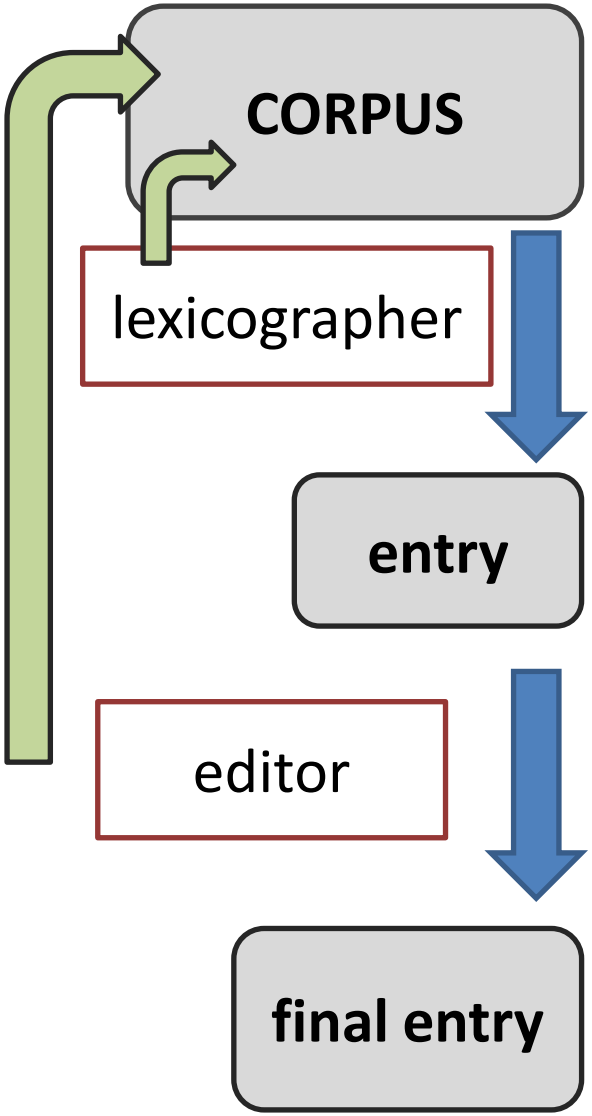
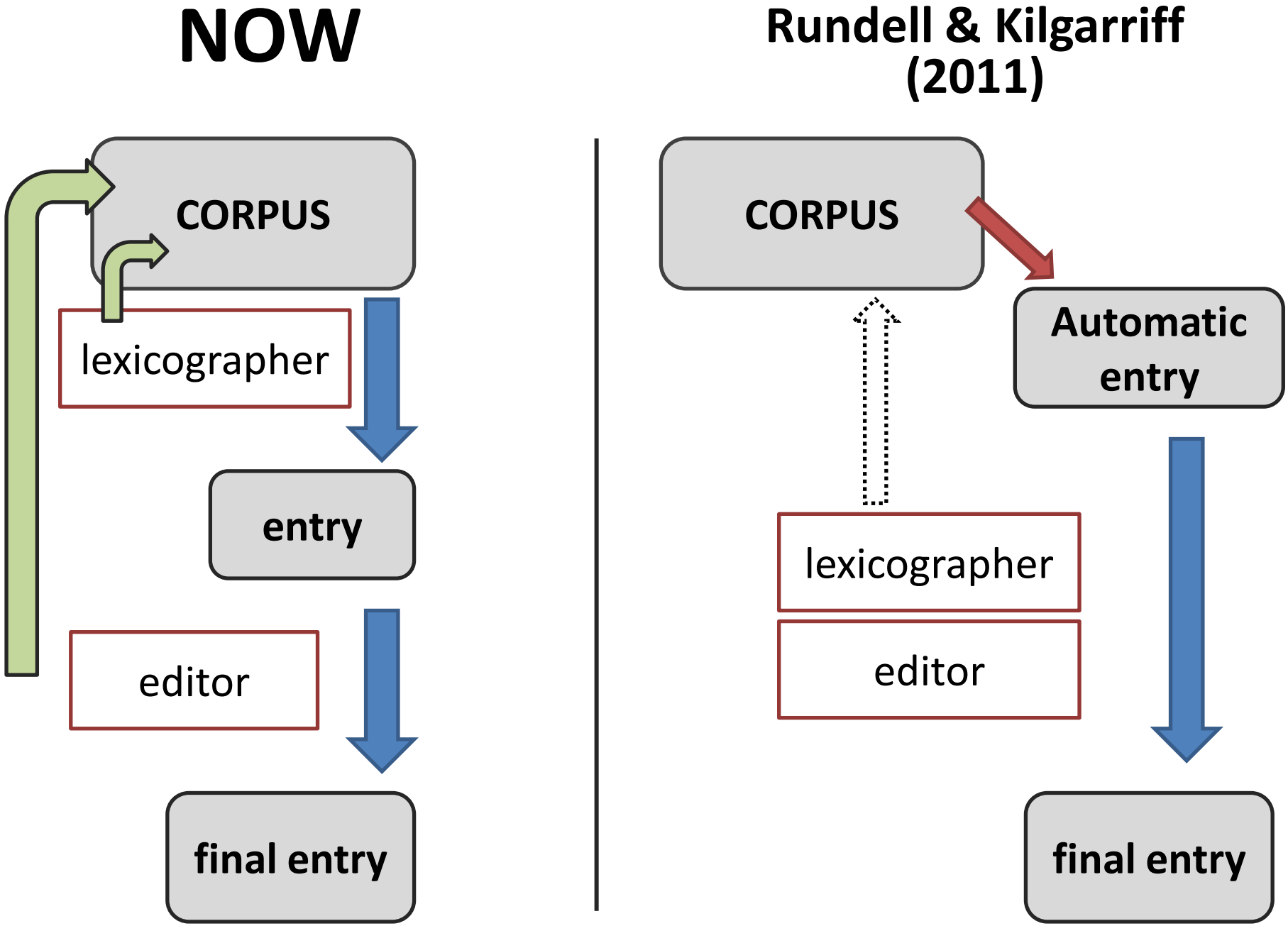
Monolingual Lexicography

Automatic extraction of data from corpora for lexicographic purposes Proposal for the Compilation of the Dictionary of Modern Slovene

Session 3

NOW

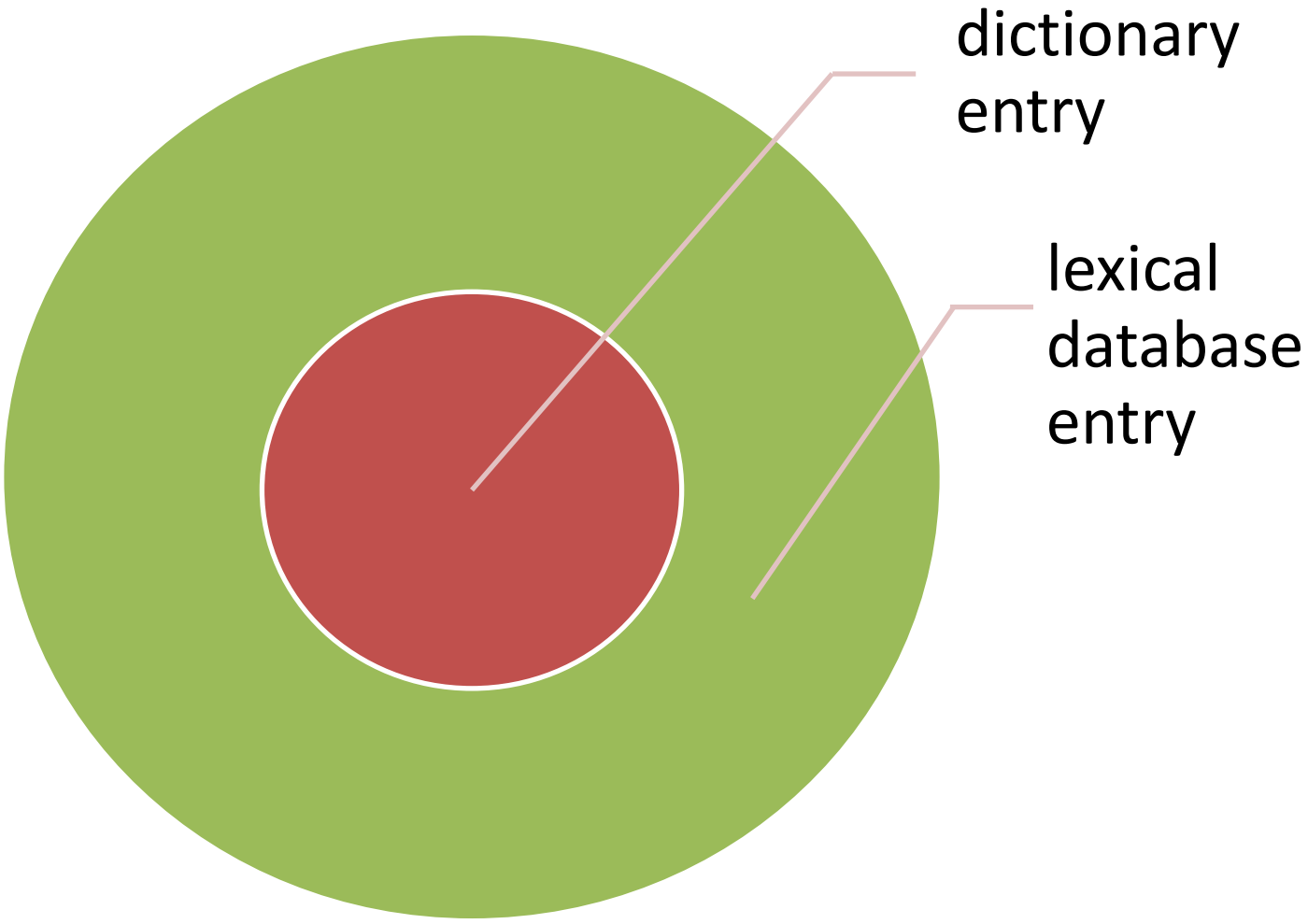



NOW
CORPUS
lexicographer
entry
editor
final entry
Rundell & Kilgarriff (2011)
CORPUS
Automatic entry
lexicographer
editor
final entry

