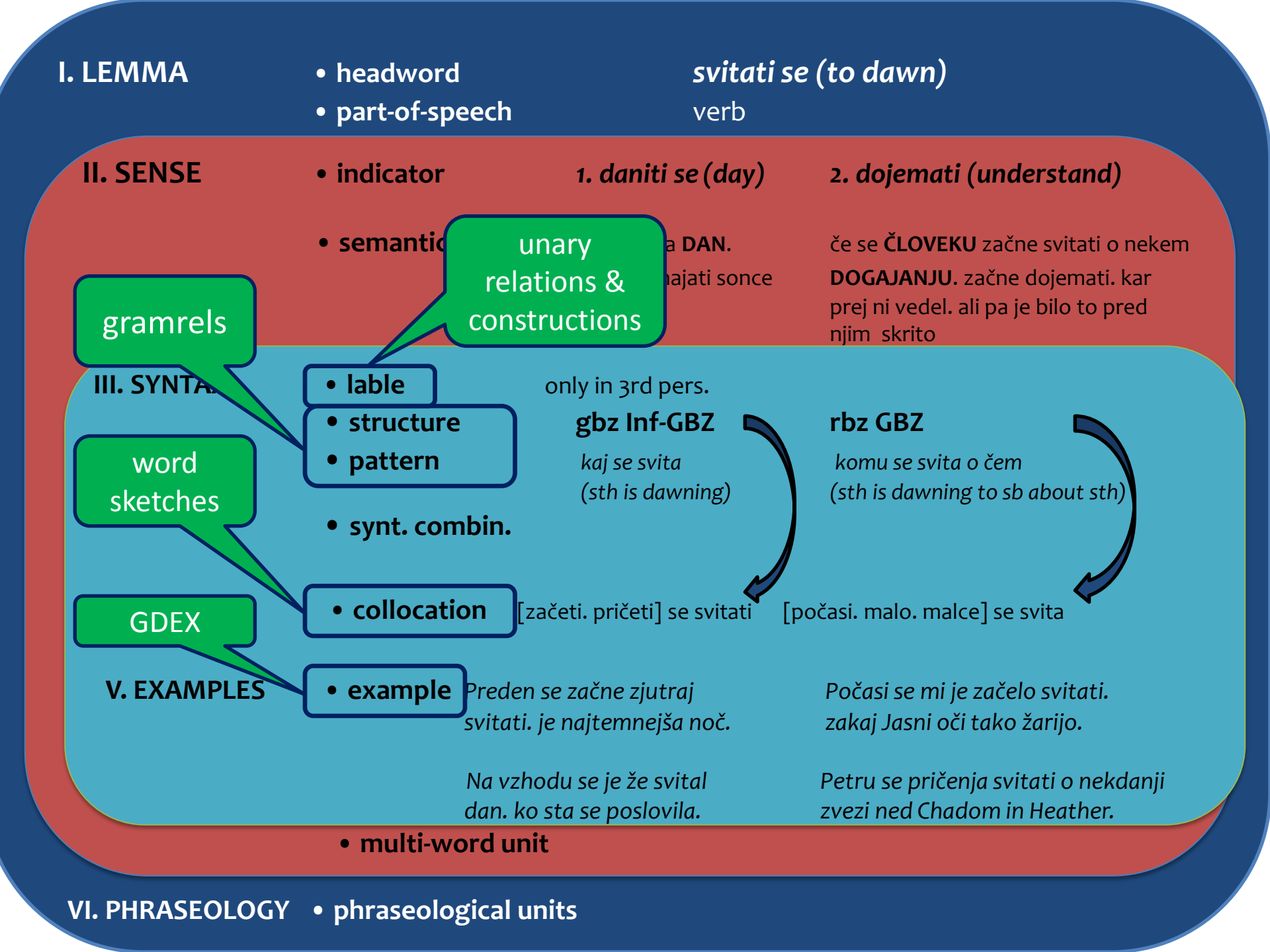
Automatic extraction of data: Slovenian case revisited

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
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
only in 3rd pers.
• structure
• pattern
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

