Developing pedagogically appropriate language corpora through crowdsourcing and gamification

Rina Zviel-Girshin[1], Tanara Zingano Kuhn[2], Ana R. Luís[3], Kristina Koppel[4], Branislava Šandrih Todorović[5], Špela Arhar Holdt[6], Carole Tiberius[7], and Iztok Kosem[8]

Abstract. Despite the unquestionable academic interest on corpus-based approaches to language education, the use of corpora by teachers in their everyday practice is still not very widespread. One way to promote usage of corpora in language teaching is by making pedagogically appropriate corpora, labelled with different types of problems (for instance, sensitive content, offensive language, structural problems), so that teachers can select authentic examples according to their needs. Because manually labelling corpora is extremely time-consuming, we propose to use crowdsourcing for this task. After a first exploratory phase, we are currently developing a multimode, multilanguage game in which players first identify problematic sentences and then classify them.

Keywords: crowdsourcing, gamification, language teaching, pedagogical corpora.

1. Introduction and background

Research on corpus-based approaches to language education has been receiving increasing interest and attention in the literature, as can be seen by the growing number of publications on the subject (e.g. Callies, 2019) as well as by the

1. Ruppin Academic Center, Emek Hefer, Israel; rinazg@ruppin.ac.il; https://orcid.org/0000-0002-7926-4476
2. CELGA-ILTEC/University of Coimbra, Coimbra, Portugal; tanarazingano@outlook.com; https://orcid.org/0000-0003-2640-5500
3. CELGA-ILTEC/University of Coimbra, Coimbra, Portugal; aluis@fl.uc.pt; https://orcid.org/0000-0002-7869-7835
4. Institute of the Estonian Language, Tallinn, Estonia; kristina.koppel@eki.ee; https://orcid.org/0000-0003-3194-9801
5. University of Belgrade, Belgrade, Serbia; branislava.sandrih@fil.bg.ac.rs; https://orcid.org/0000-0002-2714-427X
6. University of Ljubljana, Ljubljana, Slovenia; spela.arhar@cjvt.si; https://orcid.org/0000-0003-0565-0531
7. Dutch Language Institute, Leiden, Netherlands; carole.tiberius@ivdnt.org; https://orcid.org/0000-0002-7860-5427
8. University of Ljubljana & Jožef Stefan Institute, Ljubljana, Slovenia; iztok.kosem@cjvt.si; https://orcid.org/0000-0002-4282-9031

How to cite this article: Zviel-Girshin, R., Kuhn, T. Z., Luís, A. R., Koppel, K., Todorović, B. Š., Holdt, Š. A., Tiberius, C., & Kosem, I. (2021). Developing pedagogically appropriate language corpora through crowdsourcing and gamification. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds), *CALL and professionalisation: short papers from EUROCALL 2021* (pp. 312-317). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.54.1352

organisation of a highly successful conference especially dedicated to the topic (i.e. Teaching and Learning Corpora – TaLC). Overall, much current research has been jointly contributing to the continuous development of education-driven corpus tools and corpus-based teaching materials. However, it is widely known that the use of corpora by teachers is still not very widespread (e.g. Callies, 2019) due to a series of reasons, among which are lack of appropriate training and scepticism about the quality and appropriateness of the data (Kilgarriff, 2009).

There is no doubt that not all corpora are equally useful for pedagogical purposes. Authentic texts may contain inappropriate and offensive language, as well as non-standard elements, which might be problematic when presented to learners without the mediation of the teacher. Therefore, before using corpora in education, a combination of different actions must be taken, including close monitoring of the corpus content to identify possible structural (grammar and spelling) problems, and sensitive, offensive, or other inappropriate content. The creation of such a content-controlled corpus, however, is time-consuming, and often requires consulting large teams of linguists and educational experts. We thus decided to start a research and innovation project to compile pedagogical corpora through crowdsourcing. The objective of this paper is to report on the second phase of this project, i.e. the development of a multimode game, which is organised in three stages, namely, data preparation, game preparation, and training of machine learning models. In this paper, we focus on game preparation.