Slovenian situation

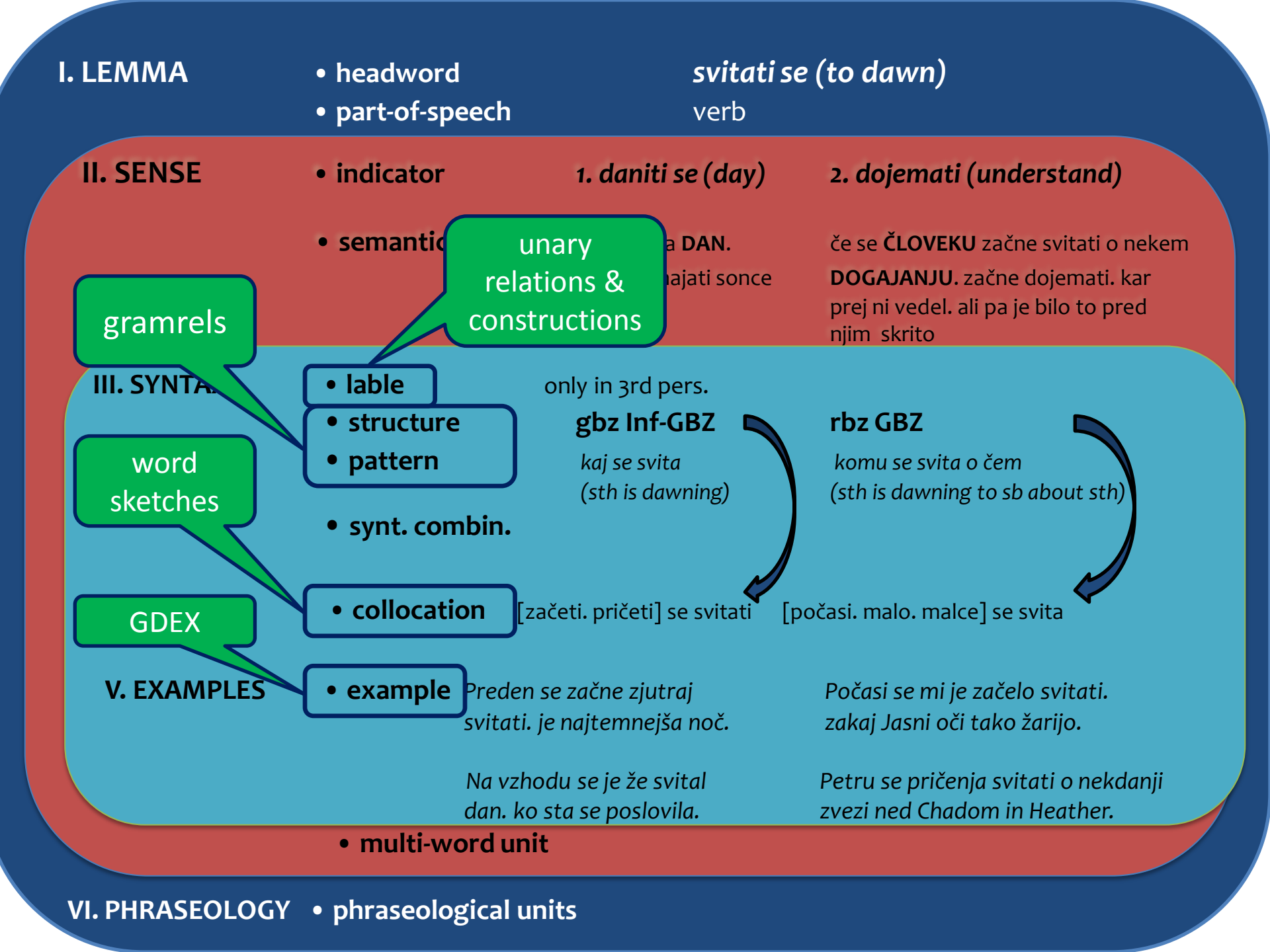
- The only (scholarly) monolingual dictionary of Slovene published in 1991 (with many entries older, from 70s and 80s)
  - Based on material from late 19th century to 1970s
  - Probems with IPR (many authors / owners of copyright)
- New edition announced for October 2014
  - will build on the old version + the supplement from 2012
  - the same IPR problem as the first edition

Slovene Lexical Database



Towards automation

- Sketch grammar
  - Version 2 (Krek, 2010): for manual analysis
  - Version 3 (2012): for automation
    - finely-grained sketch grammar based on the structures identified in the lexical database
- GDEX for Slovene
  - Version 1 (Kosem et al., 2011): for manual analysis
  - Version 2 (2012): for automation
    - different configurations for different word classes


I. LEMMA
• headword
• part-of-speech
svitati se (to dawn)
verb
II. SENSE
• indicator
1. daniti se (day)
2. dojemati (understand)
• semantic
a DAN.
ajati sonce
če se ČLOVEKU začne svitati o nekem DOGAJANJU. začne dojemati. kar prej ni vedel. ali pa je bilo to pred njim skrito
unary relations & constructions
gramrels
III. SYNTA
• lable
• structure
• pattern
only in 3rd pers.
gbz Inf-GBZ
kaj se svita
(sth is dawning)
rbz GBZ
komu se svita o čem
(sth is dawning to sb about sth)
word sketches
• synt. combin.
• collocation
[začeti. pričeti] se svitati
[počasi. malo. malce] se svita
GDEX
V. EXAMPLES
• example
Preden se začne zjutraj svitati. je najtemnejša noč.
Počasi se mi je začelo svitati. zakaj Jasni oči tako žarijo.
Na vzhodu se je že svital dan. ko sta se poslovila.
Petru se pričenja svitati o nekdanji zvezi ned Chadom in Heather.
• multi-word unit
VI. PHRASEOLOGY
• phraseological units

Sketch grammar

- regular expressions over POS tags

  =a_modifier/modifies

  2:[tag="P.*"] 1:[tag="S.*"]

- the name of the arguments (order)
- 1: 2: = words to be extracted as the first/second argument
- |, ., (), {} and * - standard metacharacters (RE)

Regular gramrels

Number	SR-14
Type	enodelna
Name	=nedoločnik
English (appr.)	**=infinitive**
Sketch Grammar	Query
1:[]	Verb, Adjective, Adverb, Noun
SLDB syntactic structure	nedoločnik
GBZ Inf-gbz	uspeti [doseči]
SBZ1 Inf-gbz	(imeti) možnost [pritožiti se]
PBZ Inf-gbz	pripravljen [oditi]; sposoben [izpeljati]
RBZ Inf-gbz	dobro [izkoristiti]

DUAL gramrels

Number	SR-01
Type	recipročna (*DUAL)
Name	=kakšen?/kdo-kaj?
English (appr.)	**=what_kind?/who-what?**
Sketch Grammar	**Query**
1: [tag="S.*"]	Noun
2: [tag="P.*"]	Adjective
SLDB structure	**kakšen? / what_kind?**
pbz SBZ1	[boleč, lep] spomin
SLDB structure	**kdo-kaj? / who-what?**
PBZ sbz1	rdeča [žoga]

TRINARY gramrels

Number	SR-06
Type	tridelna (*TRINARY)
Name	=%s
English (appr.)	**=%s**
Sketch Grammar	Query
1:[tag="G.*"]	Verb
SLDB syntactic structure	=%s
sbz1 GBZ sbz4 (**s sbz6** / med sbz6)	stisniti s [prsti, palci]
sbz1 GBZ sbz4 (s sbz6 / **med sbz6**)	stisniti med [prsti]
sbz1 GBZ (**o sbz5** | glede sbz2 | za sbz4)	pogajati se o [vdaji, izpustitvi]

Automation – Sketch grammar

- use of macros – easier to read
- direct relation between SLD elements and gramrels included in the grammar
- *SEPARATEPAGE (very complex)
- *CONSTRUCTION (very useful)
- *COLLOC (for „syntactic combinations“ in SLD)

Macros examples

- define(`nedolocnik',`[tag="G.n.*"]')
- define(`pomoznik',`[tag="Gv.*"]')
- define(`deleznik',`[tag="Gpd.*"]')
- define(`gl_nebiti',`[tag="G.*" & lemma!="biti"]')
- define(`gl_sed_3',`[tag="Gpp.t.*"]')
- define(`brez_GSVD',`[tag!="[GSVD].*" & word!="[,:;()-]"]')

Macros used in gremrels

- =predl-pred
  - 2:predlog 1:samostalnik
- =%s_s6
  - 1:samostalnik 3:predlog brez_GSVD{0,5} 2:samost_oro
- =S_V_O3_O2
  - 2:osebek brez_PSVD{0,5} 1:glagol brez_SVD{0,5} predmet_daj{1,4} brez_SVD{0,5} predmet_rod

