An Experiment on Implicitly Crowdsourcing Expert Knowledge about Romanian Synonyms from L1 Language Learners

Lionel Nicolas[1], **Lavinia Aparaschivei**[1], **Verena Lyding**[1], **Christos Rodosthenous**[2], **Federico Sangati**[3], **Alexander König**[5], **Corina Forascu**[4]

[1]Institute for Applied Linguistics, Eurac Research, Bolzano, Italy
[2]Computational Cognition Lab, Open University of Cyprus, Cyprus
[3]Cognitive Neurorobotics Research Unit, Okinawa Institute of Science and Technology, Japan
[4] Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
[5]CLARIN ERIC, the Netherlands

Abstract

In this paper, we present an experiment performed with the aim of evaluating if linguistic knowledge of expert quality about Romanian synonyms could be crowdsourced from L1 language learners, learning Romanian as their mother tongue, by collecting and aggregating their answers to two types of questions that are automatically generated from a dataset, encoding semantic relations between words. Such an evaluation aimed at confirming the viability of a fully learner-fueled crowdsourcing workflow for improving such type of dataset. For this experiment, we reused an existing open-source crowdsourcing vocabulary trainer that we designed for this very purpose and which crowdsourcing potential needed further evaluation, especially with regards to lesser-resourced languages such as Romanian. Our results confirmed that producing expert knowledge regarding Romanian synonyms could be achieved in such a fashion. Additionally, we took the occasion to further evaluate the learning impact of the trainer on the participants and gather their feedback regarding several aspects.

1 Introduction

The lack of Linguistic Resources (LRs) and the lack of exercise content are respectively two long-standing issues that are slowing down the domains of Natural Language Processing (NLP) and Computer-Assisted Language Learning (CALL). Recent efforts that implement an implicit crowdsourcing paradigm have started to tackle these issues in a concurrent fashion (Nicolas et al., 2020). Such a paradigm follows the idea that if a dataset can be used to generate the content of a specific type of exercise, then the answers to these exercises can also be used to improve back the dataset that allowed to generate the exercise content.

Among the efforts implementing this paradigm, we devised an open-source and publicly-available vocabulary trainer called v-trel (Rodosthenous et al., 2019; Lyding et al., 2019; Rodosthenous et al., 2020) in order to generate exercises from a knowledge-base called ConceptNet (Speer et al., 2017) while using the crowdsourced answers to improve ConceptNet. In the experiments we previously conducted and reported about, we provided some preliminary evidence towards its crowdsourcing potential but a more thorough investigation was still needed, especially with regards to a lesser-resourced language such as Romanian that is far less represented in ConceptNet. Furthermore, the evaluation of the learning impact of v-trel on its users also had room for further exploration. For this experiment, we aimed at filling both gaps, while taking the opportunity to gather more feedback about the vocabulary trainer.

We explain hereafter how we demonstrated that aggregating the partial and neophyte knowledge of L1 learners of Romanian[1] could be used to pro-

[1]The experiment originally targeted L2 students but the health crisis due to the Covid-19 pandemic limited our networking options and we had to rely on already established

Lionel Nicolas, Lavinia Nicoleta Aparaschivei, Verena Lyding, Christos Rodosthenous, Federico Sangati, Alexander König and Corina Forascu 2021. An experiment on implicitly crowdsourcing expert knowledge about Romanian synonyms from language learners. *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping Electronic Conference Proceedings 177: 1–14.

duce knowledge of expert quality about Romanian synonyms. We therefore explain how our experiment provides evidence that v-trel, and its underlying approach in general, can be used to devise a fully learner-powered crowdsourcing workflow for improving datasets, encoding semantic relations between words. We also explain how this experiment allowed us to gather additional insights regarding the learning impact on the participants.

This paper is organized as follows. In Section 2, we present work related to our approach and discuss similarities. Next, in Section 3, we briefly discuss v-trel and the gaps we aimed at filling with this experiment. In Section 4, we explain how we adapted v-trel for the purpose of our experiment, and in Section 5 we describe how we set up our experiment. We then discuss the results we achieved in Section 6. Finally, we explore future efforts in Section 7 and conclude in Section 8.

2 Related work

Our efforts are situated at the crossroad between crowdsourcing knowledge in order to enhance linguistic resources and automatically generating exercises for CALL purposes. Accordingly, the relevant state of the art is composed of approaches targeting only one or both of the two subjects.

With respect to the previous work related only to the automatic generation of exercises, the state of the art is composed of only a handful of approaches that generate exercises directly from linguistic resources. Most related works actually focus on the "cloze" (deletion) test, where a portion of the text has some of the words automatically removed by some NLP-based pipeline, and the learner is asked to recover the missing words (Lee et al., 2019; Hill and Simha, 2016). In Lyding et al. (2019), we confirmed the lack of automatic generation of exercises based on linguistic resources by reviewing the most recent proceedings of two CALL-oriented NLP workshops[2] and coming to the conclusion that current efforts are dedicated to other subjects such as the generation of cloze exercises, the modelling of the learner knowledge, or the detection and/or correction of mistakes in written productions. Among recent work targeting the generation of language learning exercises, we can point to Chinkina et al. (2017) and Chinkina et al. (2020), in which the authors addressed the development of a novel form of automatic generation of questions that combines a wh-question with gapped sentences. Following a small-scale pilot study, the work of Ziegler et al. (2017) provided empirical evidence supporting the effectiveness of instructional treatments, such as input enhancement, for L2 growth, as well as exploring how technological innovations could deepen our understanding of L2 acquisition. We can also refer to the work presented by De Kuthy et al. (2020), in which the authors described an automatic question generation approach to partially automate *Questions under Discussion* (QUD) annotation by generating all potentially relevant questions for a given sentence in a German text. In addition, in Fenogenova and Kuzmenko (2016), the authors proposed an approach to automated generation of English lexical exercises for learning collocations, and then compared the exercises produced to those compiled manually by language instructors.

