Using Crowdsourced Exercises for Vocabulary Training to Expand ConceptNet

Christos Rodosthenous[1], Verena Lyding[2], Federico Sangati[3], Alexander König[2,8]
Umair ul Hassan[4], Lionel Nicolas[2], Jolita Horbacauskiene[5], Anisia Katinskaia[6]
Lavinia Aparaschivei[7]

[1] Computational Cognition Lab, Open University of Cyprus, Cyprus
[2] Institute for Applied Linguistics, Eurac Research, Bolzano, Italy
[3] Orientale University of Naples, Italy,
[4] Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
[5] Kaunas University of Technology, Lithuania
[6] Department of Computer Science, University of Helsinki, Finland
[7] Faculty of Computer Science, Alexandru Ioan Cuza University of Iaşi, Romania
[8] CLARIN ERIC, the Netherlands

christos.rodosthenous@ouc.ac.cy, verena.lyding@eurac.edu, federico.sangati@gmail.com,
alex@clarin.eu, umair.ulhassan@insight-centre.org, lionel.nicolas@eurac.edu,
jolita.horbacauskiene@ktu.lt, anisia.katinskaia@helsinki.fi, aparaschiveilavinia96@gmail.com

Abstract

In this work, we report on a crowdsourcing experiment conducted using the V-TREL vocabulary trainer which is accessed via a Telegram chatbot interface to gather knowledge on word relations suitable for expanding ConceptNet. V-TREL is built on top of a generic architecture implementing the implicit crowdsourding paradigm in order to offer vocabulary training exercises generated from the commonsense knowledge-base *ConceptNet* and – in the background – to collect and evaluate the learners' answers to extend ConceptNet with new words. In the experiment about 90 university students learning English at C1 level, based on the Common European Framework of Reference for Languages (CEFR), trained their vocabulary with V-TREL over a period of 16 calendar days. The experiment allowed to gather more than 12,000 answers from learners on different question types. In this paper, we present in detail the experimental setup and the outcome of the experiment, which indicates the potential of our approach for both crowdsourcing data as well as fostering vocabulary skills.

Keywords: vocabulary trainer, commonsense knowledge, language learning

1. Introduction

Language resources (LRs), like annotated corpora or dictionaries, are needed for many data-driven NLP tasks. Yet, refined resources of wide coverage are still lacking for many languages. This situation is mainly due to the fact that LRs are expensive to create and maintain.

In order to address this persistent lack, we devised an approach for facilitating the creation and extension of LRs by means of combining the domains of language learning and crowdsourcing and implemented a generic architecture for it (Rodosthenous et al., 2019). This implicit crowdsourcing approach is unique in that it makes use of LRs to automatically generate language learning exercises, while in turn using learners' answers to these exercises to extend or correct the underlying LR (cf. Section 5). Also, the application of the approach is supported by providing an open architecture which can easily be adapted and reused for different use-cases combining various LRs and language learning exercises.

In this paper, we report on an experiment about vocabulary training of 81 university students learning English at C1 level, based on the Common European Framework of Reference for Languages (CEFR), by means of a Telegram chatbot named *LingoGameBot*, a part of the V-TREL system. V-TREL is built on top of our proposed architecture to offer vocabulary training exercises generated from the commonsense knowledge-base *ConceptNet* and – in the background – to collect and evaluate the learners' answers for extending the ConceptNet with new `RelatedTo` relations for the trained words.

ConceptNet (Speer et al., 2017) is an open multilingual semantic network that started as part of the Open Mind CommonSense project in 1999 at MIT. Since then, the project has moved on and includes content from expert-created resources, crowdsourcing, and games with a purpose. It currently holds more than 34 million assertions about words, i.e., <TermA> <Relation> <TermB>. The knowledge-base is accessible through an API, making it easier for use in applications.

The results of the experiment give positive evidence for the feasibility of the proposed approach. We show that V-TREL allowed both, to extend ConceptNet with meaningful new terms and to help language learners improve their vocabulary skills for the trained set of words.

2. The V-TREL System

The V-TREL vocabulary trainer has its foundation on a generic architecture designed to connect a vocabulary trainer with an existing language resource (Rodosthenous et al., 2019). The architecture comprises the exercise generation component which is responsible for the automatic generation of language exercises, the exercise dispatcher which is responsible for delivering the exercises to the various interfaces and receiving the answers, the evaluation component which is processing the received responses and evaluates them and last the interfaces that

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 307–316
Marseille, 11–16 May 2020

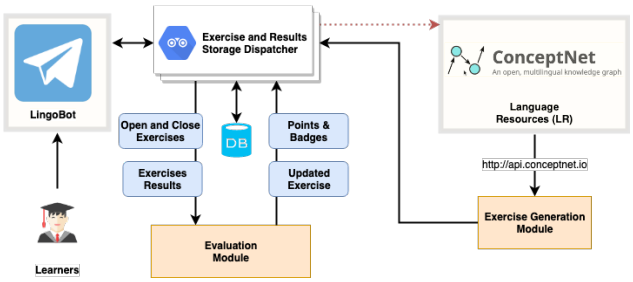

Figure 1: The architectural diagram of V-TREL, used for the LingoGameBot Telegram application.

utilize the architecture for both presenting the exercises and receiving acquired words. Figure 1 depicts the architectural schema of the instance used for V-TREL to facilitate the experiment conducted (see Section 3). Both the architecture and V-TREL are available under an open source license along with the retrieved data at `https://cognition.ouc.ac.cy/vtrel_project`.

The idea is to use the language resource (LR, for example, ConceptNet) as a source to generate exercises that are then presented to the user in a vocabulary trainer interface. The users' answers are validated against the LR and the user receives points for correct answers (that match the LR). The crowdsourcing part comes in when answers are given that are not part of the LR. These answers are stored and evaluated by combining them with answers from other students and if the evaluation is positive (e.g., a certain number of students independently have given the same answer) the answer can then be fed back into the language resource, extending and improving it in the process.

