EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Proceedings Book
Volume 2

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-2-2

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2021 Edition

Crowdsourcing Pedagogical Corpora for Lexicographical Purposes

Zingano Kuhn T., Todorović B.Š., Holdt Š.A., Zviel-Girshin R., Koppel K., Luís A.R., Kosem I.

CELGA-ILTEC/University of Coimbra, University of Belgrade/Faculty of Philology, Centre for Language Resources and Technologies/University of Ljubljana, Ruppin Academic Centre, Institute of the Estonian Language, CELGA-ILTEC/University of Coimbra, Centre for Language Resources and Technologies, University of Ljubljana/
Jožef Stefan Institute
tanarazingano@outlook.com, branislava.sandrih@fil.bg.ac.rs, spela.ArharHoldt@ff.uni-lj.si, rinazg@gmail.com, kristina.koppel@eki.ee, aluis@fl.uc.pt, iztok.kosem@cjvt.si

Corpora are valuable sources for the development of language learning materials (e.g., books, grammars, dictionaries, exercises), because they contain language as produced in natural contexts. Even though corpora are getting larger, mainly due to crawling data from the web, their pedagogical use remains rather challenging. Not all texts are appropriate for language learning or teaching purposes as they can potentially contain sensitive or offensive content, in addition to exhibit structural problems, errors, among other problems. Corpus cleaning for pedagogical purposes is however a very time-consuming task if done manually. In this paper we present a new and more effective method for creating problem-labelled pedagogical corpora for a group of languages, namely Portuguese, Serbian, Slovene, Dutch and Estonian, by means of crowdsourcing. First, we report on an experiment aimed at verifying the adequacy of crowdsourcing as a technique for corpus labelling. We then outline the lessons learned and discuss how these have led us to explore an alternative way of compiling pedagogical corpora through gamification.

Keywords: corpus creation; good example sentences; pedagogical corpora; crowdsourcing

1 Introduction

Corpora have been widely used for the development of language learning material, including learners' dictionaries, and other pedagogical resources. This is no surprise, since corpora show how language is authentically used in everyday life and thus provide valuable information for the learners' own language development. Römer (2009), Boulton (2017), and Vyatkina and Boulton (2017), to name but a few, have pointed out an impressive number of publications on corpus use for pedagogical purposes. Corpora that are built specifically for language learning purposes are usually called pedagogical corpora: "The pedagogical corpus, as its name suggests, is primarily intended to serve as a resource for teaching rather than research, although many can serve both functions" (Chambers 2016: 364). The wide variety of applications of pedagogical corpora clearly demonstrates their usefulness for language learning and teaching. According to Römer (2009) (who goes back to the distinction proposed by Leech in 1997), indirect application of pedagogical corpora refers to the work carried out by researchers and didactic material developers, while direct application involves practical activities with the corpus by learners and teachers. One of the main characteristics of a pedagogical corpus regards its design process. This is because, as Braun points out, the 'pedagogic mediation of corpora' is necessary since the structure of many existing corpora, designed with linguistic research goals in mind, conflicts with the pedagogic requirements for corpus design and use (Braun 2005). One form of pedagogic mediation of corpora is through the close monitoring of the content of the corpus to identify possible structural (grammar and spelling) problems as well as sensitive/offensive content. Although preserving the original material in the corpus can be especially useful from the point of view of authenticity, the potentially problematic examples need to be presented with some guidance from the teachers. One way to facilitate the creation of language learning materials in general, and more specifically, of lexicographical resources, as well as enhance the use of corpora in the classrooms, is by marking the potentially problematic examples in pedagogical corpora. This way, teachers, material developers, lexicographers, among others, can choose to filter out certain examples according to their needs and context of use.

The main objective of our project is to create such labelled pedagogical corpora and use them for different purposes, among which are SkELLs (Sketch Engine for Language Learning)[1] for Dutch, Estonian,[2] Serbian, Slovene, and Portuguese. SkELL (Baisa & Suchomel 2014) is a free corpus tool with a pedagogical corpus which offers selected Sketch Engine functionalities (word sketch, examples, and thesaurus). Some tailored and more learner-friendly settings[3]

[1] https://skell.sketchengine.eu

[2] One can already use SkELL for Estonian. SkELL queries from the Estonian Corpus for Learners 2020 which was built using GDEX (Kilgarriff et al. 2008). One of the classifiers of GDEX for Estonian was a blacklist of words (incl. vulgarisms, swear words, potentially sensitive/offensive words), which were all omitted in the corpus building process (Koppel 2020).

[3] The interface of the tool is easy to navigate and clear, i.e., free of complex, rarely usable features. The search functionalities are as

also make SkELL language learning suited, thus serving as a complement to learners' dictionaries. As SkELL is fully automatically created, the included corpus must already be processed for aforementioned potential problems. State-of-the-art approaches to automated corpus filtering typically include the removal of structural noise and preselected problematic lexica (i.e., with the use of blacklists, as proposed by Kilgarriff et al. 2014). However, to reach a satisfactory quality, additional and more sophisticated approaches are needed, together with a clear understanding of what types of problems need to be addressed.