Example: NOUN_1-noun_2

- VERB + prep + NOUN-gen
  „dobiti iz česa“ / to get from sth

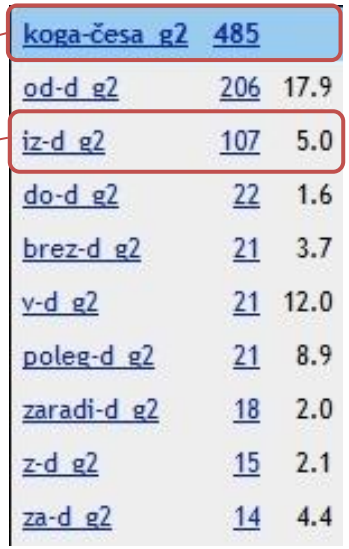
<struktura>GBZ %s sbz2</struktura>

- *SEPARATEPAGE koga-česa_g2
- *TRINARY

=%s_g2

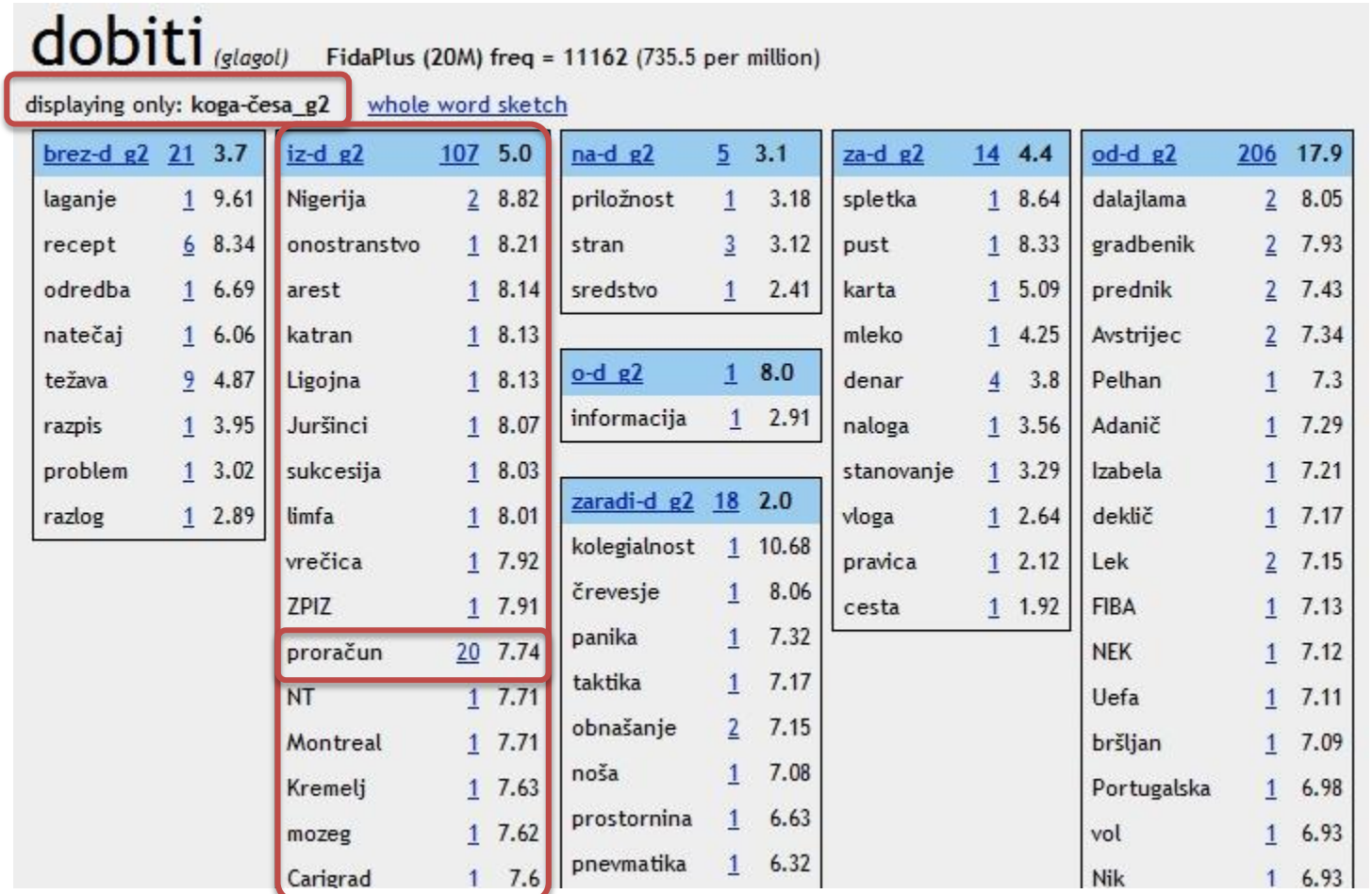
1:glagol sise{0,2} 3:predlog brez_GSVDK{0,5} 2:samost_rod

3:predlog brez_GSVDK{0,5} 2:samost_rod sise{0,1} 1:glagol



koga-česa_g2	485	
od-d_g2	206	17.9
iz-d_g2	107	5.0
do-d_g2	22	1.6
brez-d_g2	21	3.7
v-d_g2	21	12.0
poleg-d_g2	21	8.9
zaradi-d_g2	18	2.0
z-d_g2	15	2.1
za-d_g2	14	4.4

NOUN_1-noun_2



dobiti *(glagol)* FidaPlus (20M) freq = 11162 (735.5 per million)

displaying only: koga-česa_g2 whole word sketch

brez-d_g2	21	3.7
laganje	1	9.61
recept	6	8.34
odredba	1	6.69
natečaj	1	6.06
težava	9	4.87
razpis	1	3.95
problem	1	3.02
razlog	1	2.89

iz-d_g2	107	5.0
Nigerija	2	8.82
onostranstvo	1	8.21
arest	1	8.14
katran	1	8.13
Ligojna	1	8.13
Juršinci	1	8.07
sukcesija	1	8.03
limfa	1	8.01
vrečica	1	7.92
ZPIZ	1	7.91
proračun	20	7.74
NT	1	7.71
Montreal	1	7.71
Kremelj	1	7.63
mozeg	1	7.62
Carigrad	1	7.6

na-d_g2	5	3.1
priložnost	1	3.18
stran	3	3.12
sredstvo	1	2.41

o-d_g2	1	8.0
informacija	1	2.91

zaradi-d_g2	18	2.0
kolegialnost	1	10.68
črevesje	1	8.06
panika	1	7.32
taktika	1	7.17
obnašanje	2	7.15
noša	1	7.08
prostornina	1	6.63
pnevmatika	1	6.32

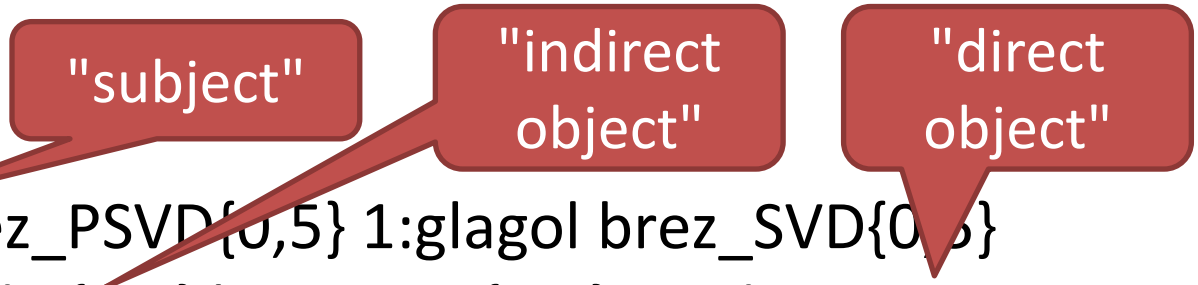
za-d_g2	14	4.4
spletka	1	8.64
pust	1	8.33
karta	1	5.09
mleko	1	4.25
denar	4	3.8
naloga	1	3.56
stanovanje	1	3.29
vloga	1	2.64
pravica	1	2.12
cesta	1	1.92

od-d_g2	206	17.9
dalajlama	2	8.05
gradbenik	2	7.93
prednik	2	7.43
Avstrijec	2	7.34
Pelhan	1	7.3
Adanič	1	7.29
Izabela	1	7.21
deklič	1	7.17
Lek	2	7.15
FIBA	1	7.13
NEK	1	7.12
Uefa	1	7.11
bršljan	1	7.09
Portugalska	1	6.98
vol	1	6.93
Nik	1	6.93

*CONSTRUCTION (very useful)

- Element <vzorci> = syntactic patterns
  - who/what does sb sth
  - who/what does sth to sb etc.
- In entries with verbs as headwords
- Under structures + collocations
- Now: examples with binary collocations
- CONSTRUCTION: examples with complete patterns

Example: S_V_O3_O4



=S_V_O3_O4

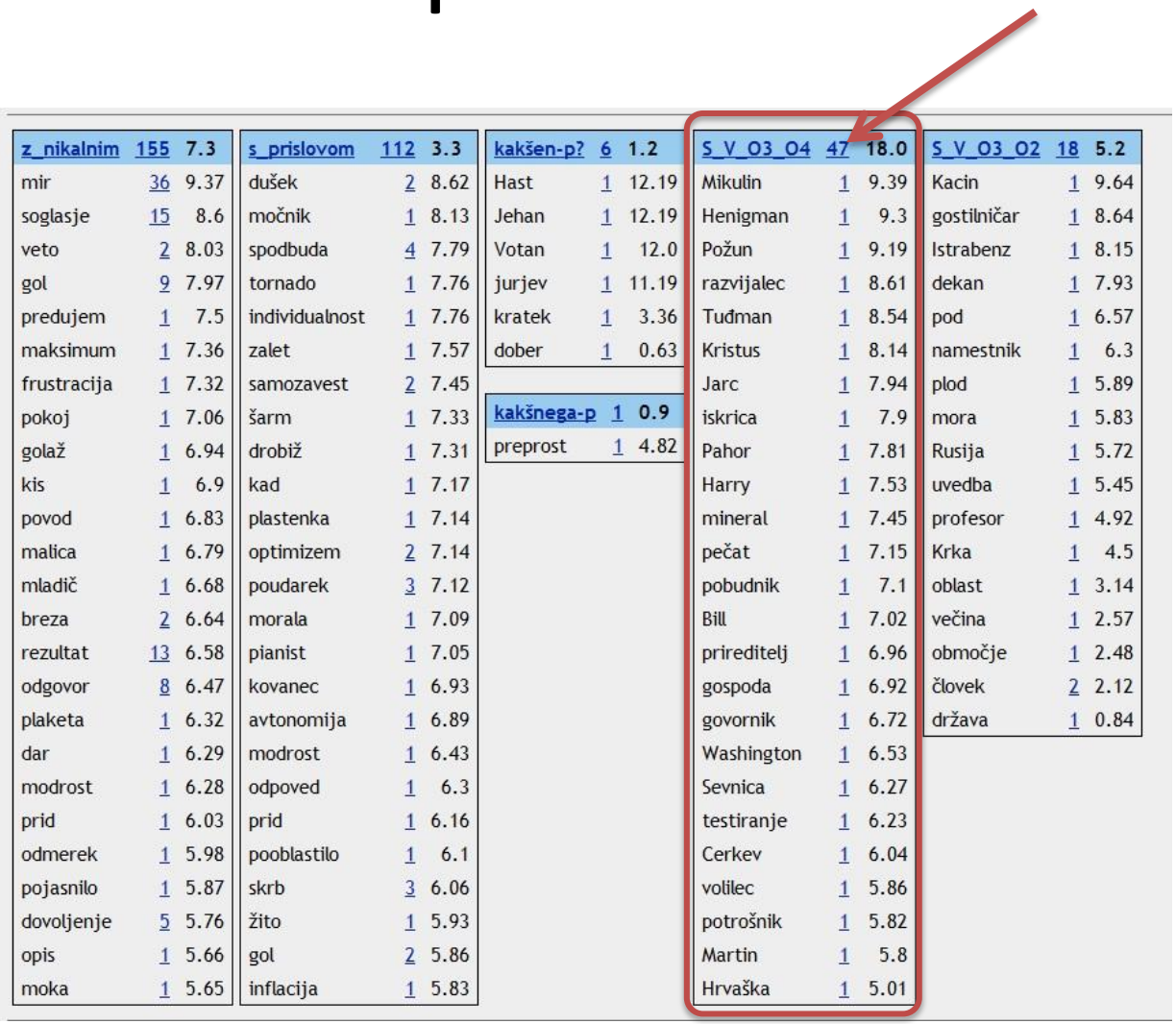
2:osebek brez_PSVD{0,5} 1:glagol brez_SVD{0,5} predmet_daj{1,4} brez_SVD{0,5} predmet_toz

2:osebek brez_PSVD{0,5} 1:glagol brez_SVD{0,5} predmet_toz{1,4} brez_SVD{0,5} predmet_daj

2:osebek brez_PSVD{0,5} predmet_daj{1,4} brez_SVD{0,5} 1:glagol brez_SVD{0,5} predmet_toz

2:osebek brez_PSVD{0,5} predmet_toz{1,4} brez_SVD{0,5} 1:glagol brez_SVD{0,5} predmet_daj

