Adapting the Sketch Engine interface for the purposes of an error corpus

Iztok Kosem

Trojina, Institute for Applied Slovene Studies
Iztok.kosem@trojina.si

Learner corpora

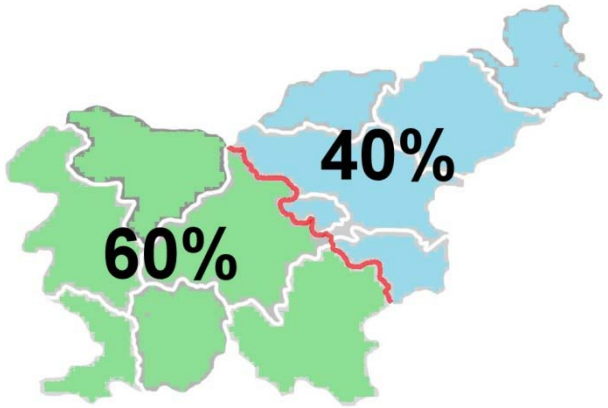
- more and more popular
- large projects (ICLE) and smaller ones (individual researchers)
- no single tool used for annotation or visualisation
- L1 “learner” corpora rare (e.g. Chyby; Busta et al., 2009)

Šolar corpus

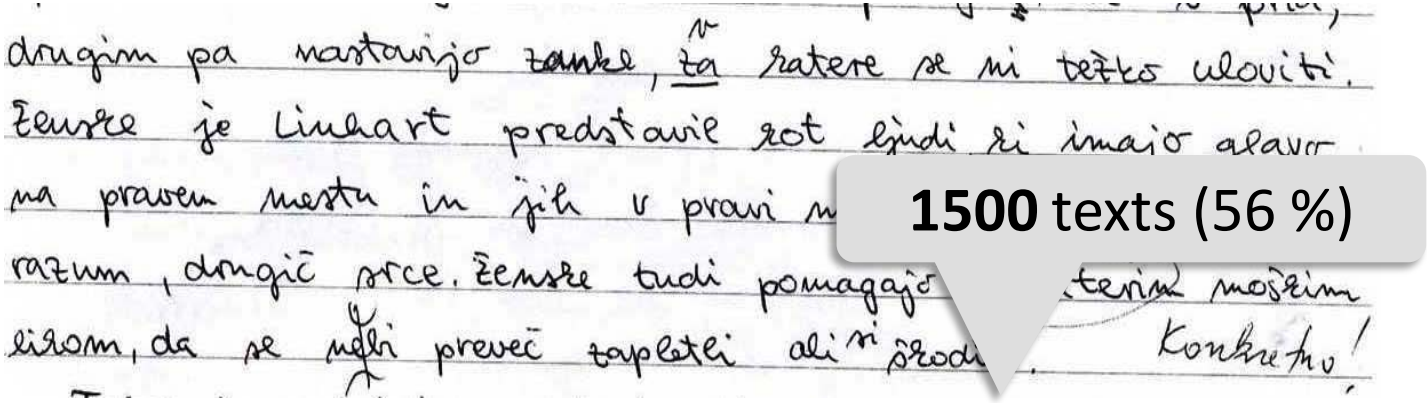
- Communication in Slovene project
  - Website: http://www.slovenscina.eu
  - Leading partner: Amebis, d. o. o., Kamnik
  - Duration: June 2008 - December 2013
  - Total value: 3,2 million Euro
- Lack of language resources in Slovene that address common language problems of students
- Identifying problems – following learner corpora example

corpus of student writing (Šolar)

- 2703 texts
- size: 963,000 words
- Balanced by region
- Texts by school subject:
  - 84,2% Slovene
  - 15,8% other subjects
- Text types:
  - essay (79,1%)
  - test (13,8%)
  - letters, complaint etc. (7,1%)