In order to make the corpus labelling process more efficient, as well as gather empirical data on the types of language examples the wider community perceives as problematic for teaching purposes, we have proposed and evaluated a method that applies crowdsourcing techniques. The goals of this paper are threefold: we report on an experiment that was performed, outline the lessons learned, and discuss how these have led us to propose an alternative way to compile pedagogical corpora with the help of the crowd.

This paper is organized as follows: Section 2 provides a brief review on the main methods applied to corpus cleaning and shows that, for our purposes, corpus labelling, rather than corpus cleaning, is required. Section 3 introduces the previous experiment that has been carried out to verify if crowdsourcing can be an adequate technique for corpus labelling, discusses the results of this experiment and presents some of the lessons learned. In Section 4, an alternative way of crowdsourcing corpus labelling, namely, through a game, is proposed. Section 5 wraps up this paper by pointing out some of the most significant challenges that have been faced so far and outlining what the next steps are.

2 From Corpus Cleaning to Corpus Labelling

Most of the literature about corpus cleaning refers to cleaning data from unnecessary information. For example, crawling data from the web implies extracting unnecessary tags, structural elements, meta-information, comments, links, commands and scripts (Spousta, Marek & Pecina 2008; Graën, Batinić & Volk 2014) or removing non-human-generated and quoted text (Styler 2011, Suchomel 2020). Many new approaches to web page cleaning were encouraged by the CLEANEVAL 2007 contest organized by ACL Web as Corpus interest group. Competitors used heuristic rules as well as different machine learning methods, including Support Vector Machines (Bauer & Knill 2007), decision trees, genetic algorithms, and language models (Hofmann & Weerkamp 2007). Another understanding of what corpus cleaning entails is cleaning the noise in the form of typos from large corpora. Reynaert (2006) talks about corpus induced corpus clean-up and presents a multilingual, language-independent and context-sensitive spelling checking and correction system, where the lexicon employed by the system is not a 'trusted' dictionary but contains noise in the form of recurrent typos found in any word type list derived from a large corpus of texts.

With regards to the compilation of pedagogical corpora specifically, many of them are carried out by linguistics institutes and university departments, often involving entire teams of linguistics experts. In this context, the main approach for creating a 'clean' corpus is to take an existing (web) corpus and select all sentences within a certain language-dependent range of words that are considered inappropriate for language learners (or a certain socio-cultural group). These are mostly swear words and sensitive words regarding politics, religion, sexuality, crime, illness, death, among others (Efthymiou, Gavriilidou & Papadopoulou 2014; Allan 2019). The goal is to exclude sentences containing these words from the (new) corpus. The easiest way of doing this is using a blacklist with 'sensitive/offensive' keywords (see Koppel et al. 2019), which is the method used for the creation of SkELL corpora, but this method has the disadvantage that either too few or too many sentences are excluded. In the first case, a relatively large new corpus still contains data that might be inappropriate for language learners in autonomous learning contexts. In the second case, a relatively small (and maybe too small) corpus is in itself appropriate but does not show the learner the subtleties of the language, since by eliminating sentences containing these words altogether, the corpus lacks representation of the neutral use of such words.[4] It is true that a careful selection of textual sources in the first place could help avoid some of the above-mentioned problems. If more reliable publishers, and possibly texts, are selected for compilation of pedagogical corpora, spelling/grammar problems might be filtered out. However, this also means that more time needs to be spent, while the variety of textual types and genres will inevitably be reduced as well. As is known, one of the advantages of using web corpora for pedagogical purposes is the wide spectrum of language use that can be found.

Although both types of corpus cleaning are unquestionably helpful, with regards to the purposes of this project, they are not sufficient, because the pedagogical corpora also need to be marked in terms of sensitive/offensive language and grammar/spelling mistakes. It should not be forgotten that in addition to the use of these corpora for publication of SkELL for Dutch, Estonian, Serbian, Slovene and Portuguese, and for dictionary making, it is expected that teachers will be able to use them for materials development. This means that all the sentences should be marked with tags for sensitivity, offensiveness, and structural mistakes, so that sentences can be filtered out according to the context of application.

For finding the balance between the two cases, manual assessment and selection seems unavoidable. Although this method should produce a pedagogical corpus with a higher level of quality, the amount of work and time that it takes for linguistics experts to create a 'clean' corpus has motivated a group of researchers within the COST action enetCollect[5] (Agerri et al. 2018; Lyding et al. 2018) to find an alternative solution by using crowdsourcing in the compilation process.

simple as possible (e.g., the search is case insensitive, it finds all parts of speech for a given word form), and the results are displayed in a readable and learning-oriented manner (e.g., instead of a list of concordances, whole sentences are displayed; the linguistic metalanguage is minimised; and special visualisations of language data are provided, such as wordclouds of similar words).

