From student writing to language tools and teaching materials: where does crowdsourcing come in?

Špela Arhar Holdt

CJVT - Centre for Language Resources and Technologies, University of Ljubljana, Slovenia
Outline

• The Communication in Slovene project
• Developmental corpus Šolar
• Pedagogical Grammar Portal
• Šolar 2.0
• Crowdsourcing tasks: Transcription, teacher corrections, grammar exercises
• Conclusion
The Communication in Slovene project

- A national endeavour aimed at establishing language resources for Slovene.
- Project leader: Simon Krek.
- Results: reference corpora of written and spoken Slovene, linguistic annotation tools (tagger, parser), a lexical database of Slovene, several language-related web portals.
- http://www.slovenscina.eu/
- Activities aimed at improving the infrastructure for teaching Slovene as L1: Developmental corpus Šolar & Pedagogical Grammar Portal. Why the focus on L1?
Developmental corpus Šolar

- Texts written by Slovene elementary and secondary school students (aged 12–18 years) as part of their coursework (essays, tests etc.) between 2009 and 2010.
- Sampled on the national level: includes all Slovene regions and different school types.
- Šolar is comprised of 2,703 texts or 939,243 words.
- Text collection was conducted in close cooperation with the teachers, who have provided photocopies of the students’ texts (65 teachers from 39 schools).
- 56% of the texts include the teachers’ feedback, namely their corrections of the students’ language errors (spelling, morphology, syntax, vocabulary, punctuation, etc.).
Pedagogical Grammar Portal

• Teacher corrections (cca. 35,000) in Šolar were manually subcategorised into (nearly 700) problem types students encounter while writing in standard Slovene.

• For a selection of the most typical problems, a set of ready-to-use corpus-based teaching materials was made available to the teachers and students in the form of an interactive multimodal online resource.

• A new approach to Slovene language didactics.
A new approach to Slovene language didactics

- Prioritizing teaching content according to the frequency of language errors in student writing;
- Conceptualizing explanations from specific language problems instead of the grammar system;
- Using authentic corpus examples to support explanations;
- Including a high number of interactive corpus-based exercises;
- Representing language use as it appears in various genres (written and spoken, standard and non-standard);
- Adopting different approaches to region-specific language problems;
- Allowing for an individualized approach to developing students’ language competence.
1. Teachers evaluated Šolar and Pedagogical Grammar Portal as very relevant and useful for their work.

2. The preparation of both resources was extremely time consuming.

Which steps in the process would benefit most from crowdsourcing?
Šolar 2.0

• Funded by the Slovene Ministry of Culture (2015-2018) with the following goals: double the size of the Šolar corpus; revise and upgrade the categorisation system for teacher corrections; optimise the corpus building procedure.

• Project leader: Iztok Kosem.

• [http://solar.trojina.si/](http://solar.trojina.si/)

• In accordance with the crowdsourcing activities in Slovene lexicography (poster session!), Pybossa was chosen as the platform to test the crowdsourcing tasks:
  • Text transcription and digitalisation.
  • Annotation and categorisation of teacher corrections.
In the first task, the texts are transcribed.

Subsequent similar tasks for adding teacher corrections and feedback.
Teacher corrections

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Error and correction</th>
<th>No. of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonants</td>
<td>Dialectics</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Redundant</td>
<td></td>
<td>166</td>
</tr>
<tr>
<td></td>
<td>Omitted syllables</td>
<td></td>
<td>549</td>
</tr>
<tr>
<td></td>
<td>Substituted</td>
<td></td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>Substituted-lgh</td>
<td></td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Substituted-mn</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Substituted-sz</td>
<td></td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>Substituted-E</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Substituted-td</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Vowels</td>
<td></td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Redundant</td>
<td></td>
<td>236</td>
</tr>
<tr>
<td></td>
<td>Omitted syllables</td>
<td></td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>Substituted</td>
<td></td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Substituted-ao</td>
<td></td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Substituted-ei</td>
<td></td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Substituted-uo</td>
<td></td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Letter combinations</td>
<td></td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>li</td>
<td></td>
<td>207</td>
</tr>
<tr>
<td></td>
<td>Nj</td>
<td></td>
<td>193</td>
</tr>
<tr>
<td></td>
<td>Redundant syllables</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Omitted syllables</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Duplicated letters</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Flip-over</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Prepositional variants</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Ikh</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Sz</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Bilabial [w]</td>
<td></td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Beginning of the word</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Middle of the word</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>End of the word</td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