Example from SkE



z_nikalnim	155	7.3
mir	36	9.37
soglasje	15	8.6
veto	2	8.03
gol	9	7.97
predujem	1	7.5
maksimum	1	7.36
frustracija	1	7.32
pokoj	1	7.06
golaž	1	6.94
kis	1	6.9
povod	1	6.83
malica	1	6.79
mladič	1	6.68
breza	2	6.64
rezultat	13	6.58
odgovor	8	6.47
plaketa	1	6.32
dar	1	6.29
modrost	1	6.28
prid	1	6.03
odmerek	1	5.98
pojasnilo	1	5.87
dovoljenje	5	5.76
opis	1	5.66
moka	1	5.65

s_prislovom	112	3.3
dušek	2	8.62
močnik	1	8.13
spodbuda	4	7.79
tornado	1	7.76
individualnost	1	7.76
zalet	1	7.57
samozavest	2	7.45
šarm	1	7.33
drobiž	1	7.31
kad	1	7.17
plastenka	1	7.14
optimizem	2	7.14
poudarek	3	7.12
morala	1	7.09
pianist	1	7.05
kovanec	1	6.93
avtonomija	1	6.89
modrost	1	6.43
odpoved	1	6.3
prid	1	6.16
pooblastilo	1	6.1
skrb	3	6.06
žito	1	5.93
gol	2	5.86
inflacija	1	5.83

kakšen-p?	6	1.2
Hast	1	12.19
Jehan	1	12.19
Votan	1	12.0
jurjev	1	11.19
kratek	1	3.36
dober	1	0.63

kakšnega-p	1	0.9
preprost	1	4.82

S_V_O3_O4	47	18.0
Mikulin	1	9.39
Henigman	1	9.3
Požun	1	9.19
razvijalec	1	8.61
Tuđman	1	8.54
Kristus	1	8.14
Jarc	1	7.94
iskrica	1	7.9
Pahor	1	7.81
Harry	1	7.53
mineral	1	7.45
pečat	1	7.15
pobudnik	1	7.1
Bill	1	7.02
prireditelj	1	6.96
gospoda	1	6.92
govornik	1	6.72
Washington	1	6.53
Sevnica	1	6.27
testiranje	1	6.23
Cerkev	1	6.04
volilec	1	5.86
potrošnik	1	5.82
Martin	1	5.8
Hrvaška	1	5.01

S_V_O3_O2	18	5.2
Kacin	1	9.64
gostilničar	1	8.64
Istrabenz	1	8.15
dekan	1	7.93
pod	1	6.57
namestnik	1	6.3
plod	1	5.89
mora	1	5.83
Rusija	1	5.72
uvedba	1	5.45
profesor	1	4.92
Krka	1	4.5
oblast	1	3.14
večina	1	2.57
območje	1	2.48
človek	2	2.12
država	1	0.84

Examples – high precision

poročil z njo. Tako združene ***moči*** so tovarni	**dale**	nov zagon in postala je najboljša tovarna
predsednik Borut ***Pahor*** je Drnovšku ponovno	**dal**	košarico ne sicer za večno, ampak vsaj
posebej zadovoljen, ker so neuvrščene ***države***	**dale**	vso prednost jedrski razorožitvi. Udeleženci
Učite se od mojstrov. " Najboljši ***govorniki***	**dajo**	svojim poslušalcem vselej občutek, kot
Vrhniki. Gostilna ***Iskrica*** je pohodnikom	**dala**	lonec pasulja, Marko Brecelj je uredil
na terenu, še preden mednarodna ***skupnost***	**da**	ZN mandat za ukrepanje, " je izjavil Anan
Jelovec, del Sredme). Sveti ***Martin*** je	**dal**	svoje ime cerkvi s prepoznavnim baročnim
privoščiš še mineralno kopel; ***minerali***	**dajo**	vodi posebno zeleno barvo. V spremstvu
premier Akajeva, je tako kot vrhovno ***sodišče***	**dal**	prednost staremu parlamentu, ki je Bakijeva
1997 je mestni svetnik Mihael ***Jarc*** sicer	**dal**	pobudo mestnemu svetu, da bi po Janezu
podaljšati - Državne ***ustanove*** so jagrom	**dale**	čepice - Zadovoljstvo v Luksemburgu - Tudi
sta) Kongresno ***testiranje*** varnosti	**dalo**	porazno sliko Kot švicarski sir
leto pa napoveduje, da bodo ***prireditelji***	**dali**	večji poudarek tudi pohodom, ki jih bodo
štela za plačano z dnem, ko bo ***potrošnik***	**dal**	nalog taki organizaciji. Ali pa, če bo
V nadaljevanju je trener Miro ***Požun*** spet	**dal**	priložnost mladim igralcem in prav vsi,
je upravičena domneva, da je ***Washington***	**dal**	Manili tiho podporo za poskus vojaške rešitve

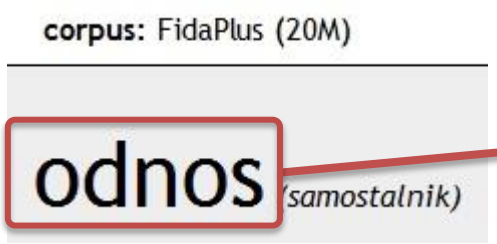
*COLLOC for „syntactic combinations“

- Element <zveza> = syntactic combinations
  - "v razmerju do" (in relation to)
- Mainly nominal headwords
- Under (sub)sense after syntactic structures as a separate category

COLLOC: d_sam_d

- =d_sam_d
- *COLLOC "%(2.lemma)_%(3.lemma)-p"
- 2:predlog 1:samostalnik 3:predlog

Example SkE: "in relation to"



d_sam_d	412	7.0
v_do	92	11.69
za_z	109	10.86
v_med	38	10.4
o_med	18	10.07
o_z	18	9.59
na_med	13	9.14
o_do	8	9.12
z_do	8	8.63
za_med	5	7.96
glede_do	3	7.86
v_z	59	7.84

avtoriteta *(samostalnik)* **Fida PLUS 620m (SLD sketch grammar) frekvenca = 9301** (12.6 na milijon)

gl-pred	5465	0.8
spodkopavati	21	27.93
spodkopati	29	27.79
rušiti	44	27.39
spodbijati	26	27.1
priznavati	57	26.73
zastaviti	61	22.89
priznati	100	19.77
podrejati	15	19.41
izžarevati	17	19.05
utrjevati	20	18.87
upirati	31	18.81
sklicevati	30	18.57
uveljaviti	46	17.71
izgubljati	34	17.68
spoštovati	39	17.07
vzpostaviti	37	16.73
omajati	11	15.69
imeti	575	15.18
uživati	46	14.94
uveljavljati	26	14.89
utrditi	20	14.57
smešiti	6	14.22
izrabljati	10	14.06
okrepiti	28	13.96
zavračati	20	13.87

>>

kakšen?	4098	3.3
moralen	337	67.97
nesporen	182	66.8
brezpriziven	16	38.76
cerkven	70	38.37
očetovski	26	38.17
starševski	33	38.09
učiteljev	23	36.58
vrhoven	57	36.21
strokoven	126	35.38
absoluten	41	35.19
visok	184	35.18
velik	348	33.94
političen	137	33.75
nezmotljiv	15	33.4
papežev	23	30.93
verski	47	30.82
nedotakljiv	14	28.19
vodilen	35	25.43
svetoven	85	25.43
lažen	18	25.15
nedvomen	10	25.11
očetov	18	24.87
znanstven	28	24.51
duhoven	22	24.21
priznan	20	23.91

>>

gl-za	3478	0.5
omajati	9	15.38
priznavati	15	14.72
spoštovati	23	14.36
razvaljati	6	13.1
vzpostavljati	9	12.91
načeti	15	12.8
spodkopati	6	12.74
podrejati	6	12.32
temeljiti	21	12.25
obvladovati	12	11.68
močiti	6	11.01
vplivati	29	10.99
uživati	21	10.51
zagovarjati	12	9.96
izhajati	16	9.88
izkazovati	6	9.28
zagotavljati	19	9.11
odločati	20	8.95
nadomestiti	13	8.73
izvirati	9	8.61
prepovedovati	6	8.51
uveljavljati	9	8.38
sklicevati	7	8.23
tikati2	8	8.13
graditi	14	8.09

>>

koga-kaj	1759	6.5
spodkopati	26	31.8
zastaviti	62	29.85
rušiti	32	29.27
spodkopavati	16	29.12
spodbijati	19	28.17
izgubljati	33	23.15
izžarevati	15	22.31
priznavati	22	21.46
vzpostaviti	31	20.72
uveljaviti	31	19.64
zavračati	21	19.36
utrjevati	13	19.03
prezirati	9	18.58
utrditi	16	17.28
okrepiti	22	17.03
zbijati	6	16.36
pridobiti	40	16.27
uživati	28	16.16
spoštovati	20	16.14
izrabljati	8	15.77
izgubiti	42	15.56
omajati	7	15.18
razvaljati	6	15.01
uveljavljati	15	14.86
priznati	30	14.36

>>

priredje	1236	2.4
ugled	85	50.25
verodostojnost	23	37.59
moč	61	34.21
karizma	14	33.61
avtoriteta	18	30.03
integriteta	12	29.9
kredibilnost	11	28.86
moč	21	27.57
hierarhija	11	25.44
legitimnost	9	25.3
oblast	25	22.33
znanje	18	19.51
odgovornost	13	18.85
zaupanje	10	18.47
samozavest	7	17.71
spoštovanje	8	16.85
pooblastilo	7	16.68
družben	11	15.91
vpliv	11	15.22
izkušnja	11	14.04
položaj	12	13.7
vloga	12	13.4
tradicija	6	12.14
red	7	10.23
podoben	6	8.56