[4] Manual analysis of the blacklists created by the Sketch Engine team for automatic creation of web corpora for Dutch, Estonian, Serbian, Slovene, and Portuguese, and which would be also used for creation of SkELL corpora for those languages, has revealed that they contain many polysemous words that have both offensive and neutral senses.

[5] https://enetcollect.eurac.edu

Thus, an experiment was set up to test the viability of such a method. This experiment will be presented in the next section.

3 The Crowdsourcing Experiment

Crowdsourcing (or citizen science) is a practice where ordinary people, i.e., the crowd, contribute to creating content, solving problems, or even to doing some research. The crowd does not necessarily need to have expertise on the subject (be Čibej, Fišer & Kosem 2015; Nicolas et al. 2020). Benjamin (2015) has pointed out two characteristic features of crowdsourcing: 1) splitting the process into microtasks that can be completed with little effort, and 2) gamification where emphasis is placed on pleasure rather than effort. Crowdsourcing has been used in lexicography, e.g., for selecting keywords, formulating definitions, cross-editing the entries, providing examples, collocations, synonyms, word associations etc. (see, e.g., Čibej, Fišer & Kosem 2015; Arhar Holdt et al. 2020; Vainik 2018). It has also been used both in building the corpora (see, e.g., Ambati & Vogel 2010; Lane et al. 2010; Post, Callison-Burch & Osborne 2012) as well as in annotating the corpora (Bontcheva et al. 2014; Gut & Bayerl 2004). But as far as we know, it has not been used to mark general web corpora with potential offensive/sensitive content and structural problems so as to create annotated pedagogical corpora that can be used for dictionary making and language learning.

The main purpose of the experiment was to have the crowd help filter out offensive sentences from web corpora. Our secondary objectives were to have the crowd identify problematic sentences in corpora that, in principle, should be offensiveness/sensitivity-free, and to learn what the crowd considers to be offensive or sensitive. These specific objectives stemmed from our knowledge that automatic extraction of sentences based on blacklists fails to filter out sensitive content, that polysemous words can have neutral and offensive senses, and that offensiveness is a subjective matter.

Pybossa[6] was chosen as the crowdsourcing platform because a) it is free and b) because the custom tasks (interface) can be written in Javascript. In addition, one of the team members of the research project has a robust experience with using Pybossa in other crowdsourcing projects (Dekker & Schoonheim 2018a, 2018b) and has direct access to a local installation (INL) which ensures that the output data can be kept safely.

A multilanguage (Portuguese, Serbian, Dutch and Slovene) crowdsourcing project[7] was created with a common landing page, where the crowd was first asked to pick their language and then was transferred to the corresponding language home page. The individual languages' homepage had all the same structure and texts, which had been written together in English and later translated to each language. In addition, the Pybossa interface, i.e., buttons, messages, etc. also needed to be translated to each of the experiment languages. This presentation page contained a short introduction to the experiment, in which the purpose of the task and justification were provided and had the purpose to motivate participation by showing the participants that their contribution would benefit the community (i.e., social motivation, Čibej, Fišer & Kosem 2015). In addition, there was an invitation to participate, which contained the following: i) an example of the task that should be performed (see Figure 1); ii) a request of how many tasks we would like them to answer and the expected time that should take; iii) information about the institutions[8] promoting the experiment and about enetCollect; iv) a disclaimer informing the anonymous status of their answers, together with an example of the type of offensive sentence they could encounter and e-mail for contact; v) an informed consent to participate. Such detailed instructions are needed because of several reasons. Firstly, anticipating what kind of contribution is expected from the participants and how long that will take them may increase engagement. Secondly, showing that known academic institutions support the experiment ensures users that this is a reliable experiment. Thirdly, a clear description of how data provided by the participants will be handled and the provision of an informed consent conveys security. Finally, an example of highly offensive content allows participants to be psychologically prepared for the task and avoids dropouts.

Probably, what is more challenging for researchers creating a crowdsourcing experiment is the formulation of the right question (microtask design, Čibej, Fišer & Kosem 2015). This question needs to encourage participants to provide only the answers the task is aiming to obtain, and nothing else, but in the most straightforward and simple way possible. Figure 1 shows the model (in English) of the task example and illustrates how it was presented on each language's home page.

[6] https://pybossa.com/
[7] https://taalradar.ivdnt.org/corpusfiltering/
[8] The Centre for the Studies of General and Applied Linguistics at University of Coimbra (CELGA-ILTEC), Portugal; the Dutch Language Institute (INT) in Leiden, Netherlands; the Society for Language Resources and Technologies in Serbia (JeRTeh); and the Centre for Language Resources and Technologies, University of Ljubljana (CJVT), Slovenia.

