EnetCollect in Italy

Lionel Nicolas[1], Verena Lyding[1], Luisa Bentivogli[2], Federico Sangati[3], Johanna Monti[3], Irene Russo[4], Roberto Gretter[5], Daniele Falavigna[5]

[1]Institute for Applied Linguistics, Eurac Research, Bolzano
[2]HLT-MT Unit, Fondazione Bruno Kessler, Trento
[3]Department of Literary, Linguistic and Comparative Studies, University of Naples "L'Orientale", Naples
[4] Institute of Computational Linguistics "Antonio Zampolli", CNR, Pisa
[5] SpeechTek Unit, Fondazione Bruno Kessler, Trento

Abstract

English. In this paper, we present the enetCollect[1] COST Action, a large network project, which aims at initiating a new Research and Innovation (R&I) trend on combining the well-established domain of language learning with recent and successful crowdsourcing approaches. We introduce its objectives, and describe its organization. We then present the Italian network members and detail their research interests within enetCollect. Finally, we report on its progression so far.

Italiano. *In questo articolo presentiamo la COST Action enetCollect, un ampio network il cui scopo è avviare un nuovo filone di Ricerca e Innovazione (R&I) combinando l'ambito consolidato dell'apprendimento delle lingue con i più recenti e riusciti approcci di crowdsourcing. Introduciamo i suoi obiettivi e descriviamo la sua organizzazione. Inoltre, presentiamo i membri italiani ed i loro interessi di ricerca all'interno di enetCollect. Infine, descriviamo lo stato di avanzamento finora raggiunto.*

1 Introduction

In this paper, we present the COST network enetCollect that aims at kick-starting an R&I trend for combining language learning with crowdsourcing techniques in order to unlock a crowdsourcing potential for all languages, consisting in learning and teaching activities. This potential will be used to mass-produce language learning material and language-related datasets, such as NLP resources.

We also present enetCollect's Italian members alongside their NLP-related interests. Indeed, NLP heavily relies on language resources and their availability is crucial for the delivery of reliable NLP solutions. Due to high costs of production, resources are often missing, especially for lesser used languages. As enetCollect researches new approaches to tackle such issues, it is a project of particular interest for the Italian NLP community.

EnetCollect connects to ongoing crowdsourcing research, including *Games With A Purpose* approaches (Chamberlain et al., 2013; Lafourcade et al., 2015) for collecting data through gamified tasks (cf. e.g. JeuxDeMots (Lafourcade, 2007), or ZombiLingo (Guillaume et al., 2016)), collaborative approaches such as *Wisdom-of-the-Crowd* initiatives (e.g. dict.cc[2], Wiktionary[3], and Duolingo (von Ahn, 2013)), or general *Human-based Computation* activities (implemented through platforms like Zooniverse[4], Crowd4u[5], etc.).

This paper aims at fostering the participation of the Italian NLP community while further allowing it to benefit from the research and collaboration opportunities enetCollect offers (e.g. research stay grants) for its remaining 2.5 years of funding. Sections 2 and 3 present enetCollect's ambition, and its organization while Section 4 introduces the Italian members and their research interests. Sections 5 and 6 report on achievements up to now and the current state of affairs.

2 Challenge, Motivation and Objectives

Started in March 2017, enetCollect will pursue, until April 2021, 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-

[1]European Network for Combining Language Learning with Crowdsourcing Techniques, Web: (EnetCollect, 2018)

[2]https://www.dict.cc
[3]https://www.wiktionary.org/
[4]https://www.zooniverse.org/
[5]http://crowd4u.org/en/

growing crowd of language learners and teachers[6] to mass-produce language learning content and, at the same time, language-related data such as NLP resources. The prospect of mass-producing language-related data can vastly impact domains such as NLP, which in turn will impact back on language learning by fostering support from various language-related stakeholders (e.g. see Section 4 for NLP-related crowdsourcing scenarios).

As intensifying migration flows (due to economical and geopolitical reasons) increase the diversification of language learner profiles and the demand for learning material, the launch of such an R&I trend is very timely. Indeed, the effectiveness of the existing material runs the risk of gradually falling behind and the varied combinations of languages taught and target groups can hardly be addressed by small-scale initiatives. EnetCollect timely kick-starts an overarching R&I trend to continuously foster various initiatives. Funding-wise, the timing is also favorable as both the increasing need for learning solutions and the problem-solving nature of crowdsourcing are widely acknowledged.

