The CrowLL project - Manually-annotated corpora for teaching and learning purposes of Brazilian Portuguese, Dutch, Estonian, and Slovene
Tanara Zingano Kuhn, CELGA-IL/TEC/University of Coimbra
Carole Tiberius, Dutch Language Institute
Spela Arhar Holdt, Centre for Language Resources and Technologies, University of Ljubljana
Iztok Kosem, Centre for Language Resources and Technologies, University of Ljubljana/Jožef Stefan Institute
Kristina Koppel, Institute of the Estonian Language
Rina Zviel-Girshin, Ruppin Academic Centre/Faculty of Engineering
Ana R. Luís, CELGA-IL/TEC/University of Coimbra

Summary
The project seeks to provide manually-annotated corpora for teaching and learning purposes of Brazilian Portuguese, Dutch, Estonian, and Slovene, as a contribution to the Manually Annotated Corpora Family available in CLARIN. Each corpus will contain 10,000 sentences annotated as appropriate or inappropriate, with categories of inappropriateness labels for inappropriate sentences. This project will also develop a crowdsourcing gamified solution for further corpus growth. The annotation methods developed in this project will be published to allow expansion to other languages. In addition to pedagogical applications, these annotated corpora can be used, within NLP, as inputs to train either a) binary machine learning models to automatically classify sentences as appropriate or inappropriate or b) multi-class classifiers to perform fine-grained annotation of inappropriate sentences.

Introduction
Evidence of authentic language use is fundamental for language learning. One way to develop authentic language learning materials is through the use of examples from corpora. However, these corpora might include sensitive content or offensive language, in addition to exhibiting structural (grammar, spelling) problems. Although such use is uncontroversially authentic, it is recommended that these corpora are carefully monitored before applied to education to flag inappropriateness, thus leaving the choice of use of certain examples to the needs and context of use of teachers and didactic material developers.

Justification
Monitoring these corpora, however, can be challenging in at least two ways:
1. Manual monitoring of large amounts of texts is extremely time-consuming, thus expensive;
2. The very nature of language limits automatization of monitoring: - many words are polysemic (shortcomings to rule-based approaches to automatically identifying offensive words); - problems identified as structural errors via automatic error detection ≠ actual mistakes, but rather spelling and grammatical variation based on the context of use; - contextual, socio-historical, and subjective aspects ≠ significant role in the determination of what sensitivity and offensiveness in language are.
As a result, a solution must be found to streamline human verification of examples.

Objectives
• Contribute to the CLARIN Manually-annotated corpora family by producing manually-annotated corpora of Brazilian Portuguese, Dutch, Estonian, and Slovene;
• Develop a crowdsourcing-based game for further corpus growth.

Manual annotation
Data preparation:
1. Source corpora:
   • Brazilian Portuguese: Timestamped JSI web corpus 2014–2021 (Trampol & Novak, 2012) – approx. 3.2 billion words (only Brasil subsorpus);
   • Dutch: Timestamped JSI web corpus 2014–2021 (Trampol & Novak, 2012) – approx. 1.3 billion words;
   • Estonian: Estonian National Corpus 2021 (Koppel & Kassev, 2022) – approx. 2.3 billion words;
   • Slovene: Gigafida 2.0 (Krek et al., 2020) – approx. 1.2 billion words;

