

# NEW SLOVENE SKETCH GRAMMAR FOR AUTOMATIC EXTRACTION OF LEXICAL DATA

Simon Krek

Amebis. d.o.o.. Kamnik. Slovenia

Jožef Stefan Institute. Slovenia

# We'll talk about...

- Slovene Lexical Database structure (brief recapitulation from SKEW-2)
- New features in the new sketch grammar for automatic extraction of data
- Transfer of information from grammar/corpus to SLD (gramrels, collocations, examples)
- Post-processing and import to iLex
- Discussion and future

## I. LEMMA

- headword
- part-of-speech

*svitati se (to dawn)*

verb

## II. SENSE

- indicator

1. *daniti se (day)*

2. *dojemati (understand)*

- semantic

unary  
relations &  
constructions

a DAN.

hajati sonce

če se ČLOVEKU začne svitati o nekem  
**DOGAJANJU**. začne dojemati. kar  
prej ni vedel. ali pa je bilo to pred  
njim skrito

gramrels

## III. SYNTAX

- 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)

- 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 pričinja svitati o nekdanji  
zvezi ned Chadom in Heather.

- multi-word unit

## VI. PHRASEOLOGY • phraseological units

# Sketch Grammar for Slovene

- ver. 15: syntactic patterns for SLB
- 32 gramrels
- 18 DUAL
- 5 TRINARY
- 1 UNARY
- 1 SYMMETRIC
- 7 regular
- ver. 16: in progress
- introduction of
  - SEPARATEPAGE
  - CONSTRUCTION
- together with the switch to Gigafida
- before summer 2011

# New features

- use of macros
  - MTE -> JOS tagset
  - 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

# Direct relation SLD-grammar (1)

- \*DUAL

=kakšen?/kdo-kaj?

2:pridevnik brez\_GSD\_pd{0,5} 1:samostalnik

- # LBS-01 #####

/1/ <struktura>pbz0 SBZ0</struktura> ||

/2/ <struktura>PBZ0 sbz0</struktura>

# Direct relation SLD-grammar (2)

- \*SEPARATEPAGE komu-čemu\_g3
- \*TRINARY
  - =%s\_g3
    - 1:glagol sise{0,2} 3:predlog brez\_GSVD{0,5}  
2:samost\_daj
    - 3:predlog brez\_GSVD{0,5} 2:samost\_daj sise{0,1}  
1:glagol
- # LBS-17 #####  
<struktura>GBZ %s sbz3</struktura>

# \*SEPARATEPAGE (very complex)

- \*SEPARATEPAGE

- \*TRINARY

- NOUN\_1-noun\_2 (5 gramrels for 5 gram. cases)

- noun\_2-NOUN\_1                    -"-

- VERB\_1-noun\_2                   -"-

- NOUN\_1-verb\_2                   -"-

- ADJ\_1-noun\_2                   -"-

- adj\_2-NOUN\_1                   -"-

- 30 gramrels in SEPARATEPAGE (naming!)

# 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

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

# NOUN\_1-noun\_2

dobiti

(glagol)

FidaPlus (20M) freq = 11162 (735.5 per million)

displaying only: koga-česa\_g2

[whole word sketch](#)

<a href="#">brez-d_g2</a>	21	3.7	<a href="#">iz-d_g2</a>	107	5.0	<a href="#">na-d_g2</a>	5	3.1	<a href="#">za-d_g2</a>	14	4.4	<a href="#">od-d_g2</a>	206	17.9
laganje	1	9.61	Nigerija	2	8.82	priložnost	1	3.18	spletka	1	8.64	dalajlama	2	8.05
recept	6	8.34	onostranstvo	1	8.21	stran	3	3.12	pust	1	8.33	gradbenik	2	7.93
odredba	1	6.69	arest	1	8.14	sredstvo	1	2.41	karta	1	5.09	prednik	2	7.43
natečaj	1	6.06	katran	1	8.13				mleko	1	4.25	Avstrijec	2	7.34
težava	9	4.87	Ligojna	1	8.13	<a href="#">o-d_g2</a>	1	8.0	denar	4	3.8	Pelhan	1	7.3
razpis	1	3.95	Juršinci	1	8.07	informacija	1	2.91	naloga	1	3.56	Adanič	1	7.29
problem	1	3.02	sukcesija	1	8.03	<a href="#">zaradi-d_g2</a>	18	2.0	stanovanje	1	3.29	Izabela	1	7.21
razlog	1	2.89	limfa	1	8.01	kolegialnost	1	10.68	vloga	1	2.64	deklič	1	7.17
			vrečica	1	7.92	črevesje	1	8.06	pravica	1	2.12	Lek	2	7.15
			ZPIZ	1	7.91	panika	1	7.32	cesta	1	1.92	FIBA	1	7.13
			proračun	20	7.74	taktika	1	7.17				NEK	1	7.12
			NT	1	7.71	obnašanje	2	7.15				Uefa	1	7.11
			Montreal	1	7.71	noša	1	7.08				bršljan	1	7.09
			Kremelj	1	7.63	prostornina	1	6.63				Portugalska	1	6.98
			mozeg	1	7.62	pnevmatika	1	6.32				vol	1	6.93
			Carigrad	1	7.6							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