In the experiment described in this paper, we have worked with version 2 of V-TREL which has a number of new features and improvements compared to the last experiment we have run (Lyding et al., 2019). The most prominent new feature is the addition of *closed questions*. While previously students were always asked *open questions* of the form "Name one word related to bird.", we added a mechanism that transforms the students' answers into closed yes/no/"I don't know" questions that can be used to evaluate this answer, resulting in questions like "Is it true that bird is related to feather?". To ensure that there are enough closed questions that have a definitive answer, there are also closed questions generated directly from ConceptNet to populate the question pool. In order to improve the user experience and avoid users getting too many questions of a certain type in a row, we have also added an option to the back-end configuration to manually define the percentage of each type of question. This makes it possible to ensure that all users will see a certain percentage of questions of each type. Finally, to improve the learner experience, we have added a feature to the open questions where users are presented a link to Wikipedia explaining the word being asked about. This way they can learn more about those words that they do not know yet or do not know very well and expand their knowledge.

Telegram interface We have implemented a Telegram chatbot as the user interface to present exercises to the students who took part in the experiment and collect their responses. The chatbot communicates via API both to the Telegram server[1] and the back-end module responsible for distributing exercises to users and storing the results. As shown in Figure 2 the interface includes a number of buttons to facilitate the navigation, while it allows textual inputs for the open questions.

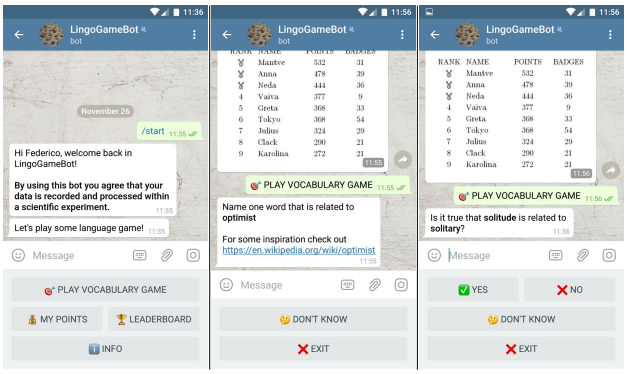

Figure 2: Screenshots of the Telegram interface.

3. Experimental Setup

The aim of the presented experiment is to evaluate the potential of the implicit crowdsourcing approach from both the NLP perspective and the language learning perspective, trying to shed some light on systems that can automatically generate exercises and at the same time acquire useful knowledge to populate an LR. By gathering user data from language students we tested the following two hypotheses:

- H1: Language learners can help to extend the relations in ConceptNet.
- H2: Automatically generated questions from ConceptNet can help in improving vocabulary of learners.

In parallel, we also evaluated the user satisfaction and engagement with the V-TREL vocabulary trainer. In crowdsourcing, it is important to keep users engaged if one is aiming to keep gathering data for a long period of time.

3.1. Participants and Experiment Setting

The experiment was carried out with three classes of Lithuanian[2] university students[3] attending English courses on C1 level. Overall 81 students, aged 19 to 21 years, were registered on the three classes (class A = 40 students, class B = 23 students and class C = 18 students) held by two teachers. The experiment was running for 16 calendar days. During this time each class had 5 training sessions[4] in which students were asked to train their vocabulary by using V-TREL for 15 minutes on their own smartphones.

The experiment was introduced by the teachers with instructions on:

[1] `https://core.telegram.org/bots/api`

[2] Mainly students of Lithuanian mother tongue with very few exceptions of Erasmus students.

[3] Students of BA study programs of technical sciences in their first and second study years.

[4] Class B had to carry out two sessions at home.

1. How to download Telegram and start the LingoGame bot, and

2. When to use the V-TREL vocabulary trainer.

No particular instructions were given on how to operate the vocabulary trainer, as the interactive application is designed to be self-explanatory. In fact, it provides short descriptions within exercises and buttons. The "(i) INFO" button instructs that "*Related-to words can be single words or multiword expressions of any word class*". Regarding the type of relation between words, no restriction was posed on the learners, i.e. as defined in ConceptNet (cf. Section 3.3) morphologically related words or semantically related words etc. count as equally valid input.
Finally, the students were asked to compile a vocabulary pre-test before the experiment, and a vocabulary post-test and a user satisfaction survey after the experiment (see Section 3.4).

3.2. Experiment Setup and Evaluation Parameters

For the experiment, the V-TREL trainer was set up to provide the user with open and closed questions in random order (ratio: 80% open questions, 20% closed questions). Both question types are automatically generated from ConceptNet (see Section 3.3). In addition, some of the closed questions are generated from the user input to open questions (see below). Open questions take the form *"Name one word that is related to 'dog'"* and ask the user to input free text. Closed questions take the form *"Is it true that 'dog' is related to 'bark'"* and ask the user to respond with "yes", "no" or "I don't know". If students respond to open questions with words, which are not part of ConceptNet, these words undergo a system evaluation[5] and, in case of a positive evaluation, they are saved as candidates for the extensions for ConceptNet. Closed questions can be of three different types and are presented in random order in different proportions. 75% of the closed questions are based on related-to words from ConceptNet, 10% of the closed questions are based on words from Conceptnet which are expected NOT to be related-to each other (see Section 3.3), and 15% are generated from the ConceptNet extension candidates based on the user input to open questions (see evaluation strategy above).