The creation of a new R&I community is addressed through formal *Research Coordination Objectives* aiming at creating a shared knowledge of the subject, at carrying out prototypical experiments and at disseminating promising results while formal *Capacity-Building Objectives* aim at creating the core R&I community, communication means and new initiatives. In Section 5, we report on progress regarding these objectives.

3 Working Groups and Coordinations

EnetCollect makes a working distinction between *explicit* and *implicit* crowdsourcing approaches: while for *explicit* crowdsourcing the crowd intentionally participates (e.g. Wikipedia), for *implicit* crowdsourcing the crowd is not necessarily aware of its participation (e.g. reCaptcha[7]). EnetCollect is organized along five **working groups (WG)** and three support groups called **coordinations**.

Whereas **WG1** focuses on *explicit* crowdsourcing approaches to create data or learning content (e.g. collaboratively creating lessons), **WG2** focuses on *implicit* crowdsourcing approaches for the same purpose (e.g. generating exercise content from language-related resources and collecting the answers to the exercises to correct and extend the resources used). **WG3** focuses on user-oriented design strategies to attract and retain a crowd (e.g. studying the relevance and attractiveness of learner profiling for vocabulary training). **WG4** focuses on studying the functional demands and the existing solutions related to language learning and crowdsourcing (e.g. technical solutions addressing the scalability need of some methods). Finally, **WG5** focuses on application-oriented questions such as ethical issues, legal regulations, and commercialization opportunities.

The five WGs are different content-wise and can be pursued in a parallel fashion. Nonetheless, they remain interdependent in the overarching objective. For example, the boundary between *explicit* and *implicit* crowdsourcing (WG1 and WG2) is sometimes difficult to draw when the crowd is explicitly involved while their actions are being implicitly crowdsourced[8]. Also, any crowdsourcing approach will fail if there is no crowd to rely on (WG3), no technical solution to support its functional needs (WG4), and no appropriate ethical or legal contexts to implement it (WG5). Alongside the WGs, three **coordination groups** on **Dissemination, Exploitation** and **Outreach** are providing standardized support for WG-transversal tasks.

4 Research Interests of Italian Members

The Italian members are currently among the most numerous and active participants to the Action and its events. In addition, the Action coordination (Chair and Grant Holder) is carried out by two Italian members from Eurac Research (see below). Being all related to NLP, enetCollect's Italian partners have a common interest in combining language learning with implicit crowdsourcing (WG2) so as to extend and correct NLP datasets. All crowdsourcing scenarios described hereafter share the same overarching approach: the NLP partner uses an NLP dataset to generate exercise content and both crowdsources and cross-matches the learners' answers in order to validate/discard the data used to generate the exercise content, just like GWAP players validate/discard data while playing. Deriving expert knowledge from cross-matched learners' answers is a challenge enetCollect aims at addressing. Relying on a crowd of

[6] 21% of the Europeans aged over 14 years (~90 millions people, Eurobarometer report, (European Commission, 2012)

[7] https://www.google.com/recaptcha

[8] E.g. crowdsourcing learner essays and their corrections by teachers to create annotated corpora.

learners is however promising in two ways. First, learners should be mostly confronted with exercise content generated from reliable NLP data so as to not undermine their efforts. Their constantly-evaluated proficiency levels thus provide a reliability score for their answers. Second, as a crowd of learners renews itself over time, the set of crowdsourced answers for each question is potentially infinite and their "inferior" reliability is thus compensated by their "superior" quantity.

The **Institute for Applied Linguistics (IAL)** of **Eurac Research** is particularly concerned with research on the three official languages of South Tyrol (Italian, South Tyrolean German and the minority language Ladin). As regards NLP, Italian is the best covered while South Tyrolean is approximated by adapting solutions for standard German and Ladin has barely any coverage. To improve this situation, the IAL aims at crowdsourcing varied NLP resources for South Tyrolean German and Ladin, starting with wide-coverage Part-of-Speech (POS) lexica. The foreseen crowdsourcing scenario is to use POS lexica to generate exercise content for widely adopted exercises such as the one for grouping words according to their properties (e.g. "select all verbs among these five words") or for identifying words within a grid of random letters (e.g. "select five adjectives in the grid". By crowdsourcing the learners' answers, the IAL aims at gradually improving the lexica while continuously adding new entries. As for the targeted crowd of learners, the IAL will build on its long-standing collaborations with schools (Vettori and Abel, 2017; Abel et al., 2014) and is considering to target the local language certification[9], an obligatory exam for public positions for which no dedicated learning tool is currently available online.