"subject"

"indirect  
object"

"direct  
object"

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

<u>z_nikalnim</u> 155 7.3	<u>s_prislovom</u> 112 3.3	<u>kakšen-p?</u> 6 1.2	<u>S_V_O3_O4</u> 47 18.0	<u>S_V_O3_O2</u> 18 5.2
mir 36 9.37	dušek 2 8.62	Hast 1 12.19	Mikulín 1 9.39	Kacin 1 9.64
soglasje 15 8.6	močnik 1 8.13	Jehan 1 12.19	Henigman 1 9.3	gostilničar 1 8.64
veto 2 8.03	spodbuda 4 7.79	Votan 1 12.0	Požun 1 9.19	Istrabenz 1 8.15
gol 9 7.97	tornado 1 7.76	jurjev 1 11.19	razvijalec 1 8.61	dekan 1 7.93
predujem 1 7.5	individualnost 1 7.76	kratek 1 3.36	Tudman 1 8.54	pod 1 6.57
maksimum 1 7.36	zalet 1 7.57	dober 1 0.63	Kristus 1 8.14	namestnik 1 6.3
frustracija 1 7.32	samozavest 2 7.45		Jarc 1 7.94	plod 1 5.89
pokoj 1 7.06	šarm 1 7.33	<u>kakšnega-p</u> 1 0.9	iskrica 1 7.9	mora 1 5.83
golaž 1 6.94	drobiž 1 7.31	preprost 1 4.82	Pahor 1 7.81	Rusija 1 5.72
kis 1 6.9	kad 1 7.17		Harry 1 7.53	uvedba 1 5.45
povod 1 6.83	plastenka 1 7.14		mineral 1 7.45	profesor 1 4.92
malica 1 6.79	optimizem 2 7.14		pečat 1 7.15	Krka 1 4.5
mladič 1 6.68	poudarek 3 7.12		pobudnik 1 7.1	oblast 1 3.14
breza 2 6.64	morala 1 7.09		Bill 1 7.02	večina 1 2.57
rezultat 13 6.58	pianist 1 7.05		prireditelj 1 6.96	območje 1 2.48
odgovor 8 6.47	kovanec 1 6.93		gospoda 1 6.92	človek 2 2.12
plaketa 1 6.32	avtonomija 1 6.89		govornik 1 6.72	država 1 0.84
dar 1 6.29	modrost 1 6.43		Washington 1 6.53	
modrost 1 6.28	odpoved 1 6.3		Sevnica 1 6.27	
prid 1 6.03	prid 1 6.16		testiranje 1 6.23	
odmerek 1 5.98	pooblastilo 1 6.1		Cerkev 1 6.04	
pojasnilo 1 5.87	skrb 3 6.06		volilec 1 5.86	
dovoljenje 5 5.76	žito 1 5.93		potrošnik 1 5.82	
opis 1 5.66	gol 2 5.86		Martin 1 5.8	
moka 1 5.65	inflacija 1 5.83		Hrvaška 1 5.01	

# 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 Breclj je uredil  
na terenu, še preden mednarodna **skupnost da** ZN mandat za ukrepanje, " je izjavil Anan  
Jelovec, del Sredme). **</p><p>** 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) **</p><p>** Kongresno **testiranje** varnosti **dalo** porazno sliko **</p><p>** Kot švicarski sir **</p>**  
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)
  - "ananas in banana" (pineapple and banana)
- 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

preposition


noun

preposition

# Example SkE: "in relation to"

corpus: FidaPlus (20M)

