Creating Expert Knowledge by Relying on Language Learners: a Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning

Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Cibej, Špela Holdt, Alice Millour, et al.

Creating Expert Knowledge by Relying on Language Learners: a Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning

Lionel Nicolas[1], Verena Lyding[1], Claudia Borg[2], Corina Forascu[3], Karën Fort[4], Katerina Zdravkova[5], Iztok Kosem[6], Jaka Cibej[6], Špela Arhar Holdt[6], Alice Millour[4], Alexander König[1,7], Christos Rodosthenous[8], Federico Sangati[9], Umair ul Hassan[10], Anisia Katinskaia[11], Anabela Barreiro[12], Lavinia Aparaschivei[3], Yaakov HaCohen-Kerner[13]

[1] Institute for Applied Linguistics, Eurac Research, Bolzano, Italy
[2] Faculty of Information & Communication Technology, University of Malta, Msida, Malta
[3] Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
[4] Sorbonne Université / STIH - EA 4509, France
[5] Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Macedonia
[6] Faculty of Arts, University of Ljubljana, Slovenia
[7] CLARIN ERIC, the Netherlands
[8] Computational Cognition Lab, Open University of Cyprus, Cyprus
[9] Orientale University of Naples, Italy,
[10] Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
[11] Department of Computer Science, University of Helsinki, Finland
[12] Human Languages Technologies Lab, INESC-ID, Lisbon, Portugal,
[13] Dept. of Computer Science, Jerusalem College of Technology (Lev Academic Center), Israel

lionel.nicolas@eurac.edu, verena.lyding@eurac.edu, claudia.borg@@um.edu.mt, corinfor@info.uaic.ro, karen.fort@sorbonne-universite.fr, katerina.zdravkova@finki.ukim.mk, iztok.kosem@ff.uni-lj.si, jaka.cibej@ff.uni-lj.si, spela.arharholdt@ff.uni-lj.si, alice.millour@etu.sorbonne-universite.fr, alex@clarin.eu, christos.rodosthenous@ouc.ac.cy, federico.sangati@gmail.com, umair.ulhassan@insight-centre.org, anisia.katinskaia@helsinki.fi, anabela.barreiro@inesc-id.pt, aparaschiveilavinia96@gmail.com, kerner@jct.ac.il

Abstract

We introduce in this paper a generic approach to combine implicit crowdsourcing and language learning in order to mass-produce language resources (LRs) for any language for which a crowd of language learners can be involved. We present the approach by explaining its core paradigm that consists in pairing specific types of LRs with specific exercises, by detailing both its strengths and challenges, and by discussing how much these challenges have been addressed at present. Accordingly, we also report on on-going proof-of-concept efforts aiming at developing the first prototypical implementation of the approach in order to correct and extend an LR called ConceptNet based on the input crowdsourced from language learners. We then present an international network called the *European Network for Combining Language Learning with Crowdsourcing Techniques* (enetCollect) that provides the context to accelerate the implementation of the generic approach. Finally, we exemplify how it can be used in several language learning scenarios to produce a multitude of NLP resources and how it can therefore alleviate the long-standing NLP issue of the lack of LRs.

Keywords: Crowdsourcing, Computer-Assisted Language Learning, Collaborative Resource Construction, COST Action

1. Introduction

The lack of wide-coverage and high-quality LRs is a long-standing issue in Natural Language Processing (NLP) that persists until nowadays despite numerous efforts aiming at addressing it such as CLARIN (Váradi et al., 2008), DARIAH (Blanke et al., 2011) or META-NET (Piperidis, 2012). While this issue is not equally problematic for all languages, it remains a crucial concern for most languages. To our knowledge, large ongoing initiatives addressing the need for LRs are currently focusing more on making the most out of the existing ones (e.g. by standardizing them and making them available to the larger public) while no initiatives of similar scale exist for creating new LRs or for improving the existing ones. In a report of 2018 (Evans, 2018), the pressing needs for LRs are highlighted together with multiple social and economic opportunities that their availability could unlock, especially for the European area. With the awareness of such needs, opportunities and large institutional support acknowledging this issue, it is unfortunate that there is no large initiative aiming at addressing the creation of new LRs or the curation of the existing ones.

The limited efforts aimed at the creation or curation of LRs are most likely due to the difficulty in automatizing these tasks which are mostly performed in a manual fashion. Such an endeavor can thus require vast amounts of expert manpower for every single LR and, consequently, creating

and curating all the LRs needed for the many languages that could benefit from NLP is simply too expensive. For example, creating a Treebank such as the Penn Treebank (Taylor et al., 2003) can require 20-25 person-years worth of expert manpower (Gala et al., 2014).

In that perspective, recent efforts have favored cost-effective approaches such as automated ones (e.g. word embeddings), or methods aiming at reducing the manpower cost through crowdsourcing (e.g. via the Amazon Mechanical Turk) or at obtaining manpower via implicit crowdsourcing (e.g. Games-With-A-Purpose approaches).

We introduce in this paper a generic approach combining implicit crowdsourcing and language learning that can be used to mass-produce in a cost-effective fashion LRs through language learning activities. The cost-effectiveness of the approach consists in exchanging the manual task of creating and curating an LR with the task of developing and maintaining a language learning service that "delegates" the creation and curation of an LR to a crowd of language learners while maintaining expert quality.

Such a generic approach relies on an implicit crowdsourcing paradigm described and discussed in Section 3.. In Section 4., we report on on-going proof-of-concept efforts in which one of the identified challenges is partially addressed while in Section 5., we report on the *European Network for Combining Language Learning with Crowdsourcing* (enetCollect) allowing us to address another identified challenge. We then discuss in Section 6. an exploratory study on the various types of exercises that can theoretically be used to crowdsource LRs, together with the implicit crowdsourcing scenarios explored by the different authors of this paper. We then finally conclude in Section 7.

