Tanara Zingano Kuhn/Špela Arhar Holdt/
Rina Zviel Girshin/Ana R. Luís/Carole Tiberius/Kristina
Koppel/Branislava Šandrih Todorović/Iztok Kosem

INTRODUCING CrowLL – THE CROWDSOURCING FOR LANGUAGE LEARNING GAME

Keywords Crowdsourcing; dictionaries; game; pedagogical corpora

In this demo, we introduce CrowLL (Crowdsourcing for Language Learning), a game with a purpose for creating pedagogical corpora of Dutch, Estonian, Serbian, Slovene, and Portuguese. CrowLL is primarily meant for the publication of SkeLL (Sketch Engine for Language Learning) (Baisa/Suchomel 2014) for these languages, but also applicable for dictionary making and teaching materials development. With this game, we propose an alternative way of creating pedagogical corpora in which corpora are not cleaned of structure and content usually considered inappropriate for learners, but rather labelled. The design process of a pedagogical corpus is characterized by the 'pedagogic mediation of corpora' (Braun 2005), which can be the close monitoring of the content of the corpus to identify possible structural (grammar and spelling) problems as well as sensitive/offensive content. One possible approach to creating pedagogical corpora consists of excluding from the corpora sentences containing words from a blacklist of taboo and swear words, e. g. using GDEX (Kilgarriff et al. 2008). However, one of the greatest disadvantages of this method regards the fact that many words from the blacklist are polysemic. That means that the neutral sense of those words is not displayed in the corpus because all sentences containing those words have been excluded. Moreover, teachers might want to address sensitive/offensive content in their lessons depending on the characteristics of their learners and the teaching context (age, group level, unit topic, etc.). One way to create pedagogical corpora that still contain potentially problematic content and structure is to label, rather than remove these sentences. The end users of these corpora, such as dictionary makers and teachers, can then filter out the sentences according to their purposes. Considering that a) the process of labelling sentences in corpora is extremely time-consuming, if done manually; b) that automatic labelling can also be challenging given the polysemic nature of words; and that c) sensitivity and offensiveness are rather subjective concepts, this project is developing a game in which the crowd helps to achieve this task. With this game, players identify and label problematic sentences automatically extracted from existing corpora. CrowLL is a multimode game available as a webpage and a mobile app. At the time of writing, the single-player mode is being finished, with the dual-player mode expected by the time of the conference. Both modes have three levels, namely, level 1 (I'm curious!), level 2 (I'm eager to help!), and level 3 (I'm feeling enthusiastic!). In level 1, players identify problematic sentences according to their judgement; in level 2, they categorise those sentences, ranging from grammar/spelling problems to offensiveness and sensitivity; and in level 3, players mark in the sentence what they consider problematic. Players can choose to play the full game cycle (levels 1, 2 and 3), a combination of two levels (level 1 and level 2 or level 2 and level 3) or only one level (either level 1, or level 2 or level 3). In addition to demoing the game, we will present the method-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ology of the game project (Zingano Kuhn et al. 2021), which is organised in three stages, namely data preparation (stage 1), game design (stage 2), and machine learning preparation (stage 3). In the third stage, the plan is to use the sentences labelled by the players as a dataset to first train a binary machine learning model that will be able to automatically classify sentences as appropriate or inappropriate, and then to train a multi-class classifier that would be able to perform fine-grained labelling of the inappropriate sentences using the same categories as used in the game.

References

Baisa, V./Suchomel, V. (2014): SkELL: web interface for English language learning. In: Horák, A./Rychlý, P. (eds.): Proceedings of the Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014. Brno, pp. 63–70.

Braun, S. (2005): From pedagogically relevant corpora to authentic language learning contents. In: ReCALL 17, pp. 47–64.

Kilgarriff, A./Husák, M./McAdam, K./Rundell, M./Rychlý, P. (2008): GDEX: automatically finding good dictionary examples in a corpus. In: Bernal, E./DeCesaris, J. (eds.): Proceedings of the XIII EURALEX International Congress. Barcelona, pp. 425–432.

Zingano Kuhn, T./Todorović, B. Š./Arhar Holdt, Š./Zviel-Girshin, R./Koppel, K./Luís, A. R./Kosem, I. (2021): Crowdsourcing pedagogical corpora for lexicographical purposes. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of the EURALEX XIX Congress. Volume 2. Alexandropoulos, pp. 771–779.

Contact information

Tanara Zingano Kuhn
Centre for the Studies of General and Applied Linguistics (CELGA-ILTEC),
University of Coimbra
tanarazingano@outlook.com

Špela Arhar Holdt
Centre for Language Resources and Technologies,
University of Ljubljana
spela.arharholdt@ff.uni-lj.si

Rina Zviel-Girshin
Ruppin Academic Center
rinazg@ruppin.ac.il

Ana R. Luís
Centre for the Studies of General and Applied Linguistics (CELGA-ILTEC),
University of Coimbra
aluis@fl.uc.pt

Carole Tiberius
Dutch Language Institute
carole.tiberius@ivdnt.org

Kristina Koppel
Institute of the Estonian Language
kristina.koppel@eki.ee

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Branislava Šandrih Todorović
University of Belgrade
branislava.sandrih@fil.bg.ac.rs

Iztok Kosem
Centre for Language Resources and Technologies, University of Ljubljana
iztok.kosem@ff.uni-lj.si

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.