3.3. Data Sourcing and Preparation

First, we retrieved 1771 words from the English Vocabulary Profile (EVP)[6]. The English Vocabulary Profile shows, in both British and American English, which words and phrases learners around the world know at each level – A1 to C2 – of the Common European Framework of Reference for Languages[7]. These words are selected by retrieving only C1 level content from the category *word* where the part-of-speech is *noun* and we excluded culturally sensitive words[8]. The retrieved words are then run against the ConceptNet 5.7 knowledge-base and we retrieve the canonical form of each of these words.
The next step includes launching V-TREL's exercise generation utility creating 1771 exercises of `RelatedTo` type and populating them with content from ConceptNet, filtered to capture English words only. For each of these exercises, we moved a step forward from previous experiments conducted (Lyding et al., 2019) and also tried to retrieve words that are not related to the subject word. To do that, we first retrieved all `RelatedTo` objects for each of the objects retrieved so far, i.e., $TermB =< subject, RelatedTo, X >$. We randomly select five of these and perform another search, i.e., $TermC =< TermB, RelatedTo, Y >$. We then perform a search for the relatedness metric (Speer et al., 2017) for each pair of $< subject, TermC >$. The relatedness metric is one of the results of the combination of ConceptNet knowledge with word embeddings acquired from distributional semantics and it is used in ConceptNet Numberbatch (Speer et al., 2017). The `NotRelatedTo` words were chosen using a relatedness metric of <0.5. This cut-off value was selected by manually inspecting a small sample of the data. An additional search to exclude all terms for which an assertion $< subject, RelatedTo, TermC >$ exists was made to safeguard the process.
For the final selection of the exercise dataset, we excluded exercises with fewer than 10 `RelatedTo` words, fewer than 10 `NotRelatedTo` words, and multi-words in the subject. From the remaining set of exercises, we randomly selected 160 exercises.
At this point, it is important for the reader to understand what the `RelatedTo` relation represents in ConceptNet. Based on ConceptNet's documentation[9], this is a general relation representing a positive relationship between a term A and a term B, but ConceptNet cannot determine what that relationship is based on the provided data. In previous versions of ConceptNet this was called `ConceptuallyRelatedTo`. The abstraction of that relation, allows students to feel more free in writing words that could in any way relate to each other.

3.4. Pre- and Post-Tests and Survey

Before and after the experimentation period of 16 days, during which students were using the vocabulary trainer, we carried out two short vocabulary tests. The vocabulary pre- and post-tests requested students to enter into a web form related words to 20 words out of the set of trained vocabulary[10]. The vocabulary tested in the pre- and post-tests was distinct (no words occurring in both the pre- and post-test).
Also, at the end of the experiment we presented the learners with a web-based questionnaire about their satisfaction with the system. The questionnaire contained questions about the vocabulary training aspects of the V-TREL trainer and

[5] As soon as five new words (k=5) have been collected for one exercise, and if at least one of the new words has been entered twice (n=2), the most frequent new word is promoted as a candidate for the extension of ConceptNet.

[6] Made freely available by Cambridge University Press, http://vocabulary.englishprofile.org/staticfiles/about.html.

[7] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

[8] The EVP provides a filter option for this.

[9] `https://github.com/commonsense/conceptnet5/wiki/Relations`

[10] 160 words of C1 level, see Section 3

overall usability of the system (see Appendix A for the list of questions).

4. Data Analysis and Results

During the 16-days experiment (see Table 1 for a recap of the experiment setup) we collected data from 92 different users (as identified by unique Telegram ids). This number is higher than the number of 81 participating students, which indicates that some other people (possibly invited by the students) have been using the freely available vocabulary trainer during the experiment period or that some students have used multiple Telegram ids.

Before analysing the data, we filtered out any Telegram accounts that belonged to the experimenters (used to verify good operation of V-TREL) and any user id that did not belong to the Telegram interface (e.g., user ids from the web application version of V-TREL).

Experiment setup	
Crowd	81 university students divided into three classes (A=40, B=23, C=18)
Setting	Classroom setup, language course, supervised by 2 teachers
Language	English, level C1
Duration	28/10-12/11 (16 days)
V-TREL training content	
Language resource	ConceptNet, version 5.7
Generated exercises	160 exercises [CEFR level C1], based on English Vocabulary Profile
Words selection	Automatic (random), excluding multiword expressions
Exercise types	Open type questions (e.g., *"Name one word that is related to 'computer'"*) Closed type questions (e.g., *"Is it true that 'money' is related to 'corruption'?"*) [80%-20% ratio]
Pre- and post-experiment evaluation	
Vocabulary tests	before and after the experiment, testing of 20 random words out of the trained vocabulary
Satisfaction survey	user survey with 15 questions on learning effect and user experience

Table 1: Overview of the experiment parameters.

For the vocabulary pre-test and post-test we received 99 and 46 answers respectively, when matching up the user ids from pre- and post-test this resulted in 39 unique users taking both tests. We assume that the rather low number of post-test results might be related to there being not enough motivation for the students to do this after the experiment had ended or not getting around to it and then forgetting. We should try to find a way to better engage the users to make sure that a larger number of them will also participate in this final test in future experiments.

For the post-experiment user survey, after removing incomplete responses, we received 36 valid responses from students that provide both quantitative and qualitative data about the usability of the V-TREL system.

4.1. Learning Effect for Learners

The evaluation of the educational value of the V-TREL vocabulary trainer is based on two types of evidence: (1) the results of the vocabulary pre- and post-tests, and (2) the results of the user satisfaction survey[11].

The pre- and post-experiment tests already showed some promising results. We have gathered completed pre- and post-tests for 39[12] users. For the evaluation we looked at how many correct answers a student gave to the 20 questions in the pre-test and compared those to the number of correct answers in the post-test. We considered every word that existed as a relation in ConceptNet as a correct answer. Even though we are aware of the sometimes doubtful quality of ConceptNet as a gold standard[13], we have assumed its correctness here nevertheless being confident that the tendency (improvement or decline) will hold in any case. For those answers that did not exist in ConceptNet we had three people independently annotate the answers as correct or incorrect and, if there were disagreements, went with the majority vote. With this method we had a number between 0 (all answers were wrong) and 20 (all answers were correct) for each pre- and post-test and could look at the difference between pre- and post-test for the same student to see whether they improved or got worse during the experiment.