2. Related Works

As far as our understanding of the state of the art goes, no previous efforts are directly comparable to ours and only a few efforts have focused on combining language learning and implicit crowdsourcing. With this specific combination in mind, we are aware of the Duolingo language learning platform which was previously used to crowdsource translations (von Ahn, 2013), of recent efforts achieved by some authors of this paper so as to extend via a vocabulary trainer the commonsense ontology *ConceptNet* (Speer et al., 2017; Rodosthenous et al., 2019; Lyding et al., 2019; Rodosthenous et al., 2020) and of two tools used in the classroom to implicitly crowdsource POS corpora (Sangati et al., 2015) and syntactic dependencies (Hladká et al., 2014).

The related state of the art also includes approaches aiming at addressing the lack of LRs. Such efforts can be grouped in three categories: the efforts aiming at automatizing the creation and curation of LRs, the efforts aiming at the creation and curation of LRs via crowdsourcing (but that are not concerned with language learning) and large initiatives focusing on making the most out of the existing LRs.

Regarding the efforts aiming at automatizing the creation and curation of LRs, the state of the art relates to numerous past works in which existing data and automatic processes are used to semi-automatically ease the manual curation and extension of an LR. In that category, we can cite previous works such as the ones targeting lexica (Nicolas et al., 2008; Cholakov and Van Noord, 2010) or treebanks (Torr, 2017; Dima and Hinrichs, 2011).

Regarding the efforts aiming at the creation and curation of LRs via crowdsourcing, they can mostly be sub-categorized into two groups: the ones relying on implicit crowdsourcing approaches (in which the crowds are not necessarily aware that their input is crowdsourced) and the ones explicitly involving a crowd through crowdsourcing platforms (in which the crowds are fully aware that their input is crowdsourced, e.g. Zooniverse[1], Crowd4u[2], Amara[3] or Amazon Mechanical Turk[4]) and confronting them with simplified tasks, often referred to as "micro-tasks", that serve more complex objectives.

Regarding implicit crowdsourcing, the related state of the art is mainly defined by Games-With-A-Purpose (GWAPs) (Chamberlain et al., 2013; Lafourcade et al., 2015). Some of the most well-known efforts are JeuxDeMots (Lafourcade, 2007) that crowdsources lexico-semantic associations between words, Phrase Detective (Chamberlain et al., 2008; Poesio et al., 2012; Poesio et al., 2013) that collects data about anaphoras, ZombiLingo (Fort et al., 2014; Guillaume et al., 2016) that crowdsources syntactic dependency relations, Wordrobe (Bos et al., 2017) that gathers varied annotations (e.g. part-of-speech, named-entities tagging etc.), Robot Trainer (Rodosthenous and Michael, 2016b) which amasses acquisition of knowledge rules, or TileAttack that compiles text-segmentation data (Madge et al., 2017).

As regards efforts involving a crowd through crowdsourcing platforms, numerous works could be cited. They 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).

Finally, with regards to large initiatives focusing on making the most out of the existing LRs, the state of the art is mostly composed of CLARIN (Váradi et al., 2008), DARIAH (Blanke et al., 2011) or META-NET (Piperidis, 2012) which tackles the lack of LRs by making the existing ones more easily maintainable, accessible and usable.

3. The Implicit Crowdsourcing Paradigm

The generic approach presented in this paper relies on the following core paradigm: ***IF*** *a specific LR can be used to generate a specific language learning exercise,* ***THEN*** *the answers collected for this exercise can be cross-matched and used to improve the LR.*

This paradigm follows the idea that the learner answers to automatically generated exercises can be either used to correct the LR (i.e. to question/validate the existing entries of the LR) or to extend it (i.e. to discard/verify new entries). This paradigm can be applied to any scenario in which LRs of one specific type can be paired with language-learning

[1] https://www.zooniverse.org/
[2] http://crowd4u.org/en/
[3] http://amara.org/
[4] https://www.mturk.com/mturk/welcome

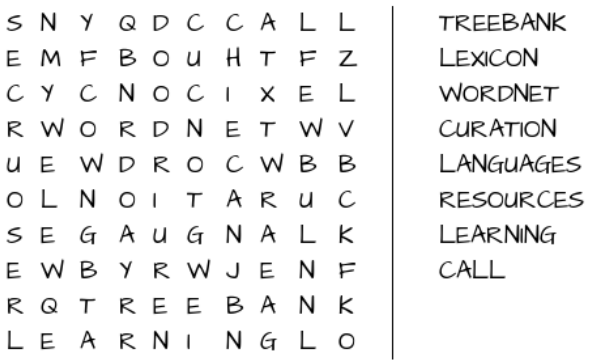

Figure 1: Letter grid exercise

exercises in a way that the exercise content can be generated from this type of LR. For example, Figure 1 displays a widely-used type of exercise that instructs learners to select words in a grid of letters that can be generated from a lexicon. From the crowdsourcing perspective, one could introduce words not recorded in a lexicon (i.e. neologisms) in the grid and see if some students do select them when being instructed to pick common nouns. If they do, this indicates that they believe it to be a common noun. The same approach could also be used with words recorded in a lexicon to double-check if students do believe them to be of a certain grammatical category.

Such a strategy can be applied to other combinations of LRs and exercises. Indeed, an exercise on verb conjugations can be generated from morphological rules and can also allow double-checking the validity of these rules. An exercise asking to transform a sentence (e.g. to the active/passive form) can be generated from a Treebank and can allow deducing what the learner believes to be the function of the words of the sentence (e.g. the subject, the object) which in turn can allow evaluating the validity of the entries of the Treebank. An exercise asking to provide words by analogy (e.g. "yellow is to banana what green is to…") can be generated from a WordNet and can also allow discovering new semantic relations or to evaluate the existing ones. As discussed in greater lengths in Section 6., a varied set of LRs and exercises can actually be considered.