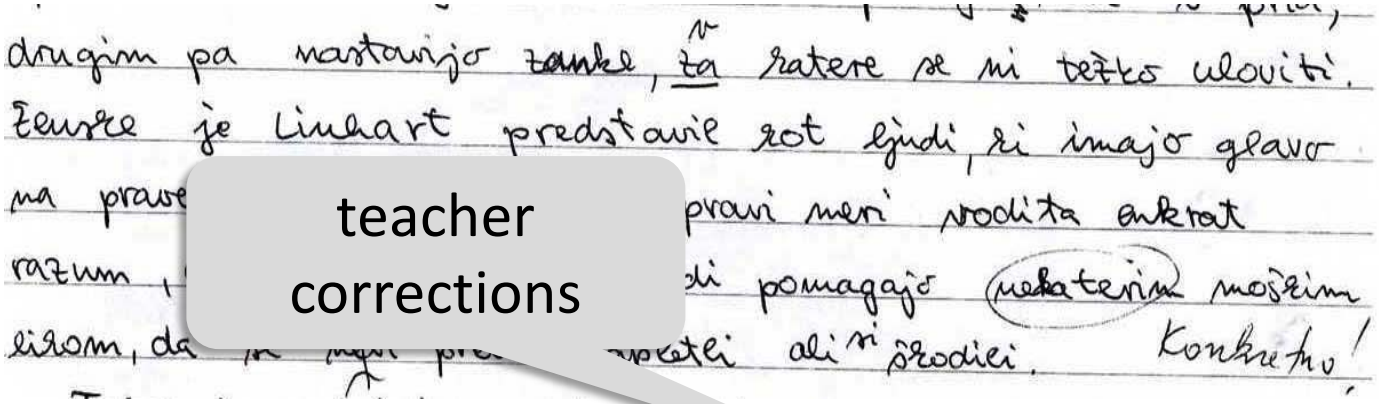


Transcription + annotation



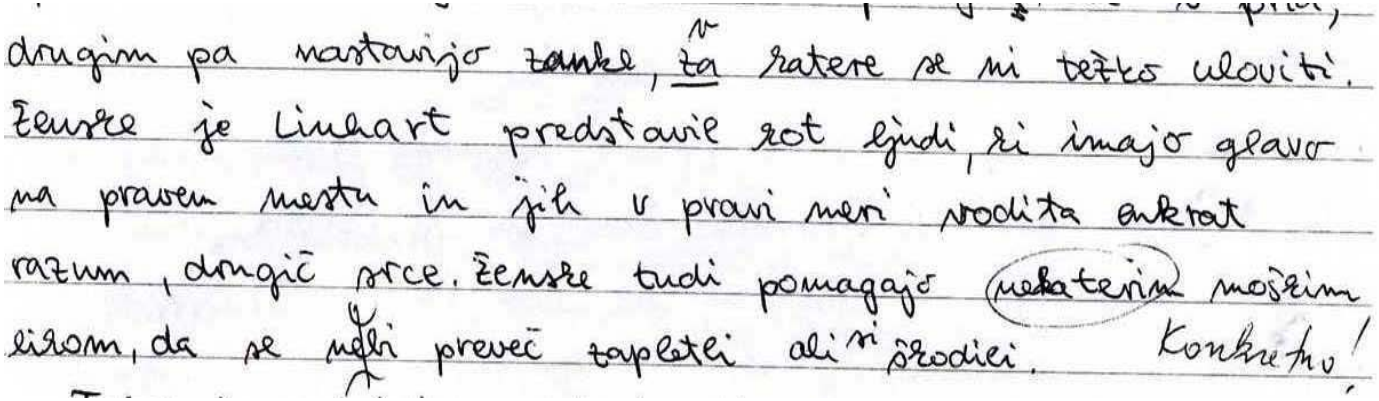
drugim pa nastavijo zanke, **za** katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce. Ženske tudi pomagajo nekaterim moškim likom, da se **nebi** preveč zapletli ali si škodili.

Transcription + annotation



drugim pa nastavijo zanke, **za** **v** katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce. Ženske tudi pomagajo nekaterim moškim likom, da se **nebi** **ne bi** preveč zapletli ali si škodili.

Transcription + annotation



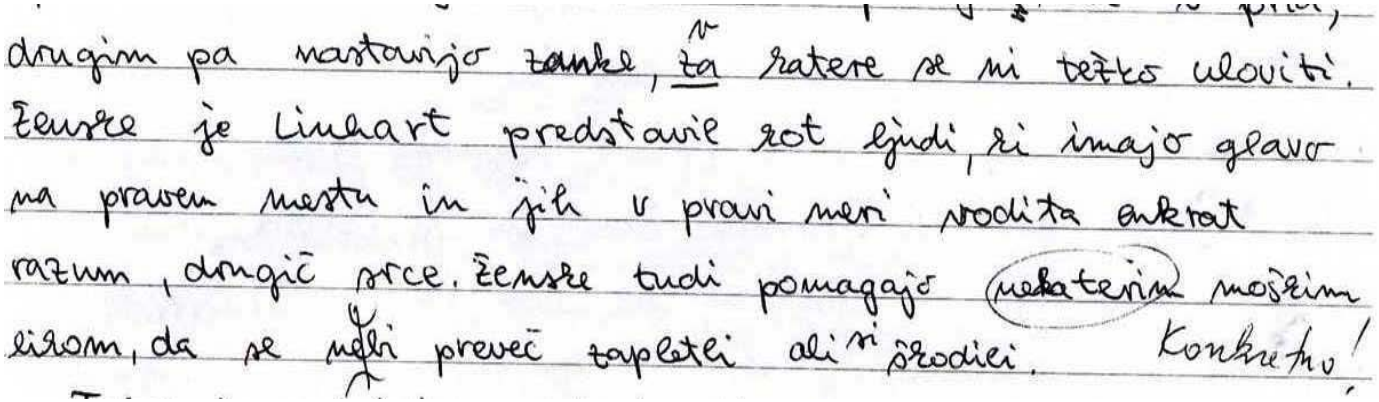
drugim pa nastavijo zanke, **za** **v** katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat raz