>>

avtoriteta *(samostalnik)* Fida PLUS 620m (SLD sketch grammar) frekvenca = 9301 (12.6 na milijon)

gl-pred	5465	0.8
spodkopavati	21	27.93
spodkopati	29	27.79
rušiti	44	27.39
spodbijati	26	27.1
priznavati	57	26.73
zastaviti	61	22.89
priznati	100	19.77
podrejati	15	19.41
izžarevati	17	19.05
utrjevati	20	18.87
upirati	31	18.81
sklicevati	30	18.57
uveljaviti	46	17.71
izgubljati	34	17.68
spoštovati	39	17.07
vzpostaviti	37	16.73
omajati	11	15.69
imeti	575	15.18
uživati	46	14.94
uveljavljati	26	14.89
utrditi	20	14.57
smešiti	6	14.22
izrabljati	10	14.06
okrepiti	28	13.96
zavračati	20	13.87
>>		

kakšen?	4098	3.3
moralen	337	67.97
nesporen	182	66.8
brezprizriven	16	38.76
cerkven	70	38.37
očetovski	26	38.17
starševski	33	38.09
učiteljev	23	36.58
vrhoven	57	36.21
strokoven	126	35.38
absoluten	41	35.19
visok	184	35.18
velik	348	33.94
političen	137	33.75
nezmotljiv	15	33.4
papežev	23	30.93
verski	47	30.82
nedotakljiv	14	28.19
vodilen	35	25.43
svetoven	85	25.43
lažen	18	25.15
nedvomen	10	25.11
očetov	18	24.87
znanstven	28	24.51
duhoven	22	24.21
priznan	20	23.91
>>		

gl-za	3478	0.5
omajati	9	15.38
priznavati	15	14.72
spoštovati	23	14.36
razvaljati	6	13.1
vzpostavljati	9	12.91
načeti	15	12.8
spodkopati	6	12.74
podrejati	6	12.32
temeljiti	21	12.25
obvladovati	12	11.68
močiti	6	11.01
vplivati	29	10.99
uživati	21	10.51
zagovarjati	12	9.96
izhajati	16	9.88
izkazovati	6	9.28
zagotavljati	19	9.11
odločati	20	8.95
nadomestiti	13	8.73
izvirati	9	8.61
prepovedovati	6	8.51
uveljavljati	9	8.38
sklicevati	7	8.23
tikati2	8	8.13
graditi	14	8.09
>>		

koga-kaj	1759	6.5
spodkopati	26	31.8
zastaviti	62	29.85
rušiti	32	29.27
spodkopavati	16	29.12
spodbijati	19	28.17
izgubljati	33	23.15
izžarevati	15	22.31
priznavati	22	21.46
vzpostaviti	31	20.72
uveljavljati	31	19.64
zavračati	21	19.36
utrjevati	13	19.03
prezirati	9	18.58
utrditi	16	17.28
okrepiti	22	17.03
zbijati	6	16.36
pridobiti	40	16.27
uživati	28	16.16
spoštovati	20	16.14
izrabljati	8	15.77
izgubiti	42	15.56
omajati	7	15.18
razvaljati	6	15.01
uveljavljati	15	14.86
priznati	30	14.36
>>		

priredje	1236	2.4
ugled	85	50.25
verodostojnost	23	37.59
moč	61	34.21
karizma	14	33.61
avtoriteta	18	30.03
integriteta	12	29.9
kredibilnost	11	28.86
moč	21	27.57
hierarhija	11	25.44
legitimnost	9	25.3
oblast	25	22.33
znanje	18	19.51
odgovornost	13	18.85
zaupanje	10	18.47
samozavest	7	17.71
spoštovanje	8	16.85
pooblastilo	7	16.68
družben	11	15.91
vpliv	11	15.22
izkušnja	11	14.04
položaj	12	13.7
vloga	12	13.4
tradicija	6	12.14
red	7	10.23
podoben	6	8.56
>>		

vezник	1086	2.4
ki	483	41.1
da	153	24.78
saj	46	23.2
in	35	22.02
zato	27	21.7
ampak	21	21.04
temveč	16	20.78
marveč	8	20.43
kar	27	19.69
kot	16	19.24
kateri	36	18.9
če	25	18.44
naj	17	16.03
kakršen	8	16.0
toda	10	15.34
ker	17	15.27
ter		frekvenca
kakor	7	14.74
kadar	6	14.54
čeprav	10	14.48
vendar	13	13.91
a	12	13.37
ali	10	13.21
drug	7	13.0
ko	13	11.1
>>		

predl-za	1056	0.8
nad	35	21.67
na	246	21.43
v	336	20.91
med	49	16.87
za	133	15.86
pri	54	15.75
znotraj	6	12.49
kot	33	11.86
glede	6	10.23
z	56	8.0
brez	8	7.21
iz	20	7.11
zaradi	9	6.11
pred	8	4.6
do	11	4.21
od	9	3.16
o	7	1.94
po	8	1.3
>>		

osebek_od	1001	2.3
priznavati	14	18.97
omajati	6	15.35
temeljiti	16	15.21
obstajati	17	12.01
izhajati	10	11.05
tikati2	7	11.03
spoštovati	7	9.56
postati	22	8.76
predstavljati	14	8.73
odločati	9	8.25
dati	29	8.22
jemati	6	8.05
ostajati	7	6.96
delovati	9	5.81
praviti	14	5.47
meniti	8	5.44
jesti	8	4.49
dejati	9	4.43
pomeniti	10	4.43
pomagati	8	4.29
imeti	39	3.52
znati	8	3.43
veljati	6	3.12
kazati	6	3.02
reči	7	2.26
>>		

predlog	965	0.6
z	321	26.72
brez	52	24.26
do	87	21.53
proti	31	19.74
zaradi	29	15.21
kot	42	14.45
pred	30	14.18
na	112	13.63
zoper	6	12.33
za	82	11.8
od	23	9.08
nad	8	8.89
k	9	7.94
pod	8	7.86
med	15	7.58
o	16	6.39
po	17	5.23
ob	6	2.97
v	41	2.57
iz	7	1.98
pri	6	1.73
>>		

oba-v-rod	923	1.6
moralen	79	51.4
nesporen	26	40.09
cerkven	22	29.73
očetovski	9	28.1
političen	43	27.73
starševski	10	26.93
nezmotljiv	6	25.06
strokoven	29	24.95
znanstven	14	22.82
učiteljev	6	22.81
lažen	9	22.31
svetoven	31	22.19
absoluten	9	21.59
priznan	9	20.39
potreben	17	20.18
verski	11	19.92
vrhoven	9	19.35
religiozen	6	19.21
velik	40	17.92
zadosten	6	17.21
intelektualen	6	17.04
pravi	12	16.86
praven	11	15.8
lasten	10	14.3
tradicionalen	6	13.17
>>		

z-d_X	697	5.5
težava	26	21.44
biti	124	19.25
človek	23	16.71
konflikt	6	16.33
obvladovati	8	13.82
trener	6	12.27
oseba	8	12.01
problem	7	11.66
mož	6	11.3
podpreti	8	9.48
poskrbeti	8	7.65
voditi	11	7.38
vplivati	7	7.28
govoriti	10	6.43
povezati	7	6.41
>>		

v_rodil-s	653	1.8
učitelj	28	30.97
stvarnik	7	26.46
starš	21	24.96
Cerkev	10	22.44
država	38	22.05
sodstvo	7	21.87
argument	10	21.32
predsednik	20	18.49
institucija	9	18.45
papež	6	17.29
sodišče	15	16.93
cerkev	10	16.38
bog	6	15.94
parlament	8	14.88
kota	8	14.64
znanje	8	14.01
vlada	11	13.36
trener	6	12.45
oče	6	12.38
vodja	6	12.25
zdravnik	6	12.0
agencija	6	11.78
položaj	7	11.4
oblast	6	11.12
zakon	6	7.88
>>		

s-koga-česa	651	1.8
priznavanje	13	33.02
pomanjkanje	23	30.01
spoštovanje	19	29.48
spodkopavanje	7	29.4
nespoštovanje	9	27.92
krepitev	7	20.75
argument	9	20.14
kriza	12	20.13
hierarhija	6	20.02
izguba	11	19.47
struktura	10	18.34
uveljavljanje	6	18.18
zloraba	7	17.93
oblika	15	17.61
pozicija	6	17.5
prava	6	16.49
moč	9	14.86
vprašanje	12	14.48
problem	8	12.96
podlaga	6	12.14
vpliv	6	12.02
>>		

koga-česa	425	7.0
priznati	48	26.89
priznavati	17	24.63
spodbijati	7	20.82
imeti	105	16.88
rušiti	6	15.58
spoštovati	9	14.27
prenesti	10	13.55
uveljaviti	7	11.58
ustvariti	8	9.98
izgubiti	6	6.58
znati	7	5.41
>>		

v-d	388	1.2
vzgoja	21	28.48
država	35	23.83
stranka	10	14.59
razred	7	14.55
svet	7	9.13
>>		

na-d	255	1.4
področje	163	55.92
igrišče	12	23.31
svet	6	9.29
>>		

komu-čemu	239	6.4
upirati	24	30.82
podrejati	12	28.71
upreti	15	23.79
ukloniti	8	23.72
podrediti	10	21.31
škoditi	7	17.09
nasprotovati	7	13.63
zaupati	8	13.48
>>		

količinski	219	4.1
dovolj	43	39.42
visoko	21	33.59
več	39	32.32
zunanje	9	28.83
malo	16	21.21
veliko	20	20.76
>>		

zanikan	188	11.5
priznati	45	30.6
priznavati	19	29.53
prenesti	12	18.19
imeti	28	9.73
>>		

za-d_X	179	1.3
veljati	55	29.32
skrivati	10	16.39
biti	19	9.71
iti	13	9.29
imeti	17	6.48
>>		

trpnik	159	1.9
omajati	9	25.63
načeti	13	23.39
priznati	10	14.54
znati	8	9.26
>>		

na-d_X	156	1.0
sklicevanje	23	48.9
sklicevati	25	33.47
napad	10	20.85
temeljiti	13	20.45
opirati	6	18.52
>>		

za-d	155	1.0
vprašanje	8	15.67
področje	7	14.85
>>		

do-d_X	123	5.2
odnos	41	38.0
spoštovanje	13	30.87
odpor	11	30.22
biti	9	6.23
>>		

kot-d_X	122	8.7
priznati	6	11.11
biti	15	9.58
>>		

s_prislovom	84	2.9
imeti	42	17.44
>>		

v-d_X	67	0.2
biti	7	6.69
>>		

z-d	62	0.4
področje	20	30.38
>>		

pred-d_X	59	3.2
strah	30	42.86
>>		

p-koga-kaj	58	3.0
sam	7	17.71
>>		

med-d	56	2.5
igralec	6	17.21
>>		

brez-d_X	46	6.6
biti	10	10.25
>>		

proti-d_X	42	6.7
upor	20	43.85
>>		

zaradi-d_X	37	3.5
biti	16	14.93
>>		

o-d_X	34	1.0
govoriti	7	13.57
>>		

Transfer of information

- API using data from Sketch Engine
- Gramrels:
  - Element <struktura> = syntactic structures
  - Element <vzorec> = syntactic patterns
  - Element <zveza> = syntactic combinations
  - Element <oznaka> = labels
- Collocations = element <kolokacija>
- Examples = element <zgled> using GDEX