Regarding the previous works related only to the crowdsourcing of linguistic resources, they can mostly be categorized into two groups aiming at curating a varied set of linguistic resources: the approaches relying on micro-task platforms (e.g. Kordoni et al. (2016), Caines et al. (2016), Lafourcade (2007), Ganbold et al. (2018), Post et al. (2012)), and the approaches implementing implicit crowdsourcing approaches that crowdsource information from a crowd that is not necessarily aware of the on-going crowdsourcing. This is usually achieved by embedding the implicit crowdsourcing approach into a workflow used for a different purpose than crowdsourcing. For example, among approaches implementing implicit crowdsourcing methods, a great share of the state of the art consists in games that implicitly crowdsource linguistic knowledge from their users while providing them entertainment. Such games are referred to as GWAPs (Games with a Purpose) and include efforts such as Lafourcade (2007), Poesio et al. (2013) or Guillaume et al. (2016)).

Finally, with regards to previous works related to both the automatic generation of language learning exercises and the crowdsourcing of linguistic resources, the state of the art contains only a limited number of efforts that combine both as we do. The most famous initiative is certainly

contacts with schools instructing L1 Romanian students that, despite being proficient, are still learning their mother tongue (see proficiency results in Section 6).

[2]Innovative Use of NLP for Building Educational Applications (Tetreault et al., 2018) and NLP for Computer Assisted Language Learning (Alfter et al., 2019).

Duolingo (von Ahn, 2013) which used to generate translation exercises and crowdsourced the answers to sell them later to third parties. Other efforts were developed in the context of the enetCollect COST Action and implement the aforementioned implicit crowdsourcing paradigm (Nicolas et al., 2020). V-trel is one of them and, as a cornerstone of our work, we discuss it in greater details in the following section. Among the other works related to enetCollect and/or the implicit crowdsourcing paradigm, we can also point the readers to Millour et al. (2019), Smrz (2019), Grace Araneta et al. (2020) and Arhar Holdt et al. (2021) that all aimed at crowdsourcing lexical knowledge. Finally, two other learning tools are also worth considering: one for crowdsourcing POS corpora (Sangati et al., 2015) and another one for crowdsourcing syntactic dependencies (Hladká et al., 2014).

3 v-trel in a nutshell

The vocabulary trainer v-trel is a prototypical language learning tool that generates vocabulary exercises from a multilingual linguistic resource called ConceptNet (Speer et al., 2017) in which words and their semantic relations to one another are recorded (e.g. translation, synonyms, hypernyms etc.) in the form of triples ($word_1$, relation, $word_2$). At the same time, v-trel crowdsources the answers with the aim of producing through aggregation an expert knowledge that can be used to enhance ConceptNet. V-trel offers exercises through a user-friendly chatbot interface accessible from the Telegram messenger[3].

V-trel generates two types of exercises: *open exercises* in which users are provided a word and asked to provide another one related to the first one by a specific semantic relation (e.g. *provide a synonym of "house"*) and *closed exercises* in which users are asked if a pair of words are related to one another according to a specific type of semantic relation (e.g., *Are "home" and "house" synonyms?*).

The version of v-trel we adapted for our experiment generates open exercises from a finite list of words and the closed exercises from both the recurrent triples suggested by learners in answers to open exercises and the existing triples already encoded in ConceptNet. By proceeding in such a fashion, the answers provided to the closed questions can be aggregated and used to, on the one hand, validate or discard triples suggested in open exercises to extend ConceptNet and, on the other hand, validate or contradict the triples already encoded. The user feedback to open questions is based both on the answer previously provided by other learners and on the existence of a matching triple in ConceptNet. User feedback to closed questions exclusively relies on the presence (or absence) of a matching triple in ConceptNet. In order to support the learners in their efforts, v-trel also implements a number of user-oriented features such as a hint feature allowing to request examples, an automatically generated link to Wikipedia[4] allowing to swiftly consult a dedicated page on Wikipedia (if any) and a point system with a functionality displaying a leaderboard that allows learners to compete among themselves.

While the experiments we described in the two last papers about v-trel (Lyding et al., 2019; Rodosthenous et al., 2020) allowed us to validate and/or enhance many relevant aspects, no extensive formal proof was made that expert knowledge could indeed be derived from the answers of the learners. This is mainly due to the fact that for the last experiment reported, while we could confirm the capacity of open questions to generate relevant triples to include in ConceptNet, we generated a large number of closed questions that diluted the set of answers crowdsourced. This setup led to an insufficient average number of answers per closed question that prevented us from performing any kind of aggregation that could produce the expert knowledge needed to validate or discard new triples or existing ones. As a fallback approach for closed questions, we manually evaluated the quality of a random sample of answers in order to demonstrate that they were on average correct for more than 50% of them and that, consequently, expert quality would statistically have been achieved by collecting more answers. Nonetheless, we discovered after the experiment a bias toward positive answers in the responses of learners that prevented us from doing so. Indeed, since the closed exercises are both mostly automatically generated from the new triples recurrently suggested in open questions and the ones available in ConceptNet, the correct answer was in most case "Yes"[5] and learners grad-

[3] https://telegram.org/

[4] E.g. https://en.wikipedia.org/wiki/House for *house*

[5] There were also a few closed questions automatically

ually understood it over time[6]. Consequently, in order to earn more points, most learners chose to always answer positively in case of doubt instead of choosing the option "I-don't-know" that allowed them to skip a question for which they were not sure of the correct answer. As a consequence, whereas the average accuracy of the answers to closed exercises where the correct answer was "yes" was far above 50%, the average accuracy of the answers for the ones where the correct answer was "no" was under 50%. This issue thus prevented us to indirectly confirm the crowdsourcing potential. Another aspect for which the evaluation of the crowdsourcing potential is further explored with this new experiment is the language targeted. Indeed, only English, the language best covered in ConceptNet has been considered so far.

Regarding the learning impact on users, we evaluated the learning impact on users by relying on pre- and post-experiment vocabulary tests that were manually revised by an expert and also some small randomly sampled sets of answers of a few students. For the last experiment described in Rodosthenous et al. (2020), while results of the pre- and post-questionnaires were not conclusive, we observed some learning impact as the average accuracy of the small randomly sampled sets of answers of the most prolific five students were slightly better for the second half of the sets than for the first. However, the difference was not vast (+4%) and the size of the sample was limited (100 answers) and only concerned five learners. We thus explore this question in order to further support our previous findings.

4 Adapting v-trel

Overall, we adapted v-trel by partially disconnecting several automatic mechanisms in order to create a more static version that allowed us to better evaluate the aspects we were interested in. In that perspective, as our main focus was not so much to produce expert knowledge in order to improve ConceptNet but to produce it for the purpose of evaluating its quality, the crowdsourcing we made was more of a simulation of crowdsourcing since we asked many questions for which we knew the answers. Regarding the evaluation of the learning impact on learners, we did not adapt v-trel in any particular way as we relied on the evolution of the accuracy of the answers provided over time. We thus relied on an intrinsic evaluation instead of using an extrinsic approach such as one with pre- and post-tests.

