Univerza *v Ljubljani*

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.

- Financed by the European Social Fund and the Slovene Ministry of Education, Science and Sports (2008-2013).

- 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.).

Available in an adapted version of the Sketch Engine concordancer: www.korpus-solar.net.

Šolar

trojína
ZAVOD ZA UPORABNO SLOVENISTIKO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST, KULTURO IN ŠPORT

Išči   Seznam besed   O korpusu   Pomoč

Shrani   Možnosti prikaza   Razvrščanje   Filter   Frekvence   Kolokacije   Vizualiziraj   ?

Iskalni niz **nebi** 278 > Premešaj 278 (240.58 na milijon) ℹ

Stran 1 od 14 Pojdi   Naslednja | Zadnja

strokovna šola, 4. letnik, Ljubljana	napisano, narisano vse mogoče. Od trgovskih izdelkov, brez katerih si življenja naj **nebi**	*ne bi* mogli predstavljati; do manekenk, ki šopirijo svoje ustnice in predstavljajo,
gimnazija, 3. letnik, Gorica	takih *primerom* | *primerov* " | „ skorajda | " ni. Svet bi lahko bil veliko lepši, če **nebi**	*ne bi* bili vsi Polikarpi. Prav zaradi denarja, je na svetu veliko vojn. Lahko
gimnazija, 4. letnik, Gorica	ali se mora prilagoditi politiki in paziti |, kaj predstavlja javnosti, da **nebi**	*ne bi* zašel v težave in navzkriž oblastem. Sam zaključi, da bi morali vsaj umetniki
strokovna šola, 1. letnik, Maribor	Poročil ju je Lorenzo |, vendar na skrivaj |, saj nobeden od staršev **nebi**	*ne bi* smel izvedeti. In to bi *jas* | *jaz* spremenil. Namesto skrite poroke |,
poklicna šola, 3. letnik, Novo mesto	*stališča* | *če* | Če bi | *jaz* imela nekoga rada in da bi bila skupaj, nikoli **nebi**	*ne bi* pustila |, da | *naju gdo* | *kdo* loči. Za ljubezen bi storila vse
strokovna šola, 2. letnik, Novo mesto	njegovo ljubeznijo. Njuna največja želja je, da bi bila lahko skupaj in da se **nebi**	*ne bi* rabila tako skrivati in da bi lahko zaživela skupaj. Ampak, da se to zgodi, morata
gimnazija, 4. letnik, Gorica	ta res manj vredna kot zgodovina?? Sama bi ravnala drgače. Po vesti in srcu. Katarine **nebi**	*ne bi* ubila. Kako sploh lahko ubiješ nekoga |, ki ga ljubiš?? Ne razumem.
strokovna šola, 4. letnik, Ljubljana	bil Simon pri tem nemočen, bi lahko Simon hitreje popustil in mogoče do te drame tudi **nebi**	*ne bi* prišlo in bi lahko Simon še zmeraj hodil, vendar je bil Simon kot oseba trmast
strokovna šola, 1. letnik, Ljubljana	tudi veliko izgubiš. Če bi najdel psa in ga prepeljal domov |, ampak mi starši **nebi**	*ne bi* dovolili |, bi se lahko upiral in s tem laho izgubil npr *nebi* | *ne bi*
strokovna šola, 2. letnik, Ljubljana	, da se *nebi* | *ne bi* nič zgodilo. Če bi jaz bil na Izidorjevem mestu |, očetu **nebi**	*ne bi* nikdar odpustil takšnega dejanja, saj bi me tako rekoč pohabil in *nebi* | *ne bi*
strokovna šola, 1. letnik, Ljubljana	Medeja je bila po mojem mnenju kruta ženska, a če dobro premislimo, mi verjetno res **nebi**	*ne bi* ubili osebe, bi pa se verjetno maščevali ter bili zelo jezni. Na svetu
gimnazija, 1. letnik, Gorica	ne bi ubil? Zakaj ni njega zaščitil? Če bi zaščitil Abela, bi bila oba živa in Kajnu **nebi**	*ne bi* bilo treba bloditi po svetu. Če je Bog lahko zaščitil Kajna, da ga *nebi* | *ne bi*
strokovna šola, 4. letnik, Ljubljana	tistem okolju in če *nebi* | *ne bi* bilo še osebje prijazno, bi se zaprla vase, oddaljila, **nebi**	*ne bi* več imela volje do življenja. Tako bi postala kot tiste osebe, ki se tam zares
gimnazija, 2. letnik, Gorica	vodita enkrat razum, drugič srce. Ženske tudi pomagajo nekaterim moškim likom, da se **nebi**	*ne bi* preveč zapletli ali si škodili. Tako je v četrtem dejanju, kjer baron odkrije
strokovna šola, 2. letnik, Ljubljana	vedno zaklepal. Nekega dne je Izidor opazil, da so vrata odprta in ni si mogel kaj, da **nebi**	*ne bi* pokukal, kaj oče tako skrbno skriva. Vstopil je v klet in si hotel vzeti en zlatnik
gimnazija, 2. letnik, Gorica	obnaša, kot da se to njemu ne dogaja. Obnaša se |, kot |, da to **nebi**	*ne bi* bilo mogoče. Ko so na sodišču |, se *prav tako* | tudi Zmešnjava
gimnazija, 1. letnik, Ljubljana	nobenega pomena več. Kreon se je trmasto odločil, da bo Antigona umrla. Kot da težav §§§e **nebi**	*ne bi* bilo dovolj |, je do njega stopil še Hajmon |, njegov sin |, |
strokovna šola, 1. letnik, Ljubljana	lahko upiral in s tem laho izgubil npr *nebi* | *ne bi* mi dovolili iti *vn* | *ven* |, **nebi**	*ne bi* smel igrati racunalnika in se kaj. Ampak *usaj* | *vsaj* obdržal bi psa in *nebi* | *ne*
strokovna šola, 4. letnik, Krško	in §§§ pomisli, da je to mogoče *zaradi avtomobilčka* | *povezano z otrokom* . Citiram: Če **nebi**	*ne bi* bilo tistega z avtomobilčkom, bi pomislil, da ima bolne oči. Moje mnenje
strokovna šola, 1. letnik, Ljubljana	so se odpravili z baklami v jamo. Med potjo so *usi* | *vsi* molili |, da se jim **nebi**	*ne bi* kaj zgodilo. Hodijo, hodijo .. Odisej seveda vodi skupino |, na