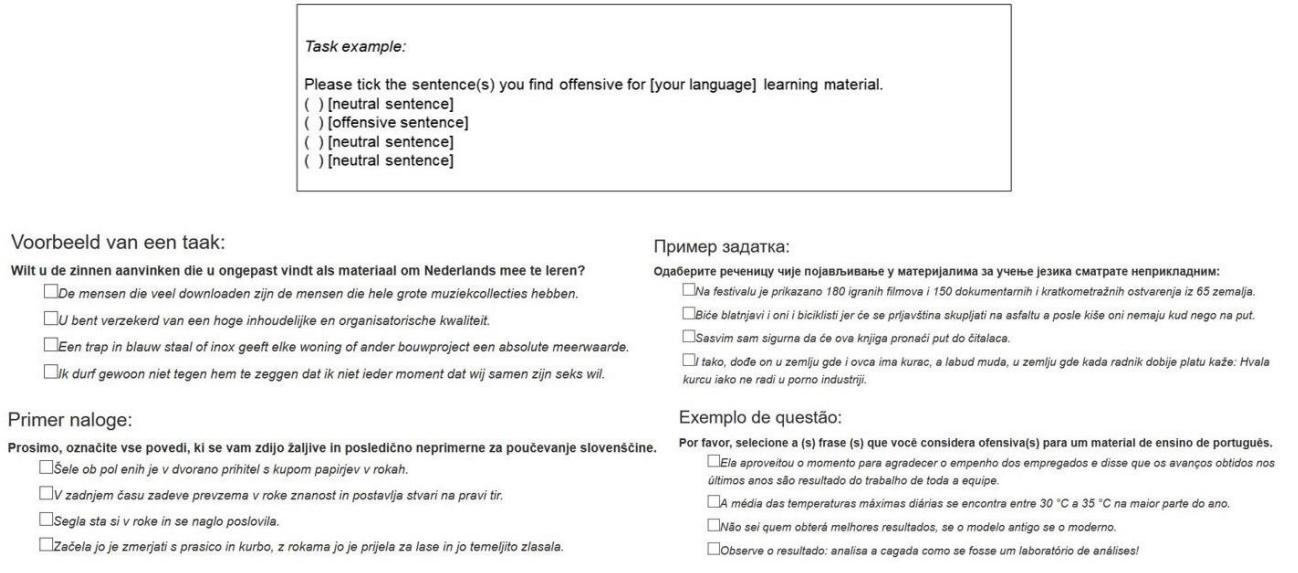

Figure 1: Task example model in English and task examples in Dutch, Serbian, Slovene and Portuguese.

The experiment was advertised via e-mail, messages, and newsletter (Dutch) to all kinds of public, from close friends and family to members of our institutions and university students, as we were not targeting language specialists only.

3.1 Methodology

The experiment followed the same methodology of data preparation and Pybossa task design for all languages. Starting with data preparation, first, a list of the 100 most frequent nouns was compiled, which was further edited according to the characteristics of each language, arriving at a list of 38 nouns (lemmas) (see Table 1). Next, sentences containing these nouns were retrieved from the correspondent corpus in Sketch Engine (see Table 1) via API. For that, two extraction processes were applied - one with the GDEX function (Kilgarriff et al. 2008) in Sketch Engine enabled and another with the GDEX function disabled - which resulted in dataset 1 and dataset 2. GDEX stands for Good Dictionary Examples and is a function in the Sketch Engine tool that, based on predefined criteria, identifies example sentences in a corpus, placing the best ones at the top of the list of concordance lines in order to facilitate the lexicographer's process of example selection. It should be mentioned that the GDEX configurations have built-in blacklists that contain malicious or offensive content. Thus, sentences in dataset 1, which were extracted via process 1, 'passed' the GDEX control and were, therefore, considered potentially good, as this functionality enabled certain structural and semantic controls. Sentences in dataset 2, which were extracted via process 2, on the other hand, had not been filtered by the GDEX function, so it was not possible to determine whether they were good or 'bad' examples. A third step was added to the data preparation: the sentences extracted with GDEX-off were further filtered by language-dependent special blacklists, created separately from those built into the Sketch Engine, which were named 'curse lists', containing only explicitly offensive or sensitive words (see Table 1). Sentences containing words or expressions from this list were then automatically annotated as potentially inappropriate and included as ground truth for further analysis (Dekker et al. 2019; Zingano Kuhn et al. 2019), comprising dataset 3.

	Dutch	Serbian	Slovene	Portuguese
lemma list	Removal of mistagged nouns, proper nouns and numerals lemmas-nl.txt	lemmas-sr.txt	100 most frequent common nouns lemmas-sl.txt	Removal of proper nouns lemmas-pt.txt
corpus	NlTenTen 2 billion[9]	srWaC[10]	Gigafida[11]	pttenten_18_fl4_50 M (50-million-word sample from PtTenTen 3.8 bi)[8]
GDEX configuration	Based on CW_minimaal SketchEngine GDEX configuration. It is a minimal configuration, which favours collocations. Replaced optimal_length (9,12) and max length 30 by a hard length limit of between 7 and 40, to match the Portuguese configuration.			Portuguese.gdex configuration available in SketchEngine
curse list	Only swear words and manually expanded with personal knowledge curselist-nl.txt	curselist-sr.txt	Internally prepared list, using words labelled as vulgar from existing Slovene dictionaries curselist-sl.txt	Only swear words, with no polysemous or cultural-related words curselist-pt.txt