Total Spelling Corrections: 3225

1.3 Ustnično-ustnični w

1.3.1 Na začetku besede

Gra ze problem, kjer učenci neustrezno črkujejo besede, ki se začnejo na u- oz. w-. Kategorijo smo uvrstili pod ’ustnično-ustnični w’, čeprav je to v fonološkem smislu morda pomenovitev, na katero moramo v literaturi opozoriti. V označenem korpusu se pojavlja bodisi zapis u- namesto w- (šol; v) ali obratno: w- namesto u- (vseb). Kombinacije z drugimi črkami se v označenem vzorcu ne pojavljajo. Če bi se, bi jih uvrstili v to kategorijo (npr. *ga je očela domov").

**CRK/W/začetek:**

- Vsedla i Usedla sva se na tla in začel je govoriti.
- Zato bi raje pustila, da meni uzelemo i vzamejo življenje, kot da ga jaz jemljem drugim.
- A na žalost je mama silila ter me napadala z vprašanj i vprašanj i v tistem trenutku sem se reja zlagala.
- Seveda pa to, da sta Antigona in Ismena pokopali Polinjek ni bilo uzele i više vladarju Kreunu.

Primere, kjer je problematični fonem izpuščen, uvrstimo samo pod izraz konsonanta ali vokala, ne pa tudi pod zapis v na začetku besede.

**CRK/KONZ/zpust:**

- V nekaj sekundah pridvi oče in prasha i vpraša mamo kaj se dogaja.

- 2018: A revised categorisation system and annotation guidelines.
- Automatic preprocessing: categories assigned automatically by a statistical model, then manually corrected if needed (facilitating a yes-no crowdsourcing task).
For the Pedagogical Grammar Portal

• Selection of examples for the grammar chapters and the language exercises.

The choice of the appropriate corpus for the specific exercise; considerations: sentence length, typicality for the language problem, content restrictions (political correctness, sensitiveness of certain topics, inclusion of personal names, terminology, level of comprehensibility and motivational value); all the correct answers to the exercise have to be provided.

• Before: The use of partially adapted GDEX-Good dictionary examples and filters in the SketchEngine tool.

• After: a series of explicit crowdsourcing tasks and/or a webpage where teachers can select and modify examples for their own need while the results are furthermore exploited for the grammar chapters.

• Very similar observations as in Pilán et. al (2017) in the context of L2 learning.
Conclusion

• Šolar and the Pedagogical Grammar Portal can be seen as the benchmark for the development of language resources and teaching materials for corpus-based L1 teaching.

• Crowdsourcing will play an important role in future development, facilitating the inclusion of the wider community as well as the optimisation of workflows for participating experts.

• Potential crowdsourcing participants for the anticipated tasks: university students, linguists; crowdsourcing can be included in coursework (elementary and high-school students); teachers can be provided with online tools designed to adapt their work for further purposes.

• Digitalisation of the entire process: writing, correcting, and feedback categorisation can take place in a digital environment that records the statistical data on student development, provides synthetic feedback, and automatically refers students to language resources and exercises that address their individual needs.

THANK YOU!
References and resources

- PILÁN, Ildikó, Elena VOLODINA, Lars BORIN. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. To appear in the Special issue of Traitement Automatique des Langues (TAL) journal, Special issue on NLP for learning and teaching. [pre-print]
- Communication in Slovene: http://eng.slovenscina.eu/
- Solar 2.0: http://solar.trojina.si/
- Pybossa: http://pybossa.com/