Gramrel to <struktura>

ADJECTIVE + NOUN

```
<skladenjska_struktura>
<struktura>kakšen?</struktura>
<kolokacije>
    <kolokacija id="839596"><k>nov</k></kolokacija>
    <kolokacija id="839746"><k>deloven</k></kolokacija>
    <kolokacija id="840017"><k>spleten</k></kolokacija>
    <kolokacija id="839637"><k>glaven</k></kolokacija>
    <kolokacija id="839725"><k>prost</k></kolokacija>
    <kolokacija id="839830"><k>parkiren</k></kolokacija>
    <kolokacija id="839601"><k>velik</k></kolokacija>
    <kolokacija id="839952"><k>vodilen</k></kolokacija>
    <kolokacija id="839625"><k>pravi</k></kolokacija>
    <kolokacija id="839814"><k>prodajen</k></kolokacija>
</kolokacije>
<zgledi>
    <zgled seek="839596" position="1">Zavod za zdravstveno varstvo Novo
    <i>mesto</i></zgled>
    <zgled seek="839601" position="1">" V glavnem v vseh večjih <i>mestih
    </i>.</zgled>
```

collocations and coresponding examples

Gramrel to <vzorec>

Construction to <vzorec>

```
<skladenjska_struktura>
<vzorec>S_V_O3_O4</vzorec>
<zgledi>
    <zgled seek="16213" position="1">Tako združene  moci
    so tovarni <i>dale</i> nov zagon in postala je
    najboljša tovarna klobukov.</zgled>
    <zgled seek="16215" position="1">Njen predsednik Borut
    Pahor  je Drnovšku ponovno <i>dal</i> košarico ne sicer
    za vecno, ampak vsaj do prvih naslednjih
    volitev.</zgled>
    <zgled seek="16215" position="2">Južnoafriški zunanji
    minister Alfred Nzo je bil po konferenci še posebej
    zadovoljen, ker so neuvrščene  države <i>dale</i> vso
    prednost jedrski razorožitvi.</zgled>
</zgledi>
```

Gramrel to label

<oblika>
  <iztocnica>mesto</iztocnica>
</oblika>
<zaglavje>
  <besvrs>samostalnik</besvrs>
  <oznaka>z_lastnim_imenom</oznaka>
</zaglavje>

unary to label: "with proper names"

Sketch Grammar for Slovene (v.16)

- 105 gramrels
- 50 macro definitions
- 25 DUAL
- 36 TRINARY
- 36 SEPARATEPAGE
- 8 UNARY
- 2 SYMMETRIC
- 19 CONSTRUCTION
- 3 COLLOC
- 18 no directive

- SEPARATEPAGE=TRINARY
- 6 CONSTRUCTION-UNARY

API and settings

- API script to extract data from word sketch information in the Sketch Engine
- a list of lemmas for extraction: lemmas with frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words)
- settings for extraction
  - lemmas divided into five frequency groups
  - different setting for each group

Selection of lemmas

- Frequent enough to offer a good-sized word sketch
  - less than 600 hits in Gigafida did not provide enough relevant data
  - we divided lemmas of each word class into five different frequency groups
- Monosemous lemmas or having up to
  - two synsets/senses in sloWNet, a Slovene version of Wordnet
  - exceptionally, in the Dictionary of Standard Slovenian (SSKJ)
- Found in sloWnet, preferably, but not in SSKJ, as we wanted to focus on new words and/or senses

Distribution of lemmas

- The final selection included
  - 515 nouns
  - 260 verbs
  - 275 adjectives
  - 117 adverbs

  - lemmas with frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words)

API script (Python)

- Python for Windows (v. 2.7.3)
- httplib2-0.7.4 library
- simplejson-2.5.2 library

- **python tblscript.py fidaplus-gf -U http://ske.slovenscina.eu/ -f 8 -l spisek_lempos.txt**

Lemmalist

- -l LEMMALIST, --lemmalist=LEMMALIST
  - The file containing a list of lemposes for which the examples are to be extracted (stdin by default).

General (Gramrellist)

- -f MINFREQ, --frequency=MINFREQ
  - Default minimum frequency of a collocate(default=0.0).
- -s MINSAL, --salience=MINSAL
  - Default minimum salience of a collocate(default=0.0).
- -F MINFREQREL, --Freqrel=MINFREQREL
  - Minimum frequency of a relation (default=25).
- -S MINSALREL, --Salrel=MINSALREL
  - Minimum salience of a relation (default=0.0).

Gramrellist

- -r GRAMRELLIST, --relations=GRAMRELLIST
  - The file containing a set of grammatical relations from a given sketch grammar for inclusion (all by default).
  - One record consists of:
    - gramrel regular expression
    - min. collocation frequency
    - min. col. salience
    - min. gramrel frequency
    - min. g. salience
    - gramrel type
  - The gramrel type should be one of: 'SVOZ' in order: 'struktura', 'vzorec', 'oznaka' and 'zveza'. If no type is provided than the first letter of gramrel name decides. For example:
    - (sub|ob)ject 3 2.5 30 20 S

Maximums & GDEX

- -n NUMBER, --number=NUMBER
  - Maximum number of sentences per collocation (default=6).
- -m MAXITEMS, --maxCollocs=MAXITEMS
  - Maximum number of collocations per grammatical relation (default 10).
- -g GDEXCONF, --gdexconf=GDEXCONF
  - Name of the gdex configuration to use.

Gramrellist example

gramrel regular expression	min. coll. freq	min. coll. salience	min. gramrel freq	min. gramrel salience	gramrel type
…					
O_tretja_oseba	8	0.5	60	0.5	O
O_z_lastnim_imenom	8	0.5	8	2.5	O
O_zanikanje	8	0.5	8	20.0	O
S_.*_p2	4	0.5	8	25.0	S
S_.*_p3	4	0.5	8	100.0	S
S_.*_p4	4	0.5	8	20.0	S
…					

We started with...

- 10 collocates per relation
- 6 examples per collocate
- Minimum salience of a relation/collocate = 0
- Minimum frequency of a collocate = 0
- Minimum frequency of a relation = 25

- Statistical & manual analysis
- identifying the lowest values where the collocation still yielded relevant results

And ended with...

- Minimum number of collocates per relation was increased to 25
- Selection of relevant collocates was ‘left’ to minimum frequency and salience settings
- Number of examples per collocate was reduced to three
- We divided lemmas into frequency groups, and prepared separate settings for each group

XML template

- DOC_TEMPLATE = ("""<?xml version="1.0" encoding="UTF-8"?>
    - <clanek>
    - <glava>
    - <oblika><zapis>%(headword)s</zapis>
    - <iztocnica>%(headword)s</iztocnica></oblika>
    - <zaglavje>
    - <besvrs>%(pos)s</besvrs>
    - """,# here come all O_"""
    - </zaglavje>
    - </glava>

Output

- ?xml version="1.0" encoding="UTF-8"?>
- <clanek>
- <glava>
- <oblika><zapis>anoreksija</zapis><iztocnica>anoreksija</iztocnica></oblika>
- <zaglavje><besvrs>samostalnik</besvrs></zaglavje>
- </glava>
- <geslo>
- <pomen>
- <indikator></indikator><pomenska_shema></pomenska_shema>
- <skladenjske_skupine><skladenjska_struktura>
- <struktura>S_predl-pred</struktura>
- <kolokacije><kolokacija kid="100344429"><k>proti</k></kolokacija></kolokacije>
- <zgledi><zgled kid="100344429" pozicija="1">Francoska manekenka, ki je leta 2007 s fotografijo v okviru kampanje boja proti anoreksiji dvignila veliko prahu, je umrla.</zgled></zgledi>

Gramrel conversion

- ?xml version="1.0" encoding="UTF-8"?>
- <clanek>
- <glava>
- <oblika><zapis>anoreksija</zapis><iztocnica>anoreksija</iztocnica></oblika>
- <zaglavje><besvrs>samostalnik</besvrs></zaglavje>
- </glava>
- <geslo>
- <pomen>
- <indikator></indikator><pomenska_shema></pomenska_shema>
- <skladenjske_skupine><skladenjska_struktura>
- <struktura>zveze s predlogi</struktura>
- <kolokacije><kolokacija kid="100344429"><k>proti</k></kolokacija></kolokacije>
- <zgledi><zgled kid="100344429" pozicija="1">Francoska manekenka, ki je leta 2007 s fotografijo v okviru kampanje boja proti anoreksiji dvignila veliko prahu, je umrla.</zgled></zgledi>

“it is more efficient to edit out the computer’s errors than to go through the whole data-selection process from the beginning”

(Rundell & Kilgarriff, 2011)

“too many choices early in the data-selection process leave more room for error”

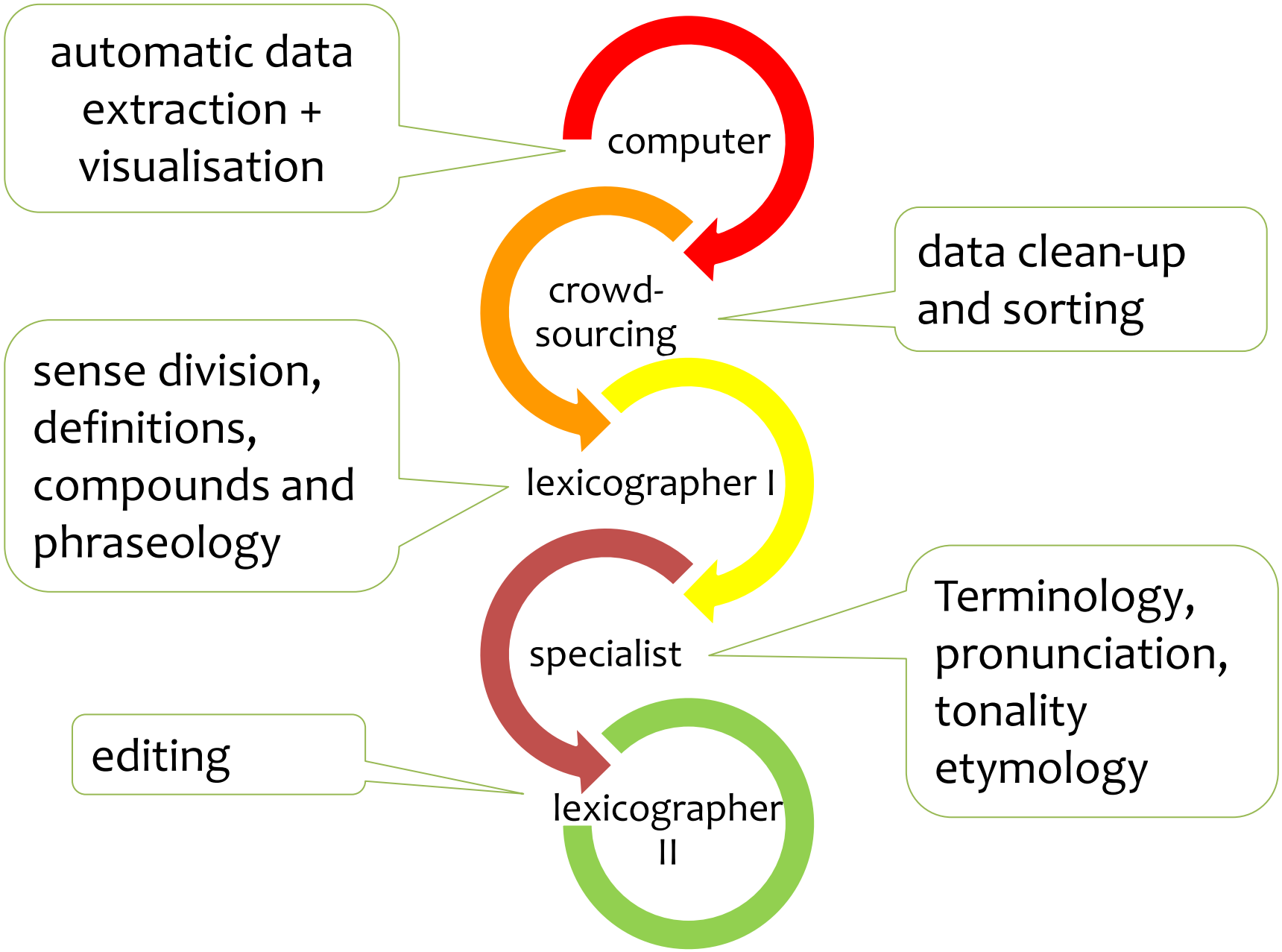
(Kosem, Gantar & Krek)

