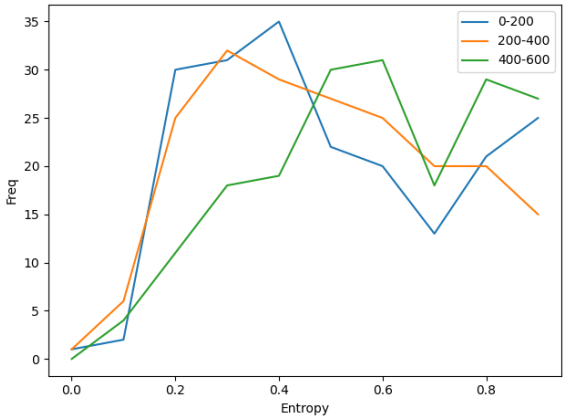

Figure 6: Histogram: entropy of activity of 600 most active users, for 3 populations: top 200 most active students (blue), 200–400 (orange), 400–600 (green). Y-axis: count of students with given entropy.

show also the *relative* percentage in the leaderboard, and check how that will influence all users: are they incentivized to move toward the top if they are told they are in the top 10%? top 20% ? top 50%?

6.4. Learner engagement across activities

For the 600 most active learners during the experimental period, we compute another indicator: the *entropy* of the distribution of the user's time across *different* activity types—namely: story exercises, flashcards, and flashcard creation by looking up new words.[5] We compute the entropy based on the distribution of time across these three classes of activity.[6] This distribution models the "probability" that the user will engage in activity i as simply $\frac{t_i}{\sum_j t_j}$, where $t(i)$ is the amount of time she spent on activity $i \in \{\text{exercise, flashcard, new word}\}$.

One possible conjecture would be that users who spend *more* time on the platform engage in—therefore, *prefer*—a more varied set of activities; that "breaking the monotony" helps the most active users keep the motivation to practice on the platform longer.

Figure 6 is a visualization of the histograms of entropies for the most active 600 users, sub-divided into 3 populations. We make some observations based on these activity entropies across the users. Recall, that the entropies are computed over three kinds of activities (at present). For the top-200 students (blue), the entropy is mostly concentrated on the left side of the graph, for students ranked 200–400 (orange), the entropy moves to the right, and for the least active it's concentrated most on the right. This suggests that the less dedicated learners—who spend less time—tend to "scatter" their time more on different activities. The "bimodal" histogram of the top-200 suggests that these users study with different styles: most focus on few activities (low entropy), while some engage in a variety of activities, spending their time more uniformly.

This also supports the conjecture in Section 6.1: that the most engaged users don't simply click around on words just to get a translation in the moment, when they encounter unfamiliar vocabulary; they actually come back to practice with their flashcards at a later time.

7. Conclusions

In this paper we discuss the range of activities and gamification features that are available at present to users of the Revita ITS. The main contribution is the presentation of our efforts to measure the impacts of the activities and gamification on the effectiveness of learning. Our experiments track a population of 600 learners using Revita at several universities. A key goal in ITS is to provide students with *personalized* learning and support their individual learning process. Achieving this goal requires strong learner engagement.

We explore how offering a variety of activities and gamification—rather than only a narrow selection of exercise types—may help learning, by keeping the learners more engaged. Most importantly, obtaining solid quantitative proof of these conjectures is not a trivial task, and requires extensive longitudinal studies with large numbers of users. Such studies require systems that are sufficiently friendly so that users would be willing to use them for many months at a time. Without actual such systems, conducting in-depth research on engagement is not possible.

In Revita, the gamification efforts are in the early stages, and currently not guided by specific theoretical or precedent-based justifications. We believe that the data we gather from these efforts will help establish new precedents and theoretical foundations.

Future work will include expanding the gamification features of Revita, and more thorough evaluations of learner engagement. We plan to track a more extensive inventory of user activities, which we hope will lead to further interesting findings.

[5] Story *reading* is not included in this calculation, because it is not directly comparable with other activities for now.

[6] Entropy in Figure 6 is normalized to be in $[0, 1]$ by using $\log_3$, since we have 3 classes—the three types of activity.