The adaptations that we performed focused mainly on the open and closed questions and are discussed hereafter. Aside from these, we localized the interface to Romanian and used synonymy as the type of semantic relations on which the learners were tested.

Indeed, in our previous experiments on v-trel, we used the "relatedTo" relation between words in ConceptNet. A closed question could have for example be "is *home* related to *family*?". From the experience we gained so far, we concluded that finding consensual answers for some of these questions was more challenging than we originally thought. We thus chose to use synonymy instead which made the task far easier. The criteria we used to further specify our notion of synonyms was that two words shall be considered as synonyms of one another if they can be exchanged/paraphrased in a sentence without altering its overall meaning. For example the Romanian words "*imagine*" ("*picture*" in English) and "*ilustraţie*" ("*illustration*" in English) can freely be exchanged in the Romanian sentence "*Profesoara le-a aratat copiilor o ilustraţie/ imagine cu o expediţie de la Polul Nord.*" (" *The teacher showed the children an illustration/picture with an expedition from the North Pole.*" in English) without altering its overall meaning. The definition of synonymy we used is thus one that also accounts for partial synonymy between words that would probably not be considered as synonyms of one another if considered outside the context in a sentence.

4.1 Adapting the open questions

The open questions and the feedback given to the learners remained globally the same. Learners thus received points if they provided an answer that matched an existing triple in ConceptNet or if they provided answers that their fellow learners provided as well a sufficient number of times. Unlike our previous experiments, we post-evaluated the answers that were given more than twice by the learners, to observe if the frequency of occurrences of an answer was correlated with its quality (see Section 6).

generated from triple encoding a relation *NotRelatedTo* for which the correct answer was "No", but they were not numerous enough.

[6]Some learners actually said it explicitly in the user questionnaire they answered after the experiment.

We limited the number of open questions so as to avoid diluting the answers of learners. The size of the set of open questions was estimated by doing a mock-up test with a few people before the experiment that allowed us to estimate the average number of answers per person and per hour. We then multiplied this number by the number of participants expected and the average number of hours we expected them to contribute to our experiment.

4.2 Adapting the closed questions

Unlike the case of open questions, our adaptations focused on avoiding two issues: a too large number of closed questions that would dilute excessively the answers of learners, as well as an imbalance between closed questions for which the correct answer was "yes" and the ones for which the correct answer was "no" (in order to avoid influencing silently the learners in answering an option more than another as it happened in a previous experiment).

We addressed the first issue by generating a finite set of closed questions. The size of this set was also estimated via the mock-up test prior to the experiment. In order to maintain the size of this set of questions, we disconnected the mechanism that automatically generates closed questions from the answers provided to open questions.

In order to address the second issue and have a balanced set between closed questions for which the correct answer was "yes" and the ones for which the correct answer was "no", we automatically generated from ConceptNet two sets of closed questions, one for each type of answer, and a single annotator manually revised them in order to ensure that our final set was indeed balanced. We thus created for our experiment a specific gold standard for the closed questions and used it afterwards to study how much the aggregated knowledge extracted from the answers of the learners was correlated with it (see Section 6).

In order to automatically generate the two sets of closed questions to revise manually, we implemented and tested mechanisms exploring ConceptNet according to two assumptions that allowed us to create and rank two different lists: a list of potential pairs of synonyms and and a list of pairs of words that could be anything but synonyms of one another.

The assumption to generate potential pairs of synonyms is a well-known one that follows the idea that *If two Romanian words A and C are translations of the same word B in a different language, then A and C might be synonyms*. This assumption thus relies on semantic relations describing translations between words that, on a conceptual level, could be considered as relations describing pairs of synonyms belonging to different languages. For example, "frumos" and "atrăgător" are synonyms and both translate to "beautiful" in English. The ranking of the pairs of words included in the list generated is then based on the number of common translations (referred to as *B* before) found in all the languages.

The assumption to generate potential pairs of words that can be anything but synonyms of one another is that *If two Romanian words A and D are respectively both translations in a different language of two words B and C that have a relation that is not a synonymy relation (e.g. antonymy or hyperonymy), then A and D might have the same relation in Romanian and are most likely not synonyms of one another*. For example "flat" is a type of "home" in English and they translate to "apartament" and "casă" respectively in Romanian. The ranking of the pairs of words included in the list generated is then based on the size of the set of pairs of translations (referred to as *B* and *C* before) found in all languages. A valuable particularity of this mechanism is that the pairs of words were meaningful as they are part of the semantic landscape of one another, as opposed to a mechanism that would randomly pick two words (e.g. *bred* and *plane*).

A single annotator then revised in an orderly fashion the two lists until our gold standard had the size we aimed at. In order to make sure that open questions and closed questions have common grounds, we used the list of words of the open questions as word *A* in the two assumptions we relied on to generate closed questions.

Creating a gold standard for the closed questions also solved another issue: the feedback provided to the student for such questions. Indeed, v-trel relies at present on ConceptNet to provide such feedback. However, ConceptNet is a dataset that contains noise that can induce improper feedback to an extent that can create distrust from the users[7]. Should v-trel become fully functional, it

[7]By browsing the online version of ConceptNet, you'll see that, for example, *school* is marked as related to *sociotem-*

will over time be capable of gradually improving ConceptNet, or some specifically-selected parts of it, and thus reduce the noise it contains while enhancing its coverage. Since our experiment aimed at demonstrating the crowdsourcing potential of v-trel, relying on a gold standard for the closed questions allowed us to circumvent this issue.

5 Experimental setup

For our experiment, we generated 750 open questions and 1792 closed questions[8].

The experiment involved three classes with a total of 48 L1 students, aged between 18 and 19 years, that were taught Romanian by two teachers that agreed to support our initiative. The students were attending two high schools with different specializations, one theoretical and the other technical, respectively referred to as school "*1*" and "*2*" in Table 1. In order to foster participation and competition between the students, a contest to win vouchers for an e-commerce for the top five ranked participants, as listed on the leaderboard (see Section 3), was organized. Out of the 48 students, 20 registered and actively participated.