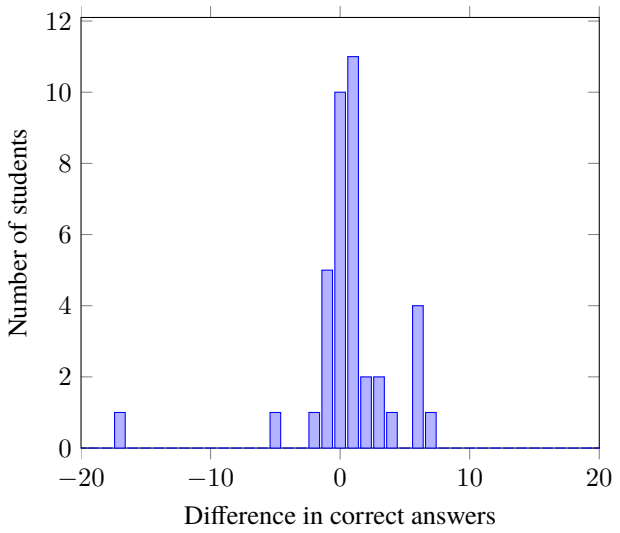

Figure 3: Changes in correct answers between pre- and post-test.

As can be seen in Figure 3 the majority of students showed no or only very little improvement or decline. Two thirds scored exactly the same or one point better or worse after the experiment. This is to be expected as vocabulary improvements take time and the experiment ran only for a very short period. But even with this in mind, the results

[11] It needs to be noted that the user survey delivered declarative statements of learners, thus providing a self-estimation and not a formal evaluation of the learners' performance.

[12] One user filled in both pre- and post-tests three times. While the post-tests were all identical, the pre-tests contained different answers and therefore we chose one of the tests at random to include in the analysis.

[13] A manual analysis of 349 `RelatedTo` relations yielded an accuracy of 79.7% (see Section 6 for details).

show a slight positive trend with a large number of students slightly improving (21 out of 39), some of them even significantly, while the number that scored significantly worse on the post-test is much smaller (7 out of 39), excluding the single "-17" value that obviously did not take the post-test seriously.

Apart from the pre- and post-tests, we additionally evaluated a randomly selected set of 100 answers from the 5 most prolific students. These 500 answers were manually evaluated by two annotators to judge whether the answer can be counted as correct or not, regardless of ConceptNet. Disregarding "I don't know answers" and a small number of cases where the annotators did not agree, we then sorted the answers by time and split them into an earlier and a later half.

Time slice	**Correct answers**	**Total answers**	**% correct answers**
earlier half	175	233	75.1%
later half	183	232	78.9%

Table 2: Number of correct answers per time slice.

As shown in Table 2, we found that there was a slight increase over time (about 4%), which also supports the hypothesis that using the vocabulary trainer actually helps to improve the students' vocabulary. Again, this result should be considered in the light of the fact that the training period was very short and a learning effect is expected to increase over time.

The user survey asked the students to evaluate the use of "open" and "closed" questions and the "hints" and "Wikipedia" features for training vocabulary skills. Overall, 86% of the respondents evaluated "open questions" as useful, 64% evaluated "closed questions" as useful, and 58% of the users felt they improved their vocabulary with V-TREL (see Figure 4 for the distribution of answers).

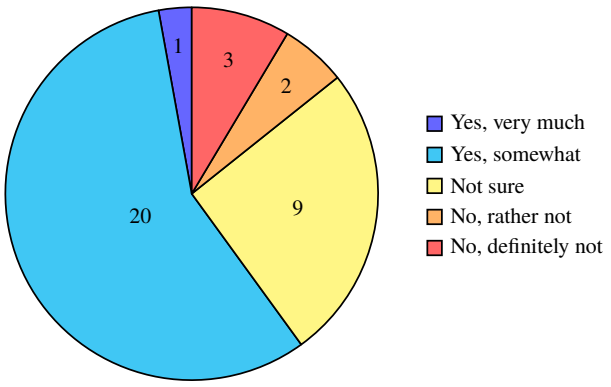

Figure 4: Evaluation of the use for training vocabulary based on responses to the question "Did you feel that the LingoGameBot helped you to improve your vocabulary?"

We also performed a manual analysis of free-text responses from participants to judge their sentiment towards open and closed questions. More than 58% of the participants (21 out of 36) expressed a positive sentiment towards the use of open questions. In general, participants liked the idea of keeping the responses open-ended and commented that it challenged their vocabulary; however, a few participants did raise the issues of difficulty of words and repetitive presentation of the same questions. In case of closed questions, more than 47% of the participants expressed a positive sentiment towards the closed questions and 27% expressed a negative sentiment. Participants pointed out that some of the relations were incorrect and several participants observed that for most of the questions "yes" was the correct answer (see Section 6 for details). 78% users said the "hint feature" helps training vocabulary, still only 19% made use of it for every second exercise at least, and some did not even notice this feature. 64% evaluated the "links to Wikipedia" as useful, and 30% actively used them for every second exercise or more. Some users suggested that it would be better to link to a different resource, e.g., a definition dictionary.

In fact, when analyzing the experiment data, we observed that the "I don't know" button ("hint feature") was clicked 898 times at least once by 58 users for a total of 148 exercises. That led to 868 contributions after the hint was presented, 436 of which were the same as the presented hint word and 432 were a different word than the hint.