2. Crowdsourcing: an alternative approach to creating corpora

A much-used method to automatically remove inappropriate content from corpora is the rule-based method, e.g. by using a blacklist of words that are considered inappropriate for learners, usually taboo words, swear words, and vulgarisms. This approach has, for instance, been employed for the creation of the SkeLL corpora (Sketch Engine for Language Learning)[9]. Although this method is fast, the main disadvantage is that it relies on quantifiable heuristics that are applied to ALL sentences in the corpus regardless of their meaning. For example, a sentence containing a word such as ‘pussy’ may be considered inappropriate for learners when it refers to a woman’s vagina, but by adding ‘pussy’ to the blacklist, sentences where it occurs in its neutral sense of ‘cat’ will also be removed from the corpus, which is undesirable. Moreover, teachers might want to work with

9. https://skell.sketchengine.eu/

offensive language or sensitive content in their classroom, depending on the unit topic and the characteristics of their students in terms of proficiency level, cultural background, and age.

Taking these two factors together – word polysemy and freedom of choice for teachers concerning the real-world examples they want to use in their classroom – we propose an alternative solution for creating corpora for pedagogical purposes. Instead of deleting sentences containing inappropriate words or sensitive content, our aim is to create problem-labelled corpora, thus allowing teachers and material developers to select the sentences according to their needs and purposes. Since this is an extremely laborious endeavour if done manually, we propose to apply crowdsourcing techniques.

Crowdsourcing (or citizen science) is a practice where members of a wider community contribute to content creation, problem solving, or even to some aspect of research. Crowdsourcing is often based on the framework of collective intelligence (Lévy, 1997) and can be thus defined as a tool to gather collective intelligence for certain tasks. Crowdsourcing in education is defined as "a type of (a) (online) activity in which (b) an educator, or an educational organization (c) proposes to a group of individuals via a flexible open call (d) to directly help learning or teaching" (Jiang, Schlagwein, & Benatallah, 2018, n.p.). Within this context, crowdsourcing activities may (1) benefit education by content, (2) provide practical experience for the participants, (3) contribute to the exchange of complementary knowledge, and (4) augment abundant feedback (evaluations) for learners (Jiang et al., 2018).

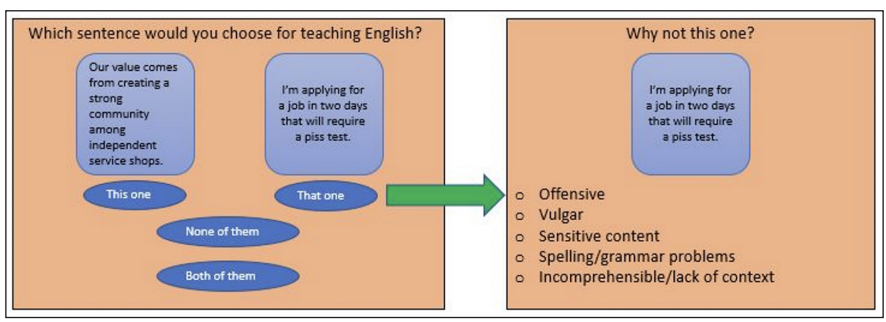
Given the clear advantages of applying crowdsourcing for the creation of language-related resources for educational purposes, we as a group of researchers within the COST Action enetCollect[10] have set up a research and innovation project entitled Crowdsourcing Corpus Filtering for Pedagogical Purposes. The main goal is to have the crowd contribute to the creation of pedagogically appropriate corpora by indicating offensive sentences in data extracted from corpora. In the first phase of this project, we ran a multilanguage crowdsourcing experiment in Pybossa[11] (Kuhn et al., 2021). Although participants' engagement was rather low and the feedback received was that the task was quite dull, analysis of the answers revealed some interesting results. We found that participants did not necessarily consider sentences with explicitly rude content inappropriate for language learners, and that

10. https://enetcollect.eurac.edu
11. Pybossa is a free, open-source crowdsourcing and microtasking platform that allows i.e. to design the crowdsourcing task, control the number of participants and save the collected data; https://pybossa.com/

although they only had to mark what was strictly 'offensive', participants often did more than this. They also marked the sentences that they found problematic, such as incomplete sentences, complex sentences, sentences containing spelling and grammar errors, or even sentences containing too many foreign terms.