Stran 1 od 14 Pojdi   Naslednja | Zadnja

Lexical Computing

2.36.5-SkE-2.151.5-3.99.3

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.

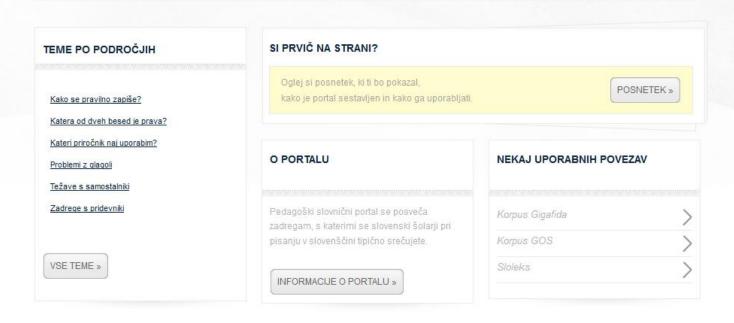- A new approach to Slovene language didactics.

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 individualised approach to developing students' language competence.
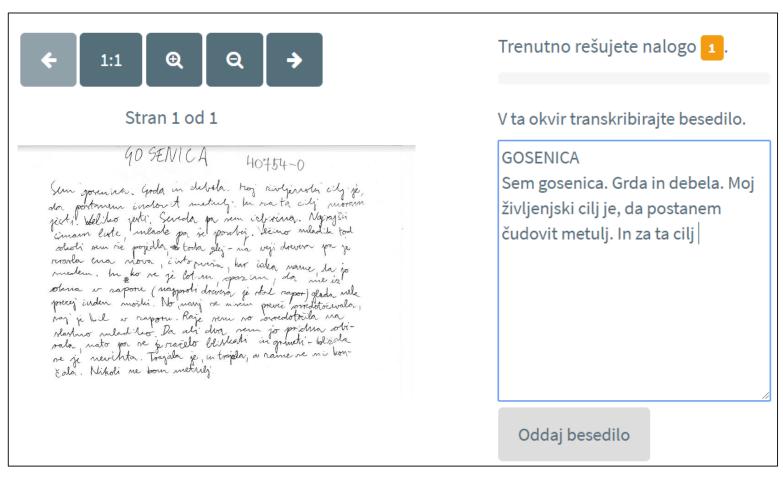
Available at: http://slovnica.slovenscina.eu/.