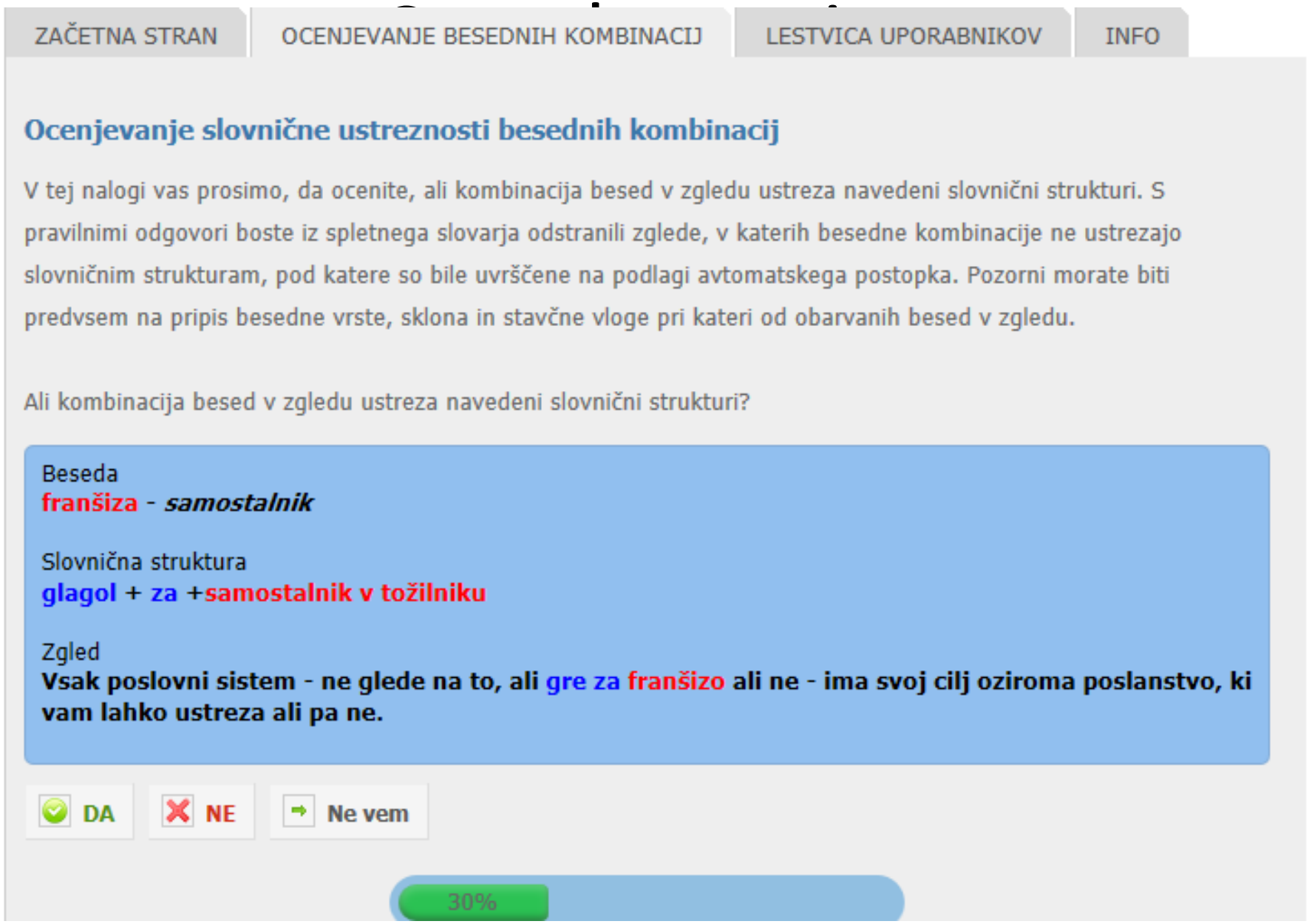
Work left for lexicographers

- Analytical
  - sense division
  - writing definitions, sense indicators
  - identification of multi-word units, phrases, pragmatics
  - adding certain labels
- Editorial
  - distributing information according to sense division
  - copying grammatical relations and collocates typical for more than one sense
  - deleting irrelevant info (collocates, examples etc.)

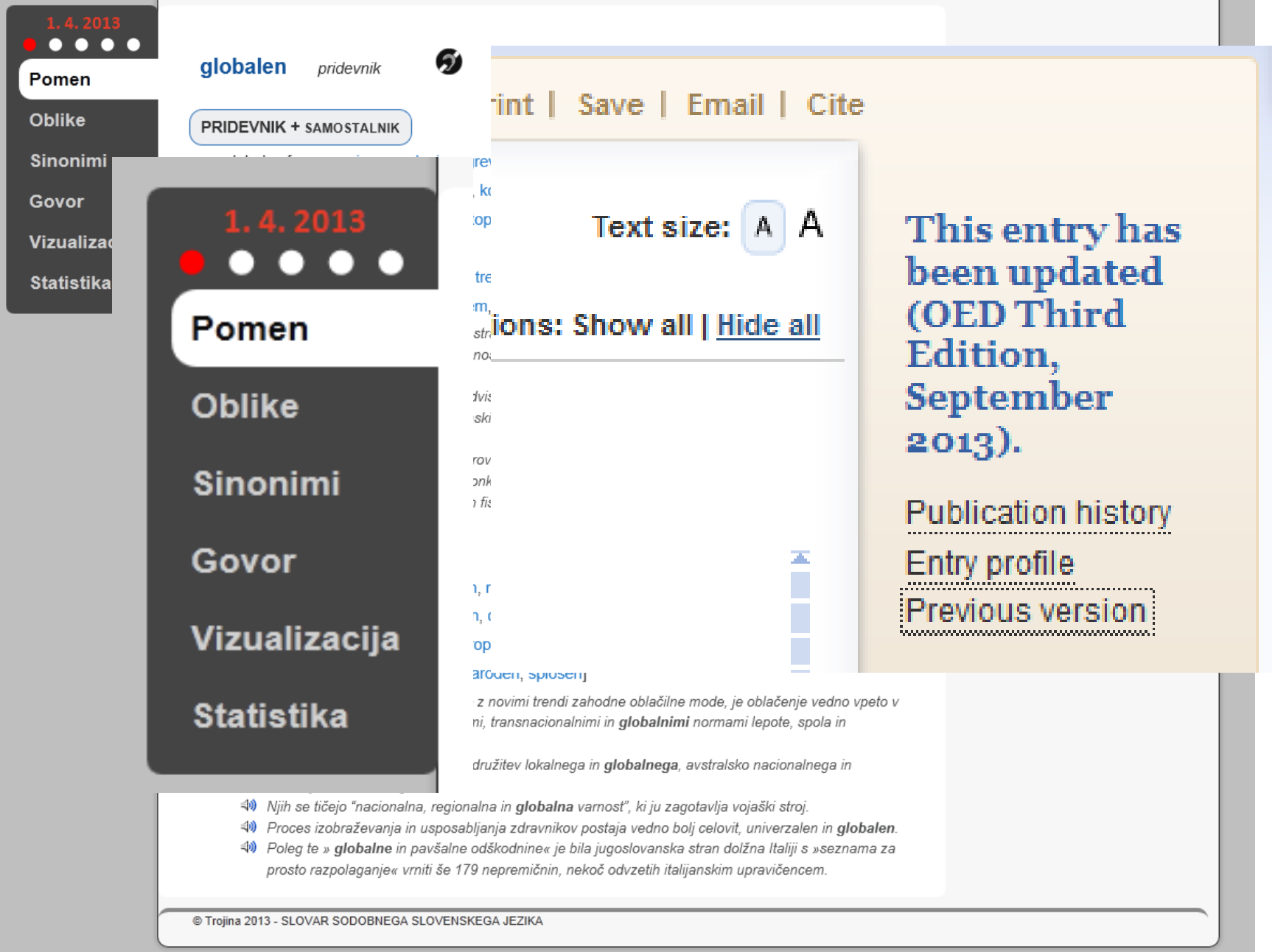
Proposal for a dictionary of contemporary Slovene

- starting from scratch
- http://www.sssj.si/)
- multi-phase dictionary compilation (and presentation)
- data extraction used for the first phase


automatic data extraction + visualisation
computer
crowd-sourcing
data clean-up and sorting
sense division, definitions, compounds and phraseology
lexicographer I
specialist
Terminology, pronunciation, tonality etymology
editing
lexicographer II


ZAČETNA STRAN
OCENJEVANJE BESEDNIH KOMBINACIJ
LESTVICA UPORABNIKOV
INFO
Ocenjevanje slovnične ustreznosti besednih kombinacij
V tej nalogi vas prosimo, da ocenite, ali kombinacija besed v zgledu ustreza navedeni slovnični strukturi. S pravilnimi odgovori boste iz spletnega slovarja odstranili zglede, v katerih besedne kombinacije ne ustrezajo slovničnim strukturam, pod katere so bile uvrščene na podlagi avtomatskega postopka. Pozorni morate biti predvsem na pripis besedne vrste, sklona in stavčne vloge pri kateri od obarvanih besed v zgledu.
Ali kombinacija besed v zgledu ustreza navedeni slovnični strukturi?
Beseda
franšiza - samostalnik
Slovnična struktura
glagol + za +samostalnik v tožilniku
Zgled
Vsak poslovni sistem - ne glede na to, ali gre za franšizo ali ne - ima svoj cilj oziroma poslanstvo, ki vam lahko ustreza ali pa ne.
DA
NE
Ne vem
30%

SLOVAR SODOBNEGA SLOVENSKEGA JEZIKA



globalen *pridevnik* P 3000

PRIDEVNIK + SAMOSTALNIK

globalno [segrevanje, zasedanje, ogrevanje]

globalna [ekonomija, kriza, delniska, konkurenčnost, recesija, raven]

globalna [razsežnost, korporacija, otoplitev, spremenljivka, vas]

globalna [pravičnost]

globalni [kapitalizem, vodič, nomad, trend, klepet]

globalni [trg, fenomen, izziv, terorizem, motorist]