The experiment ran for 17 calendar days, from 28 May 2020 to 13 June 2020. The experiment was introduced by the teachers, who were always assisted by one of the authors, with a training session tutorial that included simple installation instructions as well as some examples of how to answer questions. In order to keep students motivated, we manually crafted and sent them bot-like push messages on four occasions and wrote messages on their Facebook groups. After the experiment was concluded, we asked learners to fill a survey giving them the opportunity to provide feedback on v-trel and the overall experiment.

6 Results

6.1 Participation and expertise of the crowd

Figure 1 shows the percentages of the answers provided by the 20 learners over the 17 days of the experiment, as well as the number of learners contributing every day and the moments we sent bot-like push messages to them to keep them engaged. As one can observe, the number of answers globally increased over time while the number of learners contributing fluctuated noticeably with an average of 9,2 per day (see blue bars in Figure 1). In our opinion, the overall increase of answers contributed is partly due to the prize-winning contest we organized over the first 16 days. Overall, as it can be observed in Table 1, six students contributed for 88.27% of the answers (79.79% of answers to open questions and 91.66% of answers to closed questions). We believe that the fact that our contest offered 5 vouchers, one fewer than the number of the most active learners, is no coincidence. This is a particularly interesting fact to consider for future experiments in order to maximize participation as these learners contributed voluntarily an amount of answers that most likely required between ten to twenty hours of their time, i.e., 12037 answers for the top contributor. Such an amount of time would have cost far more than a mere 20 euros voucher if we had remunerated them per hour of participation.

In Figure 1, the bot-like push messages are depicted by black stars. They mostly served their purpose as the second, third and fourth ones did induce spikes of participation whereas the first one sent after the first day wasn't very effective. From these few observations, it is fair to say that push messages seem to be a relevant tool to foster participation.

Overall, our setting allowed us to meet our goals in terms of amount of answers crowdsourced as we obtained 17108 answers to open questions (22.8 on average) and 42610 answers to closed questions (23.8 on average), which is more than twice than our original goal of obtaining an average of 10 answers per question.

With respect to the expertise, Table 1 details the overall performances of learners in answering open and closed questions computed by confronting their answers to an improved version of our gold standard for closed questions[9] and another gold standard we compiled for open questions[10]. As one can observe, despite the fact that

poral, *austrian* and *tiger mother*, which seems incorrect outside of the context that generated these relations.

[8]We originally aimed at an equivalent number of open and closed questions but a misunderstanding with the annotator that compiled the gold standard for closed questions led to the creation of a higher number of closed questions.

[9]We manually revised the entries where the strongest disagreements between the answers or the learners and the content of the gold standard could be spotted (see further details in Section 6.2).

[10]It is worth noting that the gold standard for the answers to open questions is based on a subset of the answers which are likely of being of higher accuracy in average. The performances to open questions reported are thus over estimating the true performances of the learners (see further details in

the learners are L1 Romanian speakers, their overall performances hardly qualifies them as an expert crowd for which we would have expected performances closer to the perfection (e.g. 98% accuracy)[11]. This shows that our crowd qualifies as a non-expert crowd which skills can still be improved, even though its skill-set should be noticeably above other non-expert crowds such as L2 learners.

Finally, the noticeable variability of the performances of the learners (Min / Max 69.57% / 97.56% for open questions and 58.82% / 92.12% for closed questions) confirmed our intuition that it is worth taking performances into account when aggregating their answers.

6.2 Producing expert knowledge

6.2.1 Open questions

Within the crowdsourcing workflow of v-trel, open questions are primarily meant to extend ConceptNet by collecting triples that are not encoded in it. Recurrent triples trigger the generation of closed questions that will confirm or refute their validity[12].

A single annotator performed an evaluation after the experiment on 1640 triples out of the 2513 triples[13] that had been suggested at least twice in order to create a gold standard. We then used it to study if the number of times a triple had been suggested was correlated with its quality[14]. Figure 2 demonstrates that the answer is a firm yes. Be it by considering all answers as equally important or by attributing them a weight associated with the proficiency of the learner (computed over the average accuracy of the answers of the students for the triples present in the gold standard), the quality of a triple is clearly correlated with the number of times it has been suggested. According to our evaluation, triples that were suggested with a score of 6 or more[15] were 97% correct when considering weighted votes (around 95% with regular votes).

For our use case, we can thus confirm the crowdsourcing potential of the open questions in order to produce a knowledge worth considering for extending ConceptNet.

6.2.2 Closed questions

Within the crowdsourcing workflow of v-trel, closed questions are both meant to take expert decisions to confirm or refute the triples present in ConceptNet and accept or filter out candidate triples to extend ConceptNet that have been crowdsourced in open questions. We evaluated this crowdsourcing potential in two manners.

In order to evaluate the answers to closed questions, we first confirmed that our set of closed questions did not silently induce a bias between positive and negative answers. As we collected 51.4% (21849) positive answers with an average accuracy of 83.16% and 48.6% (20665) negative ones with an average accuracy of 83.23%, there is no reason to believe that our experimental setup induced any such bias.

We first studied if the answers provided by the learners allowed to confirm or revoke the gold standard we had compiled for our experiment. In order to do so, we revisited our gold standard for all the 1972 closed questions and took into consideration how much the answers of the learners contradicted the gold answer we had associated with the closed questions. After such reconsideration, we inverted the original decision made by the single annotator that compiled the gold standard from "yes" to "no" or vice versa for 13.3% (239 questions) out of the 1792 questions and created an enhanced version of our gold standard. This confirms that, at least for our use case, the aggregated answers to closed questions crowdsourced can indeed be used to contradict the entries of a gold standard.

We then studied the quality of the winning "yes" or "no" options to the closed questions according to the minimum margin with which a winning option wins over a losing one in terms of aggregated score. Because v-trel is still a prototype that doesn't have yet an aggregation method implemented in it for closed questions, we relied on two rather simple aggregation scores: the minimum difference between a simple majority score

Section 6.2).

[11] Even though some did achieve quite respectable performances, such as the second and fourth learners.

[12] While it is not implemented in v-trel at present, open questions are also the occasion to gather positive answers for the closed questions that are automatically generated from them.

[13] We did not evaluate all 2513 that had been suggested twice or more or the other 4179 triples that had been suggested once because of manpower constraint.

[14] It is worth noting that compiling such a gold standard was only meant to double-check this correlation. Compiling a gold standard while relying on a single annotator was thus an approach of lesser quality that still met our needs.