Table 1: Data preparation details.[12]

Moving now to the crowdsourcing experiment on Pybossa, two sets of tasks were designed and assigned randomly via the landing page. Set of tasks A contained only sentences from dataset 2, i.e., sentences that had not been GDEX-filtered. This means no pre-assumptions could be made as to their offensiveness status. Set of tasks B contained sentences from dataset 1, i.e., that had been GDEX-filtered, so were potentially good sentences, with some sentences from dataset 3, i.e., sentences that certainly contained offensive or sensitive words. Each set of tasks contained 4,560 sentences per language. Ideally, each sentence should be judged by three different people, so 13,680 judgements were needed per set of tasks. This means around 300 contributors were necessary: 150 for set of tasks A and 150 for set of tasks B. Therefore, our calculation was that each potential participant should judge around 90 sentences. Since each task contained 4 sentences, this resulted in approximately 23 tasks per participant. We estimated that this would be an optimal number of tasks per participant that would benefit the experiment, without being too time-consuming (we estimated that it would take 10 minutes to answer 23 tasks).

3.2 Results and Lessons Learned

The Pybossa output was collected after the experiment was online for two months. The level of engagement was very low for the Serbian, Slovene and Portuguese experiments (43, 12 and 32 contributors, respectively). For Dutch, numbers were more promising (131 contributors), although still far below from the total number required to have all sentences judged by three people. Despite this, the analysis of the outcome has revealed some very interesting insights.

For each sentence, we analyzed the cases in which the crowd's input contradicted our assumptions about appropriate or inappropriate content (Dekker et al. 2019; Zingano Kuhn et al. 2019). Analyzing the crowd's responses, we noticed the following cases:

- *TP (True positives)* - sentences annotated as potentially inappropriate and considered inappropriate by the crowd majority.
- *FN (False negatives)* - sentences annotated as potentially inappropriate and considered appropriate by the crowd majority.
- *FP (False positives)* - sentences annotated as potentially appropriate and considered inappropriate by the crowd majority.
- *TN (True negatives)* - sentences annotated as potentially appropriate and considered appropriate by the crowd majority.
- *UKN (Unknown)* - the number of participants who found the sentence inappropriate was equal to the number of participants who found the sentence appropriate, and vice versa.

While true positive results and true negative results confirmed our assumptions, the false negative and false positive ones

[9] NlTenTen and PtTenTen are web corpora compiled by the Sketch Engine team as part of the TenTen family (Jakubíček et al. 2013).
[10] srWaC is a Serbian corpus made up of texts collected from the Internet. https://www.sketchengine.eu/srwac-serbian-corpus/
[11] Gigafida is the reference corpus of written Slovene language. The current version 2.0 is described in Krek et al. 2020, and available at https://viri.cjvt.si/gigafida/.
[12] All input files are given on GitHub: https://github.com/Branislava/corpuscleanup_v1

have given us an opportunity to learn what participants think. The manual analysis of the *FP* sentences (*False Positives*) from each language has revealed that these sentences were mostly sophisticated (i.e., not directly formulated) cases of misogyny or religiously-offensive content. *False Positives* were also attested with sentences spreading propagation of violence towards children or containing topics related to war and politics. Since these sentences did not explicitly exhibit blacklisted words, they were not detected by the system in the first place.

Interestingly, the participants did not consider sentences with explicitly rude content necessarily inappropriate. These cases represented our false negatives. We assumed that this was due to two reasons: 1) the crowd found sentences including obscene lexica not necessarily bad learning material, and 2) some annotators were more concentrated on the structure and language accuracy, and less on the pedagogical implications of the inclusion of such sentences in language learning materials.

Another finding originated from the feedback provided by the participants after the task. The feedback was optional, and we primarily expected reports on technical problems or similar. However, among the Slovene participants, two reported that they found the task rather purposeless due to the lack of problems in the data. This indicates that for non-web corpora, as is the Gigafida reference corpus, the material for the crowdsourcing task needs to be chosen with more emphasis on the problems: their lack had a demotivational effect on the participants.

We generally concluded that participants were willing to help, but also often inclined to interpret the task in their own way. Even though in our case the experiment was specifically focused on marking what was strictly 'offensive' to the participants, they often did more than this. For example, they marked the sentences that they found inappropriate for a learner's material, such as incomplete sentences, complex sentences, sentences containing spelling and grammar errors or even sentences containing too many foreign terms.

4 Gamifying corpus labelling

Based on the modest results and on the lessons learned from the experiment, we concluded that a new way of motivating the crowd and a more specific task were required. We then opted to follow the 'Games with a Purpose' (GWAP) (von Ahn, 2006) approach, "i.e., games that are fun to play and at the same time collect useful data for tasks that computers cannot yet perform" (Hacker & Ahn 2009: 1208). GWAPs have been often designed to annotate or clear language data for the creation of various lexical infrastructures, for example JeuxDeMots (Lafourcade 2007), Phrase Detectives (Poesio et al. 2013), Wordrobe (Venhuizen et al., 2013), ZombiLingo (Guillaume, Fort & Lefebvre 2016), Game of Words (Arhar Holdt et al. 2020, Kosem et al. 2020). Thus, at this second stage of the project, a game for web corpora labelling is under development.