Ženske tudi pomagajo <u k="**Konkretno!**" l="**obkroženo**">nekaterim</u> moškim likom, d

nebi **ne bi** preveč zapleti ali si škodili.

teacher annotations and comments

Transcription + annotation



drugim pa nastavijo zanke, <u tip="B">**za** **v** </u> katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce. Ženske tudi pomagajo <u k="**Konkretno!**" l="**obkroženo**">nekaterim</u> moškim likom, da se <u tip="Z" podtip="SN">**nebi** **ne bi** </u> preveč zapletli ali si škodili.

error categories

VOCABULARY		**3807** (434)
MORPHOLOGY		**3618** (1241)
SYNTAX		**6190**
	word order	1265 (212)
	missing text	1607 (240)
	redundant text	2665 (588)
	structure	653 (61)
ORTHOGRAPHY		**21420**
	spelling	2672 (18)
	abbreviations	23
	punctuation	15371 (775)
	lower/upper case	2125 (872)
	together/apart	1179
	numeral	50

Error analysis (2010/11)

	Set
ar po eni stroni oče enako počel z njegovim očetom, zdaj pa pričakoval, da bo njegov ı. Tukaj bi lahko držal pregovor, <u1 tip="S" podtip="STR">tak kot oče, tak kot akšen sin</p1></u1>. Saj se vsi učimo obnašanja od staršev. Niti si ni zaslužil, da	primerjava
> let, je ime?la že otroka z bogatim Ožbejem, <u1 tip="S" podtip="STR">ki je bila e bila</p1></u1> zaljubljena. A <u1 tip="S" podtip="STR">od Ožbeja /p1></u1> je bil zelo žalosten, ker se bil <u1 tip="S" podtip="IZP"><p1>zaljubil</	Jezusov
drugačna<u1 tip="Z" podtip="LOC">,<p1></p1></u1> kot pri Romeu in Juliji, saj vojega očeta. A ko je izvelel za <u1 tip="S" podtip="STR">smrt svojega t</p1></u1>, je prisegel, da bo naredil vse, da <u1 tip="S" podtip="BR">bo	Jezusov
da bi sinu kupila samokres. Ko sta se vrnila domov<u1 tip="Z" podtip="LOC"><p1> ı1></u1> je Izidor našel na tleh <u1 tip="S" podtip="STR">kovanec od nec</p1></u1>. Mislil si je, da mu bo <u1 tip="S" podtip="ODV">tale<p1></p1></	Jezusov
o">te</u1> problemi začeli <u1 l="vijugasto podčrtano">večati</u1><u1 tip="Z" p1>,</p1></u1> saj je takoj po <u1 tip="S" podtip="STR">smrti njegovega :i</p1></u1><u1 tip="Z" podtip="LOC">,<p1></p1></u1> zavzel njegov prestol stric	Jezusov

Sketch Engine and learner corpora

- Cambridge University Press

yours sincerely, <#S>	**Hy**	| Hi </#S> July! I think I left my phone there
bring </#RV> anything. It doesn't <#S>	**metter**	| matter </#S> . <#UT> For <#FV><#RV> an
MP> | . </#MP> See you <#S>	**tomorow**	| tomorrow </#S> <#MP> | . </#MP>
on | at </#RT><#R> 19 | 7 </#R><#S>	**a'clock**	| o'clock </#S> , and you can bring a <#RP>
er to bring <#RD> the | a </#RD><#S>	**pensil**	| pencil </#S> and <#RD> the | a </#RD> rul
e and it finishes on 14 July. Don't <#S>	**forghet**	| forget </#S> your art book. If we
. This bed will cost three <#S>	**tousand**	| thousand </#S> and fifteen pounds. Thank
u soon My new house is <#S>	**wonderfool**	| wonderful </#S> ! It is on the beach in Calif
ornia. Here the weather is always <#S>	**beatiful**	| beautiful </#S> ! It is very big <#RP> , | ; <
) <#AGN> floor | floors </#AGN> : <#S>	**downstair**	| downstairs </#S> there <#AGV> are | is </