This paradigm exploits a win-win synergy resulting from the fact that, on an abstract level, both NLP researchers curating LRs and students learning a language have a similar goal in mind: creating and curating a language model. Indeed, while the former ones create, curate and use a language model in the form of a digital resource that "teaches" a computer how to process and/or produce a language, the later ones create, curate and use a language model in the form of personal knowledge allowing them to process and/or produce a language. By channeling through crowdsourcing the learners' efforts to complete the automatically-generated exercises, the learners thus create, as a "side-effect" of the learning process, data of primary importance for the enhancement of NLP resources.

While this approach is conceptually promising and presents several noticeable strengths, it also presents some challenges that we are currently working on addressing. In the following sections, we discuss these aspects and explain, at a later stage, how the challenges are currently tackled.

3.1. Strengths of the Approach

The presented approach exhibits five interesting strengths.

First, provided that a language learning service has the potential to continuously attract new users over time, the manpower crowdsourced is potentially endless since the crowd of language learners is naturally renewed over time.

Second, unlike GWAP approaches for which a crowd of players is involved via their interest in being entertained in a fast-evolving entertainment market where numerous competitive solutions exist, our approach targets a crowd of learners via their interest in learning in a context where the existing solutions are less numerous, fast-evolving and diversified. Because of this aspect, and the fact that the crowd targeted expects primarily to learn, we expect the crowd of learners to have clearer, more homogeneous and more confined expectations (e.g. on the user-interaction, entertainment side). The overall competition with other solutions should thus be less fierce than it is for GWAP approaches.

Third, this paradigm exploits a win-win bootstrapping strategy where the contribution on one side further fosters the contribution on the other side. In other words, the more the answers of the learners allow enhancing the LRs used to generate the exercise content, the more the exercise content itself will improve in quality and versatility and vice versa. It therefore creates a virtuous circle that progressively refines the LRs. This approach has thus the potential to create LRs that improve gradually over time and, consequently, would become of unprecedented quality and coverage.

Fourth, according to a report made for the European Union (Social, 2012), we can estimate that only the crowd of language learners aged over 14 years in Europe was at least composed of $\sim$90 millions of persons in 2012 (i.e. without considering the rest of the world, the younger population and the natural growth of this crowd over the past years). The number of language learners worldwide should thus amount to several hundreds of millions of persons while we expect the number of NLP researchers to amount to several thousands of persons worldwide[5]. This 1/10 000 $\sim$1/ 100 000 ratio between these two groups illustrates how vast the crowdsourced manpower that each NLP researcher could actually obtain by offering a language learning service to even a fraction of the language learners could be. Deducing how much manpower could be implicitly crowdsourced from such a crowd is impossible at present due to the lack of prototypical experiments. Nonetheless, so as to get an understanding of the scale of manpower it could represent, we invite the readers to consider the following assumption that is meant to be conservative: if the data crowdsourced from a single learner over the

[5]Estimation based on the attendance at NLP conferences.

course of one year would amount in average to one hour worth of expert manpower[6], then "only" ∼90 millions of European learners would allow for yearly crowdsourcing the equivalent in expert manpower of around 50,000 linguists[7]. Obtaining even a fraction of such manpower (e.g. 2% ≈ 1000 linguists) would certainly allow achieving meaningful results.

Fifth, in order to not undermine the learning efforts of the learner, the vast majority of the exercise content should provide reliable feedback. As such the exercises should mostly be generated from gold-standard data for which it is safe to base feedback on, while learners' answers on new or unreliable entries should be crowdsourced only at a moderate rate (e.g., for 5% of the questions). This characteristic limits on the one hand the amount of data we could crowdsource from each learner. On the other hand, it continuously allows evaluating the competences of learners over time.

3.2. Challenges

3.2.1. Cross-matching the Answers

Relying on a crowd of learners, who have not yet mastered a language, to perform the expert task of improving LRs can at first appear as counter-intuitive as asking your way in a city to a group of tourists. This crowd's lack of expertise can however be compensated in two ways.

On the one hand, by taking into account the performance of the learners in accomplishing a certain exercise. As mentioned previously, this evaluation of learner skills can be done whenever a learner is completing an exercise generated from existing content considered as a gold standard.

On the other hand, the crowd's lack of expertise can be compensated by cross-matching multiple answers crowdsourced from different learners for every single question. Such an aggregation strategy relies on both the classic quality/quantity trade-off for answers[8], and the possibility to decompose complex questions into smaller-grained ones that can be asked (or indirectly deduced) via Boolean questions (e.g., "Is *manger* a French verb?" or "Is *bravo* an antonym for the Italian word *cattivo*?"). Indeed, by decomposing a complex question in such way, one can ensure that the reliability of the "yes" and "no" answers will globally vary from 50% (completely random answers due to an absence of expertise for the linguistic skill targeted) to 100% (perfect understanding of the linguistic skill targeted)[9]. Since each answer with a correctness rate superior to 50% contributes to progressing towards statistical certainty[10], establishing the correct answer to the boolean question addressed is a matter of aggregating enough answers until a quality threshold is met (e.g. a reliability score above 98%).

It should be noted though that this reasoning should be valid when considering the answers of a learner as a whole whereas the performance of a learner can vary from one question to the other. For example, when aggregating learners' answers to a lexicon-generated exercise, one should take into account that not all words are equally difficult to learn and remember. As such, the performances of the learners can vary from one word to the other. The performance of a learner is also expected to evolve over time and at a varying rate depending on the language skill (e.g. vocabulary, grammar etc.) targeted by the questions. Last but not least, one should ensure that the set of learners consulted for a given question is heterogeneous enough in terms of proficiency and background to prevent the unlikely, yet possible, situation where an incorrect answer is excessively chosen by a too homogeneous set of learners with similar shortcomings. In other words, while cross-matching multiple answers crowdsourced from different learners to compensate the crowd's lack of expertise is a viable statistical strategy, implementing such a strategy requires to take into account a number of unfortunate aspects that can punctually undermine it.