The **Human Language Technology - Machine Translation (HLT-MT) research unit** of **Fondazione Bruno Kessler (FBK)** is concerned with MT technologies supporting both human translators and multilingual applications. The creation of dedicated language resources is thus a core activity. Within enetCollect, HLT-MT aims at enriching existing parallel corpora and at enhancing MT evaluation by crowdsourcing multiple translations of the same sentence (Bentivogli et al., 2018). As such translations paraphrase one another, they are also of interest for monolingual NLP purposes. Following the growing number of studies on the language learning usage of MT (Somers, 2001; Niño, 2008; Case, 2015; Dongyun, 2017), HLT-MT focuses on "post-editing" exercises fostering *correction* and *writing* skills where students are presented with a sentence and several possible translations and are asked to choose the most appropriate one and, if necessary, revise it. Existing parallel corpora and state-of-the-art MT systems trained on them will allow to test the learners' skills and generate new translations. While learning, students will thus be trained, evaluated and will sometimes be allowed to correct MT outputs and extend training corpora. For such a crowdsourcing scenario, advanced L2 learners will be targeted, especially those studying Translation Studies for Italian, English and German at partners of the Universities of Trento and Bologna.

The **PARSEME-IT research group**[10] of the **Department of Literary, Linguistic and Comparative Studies, University of Naples "L'Orientale"** aims at improving linguistic representativeness, precision, robustness and computational efficiency of NLP applications (Monti et al., 2017). It researches MultiWord Expressions (MWEs[11]), as a major NLP bottleneck, and investigates their representation in language resources and their integration in syntactic parsing, translation technology, and language learning. The possibility to enhance mono- and multilingual language resources focusing on MWEs is of particular interest, especially with regards to MWE lexica and corpora annotated with MWEs. Accordingly, a set of different exercises engaging students from different degrees (junior high, high school, and undergraduates) are envisioned. For example, exercises to improve lists of Italian MWEs and their correspondences in different languages 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. The targeted students are BA and MA students of the university L'Orientale, especially those attending the translation classes with a solid curriculum in linguistics and Translation Studies.

The **Institute of Computational Linguistics 'Antonio Zampolli' (CNR-ILC)** carries out research at the international, European, national and

[9]Exam for bilingualism, Web: (BZ Alto Adige, 2018)

[10]https://sites.google.com/view/parseme-it/home

[11]Groups of words composing one lexical unit, such as 'tirare le cuoia' (En. kick the bucket)

regional level since 1967. It participated in several EU initiatives on language resource documentation and recently took the lead of the national CLARIN-IT[12] consortium. Its main areas of competence also include Text Processing, NLP, Knowledge Extraction, and Computational Models of Language Usage. Among ILC's resources, ImagAct[13], a multimodal resource about action verbs, represents a starting point for crowdsourcing experiments, where words denoting actions could be explained through videos sharing a semantic core. Crowdsourcing could be used to build these datasets by asking learners to label actions shown in short videos. As shown with middle school pupils (Coppola et al., 2017), analyzing a video illustrating verbs and associating it with words in multiple languages reinforce metalinguistic reasoning (CARAP, 2012). Such combinations of semantic traits and action verbs can also be used for textual entailment.

The **SpeechTEK research unit** of **Fondazione Bruno Kessler (FBK)** is working on Automatic Speech Recognition (ASR) and addresses computer assisted language learning as an application field. In a first project, it aims to automatically assess children's reading capability at primary school. ASR is used to align a given text with the speech read out by a pupil, to highlight its errors and score it. A second project concerns the use of ASR and classification tools to automatically check the proficiency of Italian students aged between 9 and 16 years, in learning both English and German. Both written texts and spoken utterances have to be evaluated, using reference scores related to some proficiency indicators (e.g. pronunciation, fluency, lexical richness) given by human experts. In the first project, corrections of ASR errors can be crowdsourced and used to build more reliable models for assessing reading capabilities of children. Similarly, in the second project crowdsourcing could help both to transcribe and to score the answers uttered by the students. In both cases, crowdsourcing could allow to adapt ASR models and produce more reliable gold standards.