<b>odnos</b> (samostalnik)	
----------------------------	--



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

# 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 corresponding examples

# Gramrel to <vzorec>

Construction to <vzorec>

```
<skladenjska_struktura>
```

```
<vzorec>S_V_03_04</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>

unary to label: "with proper names"

<zaglavje>

<besvrs>samostalni</besvrs>

<oznaka>z\_lastnim\_imenom</oznaka>

</zaglavje>

# Post-processing

- Gramrels translated to standard SLD wording in structures, patterns, combinations and labels
- Inclusion of the headword in collocations in appropriate positions
- Collocations and headwords in appropriate case / number etc.
- TO BE DONE

# iLex software

iLEX v 2.3 053 Licensed to Trojina, zavod za uporabno slovenistiko. Copyright (c) 2004-2011 EMP ApS Username: SimonKrekNtb

File Edit Structure Reference List Functions Goto View Window

LBS-25-12-20... x mesto LBS-25-12-2012 (Entry Document)

Documents

Look up Design


mesto  
mesoreznica  
mesto  
mesto  
metafora  
metaforičen  
metropola  
metuljček  
meziniec  
migrirati  
mikroskop  
mimika  
mineštra  
mir  
miš  
mišica  
miška  
mlaka  
mlakuža  
mlečen  
mleko  
mleti  
mobilizirati  
močnik  
moda  
moden  
modificiran  
modificirati  
modno  
modras  
modrc  
modrček  
modrijan  
modrovati  
mojstrsko  
moka  
močlati  
moledovati

■ [pri kar]  
■ zgledi (50)  
■ bb) struktura: koga-kaj  
■ kolokacije  
■ [zasesti]  
■ [osvojiti]  
■ [razpisovati]  
■ [zasedati]  
■ [objavljati]  
■ [prevzeti]  
■ [priboriti]  
■ [zapustiti]  
■ [zagotoviti]  
■ [izboriti]


■ zgledi  
● Z zmago so si zagotovili 8. **mesto** in s tem obstanek v najmočnejši evropski skupini.  
● Imate preveč dela, da bi za nekaj trenutkov zapustili delovno **mesto** in si privoščili kakovostno malico?  
● V kvalifikacijah je s 595 krogi osvojil četrto **mesto** v svoji skupini ter prav tako četrto v celotni konkurenci.  
● Ubili ste ga, da bi prevzeli njegovo **mesto**.  
● Ko je ta po letu odšel, je zasedel njegovo **mesto**.  
● Glede na dobre rezultate si je priborila drugo **mesto** v državni reprezentanci.  
● objavlja prosto delovno **mesto** za določen čas 1 leta  
● Dejan kot vedno tudi v letošnji sezoni blesti, saj je trenutno prvi podajalec lige, po točkah pa zaseda peto **mesto**.  
● Zato razpisujemo nova delovna **mesta**.  
● S svojimi prijemi so si izborili **mesto** celo v evropski vojaški industriji.  
● Če seveda ne bodo pred tem uresničene naložbe, ki naj bi zagotovile nova delovna **mesta**.  
● Rudolf je sicer že zdavnaj zapustil revno **mesto** svoje mladosti, toda v spominu se je vse življenje vračal vanj.  
● Kljub temu padcu pa je slovenska ekipa osvojila četrto **mesto** med ekipami oziroma drugo med državami.  
● Po vojni izkušnji in boleznih je leta 1921 prevzel **mesto** profesorja na idrijski realki.  
● B, je zasedel šesto **mesto**.  
● Vsekakor pa bi si moral Maier v ekipi za svetovno prvenstvo šele priboriti svoje **mesto**, tako kot vsi drugi avstrijski reprezentantje.  
● Za delo na navedenih oddelkih objavljamo naslednja delovna **mesta**:  
● Prva dela danes kot svetnik v enem od slovenskih veleposlaništev v Evropi, druga pa zaseda zelo visoko **mesto** v Drnovškovi vladi.  
● razpisuje naslednja prosta delovna **mesta**:  
● Ob športu, ki je uspešnejši, si bo morala kultura šele izboriti svoje **mesto**, tudi s transparentnostjo in organiziranostjo tega trga.  
● Real je v tem krogu premagal doma PAOK in si je verjetno že zagotovil prvo **mesto** v skupini.  
● Ne čudi torej, da jih je upravičeno strah, da članom operativnih enot ob intervencijah v tovarnah ne bi dovolili, da bi zapustili delovno **mesto**.  
● Prvič v dveh sezonah sem osvojila prvo **mesto** v skupnem seštevku.

Introducing iLEX

**iLEX**  
Integrated Lexicography, Editing and Xml



iLEX v 2.3 053  
iLEX-XML-database (Local) v 2.0047  
java 1.6.0\_29



Copyright © 2005-2009 Erlandsen Media Publishing ApS.  
All rights reserved

iLEX Software Product License

iLEX Third Party Licences

# Discussion and future 1

- We are satisfied with the results so far
- We expect we will be able to put the data online directly for selected headwords
- Candidates for immediate online publication will be selected after we process and analyse 500 headwords

# Discussion and future 2

- The system allows for the use of the same data for further lexicographic processing, in NLP applications and for immediate use
- Final results in SKEW-4 // eLex2013
- FP7 language technology call – Beckerdee