The CrowLL project - Manually-annotated corpora for teaching and learning purposes of Brazilian Portuguese, Dutch, Estonian, and Slovene

Tanara Zingano Kuhn, CELGA-ILTEC/University of Coimbra
Carole Tiberius, Dutch Language Institute
Špela 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-ILTEC/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 datasets 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 unquestionably 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 corpus monitoring:

- many words are polysemic = shortcomings to rule-based approaches to automatically identifying offensive words
- problems identified as structural errors via automatic error detection = not 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 providing manually-annotated corpora of Brazilian Portuguese, Dutch, Estonian, and Slovene.
- Develop a crowdsourcing-based game for further corpora growth.

Manual annotation

Data preparation:

1. Source corpora:

- Brazilian Portuguese: Timestamped JSI web corpus 2014-2021 Portuguese (Trampuš & Novak, 2012) – approx. 3.2 billion words (only Brazil subcorpus)
- Dutch: Timestamped JSI web corpus 2014-2021 Dutch (Trampuš & Novak, 2012) – approx. 1.3 billion words
- Estonian: Estonian National Corpus 2021 (Koppel & Kallas, 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 formula that assigns numerical score to each corpus sentence 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 accordingly, based on weighted sum.
- For the present project, some classifiers are used in all languages, while others are language-dependent.

3. Lemma lists

- First, preparation of a list of 100 words in English, then its translation to Brazilian Portuguese, Dutch, Estonian, and Slovene.
- Lemmata 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 piss* (20 words);
  - Grey = words that are offensive or vulgar in some of the meanings, as well as words with potentially sensitive content, e.g.: *cow*, *drunk*, *suicide*, *fanatic*, *depressed*, *to molest* (60 words);
  - White = words that would typically not be considered offensive, vulgar or sensitive from the perspective of our labelling task, e.g.: *year*, *world*, *service*, *new*, *to say*, *to see* (20 words).

Data extraction:

For each language:

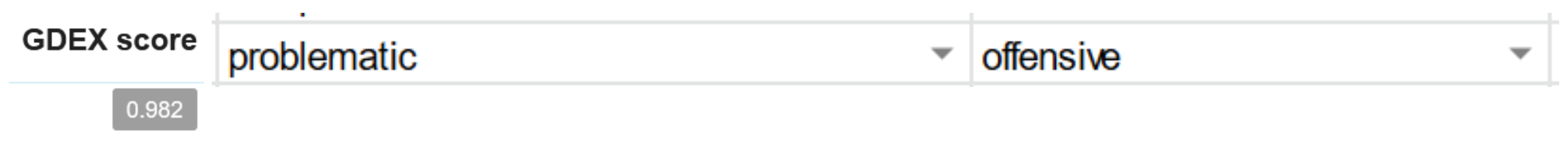
- Use GDEX configuration to extract from the source corpus the top 200 sentences per lemma of the lemma list;
- Deduplicate sentences (per lemma);
- Take the top 100 sentences (per lemma) from the remaining, totalling 10.000 sentences.

Format and Availability:

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

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

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


GDEX score
0.982
problematic
offensive

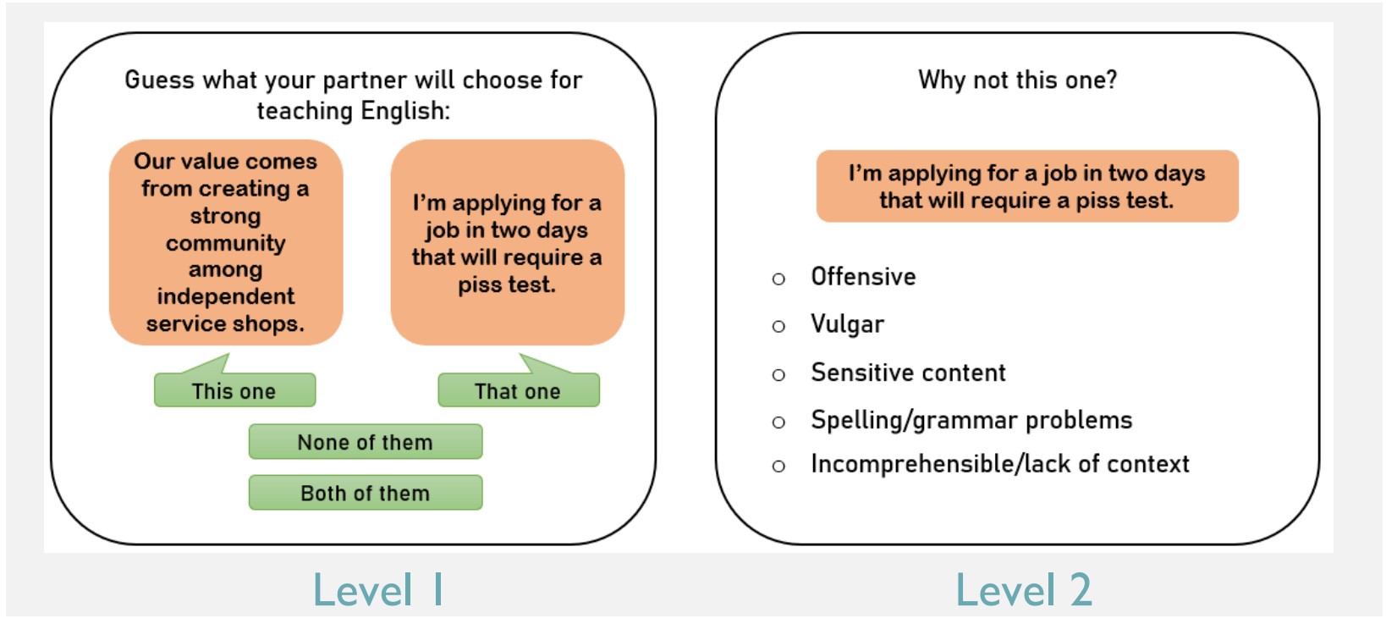
Game development

The CrowLL game:

- Game with a Purpose approach.
- Available as a webpage and mobile app.
- Single-player and dual-player mode.
- Type of crowdsourced work = crowdrating game (Morschheuser 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 enthusiastic!): players mark in the sentence what they consider to be problematic.


Guess what your partner will choose for teaching English:
Our value comes from creating a strong community among independent service shops.
I'm applying for a job in two days that will require a piss test.

Why not this one?
I'm applying for a job in two days that will require a piss test.
○ Offensive
○ Vulgar
○ Sensitive content
○ Spelling/grammar problems
○ Incomprehensible/lack of context

- Asynchronous (packages) and synchronous modes (bots)
- Scoring mechanisms: individual score from consecutive work; cooperative score based on agreement of the player in teams/partnerships.

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 *cleaned.*
- 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

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)*, 425–432.

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

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. V: Calzolari, N. (ur.). *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*. ELRA - European Language Resources Association.

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.

Trampuš, M., & Novak, B. (2012). The Internals of an Aggregated Web News Feed. *Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).*

Acknowledgements

The authors acknowledge the financial support 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 programme under grant agreement No. 731015. This study has also been supported by the CLARIN Resource Families Project Funding.