

Introducing the European NETwork for Combining Language LEarning and Crowdsourcing Techniques (enetCollect)

Verena Lyding, Lionel Nicolas, Branislav Bédi, Karën Fort

▶ **To cite this version:**

Verena Lyding, Lionel Nicolas, Branislav Bédi, Karën Fort. Introducing the European NETwork for Combining Language LEarning and Crowdsourcing Techniques (enetCollect). EuroCALL, Aug 2018, Jyväskylä, Finland. hal-01961788

HAL Id: hal-01961788
https://hal.science/hal-01961788

Submitted on 20 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Introducing the European NETwork for Combining Language LEarning and Crowdsourcing Techniques (enetCollect)

Verena Lyding[1], Lionel Nicolas[2], Branislav Bédi[3] and Karën Fort[4]

Abstract

We present enetCollect, a large European network project funded as a COST Action, that sets ground for combining crowdsourcing with IT technologies used in areas such as language learning and Natural Language Processing (NLP). This project tackles a major challenge of bringing together interdisciplinary researchers to foster language learning of all European citizens from diverse socio-demographic, cultural, educational and linguistic backgrounds. It aims at unlocking a crowdsourcing potential available for all languages, including less widely spoken languages, in order to create language resources and achieve a coverage of material for teaching the languages. It will meet its research and capacity-building goals by creating an international community of researchers that will work on producing a comprehensive theoretical framework and running prototypical experiments to benefit a wide range of users and languages, while considering ethical, legal and business issues. This article informs about its objectives, expected impact and strategic organisation that contribute to reaching its flexible and sustainable success goals.

Keywords: crowdsourcing, computer assisted language learning, language resources, COST Action.

1. Introduction

EnetCollect was founded as a large European network project through COST Association for four years starting from March 2017. The project tackles the major European challenge to foster the language learning of all citizens with diversified cultural, educational, linguistic and socio-demographic backgrounds by combining the well-established domain of language learning (Godwin-Jones, 2011) with recent crowdsourcing approaches (Brabham, 2008).

The demand for language learning is continuously increasing due to ongoing globalisation and political happenings, i.e. migration. Also, learner profiles are getting more diversified in terms of the learners' language backgrounds, daily realities and communicative needs. This calls for more adapted learning content to optimally serve the individual learner, and to increase the ecology of learning (Blewitt, 2006). At the same time, the Internet has opened new possibilities for a collaborative development; sharing and reusing content for language learning has never been easier.

2. Objectives and approach

EnetCollect aims to research and promote the possibilities that crowdsourcing offers for language learning. It can provide a long-term solution to the complex and ever-changing demands on language learning by involving the "crowd" in creating and improving language learning materials.

In CALL, Duolingo (von Ahn, 2013) is an example of a language-learning platform which follows a similar logic as enetCollect, in offering free learning services for numerous languages while crowdsourcing lessons and translations through exercises.

EnetCollect distinguishes two forms of crowdsourcing approaches, as follows. *Explicit* crowdsourcing involves a crowd that intentionally participates in data creation, i.e. by providing language content for teaching and curricula design for specific languages. It is based on the "wisdom-of-the-crowd" approach that enables stakeholders to collaboratively create resources of common interest, as implemented in the language learning platform Memrise, the online dictionaries Dict.cc, Wiktionary and Lingobee (Procter-Legg et al., 2012), and the well-known Wikipedia. *Implicit* crowdsourcing involves a crowd that is not necessarily aware of its participation. It includes generating exercises and content based on learner's achievements, e.g. Games-with-a-Purpose (GWAPs) (Chamberlain et al., 2013; Lafourcade et al., 2015). The learners' activities are monitored in terms of difficulties and errors, the efficiency in completing different exercises, and learning analytics. This involves features from intelligent CALL (iCALL)

[1] Eurac Research, Bolzano, Italy; e-mail verena.lyding@eurac.edu

[2] Eurac Research, Bolzano, Italy; e-mail lionel.nicolas@eurac.edu

[3] University of Iceland, Reykjavík, Iceland; e-mail brb19@hi.is

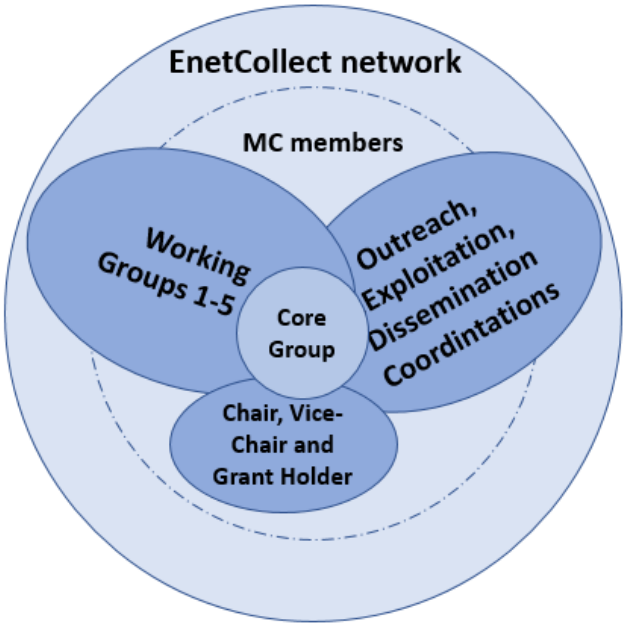
[4] Sorbonne Université, Paris, France; e-mail karen.fort@sorbonne-universite.fr

(Gamper & Knapp, 2002). Based on this, enetCollect will explore new opportunities for autonomous and life-long learning, and concentrate on the practical implementation of theories with regard to learning materials that respond not only to the learner's individual pace but also to the content and, as opposed to Duolingo, to the overall curricula design for each language individually.

