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. Eurac Research, Bolzano, Italy; verena.lyding@eurac.edu
2. Eurac Research, Bolzano, Italy; lionel.nicolas@eurac.edu
3. University of Iceland, Reykjavík, Iceland; brb19@hi.is
4. Sorbonne Université, Paris, France; karen.fort@sorbonne-universite.fr

How to cite this article: Lyding, V., Nicolas, L., Bédi, B., & Fort, K. (2018). Introducing the European NETwork for COmbining Language LEarning and Crowdsourcing Techniques (enetCollect). In P. Taalas, J. Jalkanen, L. Bradley & S. Thouësny (Eds), *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018* (pp. 176-181). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.833

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 Computer Assisted Language Learning (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.

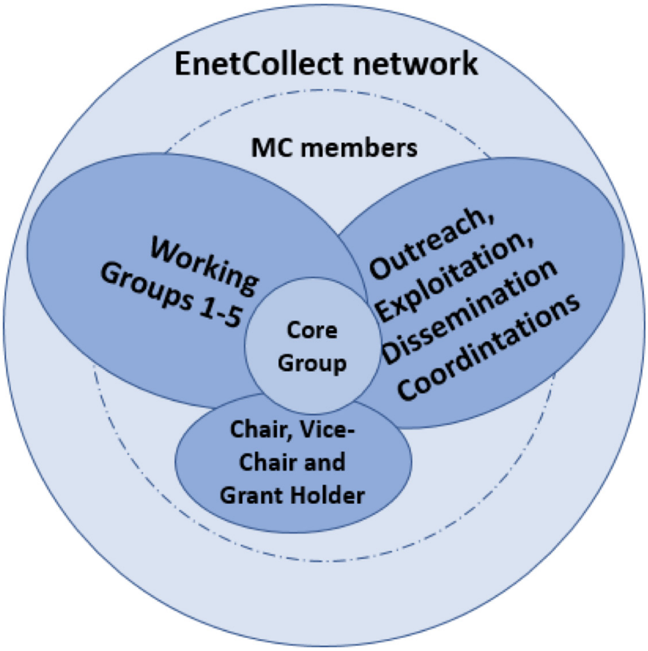
EnetCollect distinguishes two forms of crowdsourcing approaches, as follows. *Explicit* crowdsourcing involves a crowd that intentionally participates in data creation, i.e. by providing language content for teaching and curricula design for specific languages. It is based on the 'wisdom-of-the-crowd' approach that enables stakeholders to collaboratively create resources of common interest, as implemented for example in the language learning platform Memrise, the online dictionaries Dict.cc, Wiktionary, Lingobee (Procter-Legg, Cacchione, & Petersen, 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, Brun, & Joubert, 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) (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 (*explicit* crowdsourcing). WG2 is concerned with the integration of *implicit* crowdsourcing techniques into interactive learning content (e.g. gap filling exercises), which allows for the harvesting of information and data about 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.

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).

Figure 1. Organisational structure of enetCollect


EnetCollect network
MC members
Working Groups 1-5
Core Group
Outreach, Exploitation, Dissemination Coordintations
Chair, Vice-Chair and Grant Holder

Having a manpower producing language resource materials can result in high costs (Böhmová, Hajic, Hajicova, & Hladka, 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, organised 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

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

Böhmová, A., Hajic, J., Hajicova, E., & Hladka, B. (2001). The Prague dependency treebank: three level annotation scenario. In A. 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. *Convergence: The International Journal of Research into New Media Technologies, 14*(1), 75-90. https://doi.org/10.1177/1354856507084420

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., & Poesio, M. (2013). Using games to create language resources: successes and limitations of the approach. In I. Gurevych et al. (Eds), *The people's web meets NLP. Theory and applications of natural language processing* (pp. 3-44). Springer. https://doi.org/10.1007/978-3-642-35085-6_1

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

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

Lafourcade, M., Brun, N. L., & Joubert, A. (2015). *Games with a purpose (GWAPs)*. Wiley. https://doi.org/10.1002/9781119136309

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

Procter-Legg, E., Cacchione, A., & Petersen, S. A. (2012). Lingobee and social media: mobile language learners as social networkers. *Paper presented at the International Association for Development of the Information Society (IADIS) International Conference on Cognition and Exploratory Learning in Digital Age (CELDA), Madrid, Spain, Oct 19-21, 2012*. https://eric.ed.gov/?id=ED542698

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, S. (2015). *Ethical IT innovation: a value-based system design approach*. CRC Press. https://doi.org/10.1201/b19060

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* (pp. 1-2). ACM. https://doi.org/10.1145/2449396.2449398

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018
Edited by Peppi Taalas, Juha Jalkanen, Linda Bradley, and Sylvie Thouësny

Publication date: 2018/12/08

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it was not under consideration for publication elsewhere. While the information in this book is believed to be true and accurate on the date of its going to press, neither the editorial team nor the publisher can accept any legal responsibility for any errors or omissions. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover theme by © 2018 Antti Myöhänen (antti.myohanen@gmail.com)
Cover layout by © 2018 Raphaël Savina (raphael@savina.net)
Drawings by © 2018 Linda Saukko-Rauta (linda@redanredan.fi)

ISBN13: 978-2-490057-22-1 (Ebook, PDF, colour)
ISBN13: 978-2-490057-23-8 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-21-4 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, UK: British Library.
Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: Décembre 2018.