[15] 542 open questions in our gold standard met that criteria for the weighted votes and 302 for the simple votes.

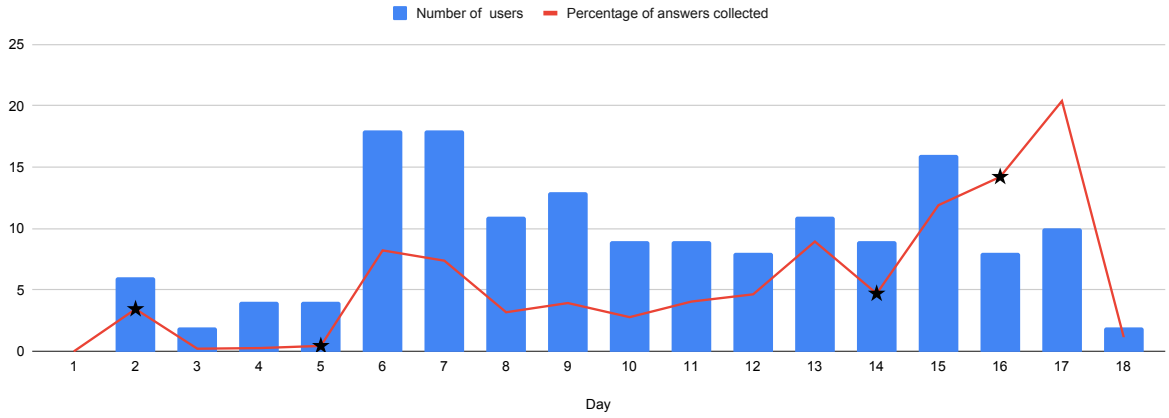

Figure 1: Percentage of the answers collected per day and numbers of contributors (stars indicate when push messages were sent)

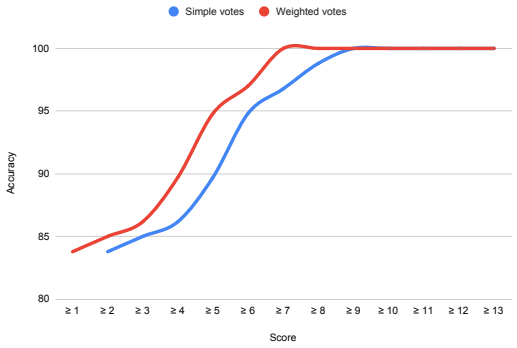

Figure 2: Accuracy of triples suggested to open question according to the number of votes.

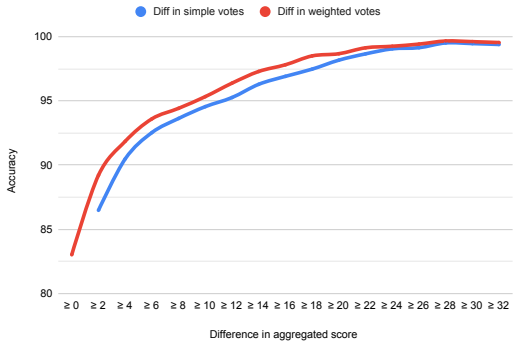

Figure 3: Accuracy of a winning option to closed questions according to the minimum difference in aggregated scores with the losing option.

and the minimum difference between a weighted majority score[16]. As can be seen in Figure 3, the greater the minimum difference between the winning option and the losing one, the higher is the accuracy of the winning option. For example, if the difference is at least of 16 points[17] then the winning option is around 98% reliable when using the weighted score, and 97% when using the simple vote[18]. This confirms once more that, at least for our use case, expert knowledge can be crowdsourced out of multiple answers provided by L1 learners to closed questions.

[16]The weight of an answer corresponded to the average accuracy of the answers of the learners according to our enhanced gold standard.

[17]639 closed questions met this criteria for the weighted scoring and 774 for the simple scoring.

[18]It should be noted that the number of answers to crowdsource for obtaining such a difference in votes depends on the triples considered.

6.3 Learning impact

In order to discuss the learning impact, we studied how the capacity of the learners in answering open and closed questions evolved over the duration of the experiment.

6.3.1 Open questions

In order to observe the learning impact regarding open questions, we reused the manual evaluation we did on the triples that were suggested at least twice by learners (see Section 6.2) and computed an average accuracy for their first 750 answers.

The reason why we only considered this set of answers is due to the fact that we had prepared 750 open questions and, since some learners provided more than 750 answers, they answered some questions several times. And when the learners were confronted with a question they had already answered, they were requested to provide an answer

Id	School	All questions		Open questions				Closed questions			
		#	% answers	#	% answers	Acc	# evals	#	% answers	Acc	# evals
1	1	12037	20.16	2821	16.49	89.66	774	9216	21.63	80.2	9190
2	2	9600	16.08	1921	11.23	88.15	852	7679	18.02	92.12	7669
3	1	9207	15.42	2023	11.82	86.2	1065	7184	16.86	87.82	7158
4	1	8589	14.38	2300	13.44	87.91	951	6289	14.76	90	6273
5	2	7994	13.39	2101	12.28	85.39	1437	5893	13.83	71.79	5880
6	2	5280	8.84	2486	14.53	85.94	1330	2794	6.56	76.81	2786
7	2	2067	3.46	1021	5.97	88.91	487	1046	2.45	75.69	1045
8	2	1070	1.79	512	2.99	79.75	237	558	1.31	74.64	556
9	1	1033	1.73	541	3.16	96.25	267	492	1.15	67.68	492
10	2	544	0.91	256	1.5	97.56	41	288	0.68	61.11	288
11	2	472	0.79	232	1.36	87.4	127	240	0.56	78.66	239
12	2	397	0.66	195	1.14	84.54	97	202	0.47	81.09	201
13	2	297	0.5	147	0.86	-	-	150	0.35	80	150
14	2	259	0.43	128	0.75	95.7	93	131	0.31	87.02	131
15	1	254	0.43	125	0.73	87.32	71	129	0.3	82.03	128
16	2	182	0.3	88	0.51	-	-	94	0.22	69.15	94
17	2	140	0.23	69	0.4	75	16	71	0.17	83.1	71
18	1	102	0.17	48	0.28	90	30	54	0.13	81.48	54
19	2	99	0.17	48	0.28	-	-	51	0.12	58.82	51
20	1	95	0.16	46	0.27	69.57	23	49	0.11	77.08	48

Table 1: Number, percentage of answers provided and accuracy of answers per learner and per type of exercises (*# evals* indicate the number of answers that matched a question in our gold standards).

different from the ones already provided. The difficulty of a question was thus increasing every time it came back. Another aspect that negatively impacted the quality of answers to questions coming back is that we did not offer them the opportunity to skip an open question. By doing so, we forced them to provide answers, including suboptimal ones, in order to be allowed to move forward. For all these reasons, observing the evolution of the performances of learners to open questions can only be performed soundly on the first 750 answers.