Overall, the users also named the following positive points regarding learning with V-TREL: *"makes you think", "broadens vocabulary", "helps practice and memorize", "useful to check one's vocabulary knowledge", "good for learning new words", "challenges ones knowledge", "really strong brainbuster"*, and criticized the following negative points: *"no explanations why something is wrong", "sometimes not accurate", "learning words without context", "approach is too broad"*.

These results indicate that V-TREL has an educational value, though at its current state it is not as strong as one would expect for an educational tool. However, we also got an indication how it could be improved, e.g., by improving the exercise material ("closed questions") or linking to more specific resources for vocabulary consultation ("word definitions").

4.2. Extension of ConceptNet

In order to evaluate the potential of V-TREL to enhance ConceptNet we analyzed in detail the data collected from the crowd of learners. From the 160 exercises, 157 have evaluated contributions through *open questions*. A total number of 727 words were contributed and evaluated for these 157 exercises. 572 of them were unique, compared to words from all exercises. On average, 4.63 words were evaluated per exercise. None of the words was found in a `RelatedTo` assertion in ConceptNet and 620 (85.28%) were completely new to ConceptNet, since no other relation between the subject was found. Moreover, 42 (5.78%) words were found with a relation in ConceptNet other than `RelatedTo` such as `Synonym`, `DerivedFrom`, `FormOf`, etc.

During the experiment, an additional 3109 words were contributed but were not positively evaluated to enter the ConceptNet knowledge-base.[14] For readers to get a better picture of the results obtained, for the exercise "Name one word that is related to 'grill'?" the words that were evaluated to enter ConceptNet are **food**, **bbq**, **meat**, **fire**, **grilled**.

[14] With additional user contributions several of these words are expected to get positively evaluated as well.

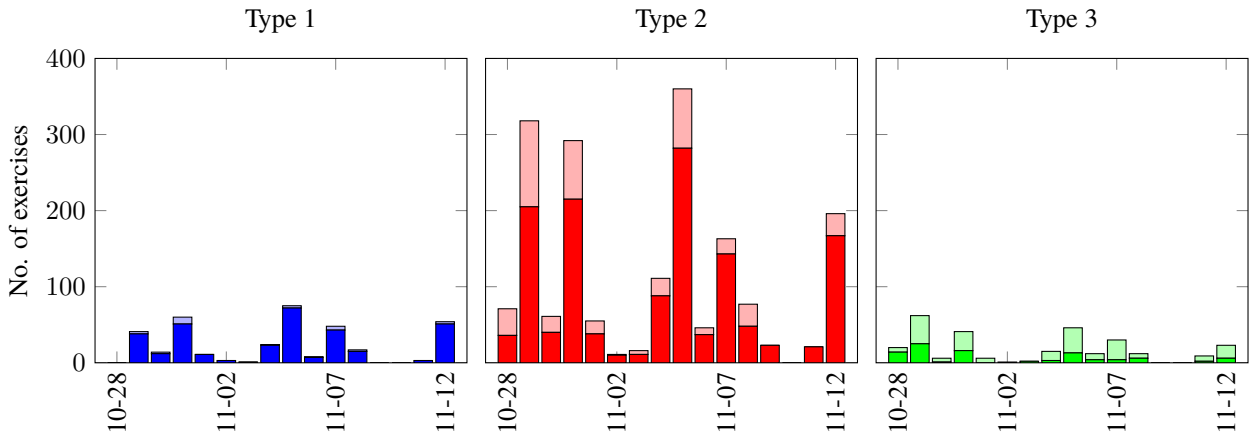

Figure 5: Number of answered exercises for the three types of closed exercises complete over time along with the share of exercises which were correctly answered.

3 of these words do not have any other type of relation in ConceptNet and 2 are linked with the `FormOf` relation.

For the whole experiment period, learners were presented with a total of 2505 closed-type questions. 371 (14.81%) of which were generated from the learners' evaluated contributions terms to open questions (Type 1), 1844 (73.61%) generated from ConceptNet's `RelatedTo` terms (Type 2) and 290 (11.58%) generated from the `NotRelatedTo` terms (Type 3). Out of these, 2471 were answered and 1795 (72.64%) were answered correctly. Finally, Table 3 reports the generated, answered and correctly answered closed exercises per type, and Figure 5 gives a graphical representation.

Closed Exercise Type	Generated (%)	Answered (%)	Answered Correctly (%)
Type 1	371(14.81%)	365(98.38%)	335(91.78%)
Type 2	1844(73.61%)	1821(98.75%)	1364(74.90%)
Type 3	290(11.58%)	285(98.28%)	96(33.68%)

Table 3: Number of closed exercises generated, answered and answered correctly per type.

4.3. User Satisfaction and Engagement

In this section, we present the results to those parts of the questionnaire related to the overall user experience[15] of the learners with the systems; furthermore, we provide an analysis of user engagement during the study.

When asked about their overall user experience with the LingoGameBot (that is the name of the Telegram interface of V-TREL), most of the participants found the system to be fun (81%) and inspiring (97%). This feedback was also evident from the comments given by participants which included the LingoGameBot potential as "*a modern way of training vocabulary*". More than 90% of the participants found the words in both open and closed questions to be relevant for vocabulary training. By comparison, some participants (14%) also considered the system to be boring and confusing. In addition, some participants commented that they did not feel motivated to use the LingoGameBot although there was an element of gamification through points and a leaderboard.

As shown in Table 4, the majority of students (80%) registered during the first four days of the study. Figure 6 shows the number of words entered by students during each day. Clearly, the number of contributions from students naturally increased with the number of classes on a day. This behavior of user engagement is also shown in Figure 6 by the number of words, provided by students, that matched the existing words in ConceptNet.