The model of the game is inspired by Matchin (Hacker & Ahn 2009). The main idea of Matchin is to elicit users' preferences about images without asking them directly, but rather by asking what their opponent player would prefer. Players are rewarded when their predictions match. Taking into consideration Hacker and Ahn's claim that "asking partners in a two-player game to guess which of two options their partner will choose represents a viable mechanism for extracting user preferences and data' (Hacker & Ahn 2009: 1208), we have decided to build on the mechanism of Matchin to collect information on corpus examples. According to Hacker and Ahn (2009), their game has been extremely successful, with tens of thousands of players. It is our hope that additional game modes and a variety of gamification elements, such as scoring, players scoreboard, avatar, etc. will contribute to motivate the crowd to play our game.

The main purpose of our game is to have players identify problematic corpus sentences (choosing between two sentences offered), and then categorize the identified sentences according to the type of problematic content. According to Sabou et al. (2014), categories should be kept between 2 to 5, to avoid cognitive overload. To define the categories, manual assessment of 100 automatically extracted sentences from corpora of Slovene and Serbian has been performed, leading to the introduction of the following five categories: offensive, vulgar, sensitive content, spelling/grammar problems, lack of context/incomprehensible. A help pop-up page will be available for players to see example sentences of each category in order to help them make categorization decisions. At the moment of writing, game modes and players interface are being developed for all languages (Dutch, Serbian, Slovene, Estonian and Portuguese).

The game development project is organized into three phases: data preparation, game preparation, and preparation of machine learning for automatic corpus labelling, as can be seen in the diagram below (Figure 2). The first phase involves preparing the datasets that will feed the game. For that, the corpora of all languages will be GDEXed with especially created pedagogically-driven GDEX configurations for each language, consisting of a common set of criteria and some language-dependent criteria. For instance, one thing that has come out from the previous experiment is that sentence length (in words) is important and has to be determined per language. These sentences will be filtered by blacklists, resulting in two types of output (i.e., the potentially good and bad sentences) that will be used as the input datasets for the game. The second phase is game preparation, which involves a) the development of different game modes, b) gamification aspects, such as scoring and motivation, and c) the player interface. In addition, a researcher interface will be built to allow easy access to the database containing the labelled sentences. In the third phase, machine learning models for all languages involved will be trained to automatically identify problematic content (manually categorized by the players) in web corpora. We expect this automatic identification of problematic content will facilitate the compilation of larger, clean corpora.

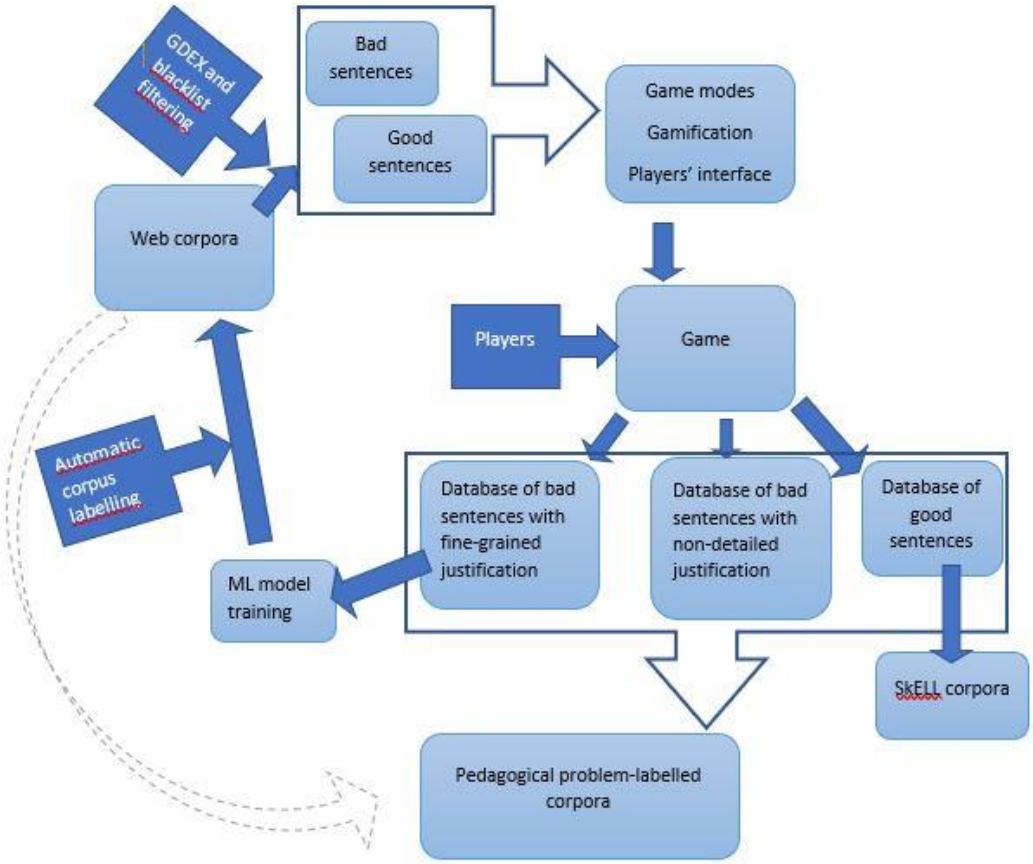

Figure 2: Game development project.