- *Danes ni več politika ali politične stranke, ki v svojih nastopih in programih ne bi omenjala tudi klimatskih sprememb in zaskrbljenosti zaradi **globalnega** segrevanja ozračja, skrbi za pitno vodo in zdravo okolje.*
- *Za podjetja je postala zaradi soodvisnosti in povezanosti **globalne** ekonomije nujna za preživetje.*
- *Spremljamo dogajanje na ljubljanski borzi, ki je zaradi **globalne** finančne krize teden začela v rdečih številkah.*
- *Brez podkupnine, ne mešajte s provizijo, v **globalnem** kapitalizmu ni več resnega posla.*
- *Zaradi pospeševanja **globalne** konkurenčnosti slovenskega gospodarstva pa bi bilo treba to strateško usmeritev uresničiti brez dodatnih fiskalnih bremen.*

PRIDEVNIK + in + PRIDEVNIK

globalen in [transnacionalen, lokalen, regionalen, univerzalen, pavšalen]

globalen in [meddržaven, nacionalen, celovit, radikalen, operativen]

globalen in [regijski, dolgoročen, evropski, konkurenčen, trajen]

globalen in [krajeven, domač, mednaroden, splošen]

- *Medtem ko lepota stalno tekmuje z novimi trendi zahodne oblačilne mode, je oblačenje vedno vpeto v zapletena pogajanja med lokalnimi, transnacionalnimi in **globalnimi** normami lepote, spola in*

1. 4. 2013

Pomen

Oblike

Sinonimi

Govor

Vizualizacija

Statistika

globalen *pridevnik* P 3000

PRIDEVNIK + SAMOSTALNIK

globalno [segrevanje, ogrevanje]

globalna [ekonomija, kriza, konkurenčnost, recesija, raven]

globalna [razsežnost, korporacija, otoplitev, spremenljivka, vas]

globalna [pravičnost]

globalni [kapitalizem, nomad, trend]

globalni [trg, fenomen, izziv, terorizem]

- *Danes ni več politika ali politične stranke, ki v svojih nastopih in programih ne bi omenjala tudi klimatskih sprememb in zaskrbljenosti zaradi **globalnega** segrevanja ozračja, skrbi za pitno vodo in zdravo okolje.*
- *Svet in z njim **globalna** ekonomija dejansko vodita planet Zemljo v fazo prekomernega izkoriščanja naravnih virov.*
- *Spremljamo dogajanje na ljubljanski borzi, ki je zaradi **globalne** finančne krize teden začela v rdečih številkah.*
- *Brez podkupnine, ne mešajte s provizijo, v **globalnem** kapitalizmu ni več resnega posla.*
- *Zaradi pospeševanja **globalne** konkurenčnosti slovenskega gospodarstva pa bi bilo treba to stratešk usmeritev uresničiti brez dodatnih fiskalnih bremen.*

PRIDEVNIK + in + PRIDEVNIK

globalen in [transnacionalen, lokalen, regionalen, univerzalen, pavšalen]

globalen in [meddržaven, nacionalen, celovit, radikalen, operativen]

globalen in [regijski, dolgoročen, evropski, konkurenčen, trajen]

globalen in [krajeven, domač, mednaroden, splošen]

- *Medtem ko lepota stalno tekmuje z novimi trendi zahodne oblačilne mode, je oblačenje vedno vpeto*

globalen *pridevnik* P 3000

1. **svetovni; mednarodni**
   1.1 splošno veljaven; razširjen
2. **zemeljski; planetarni**
3. **ki zadeva celoto; celostni**

1. svetovni; mednarodni

globalni procesi, zlasti gospodarski in politični, zajemajo ves svet

- *V New Yorku naj bi državniki in podjetniki razpravljali o **globalni** varnosti.*
- *Tudi največje svetovne firme, ki danes obvladujejo **globalni** trg, so se razvile iz malih podjetij in obratov.*
- *Motorola je zaradi **globalne** recesije v visokotehnoloških gospodarskih panogah lani odpustila 48.400 zaposlenih.*

1.1 splošno veljaven; razširjen

če postanejo neke dejavnosti ali lastnosti globalne, jih upošteva vedno več ljudi ali držav po svetu

- *Merila, kakšna ženska je lepa, postajajo vse bolj **globalna**.*

2. zemeljski; planetarni

ekologija

globalne spremembe v okolju vplivajo na celoten zemeljski planet

- *Eden najbolj preprostih in praktično izvedljivih načinov za zmanjšanje **globalnega** segrevanja je sajenje novih dreves.*
- *Krčenje ledenikov in spremembe v pokrajini sta zelo očitni in lahko razumljivi posledici **globalnega** ogrevanja.*
- *Vodik bo tudi občutno skrčil emisije ogljikovega dioksida in učinke **globalne** otoplitve.*

globalen *pridevnik* /globálen/ P 3000

1. **svetovni; mednarodni**
   1.1 splošno veljaven; razširjen
2. **zemeljski; planetarni**
3. **ki zadeva celoto; celostni**

1. svetovni; mednarodni

globalni procesi, zlasti gospodarski in politični, zajemajo ves svet

- *V New Yorku naj bi državniki in podjetniki razpravljali o **globalni** varnosti.*
- *Tudi največje svetovne firme, ki danes obvladujejo **globalni** trg, so se razvile iz malih podjetij in obratov.*
- *Motorola je zaradi **globalne** recesije v visokotehnoloških gospodarskih panogah lani odpustila 48.400 zaposlenih.*

1.1 splošno veljaven; razširjen

če postanejo neke dejavnosti ali lastnosti globalne, jih upošteva vedno več ljudi ali držav po svetu

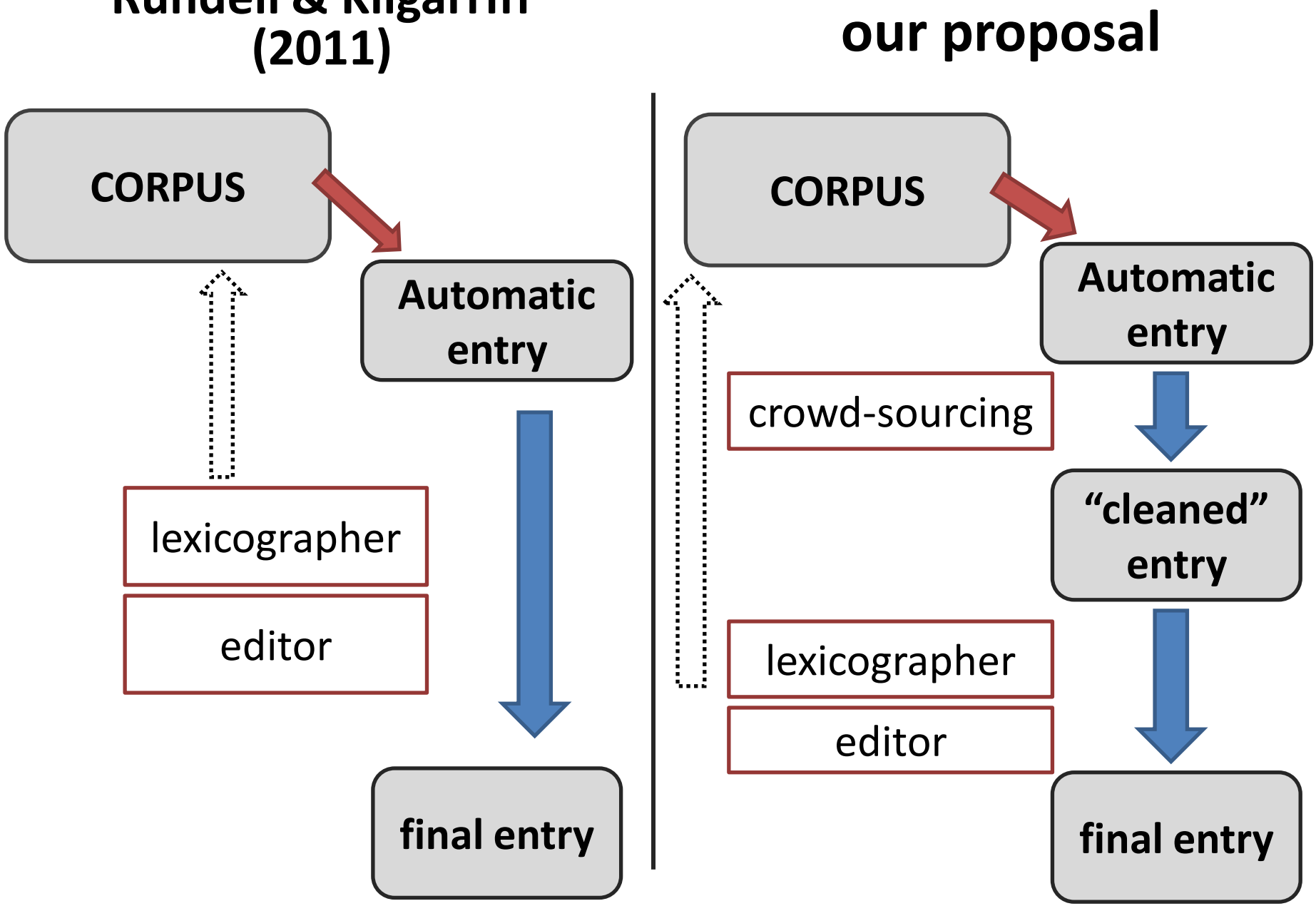
- *Merila, kakšna ženska je lepa, postajajo vse bolj **globalna**.*

2. zemeljski; planetarni

ekologija

globalne spremembe v okolju vplivajo na celoten zemeljski planet

- *Eden najbolj preprostih in praktično izvedljivih načinov za zmanjšanje **globalnega** segrevanja je sajenje novih dreves.*
- *Krčenje ledenikov in spremembe v pokrajini sta zelo očitni in lahko razumljivi posledici **globalnega** ogrevanja.*
- *Vodik bo tudi občutno skrčil emisije ogljikovega dioksida in učinke **globalne** otoplitve.*


Rundell & Kilgarriff (2011)
CORPUS
Automatic entry
lexicographer
editor
final entry
our proposal
CORPUS
Automatic entry
crowd-sourcing
"cleaned" entry
lexicographer
editor
final entry

Consortium

- University of Ljubljana
  - Faculty of Arts
  - Faculty of Social Sciences
  - Faculty of Pedagogy
  - Faculty of Computer Sciences and Informatics
  - Faculty of Electrotechnics
- University of Maribor
- University of Primorska
- "Jožef Stefan" Institute
- Amebis, Alpineon, Trojina
- http://www.cjvt.si/projekti

Plan

- 50 researchers
- Upgraded concept & proposal: spring 2015
- Already confirmed
  - crowdsourcing
  - automatic data extraction
  - multi-phase publication
- Work on
  - technical aspects
  - lexicographical aspects
  - design