The average accuracy of the subset of theses answers that had an entry in our gold standard are shown in Figure 4. As one can observe, they remained globally stable around 90% over this set of 750 answers and no progress can be observed. This is unfortunately due to another bias that this experiment allowed us to identify. Indeed, as explained earlier in Section 6.2, the more often an answer to an open questions occurs the more likely it is to be correct. As such, by not considering the answers that occurred only once and were thus not included in our gold standard, we just keep on evaluating a subset of answers for which the quality is stable over the time span of the experiment. In order to perform this evaluation, we would have needed to have a gold standard for the whole set of the first 750 answers of each of the learners and not a subset of the best ones. The increased quality of the answers can nonetheless be observed in an indirect fashion by observing the ratio over time of answers matching an entry of our gold standard vs the answers not matching any entry (that are overall of lesser quality). As observable in Figure 4, this ratio increased over time, which indirectly indicates that the accuracy of the answers provided increased, even though we can't evaluate directly to what extent.

The learning impact for open questions could thus be indirectly observed. Nonetheless, because of the many issues we listed above, its evaluation remains a subject we would need to address more conclusively in future work (see Section 7).

6.3.2 Closed questions

In order to observe the learning impact for closed questions, and instead of doing pre- and post-tests on the learner to observe the differences in performances before and after using v-trel, we chose to study the evolution of the performances of the learners over time with the idea in mind that the first and last set of answers can be seen as a form of pre- and post-tests. Figure 5 displays the average accuracy for sets of 250 answers ordered in time for the eight learners that provided more than 500 answers to the closed questions. As one can observe, the curves fluctuate greatly and do not have

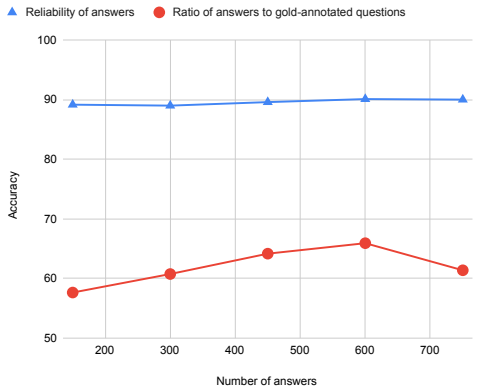

Figure 4: Accuracy and ratio of the first 750 answers to gold-annotated open questions.

the increasing direction we would have expected, with many of the curves stagnating and some even dropping. Table 2 displays for the eight learners the average accuracy of the first two hundred fifty answers, the first half of the overall answers, the second half of them and the last two hundred fifty answers provided. In that case also, our original expectations of a greater quality for the second half of the answers were not always met, with four learners performing better over time, two performing similarly and two performing worse.

We also could observe that despite using a first and last set of answers of rather large size (250 answers), the observations we could deduce regarding the learning impact on the learners from their accuracy would not always match the ones we would deduce from observing the accuracy of the larger sets consisting in the first and second half of all answers. For example, evaluating the learning impact from the first and the last sets of 250 answers or the first and second half of all answers would have led us to different conclusions for the first three learners listed in Table 2.

The fact that the four learners whose performances stagnated or decreased during the second half of their participation were part of the group that won a prize for their participation leads us to suspect that the competition among them might have had a deterring effect on the quality of their answers. We thus suspect that the strategy to earn points for these learners was to favor quantity over quality (i.e. speed over reflection). The fact that more than half of the answers were provided during the last four days of the experiment would tend to confirm our intuition (see Figure 1). If our intuition is indeed correct, while we had foreseen that such a phenomenon could happen, we underestimated its extent. In the event that we run another experiment that includes such a contest, we would need to devise strategies to prevent such a side-effect (see Section 7).

Overall, the learning impact for closed questions could not clearly be confirmed for many learners. At the same time, we could not think of, or observe, any intrinsic reason why there wouldn't be one for all learners. Confirming the learning impact of closed questions thus remains an open question to address.

6.4 User feedback

With respect to user feedback, 10 learners filled the post-experiment survey asking them questions with a free text, boolean or Likert format. During the survey, the learners were asked their thoughts on open and closed questions (free text), as well as the usefulness of these questions in vocabulary training (boolean), and the ratio of open and closed questions they prefer (Likert scale). The learners were also asked with two Likert scales how much they used the "hint" functionality and the automatically generated Wikipedia links (see Section 3) and how useful they thought it was (boolean), as well as whether they had any feedback about it (free text). They were finally asked about their overall user experience with the vocabulary trainer (Likert scale), what they liked and didn't like (free text), their thoughts on the Telegram interface and if they had any additional feedback (free text).

The students mostly gave positive feedback on the open questions, and two of them pointed out an important aspect of the Romanian language, namely the polysemy of words, which can be difficult to differentiate between the meanings of two words written identically in the absence of diacritics. All survey participants that gave a free response to the question about their thoughts on the closed questions mostly listed how simple the questions seemed at first glance, but that they took time to think of an answer. They offered a positive feedback regarding the usefulness of both types of questions for training vocabulary. Seven out of the ten survey's participants showed a preference for open questions over the closed questions.

Regarding the "hint" functionality, seven of the participants said they used it for less than half of the questions, while the rest said they used it for

User	First 250	First half	Second half	Last 250	Progress
1	84.4	84.26	76.13	93.12	worse
2	80	92.2	92.05	93.6	similar
3	88.4	88.23	87.4	80.4	similar
4	78	87.66	92.35	90.4	better
5	76.8	75.57	68	67.2	worse
6	70	74.35	79.27	81.2	better
7	54.4	68.97	82.41	80	better
8	70.8	71.48	77.78	78.4	better

Table 2: Accuracy of the answers of learners to closed questions for the first two hundred fifty, the first half, the second half and the last two hundred fifty of their answers.