Later in Section 4., we report on the first prototypical evaluation we achieved with respect to this challenge.

3.2.2. Providing a Meaningful and Competitive Service

The potential of the generic approach depends on the number of learners using the language learning service provided. Depending on the needs, this could imply the need to attract and retain several hundreds, thousands, or millions of users and thus meeting the expectations of such a crowd. This means that the language learning service should adequately be (1) didactically and content-wise relevant, (2) diversified in terms of content and exercises for the language skills considered and (3) capable of meeting the technical, ethical, legal and economic challenges that come with user bases of several hundreds, thousands, or millions of users.

With regards to the first two requirements, a reliable strategy is to adapt, whenever possible, exercises that are already used nowadays instead of devising new ones. By doing so, the expectations in terms of relevance and diversification of content are in most cases directly met while the need to train the targeted crowd is nullified (since the exercises adapted are already used in practice). As discussed later in Section 6.1., a study of existing textbooks confirmed that such a strategy could be used to devise a number of scenarios pairing a varied set of NLP resources with compatible exercises.

With regards to the third requirement, the one related to the capacity of meeting the technical, ethical, legal and economic challenges that come with user bases of several hundreds, thousands, or millions of users, the study of existing language learning services showed us that maintaining a user base of several hundreds or thousands of users can be achieved by developing a specific language learning service targeting a specific skill (e.g. a vocabulary trainer, see Sec-

[6] ∼10 seconds worth of expert manpower per day.

[7] By considering 225 working days per year and 8 working hours per day corresponding to 1800 working hours per year.

[8] A lower quality is compensated by a higher quantity.

[9] In case a learner has a reliability inferior to 50% because of a personal bias or an intent to under-perform on purpose (e.g. to undermine the crowdsourcing), since most questions are generated from gold standard data, one can automatically invert the answers of the learner to reestablish the reliability of the (inverted) answers between 50% and 100% (unless a learner undermining the crowdsourcing can tell apart the answers used for crowdsourcing.).

[10] A reliability close to 50% has however only little interest (the higher the value, the more interesting the answer).

tion 4.) and performing sufficient local networking with language teachers and/or local authorities. For example, some regions in Europe have specific language certifications[11] for which no online solution to prepare them exist. Creating specific language learning services for them should be a viable approach to secure a dedicated user base for which "one-size-fits-all" generic solutions would be didactically less relevant.

Maintaining a user base of several hundreds of thousands or millions of learners would on the other hand require to define a generic solution and compete with major actors of this market such as Duolingo. As far as we can tell, most existing NLP groups would not have the capacity nor the interest in investing in such a large endeavor. As such, only a large collaborative and coordinated effort could allow exploiting at this scale the potential of this generic approach. Later in Section 5., we report on our efforts in developing enetCollect which aims at creating the building blocks to address this challenge in the mid- to long-term.

4. A proof of concept

Preliminary efforts to implement the generic approach were started in January 2019 and were steadily continued over the course of the year (Rodosthenous et al., 2019; Lyding et al., 2019; Rodosthenous et al., 2020). Such efforts target the creation of a vocabulary trainer that crowdsources useful data to extend and correct an underlying LR: the commonsense ontology *ConceptNet* (Speer et al., 2017). In this section, we report on the latest experiments to an extent that matches the focus of this paper[12]. The vocabulary trainer currently offers two types of questions: open-ended and closed ones.

For the open-ended questions, learners are confronted with a word present in ConceptNet and are being asked to provide a word related to it (e.g. "Name one word related to *house*"). In order to help the learner, an automatically-generated link to Wikipedia is provided as well as the possibility to ask for examples. Positive feedback is provided to the student if the word inputted does have a relation in ConceptNet with the word displayed. If it does not, the learners receive "potential" points that can be converted into real points if the relation is later confirmed. At present, the relation is confirmed by means of other learners suggesting the same relation. In the future, we foresee confirming a relation by aggregating the answers crowdsourced from the closed question automatically generated from it (see after).

For the closed questions, learners are confronted with a Boolean question asking if two words are related to one another (e.g. "Is *house* related to *door*?"). Learners can answer "yes", "no" or "I don't know". Closed questions are generated in two fashions. On the one hand, they are generated from the input obtained from open-ended questions when a relation between two words is suggested by several learners often enough over a period of time. For the experiment, the criteria were for two students to suggest the same relation. At present, the expected answer for such a closed question is "yes" (we manually evaluated that this is the case for 85.9% of the closed questions we generated from the answers collected to the open-ended questions during the experiment described below). On the other hand, closed questions are also automatically generated from ConceptNet and the expected answer is the one that matches the data in it. In both types of closed questions, the feedback is provided according to the expected answer which is originally set by the generation mechanism and can thus be incorrect. In the future, we intend to dynamically validate or adjust the expected answers if the answers crowdsourced from learners suggest so. In that configuration, the generation mechanisms, be it the ones relying on ConceptNet or the one relying on the answers provided to the open-ended questions, would be considered as "artificial learners" and the expected answer they associate with a closed question would be considered as "artificial answers" (with a reliability score associated with each generation mechanisms). By doing so, we will open the possibility for regular learners to contradict or validate over time the expected answer originally set by the generation mechanisms.

During an experiment held in November 2019, 81 learners of English with high-proficiency (C1 according to the CEFR classification[13]) were involved over 16 days and informed of the prototypical nature of the experiment. A total of 9170 answers and 2471 answers to the open and closed questions was crowdsourced together with 36 valid answers to a user survey. We randomly-picked 100 answers from each of the five users that answered the largest number of closed questions. Two annotators manually evaluated the related questions in order to evaluate both the correctness of the learners' answers and of the expected answers set by the generation mechanisms. We gained the following valuable insights regarding our current implementation.

