

Automatic extraction of data: Slovenian case revisited

Simon Krek

Jožef Stefan Institute

Iztok Kosem

Trojina, Institute for Applied Slovene Studies

Polona Gantar

Fran Ramovš Institute of the Slovenian Language

We'll talk about...

- Slovene Lexical Database (SLD) structure (brief recapitulation from SKEW-2+3)
- Data extraction using Sketch Engine (API)
- 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

Procedure

- Selection of lemmas
- Finely-grained sketch grammar, designed specifically for the purposes of data extraction
- GDEX (Good Dictionary Examples) configuration(s)
- API script to extract data from word sketch information in the Sketch Engine
- Settings for extraction

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

Transfer of information

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

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

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

mesto LBS-25-12-2012 (Entry Document)

■ [pri kar]
zgledi (50)
bbj 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 boleznj 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

Visualization of SLD on the web

WEB DICTIONARY OF THE SLOVENE LANGUAGE

A DEMO VISUALIZATION OF THE SLOVENE LEXICAL DATABASE

A B C Č D E F G H I J K L M N O P R S Š T U V Z Ž



Seznam ▾

mečkati *glagol*

1 delati gube

1.1 gubati se

2 božati in objemati se

2.1 gnesti in stiskati

3 obotavljati se

1 delati gube

če ČLOVEK mečka PREDMET ali MATERIAL, ga z ROKAMI stiska tako, da postane zguban ali stisnjen

- KDO/KAJ ►
mečkati [papir]
mečkati v [rokah]
mečkati med [prsti]

- Veselo je pogledala skozi okno, Sandro pa je živčno **mečkal** neke papirje.
- Potno čelo je komaj utegnilo odkimati, tako hitro so debeli prsti **mečkali** papir med blazinicami.
- Iz žepa je potegnil še en robček, ki pa je bil že ves raztrgan, saj ga je prej ves čas **mečkal** s prsti.
- Medved se ustavi tudi pri odpiralnih za vino, stiskalnicah - ob bežnem pogledu na antično stiskanje, ko so grozdje **mečkali** z nogami in potem sok zlili v amfore, zakopane v zemlji.
- Zbrcala sem jih na kup, začela sem jih **mečkati** z rokami, potlej pa sem jih še pohodila.
- Tyrén je **mečkal** v rokah svojo kapo z napisom »OK«.
- Karel Zelden je nataknil **mečkal** v rokah časovni plan vseh opravil Bojevnikovega natovarjanja, kjer so bile odločilne operacije dodatno poudarjene z rdečo barvo.
- Ko v rokah **mečkam** medvedka, čutim ljubezen svoje hčerke

Več zgledov ...

- GLAGOL+NEOLOČNIK
[začeti] mečkati

- Marjorie je iz torbice vzela robček in ga začela **mečkati**.
- Hana je začela **mečkati** pismo, rada bi si ga ogledala na samem, ko jo je »navedek«a prestregla, rekoč: Kaj skrivaš?

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 lemmas 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).
 - The grammatical relation should be specified as a regular expression (especially '%s_s4' should be written as '.*_s4').
 - If more expressions match a given relation, then the first record is applied (hence '.*' is a fallback record for all relations).
 - 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.

Admin

- -U URL, --URL=URL
 - URL of the SkE server (path excluding run.cgi). By default: 'http://localhost/'
 - SkE server: 'http://beta.sketchengine.co.uk/bonito/')
- -u USERNAME, --username=USERNAME
- -p PASSWORD, --password=PASSWORD

Debug & some more admin

- -d, --debug
 - Prints out responses for debugging purposes.
- -t URLTMPL, --template=URLTMPL
 - Template to create SkE API query.
- -C LOGINPAGE, --CA_auth=LOGINPAGE
 - In case of using SkE server:
<https://beta.sketchengine.co.uk/login/>

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 <i id="1338652551">anoreksiji</i> 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 <i id="1338652551">anoreksiji</i> dvignila veliko prahu, je umrla.</zgled></zgledi>`

Post-processing

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

Discussion & future

- Used in the making of a new dictionary of Slovene (proposal – 4M €, <http://www.sssj.si/>)
 - five phases: data extraction used for the first phase
- Parseme (COST action)
 - PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing
- Beckerdee II – H2020?, CEF?
 - Definition extraction
 - Distributional semantics