Šolar into SkE project

- funded by the Ministry of Education, Science, Culture and Sports
- Aim: making the corpus available to teachers, makers of language materials, and researchers
- Team:
  - Trojina: Iztok Kosem, Špela Arhar Holdt
  - Lexical Computing: Vojtěch Kovář, Vit Baisa, Adam Kilgarriff
  - Amebis: Miro Romih

Annotations not included in the SkE version

- Comments:

<u k=”*Not connected to the rest of the text*.">
<u k=“*What do you mean by that?!!*”>
<u k=“*too informal!*”>

- Symbols:

<u l=”*underlined*">luksuz</u>

- Formatting correction:

<u obl=“new paragraph”>

Changes needed

- converting error tags to SkE format
- one level for categories and sub-categories

<u tip="Z" podtip="SN">**nebi** **ne bi** </u>

*<Z-SkupajNarazen>***<err>nebi </err>|<corr>ne bi</corr>** *</Z-SkupajNarazen>*

- no advanced functions (word sketches, thesaurus); *available in regular version*
- Complete localisation, including help (separate webpage set up)

How does it look?

Šolar

trojína
ZAVOD ZA UPORABNO SLOVENISTIKO
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST, KULTURO IN ŠPORT

Iskanje Seznami O korpusu Pomoč

Shrani Možnosti prikaza Razvrščanje Vzorec Filter Frekvence Kolokacije

Iskalni niz noben 34 (29.4 na milijon)

Stran 1 od 2 Pojdi Naslednja | Zadnja

Besedisce,4. letnik,Gorica	Jerneju Jerobniku je izšla tudi v knjigi | , a se ni zanjo	noben	| *nihče* zanimal. Še zadnji poizkus v njegovem življenju, ki mu
Besedisce,1. letnik,Gorica	kdor me bo našel, me bo ubil." Bog pa mu je dal znak, da ga	noben	| *nihče* ne bi ubil. Tako kot se Bog kaže v tej zgodbi,
Besedisce,1. letnik,Gorica	Eteokla. V boju sta oba brata umrla | , in ker ni bilo	nobenega	| *nikogar* drugega, ki bi bil sposoben vladati | , je
Besedisce,1. letnik,Ljubljana	da na skrivaj celo mesto žaluje za Antigono, a *se* |	noben	| *nihče* tega ne upa priznati, da ga *nebi* | *ne bi* doletela enaka
Besedisce,1. letnik,Ljubljana	*kaj* | *koga* . Težave *ponavadi* | *po navadi* rešujemo s pogovorom in	noben	| *nihče* ne umre. Vendar pa včasih reševanje *le teh* | *težav* ni lahko
Besedisce,1. letnik,Ljubljana	krono, ki jo je tudi dobil. Kralju je nalil strup v uho, vendar	noben	| *nihče* ni odkril krivca. Meni je krivico zaupal duh mojega
Besedisce,1. letnik,Ljubljana	popolnoma brez učinka, saj sta za ljubezen potrebna dva. Brez	nobenega	| *vsakega* truda se ne da priti nikamor, vendar Ofelija tega žal
Besedisce,2. letnik,Gorica	| , da odidejo , | oziramo da se vdajo. Na koncu ga	noben	| *nihče* ne zapusti in se skupaj bojujejo proti sovražniku. Pogumen
Besedisce,2. letnik,Gorica	| , kar se dogaja po svetu v določenih državah ni človeško,	noben	| *nihče* si ne zasluži takšnega obnašanja, kakršnega so deležni
Besedisce,1. letnik,Maribor	*Juliji* | *Julijo* | , ki je *mrtva ležala* | *ležala mrtva* in	noben	| *nihče* mu ni povedal | , da v resnici ni mrtva. Ko se
Besedisce,3. letnik,Ljubljana	iz človeka se lahko spremenimo v junaka, ki je glavni in mu	noben	| *nihče* nič ne more. V našem stvarnem življenju imamo ljudje
Besedisce,2. letnik,Ljubljana	njegovo sobo, ki jo je imel v kleti, kajti tja ni smel hoditi	noben	| *nihče samo* | *razen on* | *njega* . Izidor *kazen sprejme* | *je kazen sprejel*
Z-MalaVelika,2. letnik,Ljubljana	| , kajti v današnjih odnosih je čisto drugače | .	nobeden | Noben	| *Noben* oče *nebi* | *ne bi* svojemu sinu odrezal prst, ker naj bi
Besedisce,2. letnik,Ljubljana	da se nekaterim "urejenim" družinam dogajajo pretepi in noče	noben	| *nihče* priznati tega. Nasilja ne podpira nihče.
Besedisce,4. letnik,Ljubljana	biti njegovo razmišljanje | , medtem , | ko mu	noben	| *nihče* ne verjame, temveč ga vsi le obsojajo. Tega si *noben* |
Besedisce,4. letnik,Ljubljana	mu *noben* | *nihče* ne verjame, temveč ga vsi le obsojajo. Tega si	noben	| *nihče* izmed nas ne mora predstavljati, saj še nismo bili v
S-BesedniRed,2. letnik,Novo mesto	pa se je to zelo spremenilo. Nezakonska mati , | ni	noben več	| *več noben* tabu. Družba jih zelo lepo sprejme | , brez
Besedisce,2. letnik,Novo mesto	sta oba parnerja srečna. Za današnji čas ljubezen ni večna,	noben	| *nihče* se ne potrudi. tako ni samo pri Ljubezni, tako je povsod
S-Struktura,2. letnik,Novo mesto	tista prava ljubezen samo enkrat v življenju. Na koncu življenja	nima nobenega pomena	| *ni pomembno* tvoje bogastvo ali naziv, pomembno je samo to,
Besedisce,4. letnik,Krško	človek tako mirno joka. Pripovedovalec je ugotovil tudi, da se	nobeden	| *noben* drug potnik ni zmenil za to gospo , | in niti