Day	Registrations	A	B	C
2019-10-28	13	✓		
2019-10-29	38		✓	✓
2019-10-30	0			✓
2019-10-31	26	✓		
2019-11-01	1			
2019-11-02	0			
2019-11-03	0			
2019-11-04	0	✓		
2019-11-05	8	✓	✓	✓
2019-11-06	2	✓		✓
2019-11-07	4			✓
2019-11-08	4			✓
2019-11-09	0			
2019-11-10	0			
2019-11-11	0	✓		✓
2019-11-12	12		✓	

Table 4: Number of registration for each day and classes (i.e., A, B, and C) held in the day.

5. Related Work

As far as our knowledge goes, no previous efforts are directly comparable to the ones reported in this paper and only few efforts have focused on combining language learning and implicit crowdsourcing. We thus extend our description of the related state of the art to efforts aiming at, on the one hand, crowdsourcing LRs and, on the other hand, automatically generating exercise content from LRs.

[15] As opposed to the parts on the learning effect, cf. Section 4.1.

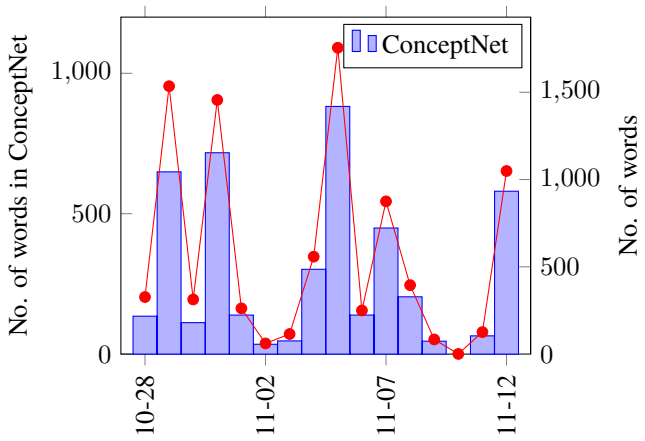

Figure 6: Daily number of words added by students against the number of words which exist in ConceptNet.

5.1. LR Improvement Using Crowdsourcing

Approaches for crowdsourcing LRs can be divided into two broad categories: crowdsourcing approaches where a crowd is explicitly engaging in a crowdsourcing campaign and implicit crowdsourcing approaches where users create relevant data as a byproduct of a specific activity.

Regarding the crowdsourcing approaches explicitly engaging a crowd, they usually rely on specific platforms (e.g., Zooniverse[16], Crowd4u[17], Amara[18] or Amazon Mechanical Turk[19]) that confront a crowd with simplified tasks, i.e., "micro-tasks", that serve more complex objectives. Related efforts in that category include, but are not limited to, efforts related to named entity annotation (Finin et al., 2010; Lawson et al., 2010; Ritter et al., 2011), transcribed speech corpora (Callison-Burch and Dredze, 2010; Evanini et al., 2010), word-sense disambiguation (Biemann, 2013), WordNets (Ganbold et al., 2018) or parallel corpora (Zaidan and Callison-Burch, 2011; Post et al., 2012). In that category of efforts, we can also cite the efforts of MacWhinney (2017) that propose a collaborative platform for collecting and sharing learner data from corpora, online tutors, and Web-based experimentation.

Regarding implicit crowdsourcing approaches, where users create relevant data as a byproduct of a specific activity, the Duolingo platform (von Ahn, 2013) used to implement a similar approach to our work as it generated language exercises allowing the crowdsourcing of translations. We can also mention two tools used in the classroom to implicitly crowdsource POS corpora (Sangati et al., 2015) and syntactic dependencies (Hladká et al., 2014). The other implicit crowdsourcing approaches we are aware of do not target language learning. They mostly are Games-With-A-Purpose (GWAP) approaches (Chamberlain et al., 2013; Lafourcade et al., 2015) where a crowd produces relevant data while playing a game. Among GWAP approaches, the JeuxDeMots game by Lafourcade2017a is especially relevant to our approach as it is designed to crowdsource data on word relations. Other GWAP approaches are TileAttack (Madge et al., 2017) which builds on player agreements to acquire textual segmentation annotations, Robot Trainer (Rodosthenous and Michael, 2016) which crowdsources knowledge rules, Zombilingo (Fort et al., 2014; Guillaume et al., 2016) which annotates syntactic dependency relations or Phrase Detectives (Chamberlain et al., 2008; Poesio et al., 2012; Poesio et al., 2013) where players contribute anaphora-related data.

[16] https://www.zooniverse.org/
[17] http://crowd4u.org/en/
[18] http://amara.org/
[19] https://www.mturk.com/mturk/welcome

5.2. Automatic Exercise Generation from LRs

The automatic exercise generation is a research subject situated within the broader domain of Computer-Assisted Language Learning (CALL). As far as our grasp of the state-of-the-art goes, only little automatic exercise generation is at present performed from LRs. Indeed most related works we came across are so-called *cloze* exercises generated from running texts where words or letters have been omitted, and learners are asked to fill these gaps, (Lee et al., 2019; Hill and Simha, 2016; Goto et al., 2010; Katinskaia et al., 2018) or exercises automatically generated from running text by remodeling sentences (Chinkina and Meurers, 2017; Lange and Ljunglöf, 2018a; Lange and Ljunglöf, 2018b). In addition to *cloze* exercises, the Revita language learning platform (Katinskaia et al., 2018) has listening exercises and an extensive module for generating multiple-choice exercises: they include lexical, grammatical, and stress exercises.

The observed sparsity of LR-based automatic generation of exercises was further confirmed by exploring the most recent proceedings of two CALL-oriented NLP workshops, namely: the *Workshop on Innovative Use of NLP for Building Educational Applications* (Tetreault et al., 2018; Yannakoudakis et al., 2019) and the *Workshop on NLP for Computer Assisted Language Learning* (Pilán et al., 2018; Alfter et al., 2019). Indeed, most works in these proceedings focus on other subjects such as the generation of *cloze* exercises, the modeling of the learner knowledge in order to predict the needs of the learner, the scoring of written production of learners, or the detection and/or correction of mistakes in written productions.