3. Structure of the network and implementation strategies

The network is structured into five specialised working groups (WGs) and three coordination groups (Outreach, Exploitation and Dissemination). It is steered by the Core Group and the Management Committee (MC) (see Figure 1). About 200 researchers from more than 40 countries representing different domains, e.g. CALL, crowdsourcing, NLP, e-lexicography, computer science, ethics, law, and business, have joined forces. WG1 deals with approaches for the collaborative creation of teaching content (*explic*it crowdsourcing). WG2 is concerned with the integration of *implicit* crowdsourcing techniques into interactive learning content (e.g. gap filling exercises), which allows to harvest information and data on learners. WG3 researches the usability and user experience of learning platforms in order to assure to attract and retain enough learners and tutors. WG4 addresses technical challenges related to an online learning environment for numerous distributed clients, such as scalability and robustness. Finally, WG5 deals with ethical, legal and commercial perspectives. The coordination groups, i.e. Outreach, Exploitation and Dissemination, support the WGs with transversal tasks enabling optimal communication and knowledge exchange within and outside of enetCollect.

Figure 1. Organisational structure of enetCollect



The objective is to carry out the groundwork for setting into motion a new research and innovation trend for the creation of online language learning solutions. The WGs will explore innovative approaches for the production and enhancement of online teaching materials, the opportunities posing the integration of crowdsourcing techniques into learning environments, and the blending of learning content with gamification and data harvesting. This effort can particularly be beneficial for less widely taught languages with smaller speaker communities that have limited language resources (Mariani, 2015), because it will enable them to gather data for teaching and provide these to learners, who would otherwise have a very limited access to such materials. While for lesser spoken languages it can be particularly challenging to involve a sufficient number of participants, the extent of the network, together with the gamification layer, comprehensiveness and online nature of its approach foster participation and support to a maximum. Depending on the type of language resources dealt with, the completion of material gathering can be achieved with semi-automatic methods (Sagot, 2010) related to *explicit* and *implicit* crowdsourcing, which will collaboratively devise lesson content (*explicit* crowdsourcing) and generate exercise content for individual learners based on their progress and feedback (*implicit* crowdsourcing).

Having a manpower producing language resource materials can result in high costs (Böhmová et al., 2001), but crowdsourcing offers a cost-effective solution. For instance, both Wikipedia, which redefined the domain of encyclopedias, and reCAPTCHA, have made the highly laborious task of manual writing and annotation possible by obtaining a workforce from the crowd. The content providers will upload and share content via telecollaboration across the boundaries of local territories, thus making it available as free, partially free, or paid, all depending on the business scheme, maintenance, and overall support of official authorities. The assurance of quality can also be part of a crowd feedback. Different types of users will evaluate the usefulness and trustworthiness of content via different reward and feedback mechanisms. The problem of ethics is also a very important in this context and will be addressed in connection with data collection and usage, and personal data protection using an ethics by design approach (Spiekermann, 2015).

4. Conclusion and future work

EnetCollect aims both at reviewing the state-of-the-art in order to gather and compile an overview of relevant approaches and techniques, and at achieving a shared understanding of the subject by creating an interdisciplinary framework for defining its terminology, key concepts, objectives and opportunities. It aims at carrying out prototypical experiments, evaluating and disseminating their results. It possesses communication means allowing to easily exchange information and reach relevant stakeholders for targeting new research stays, organized workshops, training schools, and funded initiatives. The involvement of new and current members will be pursued through promotion of the Action through relevant channels of the research domains concerned.

5. Acknowledgements

This article is based upon work from COST Action enetCollect (CA16105) supported by COST (European Cooperation in Science and Technology).

References

Blewit, J. (2006). *The ecology of learning. Sustainability, lifelong learning and everyday life.* (1st ed.). Earthscan.

Böhmová, A., Hajic, J., Hajicova, E. and Hladka, B. (2001). The Prague dependency treebank : Three level annotation scenario. In Anne Abeillé (Ed.). *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving. An introduction and cases. In *Convergence: The International Journal of Research into New Media Technologies* 14(1), 75-90. Sage Publications. DOI: 10.1177/1354856507084420

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. (Eds.), *The*

People's Web Meets NLP. Theory and Applications of Natural Language Processing, 3–44. Springer Berlin Heidelberg.

Gamper, J. and Knapp, J. (2002). A review of intelligent CALL systems. In *Computer Assisted Language Learning* 15(4), 329-342. DOI: 10.1076/call.15.4.329.8270

Godwin-Jones, R. (2011). Emerging Technologies : Mobile apps for language learning. In *Language Learning & Technology* 15(2), 2-11. http://llt.msu.edu/issues/june2011/emerging.pdf

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

Mariani, J. (2015). Technologies de la langue : Etat des lieux. In *Proceedings of the workshop on the Technologies pour les Langues Régionales de France*, Meudon, France.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In Proceedings of the *7th international conference on Language Resources and Evaluation* (LREC 2010), Valletta, Malta. May 2010.

Spiekermann, Sarah (2015). *Ethical IT Innovation: A Value-Based System Design Approach*. CRC Press.

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*, 1–2. ACM.