Issues. We identified 3 important issues. First, we overestimated the number of answers we would obtain per closed questions and only gathered an average of 0.98 answers which is clearly insufficient to perform any aggregation. We should thus either have asked fewer closed questions, or have favored some over others, or have gathered more answers to closed questions by raising the ratio of closed questions as compared to open ones or by having a longer experiment in order to crowdsource more answers for all types of questions. Second, the quality of the expected answer set by the generation mechanisms was on average too low, especially the ones generated from ConceptNet (76.6%), to provide a feedback of sufficient quality to learners. We should thus need to fine-tune these mechanisms. Third, in order to foster participation, we gamified the vocabulary trainer by allowing learners to compete between themselves with a point-based system. Points were gained by answering the expected answer whereas answering otherwise did not induce a loss of points. As mentioned by some learners in the user survey, this led learners to always answer something in case of doubt

[11] E.g. the bilingual German-Italian language certification of South Tirol, `http://www.provincia.bz.it/formazione-lingue/bilinguismo/default.asp`.

[12] More detailed explanations about the vocabulary trainer and the latest results can be found in (Rodosthenous et al., 2020).

[13] `https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages`

instead of pressing the "I don't know" option. In addition, we noticed that for 88.4% of the closed questions generated both from ConceptNet and from the answers collected with the open-ended questions the expected answer was "yes". This led students to consider "yes" as the default answer[14] which created another problematic side effect: for closed questions for which the manually-evaluated answer was "yes", the reliability of the learner answers was above 50% (i.e. 91.8%) whereas for the ones for which the manually-evaluated answer was "no", the reliability of the learner answers fell under 50% (i.e. 27.1%). The "yes" answers thus became overabundant. As it stands, the set-up could thus not rely on simple aggregation methods such as a weighted majority vote (since "yes" would probably become the final answer for all closed questions) or would first need to filter out part of the dubious "yes" answers (e.g. by considering the ones for which a higher-than-average rate of "no" answers is provided by other students).

Strengths. We identified 3 strengths. First, despite the third issue mentioned above 50% of the "no" answers and 80.8% of the "yes" answers of the students to closed questions were indeed correct. Therefore, there are good reasons to believe that if the learners were more conservative in their answers to closed questions, fewer dubious "yes" answers would be collected and the third issue could be solved. So as to achieve such a result, we should generate more closed questions for which the expected answer is "no" to not let the students consider that "yes" is the default answer and we should foster their use of the "I don't know" option by making them lose points when providing an answer that is not the expected one. We are also considering using more specific types of relations between words in ConceptNet (such as the ones denoting synonymy, antonymy, hyponymy or hyperonymy) to reduce the uncertainty of the learners. Second, the students provided in open questions 727 relations that where not present in ConceptNet, 85.9% of which were manually evaluated as correct. This confirms that the learners can be used to generate a valid set of hypotheses to extend ConceptNet. Third, the answers to the user survey allowed concluding that the user experience was fun for 81% of the learners and inspiring for 97% of them. At the same time, the competence of the learners improved during the second half (78.9% of answers manually-evaluated as correct) of the experiment when compared with the first half (75.1% of answers manually-evaluated as correct). This proves that the prototype already has an educational value and also tends to indicate that the vocabulary trainer could be used outside of the context of a research experiment.

All in all, the latest experiment allowed obtaining great insights into the current implementation. Whereas it is not yet fully functional, we would argue that the results obtained tend to confirm the viability of the approach.

[14] Also explicitly mentioned in the user survey.

5. The EnetCollect COST Action

As discussed earlier at the end of Section 3.2.2., only a large collaborative and coordinated effort could allow exploiting the potential of the generic approach presented in this paper at a scale of several hundreds of thousands or millions of learners. In this section, we report on our efforts in developing enetCollect which aims at creating the building blocks to address this challenge in the mid- to long-term.

EnetCollect[15] is an international network funded as a COST Action[16] that pursues the long-term challenge of fostering language learning in Europe and beyond by taking advantage of the ground-breaking nature of crowdsourcing and the immense and ever-growing crowd of language learners and teachers to mass-produce language learning material such as lesson or exercise content and, at the same time, language-related data such as LRs. EnetCollect was launched in March 2017 and will continue until April 2021.

5.1. Organization

EnetCollect is organized around five research-oriented working groups (WGs) that are pursued in a parallel fashion while remaining interdependent with regards to the overarching objective of creating in the mid- to long-term language learning platforms capable of attracting and retaining millions of users. Indeed, while WG3 (*User-oriented design strategies for a competitive solution*) focuses its efforts on addressing the needs for a didactically and content-wise relevant solution, WG1 (*R&I on explicit crowdsourcing for language learning material production*) and WG2 (*R&I on implicit crowdsourcing for language learning material production*) focus their efforts in creating methods to mass-produce content and exercises, whereas WG4 (*Technology-oriented specifications for a flexible and robust solution*) and WG5 (*Application-oriented specifications for an ethical, legal and profitable solution*) focus their efforts on the technical, ethical, legal and economic challenges that come with large and diversified user bases.

Discussing how the WGs progress with regards to their respective open-ended objectives and, as such, how they respond to the overall challenge of creating a meaningful and attractive learning service would go far beyond the scope of this NLP-oriented paper. In the next section, we nonetheless report key achievements demonstrating how active and productive enetCollect is and, as such, how much enetCollect allows addressing at a higher pace the challenge of creating the competitive solutions that will be needed to scale up the size of the crowds targeted.

5.2. Current Achievements of EnetCollect

By December 2019, the following main achievements can be reported in terms of participation of stakeholders, interaction among them, research outputs and both support and evaluation from the COST framework.