Exercise type	**Exercise count**	**% of correct exercises (manual gold standard)**
Type 1 "RelatedTo"	85	85.9%
Type 2 "RelatedTo"	349	79.7%
Type 3 "NotRelatedTo"	66	60.6%

Table 5: Number of exercises per type and manual evaluation of exercise correctness (Gold Standard).

6. Discussion

In this section we will briefly discuss the quality of the automatically generated exercise data and will highlight an issue with the initial exercise design.

The implicit crowdsourcing approach implemented in the V-TREL vocabulary trainer provides a prototype for a fully automated workflow, which combines automatically generated exercise content with the automated evaluation of input from learners to extend ConceptNet with new relations. In order

		Learner agreement with		Learner answers		
Exercise type	Manual gold standard	automatic gold standard	manual gold standard	"Yes"	"No"	"Don't know"
1	85.9%	90.6%	85.9%	90.6%	4.7%	4.7%
2	79.7%	80.5%	73.4%	80.5%	13.2%	6.3%
3	60.6%	13.6%	43.9%	78.8%	13.6%	7.6%

Table 6: Agreement of learners' answers with automatically generated exercises and manual gold standard.

to evaluate both the exercise quality as well as the learner reliability, we carried out a manual evaluation of a random set of 500 closed class exercises and their respective learners' responses. The random set contained `RelatedTo` exercises based on user input from the open exercises (type 1), and exercises derived from ConceptNet with `RelatedTo` terms (type 2) as well as `NotRelatedTo` terms (type 3). Table 5 gives the distribution by exercise types as well as the manual evaluation (manual Gold Standard – GS) of the correctness of each triple $< Subject, RelatedTo, TermX >$.

The manual evaluation shows that more than 85% of the type 1 exercises, which were generated from the user input, are correct. However, it also shows that the exercises generated from ConceptNet are of lower quality with error rates of about 21% for `RelatedTo` exercises (type 2) and about 40% for `NotRelatedTo` exercises (type 3). For type 3 exercises, there is an evident need to improve our initial approach for generating `NotRelatedTo` exercises (cf. Section 3.3). For type 2 exercises it shows that we need to find better ways to filter ConceptNet.

Our experiment evaluation with regard to the crowdsourced content (presented in Section 4.2) is on purpose based on a fully automated approach, assuming that ConceptNet can serve as a valid gold standard. When evaluating the learners' answers in relation to the "ConceptNet GS" and the "manual GS" (see Table 6), we observe slightly lower performances for exercises of type 1 and 2 (about 5-8% lower) and considerably higher performances for exercise type 3 (about 29% higher). Still, we observe very low learner performance on type 3 exercises (`NotRelatedTo`). In fact, the performance rate below 50% indicates a bias of learner answers towards "yes" for most questions. To investigate closer on this point we calculated the learners' reliability for exercises, in which words were `RelatedTo` each other vs. `NotRelatedTo` (see Table 7). The results show a high performance of learners on `RelatedTo` exercises (with an average of 87.5% correctness) and a very low performance on `NotRelatedTo` exercises (with an average of 24.2% correctness).

In fact, this is in agreement with results from the user survey, in which learners remarked that the correct answer mostly seems to be "yes, `RelatedTo`" and confessed that they acted alike[20]. From this situation we learned that the experiment design has to be improved in terms of offering a balanced set of exercises in future experiments.

		% of learners correctness	
Exercise type	**Ratio of "RelatedTo" vs. "NotRelatedTo" exercises**	**on "RelatedTo" exercises**	**on "NotRelatedTo" exercises**
Type 1	90.6% vs 4.7%	97.3%	16.7%
Type 2	80.5% vs 13.2%	85.3%	27.9%
Type 3	78.8% vs 13.6%	84.0%	20.0%
total	82.0% vs 11.8%	87.5%	24.2%

Table 7: Correctness rates of learners on exercises with `RelatedTo` terms and `NotRelatedTo` terms.

7. Conclusion and Future Work

In this work, we presented the continuous work on V-TREL, a vocabulary trainer used both for language learning and for updating language resources with new words using crowdsourcing. Furthermore, we concentrated on an experiment we designed and executed to verify the usefulness of our approach in updating knowledge in ConceptNet. The results showed that acquired knowledge is suitable for this purpose. To the best of our knowledge, there is no other system currently available that combines language learning using a vocabulary trainer with crowdsourcing learners' contributions for expanding a language resource and this approach can also be used for other language resources, besides ConceptNet.

In terms of the learning effect of using V-TREL for improving vocabulary of learners, the results show a very small yet positive shift of learner's vocabulary skills after using it. As we have discussed in Section 4.1, these are marginal results and cannot be conclusive. V-TREL's educational value is identified but more work is needed to improve it.

As future work, we plan to organize a shared task on exploring different aggregation methods on the collected dataset of crowdsourced answers. This shared task is being prepared by several authors of this work. We plan to collect more learners' data for that purpose.

Also, we plan to expand on the evaluation of the educational value of V-TREL in further experiments, e.g. by extending the training period, using more detailed tests and working with control groups.

Since we have access to learners of other languages, we consider adding other languages to the V-TREL, e.g., Italian and Russian.

[20] Learner's cite: "I always pressed yes, and it was most times correct."

8. Acknowledgements

This article is based upon work from COST Action enetCollect (CA16105), supported by COST (European Cooperation in Science and Technology). The work presented in this paper was started during the Crowdfest meeting organized by the Action in January 2019 in Brussels.

9. Bibliographical References

David Alfter, et al., editors. (2019). Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning, Turku, Finland, 30 September. LiU Electronic Press.