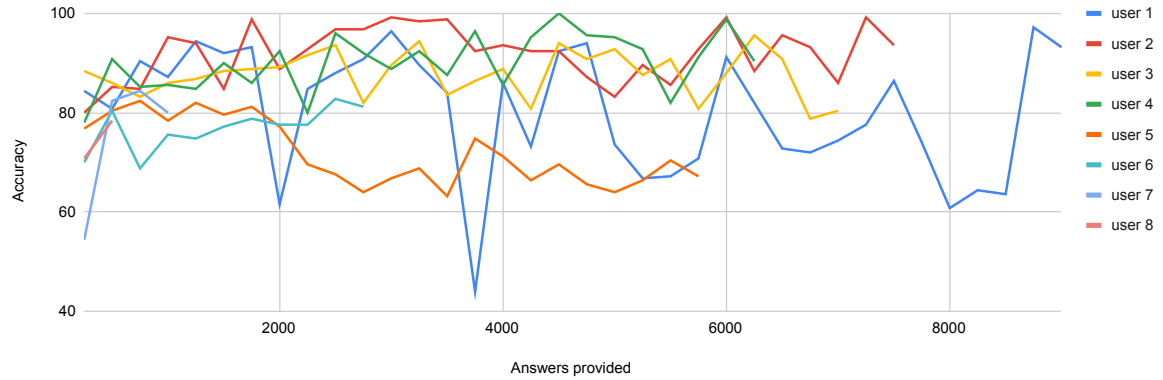

Figure 5: Average quality of the answers to closed questions over time by sets of 250 answers.

almost half of them. When asked about its usefulness, all the users found it useful. When asked about the Wikipedia links, students indicated that the links do not always correspond to the word in question or can lead to a non-existent page. Nonetheless, nine out of ten participants believed that the Wikipedia links are effective.

The participants' feedback about how much the trainer helped them to improve their vocabulary was rather positive, 8 out of 10 said that the game helped them "a lot" and the rest of them said that the game helped them to some degree. When asked about the difficulty of the words used with which they were trained, none of them considered them "too difficult", six of the participants considered them "neither too easy nor too difficult", and the four others as "mostly easy".

With respect to user experience, the vocabulary trainer seems to have met the expectations of the participants of the survey, who all indicated that it was fun to use. We can conclude that the instructions given prior to the start of the game were helpful because none of the participants expressed concern about the game being confusing or frustrating to play. Also two of the ten participants said that it was inspiring using the vocabulary trainer. When asked what they liked or disliked about this approach, the participants stated that they had a pleasant insight with the vocabulary teaching approach, and that while playing, their vocabulary skills improved. They also indicated that the competition and prizes influenced their involvement during this period. Regarding the Telegram chatbot interface, the learners claimed that they had a pleasant interaction with it. Just one person raised a concern about its instability on some occasions.

Last but not least, with respect to additional feedback, some students took the occasion to thank us for the opportunity. One student also mentioned that despite having enjoyed the game, he believed that it was better suited to middle school students.

7 Future work

Despite being satisfied with part of our results, this experiment allowed us to discover a set of short-

comings in the way we approached the experiment, on top of the challenges that we had already reported in earlier publications and that were addressed in this experiment.

Regarding the crowdsourcing potential of v-trel, we now have a dataset of real answers to closed question from learners and a refined gold standard dataset allowing us to know if the answers were correct or incorrect. As such, we have the data needed to start testing aggregation methods that could be included in v-trel. Another aspect we would like to further explore with respect to the crowdsourcing potential is to confirm its validity for other use cases relying on another type of semantic relation (e.g. hyponymy or hypernymy), for a different type of crowd or for a different language. By doing so, we would be able to see if any specific issues arise and how much our current conclusions can be extrapolated or generalized.

Regarding the learning impact, we first and foremost need to evaluate it in a more convincing fashion for both the open and closed questions. That would imply addressing the shortcomings listed in Section 6.3. For open questions, we would need to perform the post-experiment manual evaluation to build a gold standard either on the whole set of answers or a randomly picked subset. We would need to allow learners to skip questions if they have no convincing answers and would need to find means to consider all answers of learners and not only the first ones to each question. It would also be interesting to observe the learning impact for other use cases and see once again how the new results compare to the ones we obtained from this experiment. Furthermore, it would as well be interesting to compare v-trel to an equivalent solution such as the vocabulary trainers available on existing language learning solutions. However such a comparison is difficult to perform empirically on the performances of learners as it would require, first, to involve two crowds of learners that are large enough in order to ensure that any results computed are statistically relevant, second, that the two crowds are similar in terms of learners profiles in order to ensure that a tool doesn't have a more favorable crowd than the other and third, that both crowds contribute a similar amount of time. All in all, comparing v-trel to an equivalent solution in a relevant and meaningful fashion is a challenge that we do not know yet how to tackle.

Be it in terms of crowdsourcing potential or learning impact, it would be interesting to explore to which extent our results and conclusions also apply to L2 learners. Indeed, if we consider that the skills regarding language, including a mother-tongue, are a continuum, then L1 learners are among the most capable non-expert crowds we could rely on. We suspect that relying on L2 learners would not make a noticeable difference with the exception that the answers will be of lesser quality, which would certainly require us to adapt our approach to some extent.

Finally, if we were to also organize a contest to win prizes to foster participation in a future experiment, we would need to find means to mitigate the noise that we suspect such competition creates by encouraging learners to favor speed over reflection. A simple strategy could be to award an always greater amount of points for series of consecutive correct answers.

8 Conclusion

In this paper, we presented an experiment performed with the aim of evaluating if knowledge of expert quality about Romanian synonyms could be crowdsourced from language learners. Such an evaluation aimed at confirming the viability of a fully learner-fueled crowdsourcing workflow for improving such type of linguistic resources.

To perform such an experiment, we adapted an existing open-source crowdsourcing vocabulary trainer called v-trel that we designed for this very purpose. Our results clearly confirmed that such expert knowledge could indeed be produced by relying on L1 language learners and that v-trel would be a suitable tool to produce it, once some missing pieces regarding the aggregation of answers and the automatic generation of closed questions would be completed. The practical experience we obtained while running this experiment reinforced our intuition that expert knowledge about semantic relations between words other than synonymy could also be produced in an similar fashion.