With regards to participation of stakeholders and interaction among them, 260 stakeholders affiliated with institutions located in 39 out of the 40 countries participating in the COST framework, as well as institutions located in

[15] https://enetcollect.eurac.edu/

[16] https://www.cost.eu/

Canada and the United States have registered on the communication means (e.g. mailing lists). 12 different independent, co-located or co-organized meetings[17] were celebrated in 9 countries. A total of 463 members were invited to these meetings and contributed to them with overall several hundreds of presentations, training sessions and posters. Finally, 34 different research stays, lasting 402 days overall, were funded and led to profound cooperation on relevant subjects between 36 different members.

With regards to research outputs, 8 independent project proposals with objectives in line with enetCollect's goals were submitted by members, of which 5 were funded[18]. 15 publications were already accepted, despite the novelty of the efforts undertaken and thus the difficulty in gathering and achieving enough data and outputs worth publishing in the rather limited span of time since the start of enetCollect.

Finally, with respect to support and evaluation from the COST framework, there are positive signs indicating that enetCollect is considered as one of the large well-functioning COST Actions by the COST administration itself. Indeed, the yearly budget for COST Actions is mainly decided based on the number of countries they involve and their overall performance from one year to the other. For its first three years, enetCollect received a total budget superior by 62% from the average funding allocated to COST Actions to an extent that is close to the upper limit allowed by the COST framework[19] which indirectly indicates that enetCollect was meeting the expectations of the R&I experts observing its progression with respect to its objectives and providing recommendation for funding. This suspicion was confirmed by its mid-way expert report performed after the first two years by the COST administration in which enetCollect received a very favorable evaluation for which all evaluated criteria were described as being met in either a "very good" or an "excellent" fashion.

The achievements reported above tend to indicate that enetCollect allows addressing at a higher pace the challenge of creating competitive solutions that combine language learning and crowdsourcing. As COST is a suitable framework for creating large coordinated initiatives, we would argue that enetCollect is addressing the second challenge reported in Section 3. in a constructive manner and creates a suitable context for side- and follow-up initiatives to be started.

5.3. NLP within enetCollect

The NLP community has so far been the language-related R&I community most involved in enetCollect. The Action itself has been proposed and led since its start by NLP-related researchers and half of the Core Group members (CG) steering the Action are NLP-related, including the three central roles of chair, vice-chair and grant holder. Also, around 43% of the 185 members registered on the intranet have NLP-related profiles.

[17]The meetings have been of four types: three large Annual Meetings, six smaller independent Meetings, two Training Schools and one Hackathon.

[18]For an overall amount of ∼700k euros.

[19]EnetCollect received 620k euros for its first three years whereas COST Actions receive in average around 375k for their first three years (∼500k euros over four years).

6. Implicit Crowdsourcing Scenarios

In this section, we explore to what extent the generic approach described in this paper can allow addressing the long-standing issue of the lack of LRs by discussing how many types of LRs could be crowdsourced via multiple implementations. We explore this question from two angles. On the one hand, we report on a study we performed on exercises of existing textbooks (see Section 6.1.), while, on the other hand, we sketch the planned and on-going efforts of the authors to implement the approach in order to gather LRs of their own specific interest (see Section 6.2.).

6.1. A Study on Language Learning Exercises

When developing an implicit crowdsourcing tool, reusing and adapting an existing workflow can foster the adoption of the crowdsourcing tool while canceling the need to train the crowd targeted, as it is already proficient with respect to the workflow adapted. This aspect is especially relevant for well-established domains such as language learning that often-enough are less inclined to changes.

We therefore made a preliminary study[20] to estimate how many existing exercises currently used in textbooks could be automatically generated and, out of those, how many could allow crowdsourcing NLP datasets. We therefore studied the exercises from five English textbooks used nowadays in Maltese language schools and covering all together the six CEFR levels (HarperCollins, 2013a; HarperCollins, 2013b; Harvey and Rogers, 2015; Harvey, 2015; Hewings, 2015). After reviewing them, we concluded that ∼90% of the exercises could be automatically generated and half of these ∼90% could be used to crowdsource datasets relevant to the following NLP subjects: Treebanks, POS Corpora, Word-sense Disambiguation, Grammar Error Correction, Paraphrasing, Semantic Ontologies, Dialog Systems, Question Answering and Image Labelling.

We could also observe that we were not able to identify exercises for some of the LRs targeted by our members such as Multiword-expression (MWE) datasets or Morphological Rules (see next section). However, it is worth noting that our study had a preliminary nature and was limited to only five textbooks. Future efforts will cover other textbooks as well as the exercises provided by existing online language learning solutions. This study however confirmed our intuition for the NLP subjects mentioned above.

6.2. Planned and On-going Efforts

In the following paragraphs, we focus on the planned or on-going efforts of the authors of this paper to implement the generic approach in multiple ways targeting LRs of their own specific interest. This section aims at showing how diversified the interest for the generic approach already is and, as such, how likely it is to be implemented in many different fashions in the future. We thus briefly mention efforts that are currently in different stages of development, including some that are yet to be adapted to existing language learning exercises to ensure their relevance and adoption by the learners (see Section 3.2.2.).

[20]An intensive team effort of 5 people jointly working on the material for 6 hours during a meeting.

Lexical Datasets. The *Institute for Applied Linguistics (IAL)* of *Eurac Research* performs research on the three official languages of South Tyrol, namely Italian, South Tyrolean German and a minority language called Ladin. The IAL aims currently at crowdsourcing wide-coverage Part-of-Speech (POS) lexica for South Tyrolean German and Ladin via exercises asking learners to group words according to their properties (e.g. "select all verbs among these five words") or to identify words within a grid of random letters (e.g. "select five adjectives in the grid"). With regards to the crowd of learners, it is foreseen to specially craft solutions for the local language certification[21] which is mandatory for public positions and for which no dedicated online learning solution is available nowadays.