The game will be available as a webpage and as a mobile phone app. The players' answers, including the submitted labelling of sentences, will be logged and stored in a database. This way, language teachers and teaching material creators will be able to compile labelled corpora for pedagogical purposes from this output, i.e., corpora that still contain problematic sentences, but that can be used since the labels enable them to (de)select content/structure that is considered inappropriate or not (yet) suitable for the category of language learners involved. In addition, lexicographers will be able to compile filtered corpora containing only unmarked sentences, for instance, all sentences that have not been marked can be used for compiling the SkELL corpora.

5 Concluding Remarks

The progress in the field of automatic detection of good corpus examples has been considerable, and the tools have been used extensively especially in lexicography, and to a lesser extent in language pedagogy, one problem being the lack of availability of (suitable) pedagogical corpora. The approach we propose in this paper is to create pedagogical corpora from larger web corpora, using crowdsourcing. As our experiment with the labelling of corpus sentences has confirmed, crowdsourcing can be a very helpful and efficient method for these purposes. With the help of the crowd, sentences with offensive/sensitive content can be filtered out from web corpora. At the same time, the method also provides a valuable insight into what the crowd, i.e., the community, considers as (in)appropriate content.

Nonetheless, our experiment also revealed that improvements to the methodology were needed, particularly in terms of having more motivating tasks to increase the level of engagement by the participants and providing more focused questions to guarantee the input provided by the participants is relevant. The project is now exploring an alternative way of using crowdsourcing by adopting the 'Games with a Purpose' approach. In this new stage of the project, a game for web corpora labelling is under development. While the gamification approach addresses some of the issues encountered during the experiment, it brings new challenges related to game development and design, and dissemination.

If this gamification experiment turns out to be successful, it will open a new way of creating pedagogical corpora with the help of crowdsourcing. These corpora will have many different possible uses, especially in language learning, but also in other fields. For example, in lexicography, such corpora can be considered invaluable sources of good candidate examples, and on their basis, dictionary creation could become considerably faster. It is our ultimate goal to provide examples of good practice and prepare workflows that can serve as the benchmark for other languages, especially under-resourced ones.

6 References

Allan, K. (Ed.) (2019). *The Oxford Handbook of Taboo Words and Language*. Oxford University Press.

Ambati, V. & Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems?. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*.

Arhar Holdt, Š., Logar, N., Pori, E. & Kosem, I. (2020). "Game of Words": Play the Game, Clean the Database. In *Proceedings of the XIX EURALEX International Congress*.

Agerri, R., Maritxalar, M., Lyding, V., & Nicolas, L. (2018). enetCollect: A New European Network for combining Language Learning with Crowdsourcing Techniques. *Procesamiento Del Lenguaje Natural, 61*, 171-174. doi:http://dx.doi.org/10.26342/2018-61-25

Baisa, V. & Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. In A. Horák, P. Rychlý (eds) *Proceedings of the Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014*. Brno: Tribun EU, pp. 63-70.

Bauer, M.W. Knill, C. (eds.) (2007). *Management reforms in international organizations*. Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG

Behzadan, V., Aguirre, C., Bose, A. & Hsu, W. (2018). Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE.

Benjamin, M. (2015). Crowdsourcing microdata for cost-effective and reliable lexicography. In L. Li, J. Mckeown, L. Liu (eds.) *Proceedings of AsiaLex 2015, Hong Kong*. Hong Kong Polytechnic University, pp. 213-221.

Bontcheva, K., Roberts, I., Derczynski, L. & Alexander-Eames, S. (2014). The GATE crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Boulton, A. (2017). Corpora in language teaching and learning: Research timeline. *Language Teaching*, Cambridge University Press (CUP), 50 (4), pp.483-506.

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

Chambers, A. (2016). Written language corpora and pedagogic applications. In F. Farr, L. Murray (Eds.), *The Routledge Handbook of Language Learning and Technology*, pp. 362-375.

Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem, M. Jakubíček, J. Kallas, S. Krek (eds.) *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana, Brighton: Trojina, Institute for Applied Slovene Studies, Lexical Computing Ltd., pp. 70-83.

Dekker, P., & Schoonheim, T. (2018a). Crowdsourcing Language Resources for Dutch using PYBOSSA: Case Studies on Blends, Neologisms and Language Variation. In *Proceedings of the enetCollect WG3&WG5 Meeting, 24-25 October 2018*. Leiden, Netherlands.