We also took the occasion to further investigate the learning impact of v-trel on learners. On this subject our observations are far less conclusive. On the one hand, while we do believe that there has been a learning impact overall, our data does not allow us to draw any clear conclusions on this subject for all learners. On the other hand,

we observed clear shortcomings in the way we evaluated the open questions and, with respect to closed questions, we suspect that the contest to win rewards has had a deterring effect on the quality of the answers provided. In order to demonstrate the learning impact of v-trel, we thus need to first address these two issues in a follow-up experiment.

Acknowledgements. We would like to thank *Constantin Hoțoleanu* and *Daniela Pavel* from the *Liceul Teoretic Emil Racoviță Vaslui* and the *Colegiul Tehnic Ion Creangă Târgu Neamț* for supporting us in performing this experiment. This article is based upon work from COST Action enetCollect (CA16105), supported by COST (European Cooperation in Science and Technology).

References

Luis von Ahn. 2013. Duolingo: Learn a language for free while helping to translate the web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, page 1–2, New York, NY, USA. Association for Computing Machinery.

David Alfter, Elena Volodina, Lars Borin, Ildikó Pilan, and Herbert Lange. 2019. Proceedings of the 8th workshop on nlp for computer assisted language learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*.

Špela Arhar Holdt, Nataša Logar, Eva Pori, and Iztok Kosem. 2021. "Game of Words": Play the Game, Clean the Database. In *Proceedings of the 14th Congress of the European Association for Lexicography (EURALEX 2021)*, pages 41–49, Alexandroupolis, Greece.

Andrew Caines, Christian Bentz, Calbert Graham, Tim Polzehl, and Paula Buttery. 2016. Crowdsourcing a multi-lingual speech corpus: Recording, transcription and annotation of the crowdis corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2145–2152.

Maria Chinkina, Simón Ruiz, and Detmar Meurers. 2017. Automatically generating questions to support the acquisition of particle verbs: evaluating via crowdsourcing. *CALL in a climate of change: adapting to turbulent global conditions*, page 73.

Maria Chinkina, Simón Ruiz, and Detmar Meurers. 2020. Crowdsourcing evaluation of the quality of automatically generated questions for supporting computer-assisted language teaching. *ReCALL*, 32(2):145–161.

Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy, and Detmar Meurers. 2020. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5786–5798.

Alena Fenogenova and Elizaveta Kuzmenko. 2016. Automatic generation of lexical exercises. In *Proceedings of the International Conference*.

Amarsanaa Ganbold, Altangerel Chagnaa, and Gábor Bella. 2018. Using crowd agreement for wordnet localization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marianne Grace Araneta, Gülsen Eryigit, Alexander König, Ji-Ung Lee, Ana Luís, Verena Lyding, Lionel Nicolas, Christos Rodosthenous, and Federico Sangati. 2020. Substituto - A Synchronous Educational Language Game for Simultaneous Teaching and Crowdsourcing. In *Proceedings of the 9th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020)*, pages 1–9.

Bruno Guillaume, Karën Fort, and Nicolas Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3041–3052, Osaka, Japan. The COLING 2016 Organizing Committee.

Jennifer Hill and Rahul Simha. 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. Association for Computational Linguistics.

Barbora Hladká, Jirka Hana, and Ivana Lukšová. 2014. Crowdsourcing in language classes can help natural language processing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2.

Valia Kordoni, Antal van den Bosch, Katia Lida Kermanidis, Vilelmini Sosoni, Kostadin Cholakov, Iris Hendrickx, Matthias Huck, and Andy Way. 2016. Enhancing access to online education: Quality machine translation of MOOC content. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 16–22, Portorož, Slovenia. European Language Resources Association (ELRA).

Mathieu Lafourcade. 2007. Making people play for lexical acquisition with the jeuxdemots prototype. In *SNLP'07: 7th international symposium on natural language processing*, page 7.

Ji-Ung Lee, Erik Schwan, and Christian M Meyer. 2019. Manipulating the difficulty of c-tests. *arXiv preprint arXiv:1906.06905*.

Verena Lyding, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Lionel Nicolas, Alexander König, Jolita Horbacauskiene, and Anisia Katinskaia. 2019. v-trel: Vocabulary trainer for tracing word relations-an implicit crowdsourcing approach. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 674–683.

Alice Millour, Marianne Grace Araneta, Ivana Lazić Konjik, Annalisa Raffone, Yann-Alan Pilatte, and Karën Fort. 2019. Katana and Grand Guru: a Game of the Lost Words (DEMO). In *Proceedings of the ninth Language & Technology Conference*, Poznan, Poland.

Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Anisia Katinskaia, Anabela Barreiro, Lavinia Aparaschivei, and Yaakov HaCohen-Kerner. 2020. Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 268–278, Marseille, France. European Language Resources Association.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1):1–44.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409.

Christos Rodosthenous, Verena Lyding, Federico Sangati, Alexander König, Umair ul Hassan, Lionel Nicolas, Jolita Horbacauskiene, Anisia Katinskaia, and Lavinia Aparaschivei. 2020. Using crowdsourced exercises for vocabulary training to expand conceptnet. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 307–316.

Christos T. Rodosthenous, Verena Lyding, Alexander König, Jolita Horbacauskiene, Anisia Katinskaia, Umair ul Hassan, Nicos Isaak, Federico Sangati, and Lionel Nicolas. 2019. Designing a prototype architecture for crowdsourcing language resources. In *Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK 2019), Leipzig, Germany, May 21, 2019*, volume 2402 of *CEUR Workshop Proceedings*, pages 17–23. CEUR-WS.org.

Federico Sangati, Stefano Merlo, and Giovanni Moretti. 2015. School-tagging: interactive language exercises in classrooms. In *LTLT@ SLaTE*, pages 16–19.

Pavel Smrz. 2019. Crowdsourcing Complex Associations among Words by Means of A Game. In *Proceedings of CSTY 2019, 5th International Conference on Computer Science and Information Technology.*, volume 9, Dubai, UAE.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis. 2018. Proceedings of the thirteenth workshop on innovative use of nlp for building educational applications. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Nicole Ziegler, Detmar Meurers, Patrick Rebuschat, Simon Ruiz, José L Moreno-Vega, Maria Chinkina, Wenjing Li, and Sarah Grey. 2017. Interdisciplinary research at the intersection of call, nlp, and sla: Methodological implications from an input enhancement project. *Language Learning*, 67(S1):209–231.