Morphological Data. Recent efforts by the *NLP group at the University of Malta* have focused on the creation of a morphological analyzer for Maltese (Cardenas et al., 2019). Most of the information available to-date has been obtained automatically (Borg and Gatt, 2014). So as to further improve a morphological dataset, it is foreseen to use it to generate exercises where learners are instructed to inflect a lemma according to a given morpho-syntactic label (e.g. to conjugate a verb in a certain tense and form) and control if the inflected forms provided by the learners contradict or validate the existing entries of the dataset. It is foreseen to target a crowd of learners at the University of Malta who are learning Maltese as a foreign language.

MWE Datasets. The *PARSEME-IT research group of the Department of Literary, Linguistic and Comparative Studies, University of Naples "L'Orientale"* has a special focus on MWE lexica and corpora annotated with MWEs (Constant et al., 2017). As such, exercises that ask learners to identify/validate MWEs in monolingual texts and suggest possible translations or ask learners to identify/validate MWEs and their translations in parallel corpora are of special interest. The targeted students would be students of the university L'Orientale, especially those attending the translation classes with a solid curriculum in Linguistics and Translation Studies.

Lexico-morphological and Learner Datasets. The *Natural Language Processing research group of the Department of Computer Science, University of Helsinki* performs development of a language learning system called Revita[22] (Katinskaia et al., 2018). It is foreseen to crowdsource lexical and morphological data (e.g. words related to a given context or disambiguation of ambiguous grammatical forms in context) by generating exercises where students are instructed to choose one or several words from a list of words that are associated with different sets of labels. For the disambiguation task, the intersection of these sets of labels of the chosen words will then be used to reduce the ambiguity for these words in the given context. At present, most users of Revita are Italian learners of Russian and Russian learners studying Finnish.

Lexico-semantic Datasets. The *Thesaurus of Modern Slovene* was developed by the *Centre for LRs and Technologies of the University of Ljubljana* (Slovenia) (Krek et al., 2018). It is foreseen to use the Thesaurus to generate language learning materials or games in which learners can indicate semantic relations (e.g. synonymy, antonymy, hypernymy and hyponymy) for a given word, sort synonym candidates according to their headwords (when adequate) or move them into the recycle bin (when inadequate). By doing so, the students can help sort and clean automatically extracted data. Based on primary feedback from teachers, this type of crowdsourcing could be directly included in school curricula if designed didactically.

Lexical, Learner and Aggregation Datasets. The *Insight Centre for Data Analytics* in the *Data Science Institute* of the *National University of Ireland Galway* is interested in enriching existing or building new bilingual language training datasets, while developing new truth inference and aggregation methods for both closed and open-ended questions (Hassan et al., 2016). As a first step, vocabulary exercises targeting non-native learners at beginner level are being developed, especially for Gaelic languages.

Knowledge-based Datasets. Work at the *Computational Cognition Lab, of the Open University of Cyprus*[23] include efforts on GWAP for gathering background knowledge in the form of rules (Rodosthenous and Michael, 2016a; Rodosthenous and Michael, 2014) used to answer questions on stories. Relevant on-going efforts also include the creation of the vocabulary trainer discussed in Section 4., where learners are being trained on the relation between words and for which their aggregated answers will be used to improve ConceptNet (Speer et al., 2017). At present, proficient learners of English are being involved to test the vocabulary trainer and run experiments.

Lexical Datasets. The NLP group of Sorbonne Université has put efforts into designing NLP applications for French regional languages, less-resourced languages which do not benefit from a standardized spelling system. Current efforts aim at developing full-fledged games in which the resolution of the task is conceived as a side effect of the participation. Accordingly, a prototype of a role playing game fostering inter-generational linguistic transmission was developed and will be used to collect NLP resources on idioms and vocabulary (Millour et al., 2019).

Morpho-syntactic Datasets. Researchers from the *University Ss. Cyril and Methodius* in Skopje are interested in creating multilingual LRs. In spite of efforts to manually annotate the Macedonian version corpus of Orwell's 1984, complete POS tagging was only achieved recently. It is foreseen to use a crowd-oriented system where learners are being asked to pick in sentences the words belonging to a specific POS to extend the corpus. The target users consid-

[21] Exam for bilingualism, `http://www.provincia.bz.it/formazione-lingue/bilinguismo/l-esame-di-bilinguismo.asp`

[22] `http://revita.cs.helsinki.fi`

[23] https://cognition.ouc.ac.cy

ered as most suitable are the seniors from the secondary schools, who prepare themselves for the State Baccalauréat.

Acronym Datasets. The *NLP group of the Jerusalem College of Technology* has proven experience with the disambiguation of Hebrew ambiguous acronyms in various types of texts such as Jewish Law documents which usually include a relatively high rate of acronyms (HaCohen-Kerner et al., 2010; HaCohen-Kerner et al., 2013). It is foreseen to use reading comprehension exercises, to ask learners to fill in the correct long form of each acronym, when available in the text. As such, it is foreseen to target learners of varied proficiency for which such reading comprehension exercises are used.

Paraphrase Datasets. The *Human Languages Technologies Lab at INESC-ID Lisboa* has developed the eSPERTo (Mota et al., 2016) paraphrase generator to help users in improving the quality of their texts, with a special focus on Portuguese and its varieties. It is foreseen to generate exercises that allow crowdsourcing feedback on the paraphrases generated and obtain paraphrases by requesting several learners to translate (in their native language) a specific sentence. In both cases, it is foreseen to target proficient learners.

7. Conclusion

In this paper, we have presented a generic approach to combine implicit crowdsourcing and language learning in order to mass-produce LRs for any language for which a crowd of language learners can be involved. We did so by introducing the core paradigm of the approach and discussing its main strengths and challenges. We then reported on an on-going proof-of-concept implementation that partially addresses one of the main challenges and on the international network named enetCollect that tackles another main challenge. We finally discussed how varied was the set of LRs that could be crowdsourced via this approach by reporting on a preliminary study made on existing textbooks and by reporting on the planned or on-going specific efforts of the NLP members authoring this paper.