Based on the modest results of and the lessons learned from this previous experiment, it was concluded that motivation for participation should be improved, as well as a more specific task should be presented, including the possibility for the participants to point out structural problems. Thus, in the second phase of this project, we have decided to follow Von Ahn (2006) and develop a 'Game with a Purpose' (GWAP), i.e. a game that is fun to play and at the same time collects useful data for tasks that computers cannot yet perform (Hacker & Von Ahn, 2009, p. 1208). Consequently, a multimode and multilanguage (Dutch, Estonian, Serbian, Slovene, and Portuguese) game is currently under development. In this game, players will first select the sentences they consider to be inappropriate for language learning purposes, and then provide the reasons for their choice by indicating in which category or categories the selected example fits, ranging from sensitivity-related content to structural problems. Figure 1 illustrates the type of questions players will encounter in a single-player mode, however, it should be highlighted that gamification elements such as a scoring system and other engagement and motivation-enhancing features will still be added to the game. From the output of this game, i.e. the labelled sentences, corpora will be compiled that can be used by teachers, material developers, and lexicographers.

Figure 1. Illustration of one of the game modes



The game development has three stages. In the first stage, the datasets for the game have been prepared. In order to create datasets of potentially good example candidates and 'bad' example candidates, the web corpora have been automatically filtered for each language with some common and some language-related heuristics

using the GDEX function[12] in the Sketch Engine. The second stage involves game preparation, i.e. the development of varied game modes, implementation of gamification aspects such as scoring and motivation, and the design of the interface. As part of future work, the third stage will be concerned with training of machine learning models for each language based on the identification and categorisation of problematic content by the players.

3. Concluding remarks

Our work, compiling pedagogical corpora through crowdsourcing, will provide examples of good practice and a benchmark methodology, both for the preparation of corpus resources that can be more easily and freely used in the classroom as well as for the preparation of pedagogical language resources and materials. Ultimately, we would like to support teachers' awareness and usage of corpora in their teaching practice by providing them with user-friendly corpora.

4. Acknowledgements

The authors have been funded by the Horizon 2020 Framework Programme of the European Union under the enetCollect CA16105 COST Action, the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene), and the Portuguese national funding agency, FCT – Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020).

References

Callies, M. (2019). Integrating corpus literacy into language teacher education. In S. Götz & J. Mukherjee (Eds), *Learner corpora and language teaching* (pp. 245-263). John Benjamins. https://doi.org/10.1075/slcs.201.12cal

Hacker, S., & Von Ahn, L. (2009). Matchin: eliciting user preferences with an online game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1207-1216). https://doi.org/10.1145/1518701.1518882

12. GDEX stands for Good Dictionary EXamples (Kilgarriff et al., 2008). It is a system for evaluation of sentences with respect to their suitability to serve as good examples for e.g. teaching purposes. Sentences are evaluated with respect to their length, use of complicated vocabulary, presence of controversial topics (e.g., politics, religion…), among others.

Jiang, Y., Schlagwein, D., & Benatallah, B. (2018). A review on crowdsourcing for education: state of the art of literature and practice. In *Proceedings of the 22nd Pacific Asia Conference on Information Systems*. PACIS 2018 Proceedings.

Kilgarriff, A. (2009). Corpora in the classroom without scaring the students. In *Proceedings of the 18th International Symposium on English Teaching and Learning in the Republic of China*. National Taiwan Normal University. http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress* (pp. 425-432). Universitat Pompeu Fabra.

Kuhn, T. Z., Todorović, B. Š., Holdt, Š. A., Zviel-Girshin, R., Koppel, K., Luís, A. R., & Kosem, I. (2021). Crowdsourcing pedagogical corpora for lexicographical purposes. In *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. II, Democritus University of Thrace* (pp. 771-779).

Lévy, P. (1997). Collective intelligence: mankind's emerging world in cyberspace (R. Bononno. Trans.) Helix Books.

Von Ahn, L. (2006). Games with a purpose. *Computer, 39*(6), 92-94. https://doi.org/10.1109/MC.2006.196

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

CALL and professionalisation: short papers from EUROCALL 2021
Edited by Naouel Zoghlami, Cédric Brudermann, Cedric Sarré, Muriel Grosbois, Linda Bradley, and Sylvie Thouësny

Publication date: 2021/12/13

Typeset by Research-publishing.net

Cover Theme by © 2021 DIRCOM CNAM; Graphiste : Thomas Veniant
Cover Photo by © 2021 Léo Andres, Sorbonne Université
Cover Photo by © 2021 Sandrine Villain, Le Cnam
Cover Layout by © 2021 Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-97-9 (PDF, colour)

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

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2021.