Dekker, P., & Schoonheim, T. (2018b). When to use PYBOSSA? Case studies on crowdsourcing for Dutch. Presentation at *enetCollect WG1 hands-on workshop Gothenburg*, December 2018.

Dekker, P., Zingano Kuhn, T., Šandrih, B., Zviel-Girshin, R., Arhar Holdt, Š. & Schoonheim, T. (2019). Corpus filtering via crowdsourcing for developing a learner's dictionary. In *eLexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts*. Brno: Lexical Computing.

Efthymiou, A., Gavriilidou, Z. & Papadopoulou, E. (2014). Labeling of Derogatory Words in Modern Greek Dictionaries. In N. Lavidas, T. Alexiou & A. Sougari (Ed.), *Major Trends in Theoretical and Applied Linguistics 2*. Versita Ltd, 78 York Street, London W1H 1DP, Great Britain.: De Gruyter Open Poland, pp. 27-40.

Graën, J., Batinić, D. & Volk, M. (2014). Cleaning the Europarl corpus for linguistic applications. In *Konvens 2014, Hildesheim, 8 October 2014 - 10 October 2014*.

Guillaume, B., Fort, K. & Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *International Conference on Computational Linguistics (COLING)*.

Gut, U. & Bayerl, P.S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Speech Prosody 2004, International Conference*.

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*.

Hofmann, K. & Weerkamp, W. (2007). Web corpus cleaning using content and structure. In *Proceedings of the Web as Corpus Workshop (WAC3), Cleaneval Session*. Louvain-la-Neuve, Belgium.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference*. Lancaster, UK.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp.7-36.

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*. Barcelona, Spain: Documenta Universitaria, pp. 425-432.

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele* [Corpus-Based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners]. (Doktoritöö, Tartu Ülikool). Tartu: Tartu Ülikooli Kirjastus.

Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V. & Michelfeit, J. (2019). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. In: I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Janssen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds). *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*. Brno: Lexical Computing CZ, s.r.o., pp. 763−782.

Kosem, I., Martelli, F., Navigli, R., Jakubíček, M. & Kallas, J. (2020). *Crowdsourcing Module. Deliverable 4.3 of the*

European Lexicographic Infrastructure Project. https://elex.is/wp-content/uploads/2020/03/ELEXIS_D4_3_Crowdsourcing_module.pdf [25/07/2021]

Krek, S., Holdt, Š.A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. & Dobrovoljc, K. (2020). Gigafida 2.0: The Reference Corpus of Written Standard Slovene. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th international symposium on natural language processing*.

Lane, I., Eck, M., Rottmann, K. & Waibel, A. (2010). Tools for collecting speech corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with amazon's mechanical turk*.

Lyding, V., Nicolas, L., Bédi, B. & Fort, K. (2018). Introducing the European network for combining language learning and crowdsourcing techniques (enetcollect). *Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL, 2018*.

Nicolas, L., Lyding, V., Borg, C., Forăscu, C., Fort, K., Zdravkova, K., Kosem, I., Čibej, J., Holdt, Š.A., Millour, A. & König, A. (2020) Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L. & Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1), pp.1-44.

Post, M., Callison-Burch, C. & Osborne, M. (2012). Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

Reynaert, M. (2006). Corpus-Induced Corpus Clean-up. In *LREC* 2006.

Römer, U. (2009). Using general and specialised corpora in language teaching: Past, present and future. In M.C. Campoy, B. Belles-Fortuno, M. L. Gea-Valor (Eds.), *Corpus-Based Approaches to English Language Teaching*. Continuum Publishing Corporation, pp.18-35.

Sabou, M., Bontcheva, K., Derczynski, L. & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *LREC 2014*.

Spousta, M., Marek, M. & Pecina, P. (2008). Victor: the web-page cleaning tool. In *4th Web as Corpus Workshop (WAC4)-Can we beat Google*.

Styler, W. (2011). *The EnronSent corpus*. Boulder: University of Colorado at Boulder Institute of Cognitive Science.

Suchomel, V. (2020). Better Web Corpora for Corpus Linguistics and NLP. Doctoral theses. Masaryk University, Faculty of Informatics, Brno, Czech Republic.

Vainik, E. (2018). Compiling the Dictionary of Word Associations in Estonian: from scratch to the database. *Eesti Rakenduslingvistika Ühingu aastaraamat*, (14), pp.229-245.

Venhuizen, N., Evang, K., Basile, V. & Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6), pp.92-94.

Vyatkina, N., & Boulton, A. (2017). Corpora in language teaching and learning. *Language Learning and Technology*, *21*(3), 1-8.

Zingano Kuhn, T.Z., Dekker, P., Šandrih, B., Zviel-Girshin, R., Arhar, Š., Holdt, T.S & Schoonheim,T. (2019). Crowdsourcing Corpus Cleaning for Language Learning Resource Development. In *EuroCALL 2019: European Association of Computer Assisted Language Learning*.

Acknowledgements

This work was supported by the Portuguese national funding agency, FCT - Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020). The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene). Many ideas result from the framework of the CA160105 enetCollect COST Action. We thank all that made our work possible.