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/

- 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

Spelling errors in corpus Šolar

	Category	Subcategory	Error and correction	No. of occurences
1				
2	Consonants	Diactitics		90
3		Redundant		166
4		Ommited syllables		549
5		Substituted		158
6		Substituted-kgh		51
7		Substituted-mn		40
8		Substituted-sz		97
9		Substituted-šž		36
10		Substituted-td		60
11	Vowels	Redundant		75
12		Ommited syllables		236
13		Substituted		94
14		Substituted-ao		95
15		Substituted-ei		89
16		Substituted-uo		64
17	Letter combinations	ij		65
18		LJ		207
19		NJ		193
20		Redundant syllables		8
21		Ommited syllables		23
22		Duplicated letters		44
23		Flip-over		22
24	Prepositional variants	kh		22
25		sz		372
26		v		22
27	Bilabial [w]	Beggining of the word	use -- vse	26
28			ušeč -- všeč	13
29			usi -- vsi	11
30			Usak -- Vsak	8
129		End of the word		94
130		Middle of the word		54
131	TOTAL SPELLING CORRECTIONS			3225

1.3 Ustnično-ustnični w

1.3.1 Na začetku besede

Gre za problem, kjer učenci neustrezno črkujejo besede, ki se začnejo na **u-** oz. **v-**. Kategorijo smo uvrstili pod 'ustnično-ustnični w', čeprav je to v fonološkem smislu morda poenostavitev, na katero moramo v literaturi opozoriti. V označenem korpusu se pojavlja bodisi zapis **u-** namesto **v-** (*ušeč*) ali obratno: **v-** namesto **u-** (*vsesti*). Kombinacije z drugimi črkami se v označenem vzorcu ne pojavljajo. Če bi se, bi jih uvrstili v to kategorijo (npr. **ga je ozela domov*).

CRK/W/začetek:

- **Vsedla** | **Usedla** sva se na tla in začel je govoriti.
- Zato bi raje pustila , da meni **uzamejo** | **vzamejo** življenje, kot da ga jaz jemljem drugim.
- A na žalost je mama silila ter me napadala z **uprašanji** | **vprašanji** in v tistem trenutku sem se raje zlagala.
- Seveda pa to, da sta Antigona in Ismena pokopali Polinejka ni bilo **ušeč** | **všeč** vladarju Kreonu.

Primere, kjer je problematični fonem izpuščen, uvrstimo samo pod izpust konzonanta ali vokala, ne pa tudi pod zapis w na začetku besede:

CRK/KONZ/izpust:

- V nekaj sekundah pridrvi oče in **praša** | **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 futhermore exploited for the grammar chapters.

- Very similiar 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 futher 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

- ARHAR HOLDT, Špela, Iztok KOSEM, and Polona GANTAR, 2017: Corpus-based resources for L1 teaching: The case of Slovene. A. Marcus-Quinn (ed.): *Handbook on Digital Learning for K-12 Schools*. Springer International Publishing. 91–113.
- ČIBEJ, Jaka, FIŠER, Darja, KOSEM, Iztok. The role of crowdsourcing in lexicography. I. Kosem (ed,). *Electronic lexicography in the 21st century : linking lexical data in the digital age : proceedings of eLex 2015 Conference*, Herstmonceux Castle, United Kingdom. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing. 70-83.
- KILGARRIFF, Adam, Pavel RYCHLY, Pavel SMRZ in David TUGWELL. 2004. The Sketch Engine. G. Williams in S. Vessier (eds.): *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France. Universite de Bretagne-sud. 105–116.
- KOSEM, Iztok, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko.
- KOSEM, Iztok, HUSAK, Miloš, MCCARTHY, Diana, 2011. GDEX for Slovene. I. Kosem et al. (ed.). *Electronic lexicography in the 21st century: new applications for new users: Proceedings of eLex 2011*, Bled, Slovenia. Ljubljana: Trojina, Institute for Applied Slovene Studies. 150-159.
- KREK, Simon, 2012. *Slovenski jezik v digitalni dobi = The Slovene language in the digital age*, (White paper). Heidelberg [etc.]: Springer.
- LEECH, Geoffrey, 1997: Teaching and language corpora: A convergence. A. Wichmann, S. Fliegelstone, T. McEnery and G. Knowles (eds.): *Teaching and language corpora*. London: Longmann. 1-23.
- 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]
- ROZMAN, Tadeja, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

- ARHAR HOLDT, Špela, Gaja ČERV, Polona GANTAR, Iztok KOSEM, Karmen KOSEM, Irena KRAPŠ VODOPIVEC, Simon KREK, Sara MOŽE, Tadeja ROZMAN, Ana Marija SOBOČAN, Mojca STRITAR KUČUK and Ana ZWITTER VITEZ, 2013. Pedagoški slovnični portal. [Ljubljana]: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://slovnica.slovenščina.eu/
- ROZMAN, Tadeja, Mojca STRITAR KUČUK, Iztok KOSEM, Simon KREK, Irena KRAPŠ VODOPIVEC, Špela ARHAR HOLDT and Marko STABEJ, 2012. Šolar. [Ljubljana]: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://www.korpus-solar.net/
- Communication in Slovene: http://eng.slovenscina.eu/
- Šolar 2.0: http://solar.trojina.si/
- Pybossa: http://pybossa.com/