5 Progression of the Network

In this section, the most relevant achievements[14] related to the overall progression of the network are reported in relation to the formal *Research Coordination* and *Capacity-Building Objectives* outlined earlier in Section 2.[15]

Creating a core community of stakeholders. The already large initial number of 68 individual members for 34 participating countries has increased by 67% to 114 members and by 10% to 38 countries. The people subscribed to enetCollect's mailing list have increased by 149% from 79 to 197. Also, 15 financed research stays, lasting 152 days overall, led to intense cooperations.

Building the theoretical framework. The 30 presentations and 39 posters at network meetings and 15 research stays have contributed to the first building blocks of the foreseen theoretical framework, especially with regards to the state-of-the-art review. So far, 3 meetings and 1 training school were organized (168 participations in total).

Communication and outreach. EnetCollect's intranet and website are online for 9 and 7 months and host already a substantial amount of information. 11 mailing lists targeting subsets of members were created and used. 4 calls for research stays and 5 calls for meeting participation were distributed and drew attention (and members) to enetCollect. Aside from one invited talk, several early activities for publications at conferences of related research communities are ongoing.

Funding new initiatives. Funding applications were supported early on, e.g. through the advertisement of specific opportunities or dedicated internal campaigns (e.g. for Marie Sklodowska-Curie Individual Fellowships). Three applications for mid-sized projects were already submitted in the first year, of which two got positively evaluated, and one got funded by a Swiss agency.

6 Conclusion

We presented enetCollect, outlined its key aspects and introduced both its Italian members and their research interests. By harnessing even a fragment of the crowdsourcing potential existing for all languages taught worldwide, enetCollect could trigger changes of noticeable impact for language learning and language-related R&I fields, such as NLP. The fast uptake and overall progression of enetCollect within its first year indicate its relevance and the potential magnitude of its ambition.

[12] www.clarin-it.it

[13] www.imagact.it

[14] See more information on http://enetcollect.eurac.edu.

[15] We do not report on content-related results as these are too numerous and varied and, more importantly, they are (or will be) the focus of different publications authored by the members having achieved them.

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014. Koko: an l1 learner corpus for german. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2414–2421, Reykjavik, Iceland. European Language Resources Association (ELRA).

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018. Neural versus phrase-based mt quality: An in-depth analysis on englishgerman and englishfrench. *Computer Speech and Language*, 49:52 – 70.

Provincia autonoma di BZ Alto Adige. 2018. Lésame di bilinguismo. Last accessed: 2018-07-20.

Consiglio d'Europa CARAP. 2012. *Le CARAP: Un Cadre de Rfrence pour les Approches Plurielles des Langues et des Cultures, Comptences et Ressources*. Centre Europen pour les Langues Vivantes, Strasbourg Cedex.

Megan Case. 2015. Machine translation and the disruption of foreign language learning activities. *eLearning Papers*, 45:4 – 16.

Jon Chamberlain, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade, and Massimo Poesio. 2013. Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

Daria Coppola, Raffaella Moretti, Irene Russo, and Fabiana Tranchida. 2017. In quante lingue mangi? tecniche glottodidattiche e language testing in classi plurilingui e ad abilit differenziata. In Francesca Strik Lievers Giovanna Marotta, editor, *Strutture linguistiche e dati empirici in diacronia e sincronia*, Studi Linguistici Pisani, pages 199–231. Pisa University Press.

Sun Dongyun. 2017. Application of post-editing in foreign language teaching: Problems and challenges. *Canadian Social Science*, 13(7):1 – 5.

COST Action EnetCollect. 2018. Enetcollect cost website. Last accessed: 2018-07-20.

Directorate-General for Communication European Commission. 2012. Europeans and their languages. Special eurobarometer 386 report, Survey conducted by TNS Opino & Social, and co-ordinated by the European Commission.

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

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

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

Johanna Monti, Maria Pia di Buono, and Federico Sangati. 2017. Parseme-it corpus an annotated corpus of verbal multiword expressions in italian. In *Fourth Italian Conference on Computational Linguistics-CLiC-it 2017*, pages 228–233. Accademia University Press.

Ana Niño. 2008. Evaluating the use of machine translation post-editing in the foreign language class. *Computer Assisted Language Learning*, 21(1):29 – 49.

Harold Somers. 2001. Three perspectives on mt in the classroom. In *Proceedings of the eighth Machine Translation Summit (MT Summit VIII)*, Santiago de Compostela, Galicia, Spain.

Chiara Vettori and Andrea Abel, editors. 2017. *KOLIPSI II. Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale. / Die Sdtiroler SchlerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung.* Eurac Research, Bolzano/Bozen.

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*, pages 1–2. ACM.