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

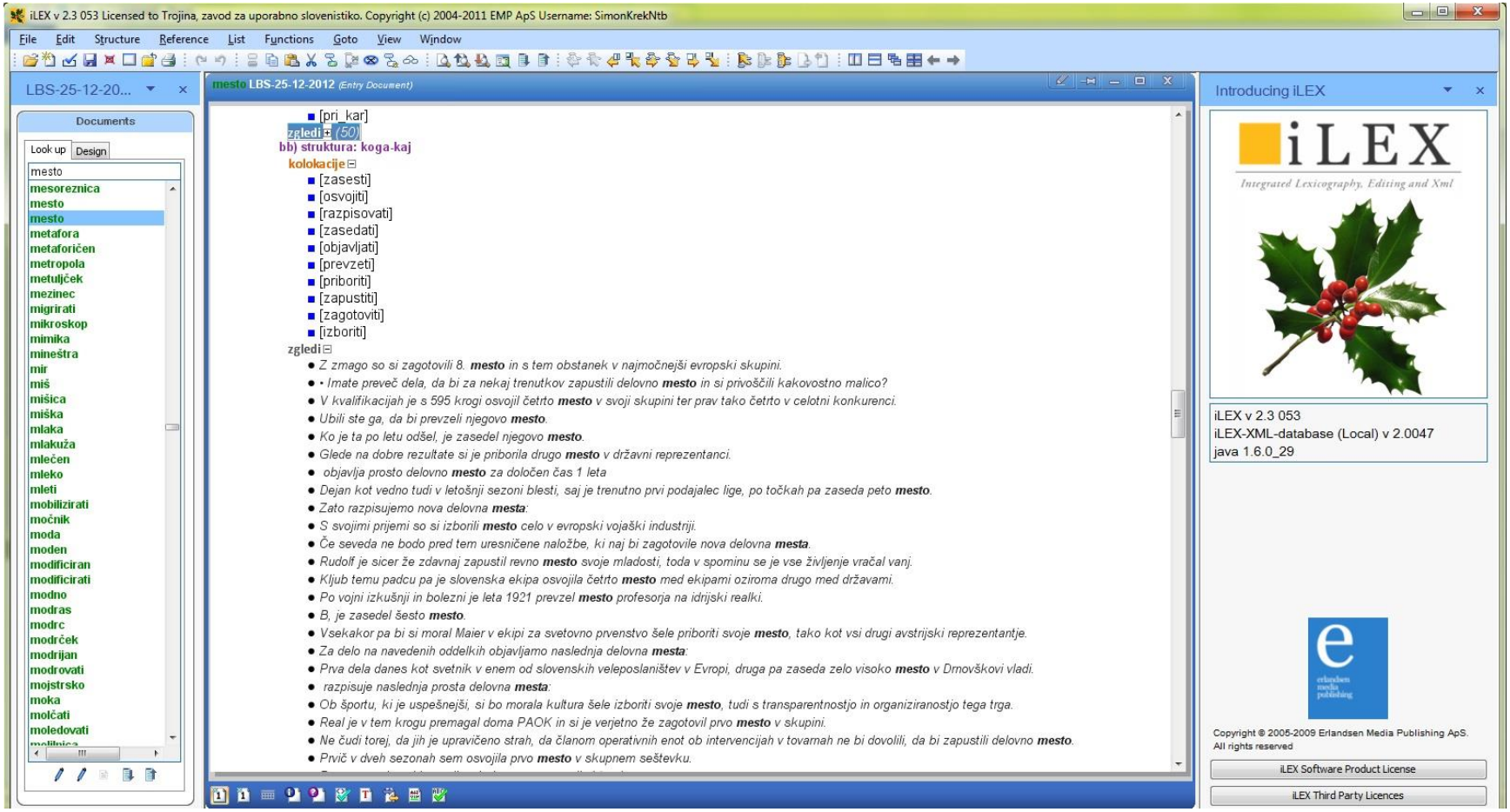
Sketch Grammar for Slovene (v.16)

- 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



Visualization of SLD on the web

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).
  - 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 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>

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