Biemann, C. (2013). Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122, Mar.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 1–12. Association for Computational Linguistics.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase detectives: A web-based collaborative annotation game. In Proceedings of the International Conference on Semantic Systems (I-Semantics' 08), pages 42–49.

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 Iryna Gurevych and Jungi Kim (Eds.), *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg, pp. 3–44.

Chinkina, M. and Meurers, D. (2017). Question generation for language learning: From ensuring texts are read to supporting learning. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 334–344, Copenhagen, Denmark, September. Association for Computational Linguistics.

Evanini, K., Higgins, D., and Zechner, K. (2010). Using amazon mechanical turk for transcription of non-native speech. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 53–56. Association for Computational Linguistics.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 80–88. Association for Computational Linguistics.

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. ACM.

Ganbold, A., Chagnaa, A., and Bella, G. (2018). Using crowd agreement for wordnet localization. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018).

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., and Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal*, 2(3):210–224.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In Proceedings of the International Conference on Computational Linguistics (COLING), Osaka, Japan.

Hill, J. and Simha, R. (2016). Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 23–30, San Diego, CA, June. Association for Computational Linguistics.

Hladká, B., Hana, J., and Lukšová, I. (2014). Crowdsourcing in language classes can help natural language processing. In Second AAAI Conference on Human Computation and Crowdsourcing.

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

Lafourcade, M., Brun, N. L., and Joubert, A. (2015). Games with a Purpose (GWAPS). Wiley-ISTWiley-ISTE, July.

Lange, H. and Ljunglöf, P. (2018a). Demonstrating the muste language learning environment. In Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018) at SLTC, Stockholm, 7th November 2018, number 152, pages 41–46. Linköping University Electronic Press.

Lange, H. and Ljunglöf, P. (2018b). Mulle: A grammar-based latin language learning tool to supplement the classroom setting. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 108–112.

Lawson, N., Eustice, K., Perkowitz, M., and Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. In Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk, pages 71–79. Association for Computational Linguistics.

Lee, J.-U., Schwan, E., and Meyer, C. M. (2019). Manipulating the difficulty of c-tests. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 360–370, Florence, Italy, July. Association for Computational Linguistics.

Lyding, V., Rodosthenous, C., Sangati, F., ul Hassan, U., Nicolas, L., König, A., Horbacauskiene, J., and Katinskaia, A. (2019). v-trel: Vocabulary trainer for tracing word relations - an implicit crowdsourcing approach. In Galia Angelova, et al., editors, Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019, pages 675–684, Varna, Bulgaria.

MacWhinney, B. (2017). A shared platform for study-

ing second language acquisition. *Language Learning*, 67(S1):254–275.

Madge, C., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2017). Experiment-driven development of a gwap for marking segments in text. In Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play, pages 397–404. ACM.

Ildikó Pilán, et al., editors. (2018). Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning, Stockholm, Sweden, November. LiU Electronic Press.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2012). The phrase detective multilingual corpus, release 0.1. In Collaborative Resource Development and Delivery Workshop Programme, page 34.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, April.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 401–409. Association for Computational Linguistics.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1524–1534. Association for Computational Linguistics.

Rodosthenous, C. and Michael, L. (2016). A hybrid approach to commonsense knowledge acquisition. In Proceedings of the 8th European Starting AI Researcher Symposium, pages 111–122.

Rodosthenous, C., Lyding, V., König, A., Horbacauskiene, J., Katinskaia, A., ul Hassan, U., Isaak, N., Sangati, F., and Nicolas, L. (2019). Designing a prototype architecture for crowdsourcing language resources. In Thierry Declerck et al., editors, Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK 2019), pages 17–23. CEUR.

Sangati, F., Merlo, S., and Moretti, G. (2015). Schooltagging: interactive language exercises in classrooms. In LTLT@ SLaTE, pages 16–19.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge.

Joel Tetreault, et al., editors. (2018). Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, Louisiana, June. Association for Computational Linguistics.

von Ahn, L. (2013). Duolingo: learn a language for free while helping to translate the web. In Proceedings of the 2013 international conference on Intelligent user interfaces, pages 1–2. ACM.

Helen Yannakoudakis, et al., editors. (2019). Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy, August. Association for Computational Linguistics.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 1220–1229. Association for Computational Linguistics.

A. Post-study Survey

The following questions were asked as part of the post-study survey of user satisfaction and engagement.

Q1. What did you think of the "open" questions which asked you to input one word related to a specific word provided by the game (e.g., "name one word related to 'house'")?

Q2. Would you say that "open" questions are useful for training vocabulary?

Q3. What did you think of the "closed" questions which asked you if two words were related or not related (e.g., "Is 'house' related to 'home'")?

Q4. Would you say that "closed" questions are useful for training vocabulary?

Q5. What ratio of "open" and "closed" questions would you recommend/prefer?

Q6. How often did you make use of the "hint" feature where an example is provided when you press the "I don't know" button when answering an open question (e.g., "a possible response was 'home'")?

Q7. Do you think that the "hint" feature is useful for training vocabulary?

Q8. Do you have any comment on the "hint" feature?

Q9. How often did you make use of the links to wikipedia which were provided when you were asked to answer an open question?

Q10. Do you think the links to wikipedia are useful for training vocabulary?

Q11. Do you have any comment on the links to Wikipedia?

Q12. Did you feel that the LingoGameBot helped you to improve your vocabulary?

Q13. What did you think of the words that you were trained on?

Q14. How was your overall user experience with the LingoGameBot vocabulary trainer?

Q15. What did you like and/or dislike about this approach to train vocabulary?

Q16. Did you notice anything particular regarding the LingoGameBot interface on telegram? Was it pleasant to use? Have you encountered some bug?

Q17. Any other comments: