Data preparation in crowdsourcing for pedagogical purposes: the case of the CrowLL game

Tanara ZINGANO KUHN
Research Centre for General and Applied Linguistics, University of Coimbra

Špela ARHAR HOLDT
Faculty of Arts, University of Ljubljana; Faculty of Computer and Information Science, University of Ljubljana

Iztok KOSEM
Faculty of Arts, University of Ljubljana; Jožef Stefan Institute

Carole TIBERIUS
Dutch Language Institute

Kristina KOPPEL
Institute of the Estonian Language

Rina ZVIEL-GIRSHIN
Ruppin Academic Center

One way to stimulate the use of corpora in language education is by making pedagogically appropriate corpora, labeled with different types of problems (sensitive content, offensive language, structural problems). However, manually labeling corpora is extremely time-consuming and a better approach

Zingano Kuhn. T., Arhar Holdt, Š., Kosem, I., Tiberius, C., Koppel, K., Zviel-Girshin, R.: Data preparation in crowdsourcing for pedagogical purposes: the case of the CrowLL game. Slovenščina 2.0, 10(2): 62–100.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.2.62-100

should be found. We thus propose a combination of two approaches to the creation of problem-labeled pedagogical corpora of Dutch, Estonian, Slovene and Brazilian Portuguese: the use of games with a purpose and of crowdsourcing for the task. We conducted initial experiments to establish the suitability of the crowdsourcing task, and used the lessons learned to design the Crowdsourcing for Language Learning (CrowLL) game in which players identify problematic sentences, classify them, and indicate problematic excerpts. The focus of this paper is on data preparation, given the crucial role that such a stage plays in any crowdsourcing project dealing with the creation of language learning resources. We present the methodology for data preparation, offering a detailed presentation of source corpora selection, pedagogically oriented GDEX configurations, and the creation of lemma lists, with a special focus on common and language-dependent decisions. Finally, we offer a discussion of the challenges that emerged and the solutions that have been implemented so far.

Keywords: crowdsourcing, game with a purpose, example sentences, pedagogical corpus

1 Introduction

Evidence of authentic language use is fundamental for language learning. One way to access this evidence is through the use of examples from corpora, i.e., large collections of texts produced in natural contexts, saved in electronic form. However, these corpora may include sensitive content or offensive language, in addition to exhibiting structural problems. While such use is unquestionably authentic, some teachers or material developers might consider it to be inappropriate for their needs, thus finding it necessary to manually filter the corpus before applying authentic examples to pedagogical contexts, which is a laborious task.

To facilitate and stimulate the use of corpora in education we propose creating problem-labeled pedagogical corpora. This way, the process of example selection could be significantly streamlined. At the same time, instead of deleting potentially problematic content from the corpus we will label it, thus leaving the choice of the use of certain

examples dependent on the needs and contexts of use of teachers and didactic material developers. The types of problems to be labeled are: vulgar, offensive, sensitive content, grammar/spelling problems, incomprehensible/lack of context.

Creating such corpora is challenging due to at least three reasons. Firstly, the process of labeling sentences in corpora is extremely time-consuming, if done manually. Secondly, automatic labeling can also be demanding given the polysemic nature of words. Thirdly, sensitivity and offensiveness are rather subjective concepts. Our proposal is thus to use the help from the crowd to achieve this task. For that, we are currently developing CrowLL – Crowdsourcing for Language Learning,[1] a multi-mode, multi-language (Dutch, Estonian, Slovene, and Portuguese) digital game. In this game, the players will be offered two examples (automatically extracted from existing corpora) and prompted to choose one (or both, or even none) that they consider to be appropriate for language teaching purposes. They will be asked to categorize the problem(s) of the example that has not been chosen and point out the constituent parts of the sentence that they consider to be problematic. With the output obtained from the players, we will compile problem-labeled pedagogical corpora for the languages mentioned above. These corpora can be used for the development of auxiliary language learning resources, such as Sketch Engine for Language Learning – SKELL (Baisa and Suchomel, 2014),[2] dictionaries and teaching materials; and, within Natural Language Processing, for the creation of datasets aimed at training machine learning algorithms for the compilation of larger pedagogical corpora.

Data preparation plays a crucial role in any crowdsourcing project that deals with the creation of language learning resources. Indeed, the quality and structure of the input data, together with the type of

1 The research group carrying out the Crowdsourcing Corpus Filtering for Pedagogical Purposes project, within which the Crowdsourcing for Language Learning (CrowLL) game is being developed, originated under the umbrella of the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect) COST Action (CA 16105). It is currently composed of seven members from six countries (Brazil, Estonia, Israel, Netherlands, Slovenia, and Portugal) and encompasses four languages (Dutch, Estonian, Slovene, and Portuguese). See https://ucpages.uc.pt/celga-iltec/crowll/ for further information on the project.

2 SKELL is a free language learning tool that provides automatic summaries of corpus data, namely, examples, collocations and thesaurus. Available at https://skell.sketchengine.eu (30. 8. 2022).

task, have a direct impact on the quality of the output. Consequently, our research question in this paper is: What is the methodology of data preparation that is required to attend to the needs of a crowdsourcing game dealing with identification of offensive language, sensitive content and structural problems in authentic language material? We present the steps taken, the decisions made, the challenges faced and the solutions found to create the methodology for preparing a dataset of 10,000 sentences per language to develop and internally test the CrowLL game. For that, we use three key elements: source corpora, from where the sentences to be labeled by the players will be extracted; Good Dictionary Examples – GDEX (Kilgarriff et al., 2008) configurations, which automatically identify more pedagogically-suited examples in the source corpora and assign scores to the sentences; and lemma lists, which define the sentences to be extracted from the corpora. After the game is developed and tested with real users, the methodology of data preparation itself can also be evaluated.

The paper builds on our previous work within the enetCollect COST action.[3] We have previously established the motivation for a gamified approach to the labeling of examples in pedagogical corpora. We have developed the idea, formulated research questions, conducted initial tests with the crowd to establish the suitability of the crowdsourcing task, and used the lessons learned to design both the game flow and a work plan for the implementation. We have presented different stages of this work at conferences, as available in Kuhn et al. (2021) and Zviel-Girshin et al. (2021). In this paper, we focus on the newest development, namely on the first stage of the game preparation that primarily addresses issues related to the (corpus) data needed for the game. While the paper builds upon our previous work, it also presents a new, summative view and describes various applicative methodological decisions that were tested on different languages to ensure further usability of our proposed model, both by other languages and for purposes other than the CrowLL game development.

This paper is structured as follows. Section 2 reviews different approaches to the identification of good examples for the creation of pedagogical corpora. Section 3 introduces crowdsourcing and gamification,

3 https://www.cost.eu/actions/CA16105/ (28. 10. 2022)

specifically within the context of language learning. Section 4 presents the CrowLL game, firstly reporting on our previous crowdsourcing experiment, whose results have led to the adoption of the Games with a Purpose (von Ahn, 2006) approach. Section 5 describes the methodology for data preparation in detail, and Section 6 analyzes and discusses the results.

2 Pedagogical corpora and language examples

Text corpora are collections of authentic (written or spoken) texts in electronic form, sampled to represent a specific type of language use (e.g. Gries, 2009; Sinclair, 2005). Corpus texts are typically equipped with metadata and linguistic information on different levels, increasing their value for different purposes in applied linguistics, natural language processing, and other fields that benefit from analyzing language data. In this paper, we focus on the field of language education, where the importance and value of corpora have been firmly established (Boulton, 2017; Callies, 2019; Römer, 2009; Vyatkina and Boulton, 2017). Corpora can be used by researchers and teachers for the creation of teaching and testing materials, language resources (such as learners' dictionaries), or directly by students, as classroom work with authentic language facilitates bottom-up language learning (Osborne, 2002).

It has been established (e.g. Callies, 2019) that direct use of corpora for teaching purposes is still not very widespread for a series of reasons, among which is skepticism about the quality and appropriateness of the data, especially because corpora are usually compiled for carrying out research, not for language teaching. Attempts to address this problem and promote the use of corpora for teaching have led to the emergence of specialized *pedagogical corpora*, i.e., corpora prepared specifically for language learning purposes (Chambers, 2016, p. 364). One of the main characteristics of a pedagogical corpus is the need for "pedagogic mediation" (Braun, 2005), which takes into consideration a set of factors related to the learners and the learning context. For purposes of good example selection, for instance, we argue that one type of monitoring could focus on identification of possible

structural (grammar and spelling) problems as well as sensitive/offensive content, which might be problematic when presented to learners without the mediation of the teacher.

The creation of pedagogical corpora is a costly and time-consuming endeavor; however, the process can be supported by the automatization of certain procedures. One possible approach is to clean elements considered to be problematic for pedagogical purposes from existing corpora, such as offensive words and structural errors (misspellings, grammar errors).

In reference to the former, one area that has invested extensively in the identification of offensive language is natural language processing (NLP), mainly with research on the automatic detection of hate speech, with the aim to contribute to monitoring abusive behavior on the internet (e.g., social media, comments on media channels). Some examples of efforts on this topic are specific evaluation tasks at SemEval (International Workshop on Semantic Evaluation),[4] such as OffensEval[5] (Zampieri et al., 2019; Zampieri et al., 2020), and the Workshop on Online Abuse and Harms (WOAH),[6] currently in its 6th edition (2022). An impressive amount of research on the subject has been carried out in NLP, as can be seen, for example, in Poletto et al. (2020). This survey presents an up-to-date, systematic review of the available resources on hate speech, with detailed analysis, some of the current weaknesses, and goals for improvement. According to the authors, it is a complement to previous surveys, in particular, Lucas (2014), Wiegand and Schmidt (2017), and Fontana and Nunes (2018) (Poletto et al., 2020, p. 479). Datasets, such as the ones available on the dedicated webpage Hate Speech Dataset Catalogue (Vidgen and Derczynski, 2020),[7] and lexica, such as HurtLex (Bassignana et al., 2018), are some of the resources developed in NLP that could be used as a source of keywords for corpus cleaning. This approach consists of using blacklists containing swear words, vulgarisms, and words related to sensitive content in order to remove from the corpus sentences where these words occur (see below for a combined use of blacklists and GDEX). That means

4 https://semeval.github.io/ (28. 10. 2022)
5 https://sites.google.com/site/offensevalsharedtask/home (28. 10. 2022)
6 https://www.workshopononlineabuse.com (28. 10. 2022)
7 https://hatespeechdata.com/ (28. 10. 2022)

the "clean" corpus would not contain any sentences with those words. Another contribution from NLP to corpus cleaning would be through the application of offensive identification models at the sentence level, thus eliminating from the source corpus sentences automatically identified as offensive. However, one of the challenges in computational approaches to this subject is that other aspects, above and beyond the linguistic surface, have a crucial influence in the determination of what offensiveness is. Schmidt and Wiegand (2017) present a few works that seek to incorporate context to hate speech detection, but acknowledge that in certain difficult cases the method fails, so more investigation is needed. Relatedly, Poletto et al. (2020) point out a shortcoming of not considering the pragmatic aspects of swearing when evaluating hate speech – the production of false positives.

Whatever perspective is adopted with regard to identifying offensiveness, either at a word or sentence level, we have argued (Kuhn et al., 2021) that the total elimination of sentences from the corpus should be avoided because: 1. very few words are problematic in all of their senses and contexts, and 2. teachers and didactic material developers should be free to use whatever examples they find useful for their various needs. We thus propose to label potentially problematic data in pedagogical corpora instead of removing it.

For structural errors, automatic error detection (following different methods), has been widely adopted. For instance, Reynaert (2006) adopts a corpus-induced corpus clean-up approach to detect typos in texts. Rather than dictionaries, the lexicon used in the clean-up process consists of typos found in large corpora. However, Xu and Chamberlain (2020) have shown that some problems identified as structural errors by automatic error detection methods might not be actual mistakes, but rather spelling and grammatical variations based on the context of use. They argue that humans are still required to perform the clean-up task, and thus developed a game (Cipher) in which players are asked to identify different types of errors in texts and annotate them.

A more lexically-oriented approach to the compilation of pedagogical corpora refers to the adoption of sophisticated methods that automatically analyze texts according to several criteria to identify good examples. These good examples can then be gathered in a pedagogical

corpus. The current state-of-the-art in corpus linguistics is Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008), available as a feature in the Sketch Engine (Kilgarriff et al., 2004, 2014) corpus query system. The general idea of GDEX is to provide a list of suitable, good-quality candidate corpus sentences that lexicographers can directly add into the dictionary as illustrative examples. At the heart of GDEX is a rule-based formula that assigns a numerical score to each corpus sentence based on how well it meets the pre-defined criteria. The criteria can determine, for instance, the length of the sentence, the number of words in the sentence, the frequency of word forms or lemmas in the corpus, the presence or absence of certain elements in the sentence, and so on. The scoring formula (with additional parameters) constitutes a so-called GDEX configuration. There are two groups of classifiers used in the configuration: hard and soft. Hard classifiers include a very high penalty giving sentences a very low score, resulting in pushing them to the bottom of the candidate list. Soft classifiers either penalize sentences or award bonus points, helping to rank good dictionary example candidates. As a result, GDEX lists all example candidates in descending order and can also be used to filter out all sentences below a certain threshold (Kosem et al., 2019).

A GDEX-based methodology has already been used to create pedagogical SKELL (Sketch Engine for Language Learning) corpora for Russian, Estonian (Koppel, Kallas, et al., 2019), English, German, Italian and Czech. This entails filtering a source corpus with a GDEX configuration, leaving only the sentences that meet all the criteria of good dictionary examples and removing the rest. But creating corpora by eliminating data brings out the shortcomings we mention earlier in this paper. The English noun *ass*, for example, can refer either to a body part, a donkey or a stupid/annoying person.[8] Since in some instances it may be considered problematic, it might be added to the blacklist. In that case, all sentences containing the word *ass* are removed from the corpus regardless of the word's meaning. This is not ideal for either lexicographers, who want to illustrate all the meanings of a word in a dictionary, or teachers, who should be given the choice to decide what they want to use for teaching, considering the

8 https://www.macmillandictionary.com/dictionary/british/ass (30. 8. 2022).

students' characteristics, such as level, age, and background and relevance to the course topic.

Building on GDEX, Stanković et al. (2019) adopted machine learning to identify good candidate examples for Serbian. First, they analyzed lexical and syntactic features in a corpus compiled of illustrative examples from the five digitized volumes of the Serbian Academy of Sciences and Arts (SASA) dictionary. They then identified 14 features relevant for the task (character-based, token-based and syntactic features) and prepared a gold dataset of good examples. Sentences from the prepared dataset, represented as feature-vectors, were used for a supervised machine learning model, which was then used in a GDEX classifier for contemporary Serbian sentences. A decision-tree classifier trained on the data predicted whether a certain corpus sentence is a good candidate for an illustrative example for the given dictionary headword or not, with an accuracy of 83% for both positive and negative samples (Šandrih, 2020).

Another tool to automatically identify good examples based on a series of criteria and using both rule-based and machine-learning approaches is HitEx. The combined approach was designed to assess the readability and suitability of (initially coursebook) material for teaching Swedish as L2 (Pilán et al., 2013, 2014; Pilán et al., 2016). For this task, 61 features of different types were used: length-based (e.g. number of tokens and characters), lexical (e.g. CEFR[9]-annotated wordlists), morphological (e.g. part-of-speech), syntactic (dependency relation tags), and semantic features (e.g. number of senses of a specific word). Candidate sentences were first ranked according to these features, and the 100 highest-ranked sentences were given to the machine-learning model for classification. The sentences were classified according to their proposed suitability for students at a certain CEFR level, and returned in the order of their heuristic ranking. Using the complete feature set at the document level, the tool obtained 81% accuracy, however, the classification accuracy for sentences was only 63.4%, presumably because the amount of context was too limited for the features to capture differences between the sentences.

9 Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press (2001).

Taken together, it can be concluded that the creation of pedagogical corpora can be challenging in at least two ways: 1. manually monitoring large amounts of texts is extremely time-consuming, and consequently, expensive; and 2. automatization of processes to support compilation has limitations due to the very nature of language. As mentioned above, one of the main shortcomings of rule-based approaches to automatic corpus cleaning, such as the method used for the development of SKELL corpora, lies in the fact that many of the words in the blacklists used as a reference to exclude sentences from a corpus are polysemic. Moreover, the automatic identification of structural problems does not take into consideration language variation. Finally, the NLP field has acknowledged that further investigation and development are needed in order to include contextual aspects to automatic offensiveness identification, with current methods still falling short.

As a result, human verification of sentences is required. More importantly, from our perspective, pedagogical corpora should be *labeled* for potentially problematic content rather than *cleaned* from it. In order to streamline the verification of the sentences for the creation of problem-labeled pedagogical corpora, we have decided to ask the crowd for help. It was in this context that the Crowdsourcing Corpus Filtering for Pedagogical Purposes project was created.

3 Crowdsourcing and gamification

Crowdsourcing is a technique for gathering data or performing large-scale tasks which is often based on the framework of collective intelligence (Lévy, 1997). Concepts related to crowdsourcing include co-creation, open innovation, and user innovation (Chesbrough, 2006; Prahalad and Ramaswamy, 2000; Von Hippel and Katz, 2003). The benefits of crowdsourcing have been thoroughly established (Aitamurto et al., 2011; Buecheler et al., 2010; Lew, 2014; Morschheuser et al., 2017; von Ahn and Dabbish, 2008), and success stories can be found in various fields, from astronomy (e.g. Zooniverse; Simpson et al., 2014) to business. Language-related use of crowdsourcing is found in NLP (e.g. for tasks such as named entity recognition and entity linking), but

also in fields such as lexicography (e.g. Arhar Holdt et al., 2018; Kosem et al., 2018) and more recently in language learning.

The role of crowdsourcing and its potential in language education has been investigated by enetCollect (the European Network for Combining Language Learning and Crowdsourcing Techniques), a large European network project funded as a COST action. The action addressed the pan-European challenge of fostering the language skills of all citizens regardless of their social, educational, and linguistic backgrounds. Its focus was on exploring the possibilities of how to use crowdsourcing to enhance the production of learning materials to cope with both the increase in demand for learning a second language (for migration, business, and tourism purposes), and the demand for more accessible materials in the many languages that are of interest to learners.

As the enetCollect research has confirmed, combining crowdsourcing and language learning is not a new undertaking, and it is possible to merge them to mass-produce language resources for any language in which a crowd of language learners can be involved (Arhar Holdt et al., 2021; Bédi et al., 2019; Lyding et al., 2018; Nicolas et al., 2020). Several language learning portals based on crowdsourcing have gathered huge multilingual audiences. Although this paper is not the platform for a detailed presentation of any of these portals, we offer some data to provide an insight into the scale of the crowd they were able to reach between 2017–2018 (Gorovaia, 2018). Rosetta Stone, the oldest of the portals and founded in 1992, attracted 75,720,000 users. Babbel, which opened in 2007, gathered 20,000,000 users. Mango Languages, launched in 2007, attracted 300,000 users. LiveMocha, which began in 2007, had 12,000,000 users in 2016. Busuu, which started in 2008, reached an audience of 70,000,000, while Duolingo, launched in 2011, had 300,000,000. Duolingo is notable for having built one of the world's most popular language-learning apps while hiring only a handful of language experts. Each day, it provides millions of sentence examples and exercises to users, almost all of them created by its 300 million or so volunteers. All of these portals are educational business entities, which confirms that educational businesses are able to attract users. The content they provide may facilitate and improve teaching, and

crowdsourcing may be used to help to create resources for additional educational areas or new languages.

An important aspect of crowdsourcing is crowdsourcer motivation, i.e. finding the best method for a specific crowdsourcing task that will attract enough people and ensure their participation until the end of the task. Lew (2014) states there are three types of motivation: psychological, social, and economic. Psychological motivation is driven by the expectation that participants will find the task psychologically satisfying or personally fulfilling. Social motivation relies on the desire of individuals to interact with others who share similar interests, contribute to the community, or improve a certain skill. Economic motivation involves financial benefits for the participants who can, for example, receive micropayments for successfully completed tasks (see Rumshisky, 2011).

A method that relies heavily on the psychological motivation of the participants, and aims to make completing the task pleasurable, is a game with a purpose (GWAP). GWAPs are "games that are fun to play and at the same time collect useful data for tasks that computers cannot yet perform" (Hacker & von Ahn, 2009, p. 1208). They have been increasingly used to crowdsource data to create lexical infrastructures of different types, and examples of GWAP include Dodiom (Eryiğit et al., 2022), Jeux de Mots (Lafourcade, 2007), Phrase Detective (Chamberlain et al., 2008), ZombiLingo (Guillaume et al., 2016), Jinx (Seemakurty et al., 2010), Game of Words (Arhar Holdt et al., 2020), and Cipher (Xu and Chamberlain, 2020).

In sum, when applied in the right circumstances, to the right crowd, and using a method and motivation best suited for a specific task, crowdsourcing can deliver very useful outcomes. It is, however, important to note that successful completion of a crowdsourcing task also requires a careful analysis of the related goals, the problem-solving environment, the expertise required, complementary activities and capabilities, and the competitive environment (Aitamurto et al., 2011; Morschheuser et al., 2017; Pe-Than et al., 2015).[10]

10 There is evidence that crowdsourcing tasks are sometimes not well-defined, or are given to the "wrong" unskilled/untrained crowd that cannot complete the task.

4 The crowdsourcing for language learning game – CrowLL

4.1 Background

In 2019 we carried out an experiment on the use of crowdsourcing for corpus filtering in which we asked the crowd to identify offensive sentences for pedagogical purposes (Kuhn et al., 2021). The sentences to be judged were automatically extracted from corpora of Brazilian Portuguese, Dutch, Serbian, and Slovene, and the participants were from Brazil, Netherlands, Serbia, and Slovenia, respectively. This study has revealed that the crowd considered to be offensive sentences which, although not directly formulated as such, expressed misogyny, religiously-offensive content, violence towards children, or contained topics related to war and politics. The study has also shown that sentences with explicitly rude content were not necessarily considered to be inappropriate.

These revealing results support our understanding that offensiveness and sensitivity are subjective and that their expression through language involves mechanisms that go beyond the explicit use of swear words. The findings of the experiment have also indicated that crowdsourcing seems to be an adequate technique to deal with such a contentious topic. Nevertheless, the traditional approach used in the experiment, namely, via the Pybossa crowdsourcing platform,[11] was considered to be rather unappealing by the participants, and thus we decided to experiment with the Games with a Purpose approach. This has also been adopted to address a similar topic by High School Super Hero (Bonetti and Tonelli, 2020, 2021), a game currently under development that focuses on the linguistic annotation of abusive language to collect data for hate speech detection. However, while GWAPs have been used for various purposes in different fields (cf. section 3), the use of games to monitor offensiveness and sensitive content in authentic examples is still in its infancy.

One additional point should be made. Given that some participants in our experiment considered sentences with structural problems inappropriate for language learning, we decided to include this type of problem in the game, in addition to offensiveness and sensitive content.

11 https://pybossa.com (30. 8. 2022)

4.2 CrowLL

The Crowdsourcing for Language Learning (CrowLL) game is under development for Brazilian Portuguese,[12] Dutch, Estonian, and Slovene. The idea for CrowLL was originally inspired by the Matchin game (Hacker and von Ahn, 2009). In this, two players compete with each other to guess which of the two pictures that are shown to them their opponent will choose. If their predictions match, they score points. According to Hacker and von Ahn (2009), this game mechanism can be used to elicit user preferences. Harris (2014) has also shown that asking about the partner's opinion leads to better results with regard to both parties giving the same answers than when the players make decisions based on their own opinions. Given that our interest in the game is to find out what examples players consider to be offensive, have sensitive content or have structural problems, this in fact includes asking players to make judgements that can vary from one person to another. Thus, the selection of a game mechanism that elicits the users' opinions and preferences seems to be a viable solution.

Nevertheless, we have also opted to offer a single-player mode. Although with this mode, the game might not benefit from the advantages put forth by the dual-player mode, the organizational factors have led us to opt to start with the development of the solo mode. Namely, the computational implementation of the solo mode requires less time and is, consequently, less expensive.

In terms of the type of crowdsourced work, Morschheuser et al. (2017) propose a categorization of crowdsourcing types based on the framework presented by Geiger and Schader (2014). Based on this, we consider CrowLL as a crowdrating game, given that "crowdrating systems commonly seek to harness the so-called *wisdom of crowds* (Surowiecki, 2005) to perform collective assessments or predictions. In this case, the emergent value arises from a huge number of homogeneous 'votes'" (Morschheuser et al., 2017, p. 27).

With CrowLL, the definition of whether a sentence is problematic or not, to which category of problem it belongs, and what constituent part of the sentence is problematic will emerge from the majority consensus.

12 European Portuguese will be included later.

CrowLL will be a collaborative game with three levels. In level 1 (I'm curious!), players identify appropriate sentences for language teaching (Figure 1). In level 2 (I'm eager to help!), they categorize the sentences that have not been chosen (i.e., considered to be inappropriate), ranging from grammar/spelling problems to issues of offensiveness and sensitivity (Figure 2). In level 3 (I'm feeling enthusiastic!), players mark in the sentence what they consider to be problematic. Players can choose to play the full game cycle (all levels), a combination of two levels, or only one level.

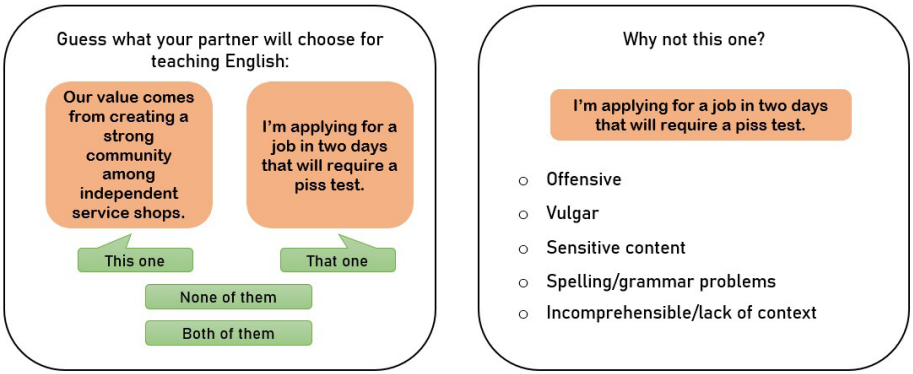

Figure 1: Levels 1 and 2 of CrowLL.

Initially, the dual-player mode should involve two human players. However, 'the cold-start problem', i.e., the lack of an opponent to start a game (Dulačka et al., 2012 as cited in Pe-than et al., 2015) has made us think of alternatives. Indeed, it can be a challenge to find a playing partner at any given time, especially in the case of small language communities such as some of those for which this game is being developed. Therefore, we propose two solutions, a synchronous and an asynchronous mode. In the synchronous mode, players will play against bots with pre-recorded answers. Players are rewarded when their predictions match the pre-recorded answers. With the asynchronous approach, we will offer delay mechanics (Pe-Than et al., 2015). Here, players will choose packages containing sentences previously judged by others, and players will be rewarded once their answers are confirmed by others at a later time. Depending on whether the game is

played in single- or dual-player mode, some of the questions will have to be changed and the scoring will also be different.

We have several ideas with regard to incentive through scoring mechanisms, ranging from offering an individual score that stems from consecutive work, to keeping a record of a cooperative score that shows the agreement of the player in teams/partnerships (so-called normalization motivation, according to Preist et al., 2014), including displaying scoreboards of the player's country's ranking position in comparison to the other countries (Olympic Games style). In this way, the game can be competitive on an individual level, while at the same time cooperative on the team level.

5 Methodology of data preparation

In order to start testing the game so that adjustments and development can be made before the official public release, we have decided to create an initial dataset of 10,000 sentences per language. The data extraction procedure involves – from each of the source corpora – the use of GDEX and a lemma list to extract the sentences. However, before proceeding with the extraction, a series of actions are required:

1. Definition of the source corpora from which sentences will be extracted;
2. Provision of pedagogically oriented GDEX configurations;
3. Creation of lemma lists to extract sentences from the corpora.

Next, we will explain each action in more detail.

5.1 Source corpora

One of the crucial guidelines for choosing our source corpora was that they were at least in some part openly available. This way, the resulting labeled datasets can be shared with and used by others. This decision aims at contributing to overcome one of the main problems in the area of language resource development, namely the lack of open-source data for many languages, as noted, for example, by Vajjala (2022) with regard to research on automatic readability assessment.

For Dutch and Brazilian Portuguese, we use the respective corpora of the Timestamped JSI web corpus, which is a family of web corpora created from IJS newsfeed by the Jozef Stefan Institute, in Slovenia, for 18 languages (Trampuš and Novak, 2012). Corpora in this family comprise news articles continuously crawled from RSS feeds. Both corpora are available in Sketch Engine. The Dutch corpus covers texts originating from the Netherlands and Belgium from 2014 to 2021. The whole corpus, totaling approximately 1.3 billion words, will be used. The Portuguese corpus covers texts from 2014 to 2021, published online in different countries, totaling over 4.5 billion words. As we are first developing CrowLL for Brazilian Portuguese, we only used texts marked with Brazil as a source country, thus making a subcorpus of 3,202,820,993 words.

For Estonian we use the Estonian National Corpus 2021 (Koppel and Kallas, 2022), which is the latest and largest corpus of written texts of modern Estonian. The texts span the period from 1990 to 2021. The most extensive part of the Estonian National Corpus 2021 is the Estonian Web Corpora, i.e. texts crawled from the web. It contains eleven sub-corpora (i.e. Web 2013, Web 2017, Web 2019, Web 2021, Feeds 2014-2021, Wikipedia 2021, Wikipedia Talk 2017, the Open Access Journals (DOAJ), Literature, Balanced Corpus, and the Reference Corpus) totaling 2.3 billion words.

For Slovene we use Gigafida 2.0 (Krek et al., 2020), the most recent version of the reference written corpus of Slovene. It contains 38,310 texts and 1,134,693,333 words. The texts span the period from 1991 to 2018, and cover newspapers, internet resources (the texts collected using the IJS Newsfeed service; Trampuš and Novak, 2012), magazines, fiction, non-fiction (such as textbooks), and various other texts. Newspaper texts represent nearly half of the corpus (47.8% of tokens), followed by internet texts (28%) and magazines (16,5%).

5.2 Pedagogically oriented GDEX configurations

In section 2, we introduced GDEX (Good Dictionary Examples) (Kilgarriff et al., 2008). While the Sketch Engine team has made general GDEX configurations for a number of languages available on their

platform, GDEX configurations can be specially devised to better fit specific purposes, depending on the objectives of the project at hand. As the objective of the CrowLL game is to have the crowd help to create problem-labeled corpora for language learning, the sentences to be presented to the crowd for labeling have to be previously prepared to fit the pedagogical purpose. In order to do this automatically, we have opted to use pedagogically oriented GDEX configurations.[13] Slovene and Estonian have adopted configurations that have been previously devised for pedagogical purposes, while Dutch and Portuguese have built on existing pedagogically oriented configurations.

The Slovene GDEX configuration was originally devised for lexicographic projects at the Centre for Language Resources and Technologies, and more specifically this includes the Slovene Lexical Database and Collocations Dictionary of Modern Slovene (Gantar et al., 2016; Kosem et al., 2011; Kosem et al., 2012; Kosem et al., 2013). The initial lexicographically oriented GDEX configuration was also used for pedagogical purposes, i.e. in the preparation of examples for exercises in the Pedagogical Corpus Grammar (Arhar Holdt et al., 2011; Arhar Holdt et al., 2017).

The Estonian configuration was originally devised for extracting examples for the Estonian Collocations Dictionary (Kallas et al., 2015) aimed at learners of Estonian as a foreign language on the B2-C1 level. The configuration was later used to create a corpus – the etSkELL corpus – that only includes sentences that meet all the pre-defined criteria (i.e. have a GDEX score above 0.5). The etSkELL corpus is now also used as a source corpus in the Estonian SKELL, as well as in the language portal Sõnaveeb for presenting the users a set of authentic corpus examples (Koppel, Kallas et al., 2019; Koppel, Tavast et al., 2019; Koppel, 2020).

For Dutch, special GDEX configurations were developed in the context of the project Woordcombinaties[14] (Word combinations) which is

13 While we are aware that some fields of the NLP area are devoted to related issues that could potentially contribute to the automatic identification of pedagogical sentences or even to enhancing GDEX configurations, such as automatic normalization, automatic error detection, and readability assessment, a decision was made to adopt or adapt existing versions of GDEX configurations as a first step towards identifying candidate sentences for pedagogical purposes. Moreover, and relatedly, it is outside the scope of this paper to explore other approaches to further enhance GDEX configurations.

14 https://woordcombinaties.ivdnt.org/

targeted at advanced language learners (Colman and Tiberius, 2018). For this project, a minimal configuration was defined only using the classifiers not surrounded by round brackets in Table 1, as well as a more restrictive configuration also incorporating the classifiers in between brackets. Lexicographers in the project Woordcombinaties have access to both configurations, and both are being used. For the initial dataset for CrowLL a combination of the two configurations will be used, to bring the Dutch configuration more in line with the configurations for the other languages.

The GDEX configuration that was devised for academic Portuguese in the context of a design of a dictionary for university students (Kuhn, 2017) is the basis for the development of the configuration for data extraction. Given the pedagogical aspect of the academic configuration, adjustments were mostly made according to the characteristics of the type of language, i.e., from academic to general language. Additional development might take place in the future.

Out of the four languages, Estonian has carried out a study especially developed to evaluate its GDEX configuration, while the other languages have relied on the successful and extensive use of the configurations by lexicographers and other users. The output of the Estonian GDEX configuration has been assessed by lexicographers and L2 learners of Estonian. The two types of annotators performed a task to determine whether authentic and unedited corpus sentences would be suitable as example sentences for learners' dictionaries on the B2-C1 level. The results of the assessment showed that both types of annotators considered as many as 85% of the corpus sentences chosen by the Estonian GDEX configuration as good examples, confirming the premise that the methodology GDEX uses to select the examples is reliable (Koppel, 2019). The pre-existing Slovene GDEX configuration adopted in our methodology has been widely tested by lexicographers and successfully implemented in the development of other resources, such as a pedagogical grammar, as noted above. For Dutch, the configuration used is a combination of two configurations that have been tested extensively by a team of lexicographers within the Woordcombinaties project. The Portuguese GDEX configuration for the game is actually the only one that has not been previously tested, as it consists

of an adaptation of an existing configuration. However, the configuration that was used as the basis has been carefully devised and used by other users (for example, when integrated in the Sketch Engine tool).

As mentioned in section 2, GDEX configurations consist of two types of classifiers: hard and soft. Sentences are evaluated against those classifiers and scores are calculated accordingly, based on the weighted sum. Hard classifiers serve to severely penalize sentences, separating the good from the (really) bad ones. Soft classifiers, on the other hand, penalize or give bonuses to the sentences, thus contributing to ranking qualitatively more similar sentences. For the present project, some classifiers are used in all languages, while others are language-dependent. Table 1 provides an overview of the classifiers used in the configurations of the four languages of the game.

Hard classifiers (in bold in Table 1) mean that the evaluation of these features in the sentences weighs heavily on their score. A sentence must start with a capital letter and finish with a period, an exclamation mark or a question mark to be considered a **whole sentence**. For pedagogical purposes, it is crucial that only whole sentences are extracted from the source corpora. The **blacklist – illegal characters** classifier is used to detect the sentences containing strings with unwanted characters such as parts of the program code (<tag>) or URLs (//), because such sentences are not wanted in pedagogically oriented content. Spam texts are usually machine-generated, and thus are not appropriate for language learning. With the **blacklist – spam** classifier, sentences containing words in this blacklist get a very low score. In addition to spam texts, other characteristics of texts found on the web can be counterproductive for pedagogical purposes, such as the presence of typos and misspellings. In order to filter those sentences out, a **minimum frequency for tokens** is established. Another aspect to be considered in a pedagogical example is its length. Very long sentences can compromise intelligibility, i.e., "examples that are intelligible (to the users) are those that are not too long and do not contain complex syntax or rare or specialized vocabulary" (Kosem et al., 2019, p.120), while very short sentences might lack context and lose informative value (ibid.). Thus, sentences that do not fit between the **minimum and maximum sentence length** values get a high penalty.

Table 1: *Overview of the classifiers used in pedagogically oriented configurations for Slovene, Dutch, Estonian and Brazilian Portuguese (adapted from Kosem et al., 2019)*

Classifier	Slovene	Dutch	Estonian	Brazilian Portuguese
whole sentence	X	X	X	X
blacklist - illegal characters	X	X	X	X
blacklist - spam	X		X	X
minimum frequency for tokens	X (3)	X (20)	X (5)	X (5)
minimum and maximum sentence length	X (7 and 60)	X (<30)	X (4 and 20)	X (7-30)
graylist – bad words	X	(X)	X	X
optimal sentence length	X (15-40 tokens)	X (9-12 tokens)	X (6-12 tokens)	X (10-18 tokens)
penalty for long words	X (longer than 12 characters)			X (longer than 12 characters)
penalty for rare characters	X	X	X	X
penalty for capital letters	X	X (part of rare characters)	X	
penalty for tokens with mixed symbols	X	X	X	X
penalty for proper nouns	X	(X)	X	X
penalty for pronouns	X		X	
penalty for sentence initial words	X (list of words provided)	(X)	X	
penalty for sentence initial phrase	X	(X)	X	
penalty for sentence initial tags		(X)	X	
penalty for rare words	X (fewer than 1,000 hits in the corpus)	(X)	X (fewer than 1,000 hits in the corpus)	X (fewer than 500 hits in the corpus)
penalty for commas	X (3 or more)		X (2 or more)	X (2 or more)
penalty for abbreviations		(X)		
penalty for sentences without a finite verb			X	
penalty for more than two occurrences of *que* (that, which)				X

As can be seen in Table 1, the use of soft classifiers (in non-bold in Table 1) varies among the languages, with optimal sentence length and graylist – bad words being used in all of them. Sentences within the **optimal sentence length** get a higher score than the other sentences outside this interval, and are thus ranked higher up among all the sentences. Length values vary from language to language, and have been defined based on what each language considered to be the optimal sentence length interval for pedagogical purposes.[15]

Words in the **graylist – bad words** are compared against the sentences in a corpus and, if any word is found, the sentence is penalized. Evaluation of the settings has shown that this penalization is enough to push such potentially problematic sentences lower down the ranking, but still not too low in case the penalization is unjustified (polysemous words, etc.). This means that sentences with higher scores (in the upper part of the list) will probably not contain explicitly offensive words, that sentences with very low scores (at the bottom of the list) will probably contain offensive words, and that the ones in the middle might or might not contain them. While we want the players to assess the sentences from the upper and lower parts and possibly confirm that they are non-problematic and problematic, respectively, one of the most interesting contributions from the players will be the evaluation of sentences pertaining to exactly this grey, middle area, where one can expect to find explicitly offensive lemmas, offensive lemmas that are polysemous and not being used in an offensive manner, offensive sentences with no overtly offensive lemmas, and sentences with sensitive content. This type of evaluation is still not well performed by computers, so we need humans to do it.

The Slovenian graylist contains 1,909 words (nouns, adjectives, verbs and adverbs) that were identified in several lexicographic and linguistic projects as vulgar or (potentially) offensive. For Portuguese, there are two graylists of explicitly offensive and vulgar items (nouns, adjectives and verbs), one consisting of lemmas and another one of word forms and strings (e.g., *fodid.+*), totaling 91 items. These lists result from manual

15 It was observed that different languages differ in the average sentence length due to various reasons such as word formation (e.g. compounds in Estonian are mainly written as one word, as opposed to two or more words in Slovene), existence of articles etc.

evaluation and editing of the list of taboo lemmas and word forms created by the Sketch Engine team for the default Portuguese GDEX that they have devised. Words related to cultural aspects, such as those related to religion or nationalities, that were not offensive or vulgar but had probably been included because of their potential to spark hate speech, were discarded. In addition, new offensive or vulgar items were added, but further editing can be carried out if necessary. The Estonian graylist contains 1,472 words (nouns, adjectives, verbs), consisting of words tagged as vulgar, offensive, colloquial, and slang in the EKI Combined Dictionary (Langemets et al., 2022), swear words in foreign languages (e.g. *fuck*), their adapted variants (e.g. *fakk*, *pohui* 'похуй'), and words written differently from the written language norm. The Dutch configuration uses a graylist of 93 words which is based on words labeled as vulgar or offensive in the Algemeen Nederlands Woordenboek.[16] If needed, the Dutch graylist will be further refined in the future.[17]

Other classifiers relevant in the context of language learning are **penalties for long words**, **rare characters**, **tokens with mixed symbols** and **capital letters**. This is based on the assumption that longer words, too many rare characters and capital letters as well as the occurrence of non-words have an impact on reading complexity. For pedagogical purposes, a penalty can also be given to **proper nouns** in order to give priority to sentences without (or with few) of these, as in many cases the named entities in those sentences might not be known to the learners. The same applies to **abbreviations** which learners may not necessarily be familiar with. Penalizing **pronouns** can also help, as sentences with many pronouns are often too anaphoric and lack context for proper understanding.

16 https://anw.ivdnt.org/search (30. 8. 2022). Note the ANW is a dictionary under construction, and thus new words (including words labelled as vulgar or offensive) are continuously being added. The current GDEX configuration for Dutch uses the words labelled as vulgar or offensive in the ANW at the time the GDEX configuration was defined for the project Woordcombinaties.

17 As can be noticed, there is a considerable difference between the number of lemmas in the graylists for different languages. More thorough studies on problematic vocabulary were conducted for Slovenian and Estonian, and more extensive word lists were obtained as a result. It should be noted that these graylists contain lemmas that are problematic only in part, e.g. in one of their senses. Consequently, the penalization of sentence(s) containing the word(s) is milder. Using different approaches to graylists will open possibilities to compare them at the end of the study.

Another type of classifier uses lists containing **words** and **phrases** that should not occur **in** a **sentence-initial position**. These words and phrases are heavily penalized because in previous manual evaluations of extracted sentences for Slovene, Estonian and Dutch, several sentence-initial words and phrases were identified that are a good signal that the sentence is contextually dependent on the previous sentence(s), and is thus less suitable to be used as a standalone component for pedagogical purposes. Similarly, certain **sentence-initial tags** can be penalized, e.g. conjunctions, because sentences starting with conjunctions are often anaphoric.

Furthermore, sentences containing less frequent words tend to be considered inadequate to serve as examples of language use in pedagogical contexts, as such words are likely not known to the learners and might act as a distraction. The **penalty for rare words** classifier penalizes sentences with words whose frequency is below a certain threshold, so these sentences get lower scores. The use of too many commas in a sentence might be indicative of complexity, so the **penalty for commas** is a classifier that penalizes sentences if they have more than a defined number. A **penalty for sentences without a finite verb** can help to filter out less typical sentences. The grammar of the Estonian language (Erelt and Metslang, 2017), for instance, states that a typical sentence contains a finite verb and phrases (collocations) that go with the verb.

Portuguese adopts a separate **penalty for more than two occurrences of *que*** (*that*, *which*). This classifier has been created to avoid sentences with too many subordinate or relative clauses, because high syntactic complexity makes understanding more difficult, which is something to be avoided in pedagogical examples.

5.3 Lemma lists

To ensure at least partial comparability of the multilingual results, we decided to extract the data using lemmata, comparable across the participating languages. For this purpose, we first prepared a list of 100 words in English using the criteria described below. In the second step, we translated the list to Slovene, Brazilian Portuguese, Dutch, and

Estonian, reporting on problems with translation equivalents, as well as their frequency in the corresponding source corpora. We discuss some of these issues in Section 6.

We wanted to include lemmata that were of different relevance for labeling in the context of the CrowLL task: (a) words that were clearly (on the surface and in the vast majority of the meanings) offensive or vulgar, for example: *nigger, whore, bitch, retarded, to fuck, to piss*; (b) words that were offensive or vulgar in some of the meanings, as well as words with potentially sensitive content, for example: *cow, drunk, suicide, fanatic, depressed, to molest*; (c) words that would typically not be considered offensive, vulgar or sensitive from the perspective of our labeling task, for example *year, world, service, new, to say, to see.* Vocabulary from the first group would typically make it to blacklists, and thus a blacklist-based methodology would automatically filter out corpus occurrences with these words before they would be included in any teaching material. Here, we are including it to test the hypothesis that these corpus occurrences would also be marked as inappropriate by the crowd. On the other hand, non-problematic words are included to test the complementary premise. The most interesting for our task, however, are words in group (b). The lemmata list thus includes 20 words from groups (a) and (c) and 60 words from group (b).

The seed lemmata were selected using the translation into English of a list of words that were identified during the creation of a GWAP called Game of Words (Arhar Holdt et al., 2021). This game prompts the players to provide synonyms and collocations for different Slovene words, with the implicit purpose to clean the noise from two automatically created databases comprising openly available lexical information for Slovene. As the game is aimed at young(er) users, not only vulgar and offensive words were removed from the list of potential prompts, but also words with sensitive content that could cause the player unnecessary discomfort. The criteria for removal were based on existing resources, such as dictionaries, and privately compiled lists by researchers or journalists (ibid., p. 43). Semantically, the removed words covered a) human features, such as race, nationality, gender, age, sexual orientation, religious and political beliefs, migration status, social status, education, handicap, bodily and mental features etc., as

well as b) sensitive topics, such as violence, illness, death, addiction, sex, excretions, etc. Offensive, vulgar, and potentially sensitive words for CrowLL were selected based on these categories, while non-problematic words were chosen from the most frequent words in English Web 2020, available on Sketch Engine.

The majority (50) of the included lemmata are nouns, 25 are verbs and 25 are adjectives. An example of seed lemmata with labels and translations is provided in Table 2.

Table 2: *Common lemma list and its translations to Slovene, Estonian, Brazilian Portuguese and Dutch*

Category	Type	English	POS	Slovene	Estonian	Brazilian Portuguese	Dutch
Race	B	black-skinned	A	temnopolt	mustanahaline	negro	zwart
Race	B	native	N	domorodec	pärismaalane	índio	autochtoon
Race	B	racist	A	rasističen	rassistlik	racista	racistisch
Race	A	nigger	N	črnuh	neeger	crioulo	neger
sexual orientation	B	homosexual	A	homoseksualen	homoseksuaalne	homossexual	homoseksueel
sexual orientation	B	straight	A	heteroseksualen	heteroseksuaalne	heterossexual	heteroseksueel
sexual orientation	B	lesbian	A	lezbičen	lesbiline	lésbica	lesbisch
sexual orientation	A	faggot	N	peder	pede	bicha	flikker
violence	B	to murder	V	umoriti	mõrvama	assassinar	vermoorden
violence	B	brutal	A	brutalen	brutaalne	brutal	brutaal
violence	B	to bully	V	ustrahovati	kiusama	intimidar	intimideren
violence	B	to torture	V	mučiti	piinama	torturar	martelen
violence	B	to rape	V	posiliti	vägistama	estuprar	verkrachten
violence	B	to beat	V	pretepati	peksma	bater	slaan
violence	B	to molest	V	zlorabljati	ahistama	molestar	lastigvallen
violence	B	to shoot	V	ustreliti	tulistama	atirar	schieten
non-problematic	C	time	N	čas	aeg	tempo	tijd
non-problematic	C	way	N	način	viis	maneira	manier
non-problematic	C	to include	V	vključiti	sisaldama	incluir	omvatten
non-problematic	C	good	A	dober	hea	bom	goed

As mentioned at the outset of this section, the data extraction procedure to obtain 10,000 sentences is meant for game development and initial tests. More specifically, the procedure will be performed as follows. We will use GDEX configurations to extract the top 200 sentences per lemma of the lemma list so that we have a buffer in case of duplicates. We will then verify those 20,000 sentences and reduce them to 10,000 sentences per language.

Once we have this data, we will proceed with manual annotation of the sentences with the labels from the game (non-problematic/problematic; category of the problem), which will allow us to evaluate the labeling system and the quality of the input data, and propose adjustments to the resources and the game if necessary. These annotated sentences will comprise manually annotated pedagogical corpora, and will be available as part of the CLARIN Language Resources Family. They will also be fed into the game to be used for scoring mechanism development, such as the scores given by comparison with other players and asynchronous play, for implementation of the dual-player mode, as prerecorded answers for a bot, and as input data for the game.

When the game is launched, additional data will be required as input. The extraction of this data will follow a slightly different approach, given that we want the crowd to label as many sentences from the source corpora as possible. With the source corpora, pedagogically oriented GDEX configurations, and tested labeling system and gameplay, data input for the game will be extracted as follows. First, we will GDEX the corpus, i.e., run the GDEX configuration to assign GDEX scores to all sentences in the corpus. We will then extract sentences in batches, with varying GDEX scores, i.e., a certain number of sentences with the highest scores, medium scores and low scores. These sentences will be input into the game for players to play. Once the game is tested with actual players, an evaluation of the methodology of data preparation can be carried out.

6 Analysis and discussion

One of the main aspects that might have an impact on the results of the initial test with annotation of 10,000 sentences is that the resources

that were used for data preparation present different levels of development. While Estonian and Slovene use source corpora that have been carefully compiled in the context of other projects, with rich metadata and advanced annotation, Dutch and Portuguese use automatically compiled web corpora with no human curation and POS-tagged by the Sketch Engine team. It should be acknowledged that these differences in the development of the resources might influence the quality of the input data (extracted sentences), with consequent reflection on the quality of the output data (annotated sentences).

Preparing the common lemma list posed many challenges, becoming an iterative process in which English words were proposed, translated to the target languages and then – based on the suitability of the translation equivalents – accepted or replaced. A discussion was needed if for one or more target languages a translation equivalent was not suitable from the perspective of form, meaning, connotation or frequency.

To ease the data extraction, we aimed for a list of single-word lemmata for all target languages. We thus avoided English prompts that would require multiword translations. For example, for the English verb *to fuck off* not all languages had single-word translations (Slovene: *odjebati,* Estonian: *perse käima*, Portuguese: *ir se foder,* Dutch: *opsodemieteren*), therefore we replaced it with the verb *to fuck* (Slovene: *jebati,* Estonian: *keppima*, Portuguese: *foder*, Dutch: *neuken*). More permissive were our decisions when it came to the part-of-speech of the translation equivalents. For most of the cases, providing translation equivalents of the same POS was unproblematic. In rare instances where the POS of otherwise the most suitable translation candidate did not match, we kept it on the list. For example, some English adjectives in Estonian are actually case forms of a noun, e.g. *depressioonis* 'in depression' (not 'depressed'). When examining the occurrences of the lemmata in the source corpora, we also noticed that some POS differences stemmed from the features of the taggers used to annotate the data (e.g., the Portuguese equivalent *retardado* for the English adjective *retarded* occurs erroneously tagged as verbs (participle) in the Portuguese corpus). While such problems would have to be considered when extracting the data, they did not influence the selection of the candidates for the common lemma list.

Important for the list was the connotation of the translation equivalents. When the target language did not have a translation equivalent with comparable sensitivity, the English word was replaced. For example, the English noun *bimbo* for an 'attractive but unintelligent or frivolous young woman' did not have a suitable single-word translation in Portuguese, so we replaced it with a (more offensive) *slut* (Slovene: *cipa*, Estonian: *libu*, Portuguese: *vagabunda*, Dutch: *slet*). Other semantic differences, such as nuances in the meaning(s) of the translated words were accepted, as we did not want to create a list that would be overly curated, artificial, and methodologically difficult to expand with further lemmata and to other languages. In situations where more semantically suitable translation equivalents were possible, we opted for the one that was less polysemic (for example, for the English noun *corpse,* we chose the Portuguese *cadáver* and not *corpo* which has a wider use).

Finally, the translation equivalents were checked for their frequency in the corresponding source corpora. According to our methodology, we needed at least 100 heterogenous corpus examples per lemma, but to have enough data to select from we aimed to extract 200. Especially in "cleaner" corpora, such as the Slovene source corpus Gigafida, the offensive and vulgar words were rare, but nearly all proposed lemmas had over 200 occurrences. We decided to keep the noun *asshole* with a Slovene translation *pezde* (198 occurrences in the Slovene source corpus) and replace the adjective *transsexual* (less than 10 occurrences in the Dutch source corpus) with a more frequently occurring *transgender*.

Once the game is fully operational, a series of issues need to be considered. For example, it is important to ensure the rapid implementation of the game's results into practice. This requires both a set of clear parameters on what a minimum number – as well as a maximum number – of user responses per example is, what level of agreement is required, etc., as well as automatic tools or algorithms for regular data analysis and summarization. All this helps to increase the quantity of crowdsourced data, as more examples can be added to the game (and at the same time the sufficiently examined ones removed) on a regular basis. Technical aspects should also be paid enough attention,

meaning the server should have enough capacity and storage space to cater for heavy usage, which can partly be addressed by conducting rigorous stress tests before the launch of the game. Last but not least, a detailed promotion plan needs to be prepared in advance, including the steps on how to not only attract users, but also keep them long term.

7 Conclusions

In this paper, we proposed a methodology of data preparation for the development of the Crowdsourcing for Language Learning (CrowLL) game, from which data will be collected through crowdsourcing to create problem-labeled pedagogical corpora for Dutch, Estonian, Slovene, and Brazilian Portuguese. For this process a series of decisions had to be made, from the choice of source corpora, to GDEX configuration development and lemma list creation. By describing the methodology and reflecting on the challenges posed and solutions found, it is our intention to provide researchers sharing common interests with a model that can be applied to other languages, and potentially to other purposes.

The next steps of our project involve the extraction of sentences for the game, full implementation of the game, collection of answers (from actual players), statistical analysis of labeled data, and design and administration of a user survey to evaluate the game design and user experience. With the players' answers, we will compile problem-annotated corpora and develop other auxiliary language learning resources, such as SKELL for all the languages. After that, we plan to start the third stage of the project, in which we will use the problem-labeled corpora to create the basis for the future development of machine-learning training models to automatize identification and labeling of problematic content, thus contributing to the further and faster creation of pedagogical corpora.

Acknowledgments

The authors acknowledge the financial support received from the Portuguese national funding agency, FCT – Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020) and the Slovenian Research Agency

(research core funding No. P6-0411, Language Resources and Technologies for Slovene, and project funding No. J7-3159, Empirical foundations for digitally-supported development of writing skills). The research received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 731015. This study has also been supported by the CLARIN Resource Families Project Funding.

References

Aitamurto, T., Leiponen, A., & Tee, R. (2011). *The promise of idea crowdsourcing–benefits, contexts, limitations* [White paper]. Nokia Ideas project.

Arhar Holdt, Š., Kosem, I., & Gantar, P. (2017). Corpus-based resources for L1 teaching: The case of Slovene. In *Handbook on digital learning for K-12 schools* (pp. 91–113). Springer, Cham. doi: 10.1007/978-3-319-33808-8_7

Arhar Holdt, Š., Kosem, I., Krapš Vodopivec, I., Ledinek, N., Može, S., Stritar Kučuk, M., Svenšek, T., & Zwitter Vitez, A. (2011). *Pedagoška slovnica pri projektu Sporazumevanje v slovenskem jeziku: K16 – Standard za korpusno analizo slovničnih pojavov*. Ljubljana: Ministrstvo za šolstvo in šport: Amebis. Retrieved from http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik16/Kazalnik_16_Pedagoska_slovnica_SSJ.pdf

Arhar Holdt, Š., Logar, N., Pori, E., & Kosem, I. (2021). "Game of Words": Play the game, clean the database. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Proceedings of the EURALEX XIX congress: Lexicography for inclusion, 7–11 September, Aleksandroupolis, Greece* (Vol I., pp. 41–49). Retrieved from https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol1-p041-049.pdf

Baisa, V., & Suchomel, V. (2014). SkELL: Web interface for English language learning. *Proceedings of the eighth workshop on recent advances in Slavonic natural language processing, RASLAN 2014* (pp. 63–70). Retrieved from https://nlp.fi.muni.cz/raslan/2014/12.pdf

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. *CEUR Workshop proceedings,* 1–6. Retrieved from http://ceur-ws.org/Vol-2253/paper49.pdf

Bédi, B., Chua, C., Habibi, H., Martinez-Lopez, R., & Rayner, M. (2019). Using LARA for language learning: a pilot study for Icelandic. In F. Meunier, J. van de Vyver, L. Bradley & S. Thouësny (Eds.), *CALL and complexity: short papers from EUROCALL 2019* (pp. 33–38). Research-publishing.net. doi: 10.14705/rpnet.2019.38.982

Bonetti, F., & Tonelli. S. (2020). A 3D role-playing game for abusive language annotation. *Workshop on games and natural language processing* (pp. 39–43). Retrieved from https://aclanthology.org/2020.gamnlp-1.6

Bonetti, F., & Tonelli. S. (2021). Challenges in designing games with a purpose for abusive language annotation. *Proceedings of the first workshop on bridging human–computer interaction and natural language processing* (pp. 60–65). https://aclanthology.org/2021.hcinlp-1.10

Boulton, A. (2017). Corpora in language teaching and learning: Research timeline. *Language Teaching*, *50*(4), 483–506. doi: 10.1017/S0261444817000167

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, *17*(1), 47–64. doi: 10.1017/S0958344005000510

Buecheler, T., Sieg, J. H., Füchslin, R. M., & Pfeifer, R. (2010). Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework. In H. Fellermann, M. Dörr, M. Hanczyc, L. L. Laursen, S. Maurer, D. Merkle, P-A. Monnard, K. Stoy, S. Rasmussen (Eds.), *Artificial live XII: proceedings of the twelfth international conference on the synthesis and simulation of living systems* (pp. 679–686). MIT Press. doi: 10.21256/zhaw-4094

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 Publishing Company. doi: 10.1075/scl.92.12cal

Chamberlain, J., Poesio, M., & Kruschwitz, U. (2008). Phrase detectives: A web-based collaborative annotation game. *Proceedings of the international conference on semantic systems (I-Semantics'08)* (pp. 42–49). Retrieved from https://www.jonchamberlain.com/media/doc/Chamberlain-2008Phrase.pdf

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). Routledge. doi: 10.4324/9781315657899.ch26

Chesbrough, H. W. (2006). *Open innovation: The new imperative for creating and profiting from technology.* Harvard Business School Press.

Colman, L., & Tiberius C. (2018). A good match: A Dutch collocation, idiom and pattern dictionary combined. *Proceedings of the XVIII EURALEX international congress: Lexicography in global contexts* (pp. 233–246). Retrieved from https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2952-1-10-20180820.pdf

Erelt, M., & Metslang, H. (2017). *Eesti keele süntaks.* Eesti keele varamu III. Tartu Ülikooli Kirjastus. Retrieved from https://dspace.ut.ee/handle/10062/70510

Eryiğit, G., Şentaş, A., & Monti, J. (2022). Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering* (pp. 1–33). doi: 10.1017/S1351324921000401

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering automated lexicography: The case of the Slovene lexical database. *International Journal of Lexicography*, *29*(2), 200–225. doi: 10.1093/ijl/ecw014

Gorovaia, N. (2018). *Behavior of users on the crowdsourcing platforms*. [Poster session]. EnetCollect WG3/WG5 meeting, October 24–25, Leiden, Netherlands.

Gries, S. (2009). What is corpus linguistics*? Language and Linguistics Compass*, *3*, 1–17. doi: 10.1111/j.1749-818X.2009.00149.x

Guillaume, B., Fort, K., & Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 3041–3052). Retrieved from https://aclanthology.org/C16-1286

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

Harris, C.G. (2014). The beauty contest revisited: measuring consensus rankings of relevance using a game. *Proceedings of the first international workshop on gamification for information retrieval – GamifIR@ECIR '14* (pp. 17–21). doi: 10.1145/2594776.2594780

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. *Proceedings of the eLex 2015 conference* (pp. 1–20). Retrieved from https://elex.link/elex2015/proceedings/eLex_2015_01_Kallas+etal.pdf

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), 7–36. doi: 10.1007/s40607-014-0009-9

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX international congress* (Vol. 1, pp. 425–432). https://tinyurl.com/yckr9w8s

Kilgarriff, A., Rychlý, P., Smrz, P., & D. Tugwell (2004). The Sketch Engine. *Proceedings of the eleventh EURALEX international congress, EURALEX 2004* (pp. 105–116). Retrieved from https://tinyurl.com/mvrp4ymy

Koppel, K. (2019). Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks. *Lähivõrdlusi. Lähivertailuja, 29*, 84–112. doi: 10.5128/LV29.03

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele*. Doktoritöö, Tartu Ülikool. Retrieved from https://dspace.ut.ee/handle/10062/67138

Koppel, K., & Kallas, J. (2022). *Eesti keele ühendkorpus 2021*. doi: 10.15155/3-00-0000-0000-0000-08D17L

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. *Proceedings of the eLex 2019 conference* (pp. 763–782). Retrieved from https://zenodo.org/record/3612933#.Yywd1XZBy70

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. *Proceedings of the eLex 2019 conference* (pp. 434–452). Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_24.pdf

Kosem, I. (2012,). Using GDEX in (semi)-automatic creation of database entries [Conference presentation]. *SKEW-3, 3rd international Sketch Engine workshop, 21–22 March, 2012.*

Kosem, I., Gantar, P., & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. *Proceedings of the eLex 2013 conference* (pp. 32–48). Retrieved from http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf

Kosem, I., Husák, M., & McCarthy, D. (2011). GDEX for Slovene. *Proceedings of eLex 2011* (pp. 151–159). Retrieved from http://www.dianamccarthy.co.uk/files/Kosemetal-paper.pdf

Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J., & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, *32*(2), 119–137. doi: 10.1093/ijl/ecy014

Krek, S., Arhar Holdt, Š., 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. *Proceedings of the twelfth language resources and evaluation conference* (pp. 3340–3345). Retrieved from https://aclanthology.org/2020.lrec-1.409

Kuhn, T. Z. (2017). *A design proposal of an online corpus-driven dictionary of Portuguese for university students* [Doctoral dissertation, Universidade de Lisboa]. Retrieved from http://hdl.handle.net/10451/32013

Kuhn, T. Z., Šandrih Todorović, B., Holdt, Š. A., Zviel-Girshin, R., Koppel, K., Luís, A.R., & Kosem, I. (2021). Crowdsourcing pedagogical corpora for lexicographical purposes. *Proceedings of the XIX EURALEX congress: Lexicography for inclusion* (Vol. II., pp. 771–779). Retrieved from https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol2-p771-779.pdf

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. *Proceedings of SNLP'07: 7th international symposium on natural language processing*. Retrieved from https://hal-lirmm.ccsd.cnrs.fr/lirmm-00200883

Langemets, M., Hein, I., Jürviste, M., Kallas, J., Kiisla, O., Koppel, K., Leemets, T., ..., & Tubin, V. (2022). *EKI ühendsõnastik 2022*. doi: 10.15155/3-00-0000-0000-0000-08C0AL

Lévy, P. (1997). *Collective intelligence: Mankind's emerging world in cyberspace*. Plenum Trade. New York.

Lew, R. (2014). User-generated content (UGC) in online English dictionaries. *OPAL, 4*, 8–26. Retrieved from https://pub.ids-mannheim.de//laufend/opal/opal14-4.html

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* (pp. 176–181). Research-publishing.net. doi: 10.14705/rpnet.2018.26.833

Morschheuser, B., Hamari, J., Koivisto, J., & Maedche, A. (2017). Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies*, *106*, 26–43. doi: 10.1016/j.ijhcs.2017.04.005

Nicolas, L., Lyding, V., Borg, C., Forăscu, C., Fort, K., Zdravkova, K., Kosem, I., ..., & HaCohen-Kerner, Y. (2020). Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. *Proceedings of the 12th language resources and evaluation conference* (pp. 268–278). Retrieved from https://aclanthology.org/2020.lrec-1.34

Osborne, J. (2004). Top-down and bottom-up approaches to corpora in language teaching. language and computers. In U. Connor, T. A. Upton (Eds.),

Applied Corpus Linguistics. A Multidimensional Perspective (pp. 251–265). Brill. doi: 10.1163/9789004333772_015

Pe-Than, E. P. P., Goh, D. H. L., & Lee, C. S. (2015). A typology of human computation games: an analysis and a review of current games. *Behaviour & Information Technology*, *34*(8), 809–824. doi: 10.1080/0144929X.2013.862304

Pilán, I., Vajjala, S., & Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. ArXiv. doi: 10.48550/arXiv.1603.08868

Pilán, I., Volodina, E., & Johansson, R. (2013). Automatic selection of suitable sentences for language learning exercises. *20 Years of EUROCALL: Learning from the past, looking to the future: 2013 EUROCALL Conference Proceedings* (pp. 218–225). Retrieved from https://aclanthology.org/W14-1821.pdf

Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 174–184). Retrieved from https://aclanthology.org/W14-1821

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources & Evaluation*, *55*(2), 477–523. doi: 10.1007/s10579-020-09502-8

Prahalad, C. K., & Ramaswamy, V. (2000). Co-opting customer competence. *Harvard Business Review*. Retrieved from https://hbr.org/2000/01/co-opting-customer-competence

Preist, C., Massung, E., & Coyle, D. (2014). Competing or aiming to be average? Normification as a means of engaging digital volunteers. *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing (CSCW '14)* (pp. 1222–1233). doi: 10.1145/2531602.2531615

Reynaert, M. (2006). Corpus-induced corpus clean-up. *Proceedings of the fifth international conference on language resources and evaluation* (pp. 87–92). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/229_pdf.pdf

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* (pp.18–35). Continuum Publishing Corporation.

Šandrih Todorović, B. (2020). *Impact of text classification on natural language processing applications*. [Универзитет у Београду].

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10). doi: 10.18653/v1/W17-1101

Seemakurty, N., Chu, J., von Ahn, L., & Tomasic, A. (2010). Word sense disambiguation via human computation. *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 60–63). doi: 10.1145/1837885.1837905

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: observing the world's largest citizen science platform. *Proceedings of the 23rd international conference on world wide web*, 1049–1054. doi: 10.1145/2567948.2579215

Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxbow Books. Retrieved from https://users.ox.ac.uk/~martinw/dlc/chapter1.htm

Stanković, R., Šandrih, B., Stijović, R., Krstev, C., Vitas, D., & Marković, A. (2019). SASA dictionary as the gold standard for good dictionary examples for Serbian. *Proceedings of the eLex 2019 conference* (pp. 248–269). Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_14.pdf

Trampuš, M., & Novak, B. (2012). The internals of an aggregated web news feed. *Proceedings of 15th multiconference on information society 2012 (IS-2012)*. Retrieved from http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf

Vajjala, S. (2022). Trends, limitations and open challenges in automatic readability assessment research. *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5366–5377). Retrieved from https://aclanthology.org/2022.lrec-1.574

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE, 15*(12): e0243300. doi: 10.1371/journal.pone.0243300

von Ahn, L. (2006). Games with a purpose. *Computer*, *39*(6), 92–94. Retrieved from https://www.cs.cmu.edu/~biglou/ieee-gwap.pdf

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67. doi: 10.1145/1378704.1378719

Von Hippel, E., & Katz, R. (2002). Shifting innovation to users via toolkits. *Management science*, *48*(7), 821–833.

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

Xu, L., & Chamberlain, J. (2020). Cipher: a prototype game-with-a-purpose for detecting errors in text. *Workshop games and natural language processing* (pp. 17–25). Retrieved from https://aclanthology.org/2020.gamnlp-1.3

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings of the 13th international workshop on semantic evaluation (SemEval-2019)* (pp. 75–86). doi: 10.18653/v1/S19-2010

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, C. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *Proceedings of the 14th international workshop on semantic evaluation.* Retrieved from https://arxiv.org/abs/2006.07235

Zviel-Girshin, R., Kuhn, T. Z., Luís, A. R., Koppel, K., Šandrih Todorović, B., 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). doi: 10.14705/rpnet.2021.54.1352

Priprava podatkov pri množičenju v pedagoške namene: primer igre CrowLL

Eden od načinov za spodbujanje uporabe korpusov pri jezikovnem izobraževanju je izdelava pedagoško primernih korpusov, označenih z različnimi vrstami problematik (občutljiva vsebina, žaljiv jezik, strukturne težave). Ker je ročno označevanje korpusov zelo časovno potratno, je potrebno poiskati boljši pristop. Predlagamo kombinacijo dveh pristopov k oblikovanju problemsko označenih pedagoških korpusov nizozemščine, estonščine, slovenščine in brazilske portugalščine: uporabo iger z namenom množičenja. Z udeleženci smo izvedli začetne poskuse, da bi ugotovili, če je naloga množičenja ustrezna, pridobljene izkušnje pa smo uporabili za oblikovanje igre *Crowdsourcing for Language Learning (CrowLL)*, v kateri igralci prepoznavajo problematične povedi in segmente ter jih razvrščajo. V prispevku se osredotočamo na pripravo podatkov, saj ima ta korak ključni pomen pri vsakem projektu množičenja, ki obravnava ustvarjanje jezikovnih učnih virov. Predlagamo metodologijo za

pripravo podatkov, podrobno predstavljamo izbiro izvornih korpusov, pedagoško usmerjene konfiguracije GDEX in oblikovanje seznamov lem, s posebnim poudarkom na pogostih in od jezika odvisnih odločitvah. Za konec ponujamo razpravo o izzivih, ki smo jih zasledili, in o rešitvah, ki smo jih do sedaj že uvedli.

Ključne besede: množičenje, igra z namenom, vzorčni stavki, pedagoški korpus