While the ideas and achievements reported in this paper are the results of an already noticeable shared effort, especially with regards to the overall enetCollect initiative, there are still many aspects to implement, double-check and evaluate more precisely in order to fully understand the viability and potential of this generic approach[24]. Nonetheless, given the large NLP support received by enetCollect, and for this subject in particular, we can assume that the approach is already perceived as viable and worth investing efforts by many NLP stakeholders. Hopefully, initiatives aiming at disseminating the approach, such as this paper, will further raise awareness and accelerate its development.

8. Acknowledgments

This paper is based upon work from the enetCollect COST Action, supported by COST (European Cooperation in Science and Technology).

[24]E.g. how much expert manpower can we crowdsource from a group of learners? How many learners one can involve?

9. Bibliographical References

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

Blanke, T., Bryant, M., Hedges, M., Aschenbrenner, A., and Priddy, M. (2011). Preparing dariah. In *2011 IEEE Seventh International Conference on eScience*, pages 158–165. IEEE.

Borg, C. and Gatt, A. (2014). Crowd-sourcing evaluation of automatically acquired, morphologically related word groupings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3325–3332, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Bos, J., Basile, V., Evang, K., Venhuizen, N., and Bjerva, J. (2017). The groningen meaning bank. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

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.

Cardenas, R., Borg, C., and Zeman, D. (2019). CUNI–Malta system at SIGMORPHON 2019 Shared Task on Morphological Analysis and Lemmatization in context: Operation-based word formation. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 104–112, Florence, Italy, August. 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 et al., editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

Cholakov, K. and Van Noord, G. (2010). Using unknown word techniques to learn known words. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 902–912.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Dima, C. and Hinrichs, E. (2011). A semi-automatic, iterative method for creating a domain-specific treebank. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 413–419, Hissar, Bulgaria, 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.

Evans, J. (2018). Report on language equality in the digital age.

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.

Gala, N., Rapp, R., and Bel-Enguix, G. (2014). *Language Production, Cognition, and the Lexicon*. Springer Publishing Company, Incorporated.

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)*.

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.

HaCohen-Kerner, Y., Kass, A., and Peretz, A. (2010). Haads: A hebrew aramaic abbreviation disambiguation system. *Journal of the American Society for Information Science and Technology*, 61(9):1923–1932.

HaCohen-Kerner, Y., Kass, A., and Peretz, A. (2013). Initialism disambiguation: Man versus machine. *Journal of the American Society for Information Science and Technology*, 64(10):2133–2148.

HarperCollins. (2013a). *Work on your vocabulary, hundreds of words to learn and remember, Elementary*. HarperCollins.

HarperCollins. (2013b). *Work on your vocabulary, hundreds of words to learn and remember, pre-intermediate*. HarperCollins.

Harvey, A. and Rogers, L. (2015). *B1 Workbook Beyond*. Macmillan education.

Harvey, A. (2015). *B2 Workbook Beyond*. Macmillan education.

Hassan, U. U., Curry, E., Zaveri, A., Marx, E., and Lehmann, J. (2016). ACRyLIQ: Leveraging DBpedia for Adaptive Crowdsourcing in Linked Data Quality Assessment. In *20th International Conference on Knowledge Engineering and Knowledge Management (EKAW), November 19-23, 2016, Bologna, Italy*, EKAW 2016.

Hewings, M. (2015). *Advanced grammar in use, a self-study reference and practice book for advanced learners of English*. Cambridge University Press.

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 Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B., and Dobrovoljc, K. (2018). Thesaurus of modern slovene 1.0. Slovenian language resource repository CLARIN.SI.

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

Lafourcade, M. (2007). Making people play for lexical acquisition. In *7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December.

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.

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.

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.

Millour, A., Grace Araneta, M., Lazić Konjik, I., Raffone, A., Pilatte, Y.-A., and Fort, K. (2019). Katana and Grand Guru: a Game of the Lost Words (DEMO). In *9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'19)*, Poznań, Poland, May.

Mota, C., Barreiro, A., Raposo, F. A., Ribeiro, R., dos Santos Lopes Curto, S., and Coheur, L. (2016). eSPERTo paraphrastic knowledge applied to question-answering and summarization. In *NooJ 2016: Automatic Processing of Natural-Language Electronic Texts with NooJ*, volume 667 of *Communications in Computer and Information Science*, pages 208–220. Springer International Publish, June.

Nicolas, L., Sagot, B., Molinero, M. A., Farré, J., and de La Clergerie, É. (2008). Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 633–640. Association for Computational Linguistics.

Piperidis, S. (2012). The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 36–42, Istanbul, Turkey, May. European Language Resources Association (ELRA).

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. T. and Michael, L. (2014). Gathering Background Knowledge for Story Understanding through Crowdsourcing. In *Proceedings of the 5th Workshop on Computational Models of Narrative (CMN 2014)*, volume 41, pages 154–163. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, July–August.

Rodosthenous, C. and Michael, L. (2016a). A Hybrid Approach to Commonsense Knowledge Acquisition. In *Proceedings of the 8th European Starting AI Researcher Symposium (STAIRS 2016)*, volume 284, pages 111–122. IOS Press, August.

Rodosthenous, C. and Michael, L. (2016b). 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.

Rodosthenous, C., Lyding, V., Sangati, F., König, A., ul Hassan, U., Nicolas, L., , Horbacauskiene, J., Katinskaia, A., and Aparaschivei, L. (2020). Using crowdsourced exercises for vocabulary training to expand conceptne. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference, LREC 2020*, Marseille, France.

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

Social, T. O. . (2012). Europeans and their languages.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer.

Torr, J. (2017). Autobank: a semi-automatic annotation tool for developing deep minimalist grammar treebanks. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–86, Valencia, Spain, April. Association for Computational Linguistics.

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M., and Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

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.

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.