2. Pedagogically-oriented GDEX configurations for each language:
   • GDEX (Kilgarriff et al., 2008): a rule-based formal that assigns numerical score to each corpus sentence sensitive based on how well it meets the pre-defined criteria.
   • Hard classifiers: severely penalise sentences, separating the good from the (really) bad ones. E.g., whole sentence, minimum and maximum sentence length;
   • Soft classifiers: penalise or give bonus to the sentences, thus contributing to ranking qualitatively more similar sentences. E.g., greylist bad words, optimal sentence length;
   • Sentences are evaluated against those classifiers and scores are calculated according to weight sum.
   • For the present project, some classifiers are used in all languages, while others are language-dependent.
   3. Lemma list:
   • First, preparation of a list of 100 words in English, then its translation to Brazilian Portuguese, Dutch, Estonian, and Slovene;
   • Lemmas of different relevance for labelling in the context of the CrowLL task:
     • Black = clearly (on the surface and in the vast majority of the meanings) offensive or vulgar words, e.g.: nigger, whore, bitch, retarded, to fuck, to jizz (20 words);
     • Grey = words that are offensive or vulgar in some of the meanings, as well as words with potentially sensitive content, e.g.: drunk, suicide, fanatic, depraved, to molest (50 words);
     • White = words that typically not be considered offensive, vulgar or sensitive from the perspective of our labelling task, e.g.: your, world, service, man, to say, to is (20 words).

Data extraction:
For each language:
• Use GDEX configuration to extract from the source corpus the top 200 sentences per lemma of the lemma list;
• Deduplicate sentences (per lemma);
• Take the top 100 sentences (per lemma) from the remaining, totalling 10,000 sentences.

Format and Availability:
The manually annotated corpus will be distributed in tab-separated value (TSV) format with the following headers:

<table>
<thead>
<tr>
<th>Header</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Brazilian Portuguese, Dutch, Estonian, Slovene</td>
</tr>
<tr>
<td>Sentence</td>
<td>the extracted corpus sentence</td>
</tr>
<tr>
<td>Sentence ID</td>
<td>a unique identifier for the sentence in the manually-annotated corpus</td>
</tr>
<tr>
<td>GDEX score</td>
<td>score assigned to the sentence by GDEX function in Sketch Engine</td>
</tr>
<tr>
<td>Seed Lemma</td>
<td>the lemma used as seed for automatic sentence extraction</td>
</tr>
<tr>
<td>Part of Speech</td>
<td>the part of Speech tag of the seed lemma, i.e. adjective, noun, verb</td>
</tr>
<tr>
<td>Lemma Type</td>
<td>the type of the seed lemma, i.e. black, grey, white</td>
</tr>
<tr>
<td>Label</td>
<td>the label assigned by the annotator indicating whether the sentence is problematic or non-problematic</td>
</tr>
<tr>
<td>Problem</td>
<td>the problem category label assigned by the annotator, i.e. offensive; inappropriate content; spelling problems; grammar problems; wrong grammar; lack of content/incomprehensible</td>
</tr>
<tr>
<td>Annotator ID</td>
<td>the unique identifier for the annotator</td>
</tr>
</tbody>
</table>

Examples:
Uma cozinheira diz que foi chamada de crioula durante uma discussão no restaurante em que trabalha. A week says she was called a nigger during an argument at the restaurant where she works.

Game development
The CrowLL game:
• Game with a Purpose approach.
• Available as a webpage and mobile app.
• Single-player and dual-player mode.
• The goal of the game is to label words = crowdsourcing game (Monteith et al. 2017), i.e., majority consensus.

Game mechanics:
• Collaborative game with three levels.
• Level 1 (I’m curious!): players identify problematic sentences according to their judgement.
• Level 2 (I’m eager to help!): players categorise those sentences, ranging from grammar/spelling problems to offensiveness and sensitivity.
• Level 3 (I’m feeling enthusiastically!): players mark in the sentence what they consider to be problematic.

Concluding remarks
• We propose an alternative way of creating pedagogical corpora in which structure and content usually considered inappropriate for learners is labelled rather than censored.
• The resulting corpora can be used in the classroom and for research as well as for training data for machine learning algorithms.
• 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.

References

Acknowledgements
The authors acknowledge the financial support from the Portuguese national funding agency, FCT - Foundation for Science and Technology, LDA, grant number ‘UID/CLC/04439/2020’ and the Slovenian Research Agency (research core funding P1-0041, Language Resources and Technologies for Slovene, and project funding No. J2-4539, Empirical foundations for dignity-supported development of writing skills). The research received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 750103. This study has also been supported by the CLARIN Resource Families Project Funding.