Stran 1 od 2 Pojdi Naslednja | Zadnja

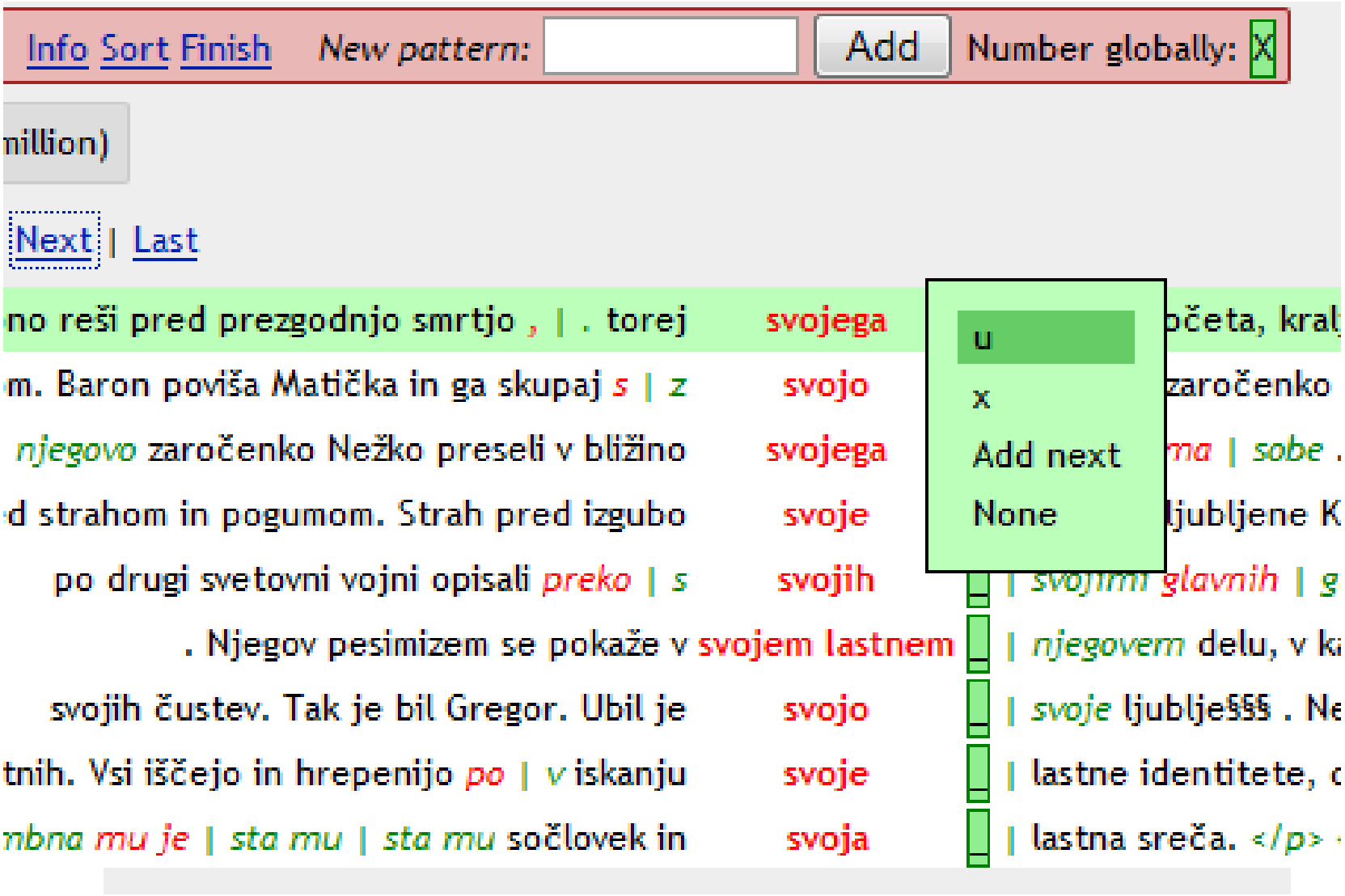
www.korpus-solar.net

sta oba brata umrla | , in ker ni bilo	nobenega	| *nikogar* drugega, ki bi bil sposoben vl
celo mesto žaluje za Antigono, a *se* |	noben	| *nihče* tega ne upa priznati, da ga *neb*
| *po navadi* rešujemo s pogovorom in	noben	| *nihče* ne umre. Vendar pa včasih reše
bil. Kralju je nalil strup v uho, vendar	noben	| *nihče* ni odkril krivca. Meni
aj sta za ljubezen potrebna dva. Brez	nobenega	| *vsakega* truda se ne da priti nikamor,
, | oziramo da se vdajo. Na koncu ga	noben	| *nihče* ne zapusti in se skupaj bojujej
vetu v določenih državah ni človeško,	noben	| *nihče* si ne zasluži takšnega obnašanj
, ki je *mrtva ležala* | *ležala mrtva* in	noben	| *nihče* mu ni povedal | , da v resnici n
emenimo v junaka, ki je glavni in mu	noben	| *nihče* nič ne more. V našem stvarnem
e imel v kleti, kajti tja ni smel hoditi	noben	| *nihče samo* | *razen on* | *njega* . Izido
anašnjih odnosih je čisto drugače | .	nobeden | Noben	| *Noben* oče *nebi* | *ne bi* svojemu sinu
n" družinam dogajajo pretepi in noče	noben	| *nihče* priznati tega. Nasilja

www.korpus-solar.net

What could be improved?

- Technical:
  - Visualisation of nested errors (example)
  - Search by error category
- Summarization of results:
  - Visualisation of frequency
  - Sample searches prepared and offered as **links** (perhaps a new SkE feature?)
    - In the case of Šolar, this would require annotating lower-level categories (in normal installation of SkE?)
    - **Linking** “analysis corpus” and “free” corpus?


Info Sort Finish
New pattern:
Add
Number globally:
million)
Next | Last
no reši pred prezgodnjo smrtjo , | . torej svojega
m. Baron poviša Matička in ga skupaj s | z svojo
njegovo zaročenko Nežko preseli v bližino svojega
d strahom in pogumom. Strah pred izgubo svoje
po drugi svetovni vojni opisali preko | s svojih
. Njegov pesimizem se pokaže v svojem lastnem
svojih čustev. Tak je bil Gregor. Ubil je svojo
tnih. Vsi iščejo in hrepenijo po | v iskanju svoje
nbna mu je | sta mu | sta mu sočlovek in svoja
u
x
Add next
None
očeta, kral
zaročenko
ma | sobe .
ljubljene K
| svojimi glavnih | g
| njegovem delu, v k
| svoje ljublje§§§ . Ne
| lastne identitete, c
| lastna sreča. </p>

Prospects

- Using annotation option for analysis;
  - advantage: allows for different classifications of errors
  - downside: annotations not recorded permanently
  - Solution: linking to CPA?
- Next project: Proof-reading corpus
  - corrected texts in Word – using Track changes
  - exporting data into xml
  - conversion into SkE compatible format