Acknowledgements

This work was supported in part by the Academy of Finland, Helsinki Institute for Information Technology (HIIT), BusinessFinland (Grant "Revita", 42560/31/2020), and Future Development Fund, Faculty of Arts, University of Helsinki.

8. Bibliographical References

Banks, S., Rafter, R., and Smyth, B. (2015). The recommendation game: Using a game-with-a-purpose to generate recommendation data. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 305–308.

Cen, H., Koedinger, K., and Junker, B. (2006). Learning factors analysis – a general method for cognitive model evaluation and improvement. In Mitsuru Ikeda, et al., editors, *Intelligent Tutoring Systems*, pages 164–175, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using Games to Create Language Resources: Successes and Limitations of the Approach. In Gurevych, et al., editors, *Theory and Applications of Natural Language Processing*, page 42. Springer, January.

Chen, X. and Meurers, D. (2016). Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94.

Corbett, A. T. and Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278.

Csikszentmihalyi, M. (1991). *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, NY, March.

Dubay, W. (2009). Unlocking Language: The classic readability studies. *Professional Communication, IEEE Transactions on*, 51:416 – 417, 01.

Flesch, R. (1979). How to write plain English: Let's start with the formula. *University of Canterbury*.

Fort, K., Guillaume, B., and Chastant, H. (2014). Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 2–6.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79. Association for Computational Linguistics.

Herdagdelen, A. and Baroni, M. (2010). The concept game: Better commonsense knowledge extraction by combining text mining and a game with a purpose. In *2010 AAAI Fall Symposium Series*.

Herdağdelen, A. and Baroni, M. (2012). Bootstrapping a game with a purpose for commonsense collection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–24.

Ho, C.-J., Chang, T.-H., Lee, J.-C., Hsu, J. Y.-j., and Chen, K.-T. (2009). KissKissBan: A competitive human computation game for image annotation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 11–14.

Hou, J., Koppatz, M. W., Quecedo, J. M. H., Stoyanova, N., Kopotev, M., and Yangarber, R. (2019). Modeling language learning using specialized Elo ratings. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications, ACL: 56th annual meeting of the Association for Computational Linguistics*.

Katinskaia, A., Nouri, J., and Yangarber, R. (2017). Revita: a system for language learning and supporting endangered languages. In *6th Workshop on NLP for CALL and 2nd Workshop on NLP for Research on Language Acquisition, at NoDaLiDa*, Gothenburg, Sweden.

Katinskaia, A., Nouri, J., and Yangarber, R. (2018). Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel.

Klinkenberg, S., Straatemeier, M., and van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on-the-fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824.

Kopotev, M. (2010). Система прогрессивного тестирования Karttu: описание и первые результаты (The Karttu system for progressive testing: description and initial results). Русский язык за рубежом *(Russian language abroad)*, (3):23–29.

Kopotev, M. (2012). Karttu: результаты языкового тестирования в школе и вузе (Karttu: results of language testing in schools and universities). Формирование и оценка коммуникативной компетенции билингвов в процессе двуязычного образования (Formation and assessment of communicative competency of bilinguals in bilingual education), pages 312–339.

Koster, R. (2004). *A Theory of Fun for Game Design*. Paraglyph Press.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Madge, C., Bartle, R., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2019a). Making text annotation fun with a clicker game. In *Proceedings of the 14th In-*

ternational Conference on the Foundations of Digital Games, New York, NY, USA. Association for Computing Machinery.

Madge, C., Bartle, R., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2019b). Incremental game mechanics applied to text annotation. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 545–558.

Pavlik Jr, P. I., Cen, H., and Koedinger, K. R. (2009). Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*.

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, 98:169–179.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*, volume 9. Springer Science & Business Media.

Slavuj, V., Kovačić, B., and Jugo, I. (2015). Intelligent tutoring systems for language learning. In *Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE.

Sweetser, P. and Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):3, July.

von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, August.

von Ahn, L., Liu, R., and Blum, M. (2006). Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 55–64, New York, NY, USA. Association for Computing Machinery.

von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6):92–94.

Walsh, G. and Golbeck, J. (2010). Curator: a game with a purpose for collection recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2079–2082.