Slovene Lexical Database
automatic extraction and crowdsourcing

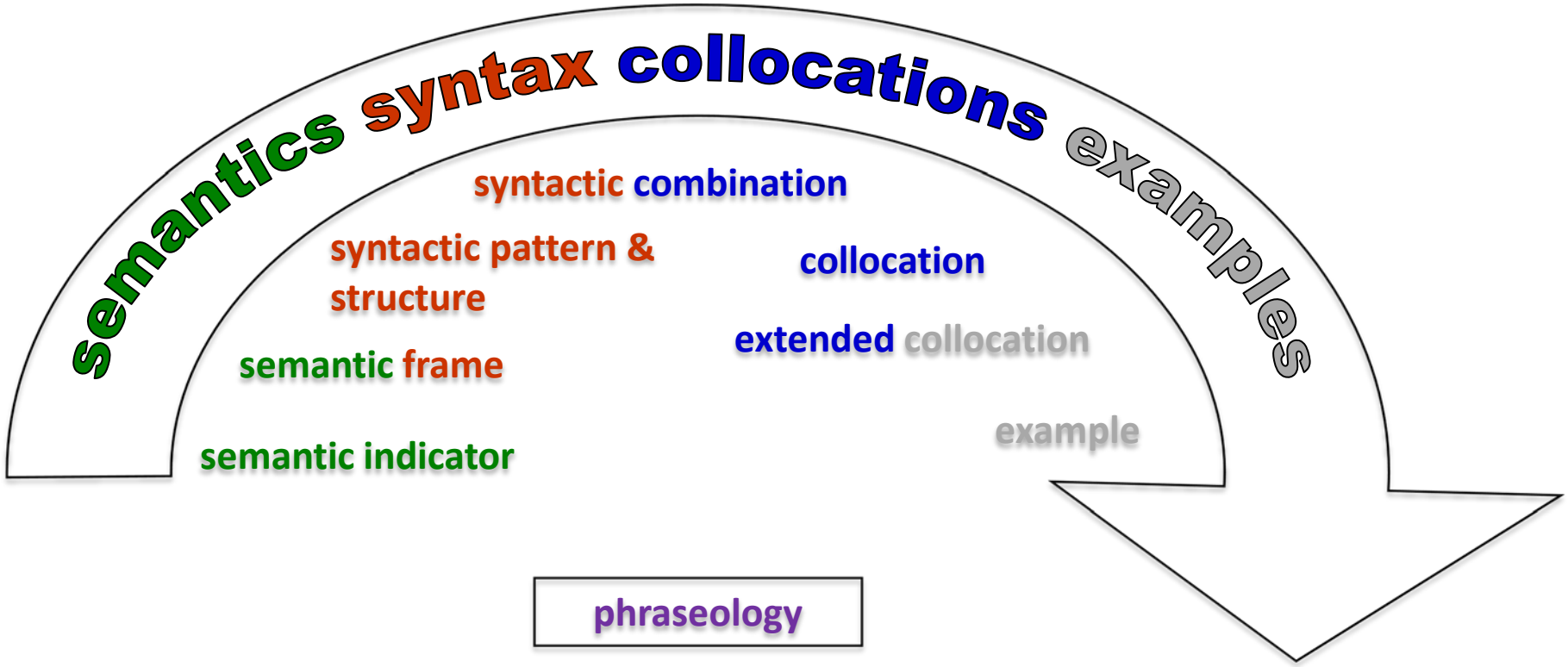
Simon Krek
„Jožef Stefan” Institute
Iztok Kosem
Trojina, Institute for Applied Slovene Studies
Polona Gantar
Fran Ramovš Institute of the Slovenian Language

Plan

- Slovene Lexical Database
- Extraction of data (Sketch Engine)
    - Sketch Grammar
    - GDEX (Good Dictionary EXamples)
- Workflow / crowdsourcing
- ACDC (Automatically Constructed Dictionary Content)

SLD Basics

- corpus data analysis
- lexicogrammatical approach
  - semantics and syntax are not separated
- meaning = meaning potential
  - is not stable (norms & exploitations)
- lumpers vs. splitters = splitters
- lexicography first, NLP second


semantics syntax collocations examples
syntactic combination
syntactic pattern & structure
collocation
extended collocation
semantic frame
example
semantic indicator
phraseology

I. LEMMA

- **headword** — ***svitati se (to dawn)***
- **part-of-speech** — verb

II. SENSE

- **indicator** — ***1. daniti se (day)*** | ***2. dojemati (understand)***
- **semantic frame** — ko se svita **DAN**. začne vzhajati sonce | če se **ČLOVEKU** začne svitati o nekem **DOGAJANJU**. začne dojemati. kar prej ni vedel. ali pa je bilo to pred njim skrito

III. SYNTAX

- **lable** — only in 3rd pers.
- **structure** — **gbz Inf-GBZ** | **rbz GBZ**
- **pattern** — *kaj se svita (sth is dawning)* | *komu se svita o čem (sth is dawning to sb about sth)*
- **synt. combin.**

IV. COLLOC.

- **collocation** — [začeti. pričeti] se svitati | [počasi. malo. malce] se svita

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**

I. Lexical Unit

- link to the lexicon
  - morphosyntactic information
  - corpus frequency
  - pronunciation etc.
- additional grammatical information
  - un/countability, part-of-speech subtypes etc.

II. Semantic Level

- Semantic Indicators
  - simple EFL-like explanations or synonyms forming a sense menu
  - self-explanatory in relation to each other
- Semantic Frames
  - COBUILD / FrameNet / Corpus Pattern Analysis
  - combination of the systems

Semantic Indicators – koža (skin)

koža *samostalnik*

1. vrhnji del telesa

1.1 pri človeku

1.2 pri živali

2. odstranjen vrhnji del živalskega telesa

3. ovoj ali lupina

Semantic Frames

- identification of verb/semantic arguments
  - prototypical pattern – “the norm” (Hanks)
  - the headword in its syntactic environment
- identification of semantic types in particular syntactic positions
- the semantic scenario
  - a full-sentence definition making a link between the arguments and the situation (FN) typical for a particular sense

Semantic Frame

2. dojemati

2.1 nekaj vedeti

če se ČLOVEKU svita o nekem DEJSTVU, potem o tem nekaj ve ali sluti

– semantic types in capital letters (ID-ed)
– linked with collocates via syntax

III. Syntactic Level

- **syntactic structures** (formal)
  - clause and phrase level (all POS; only for NLP)
  - the number of syntactic structures is finite
  - source: word sketches (Sketch Engine)
- **syntactic patterns**
  - valency (mainly verbs; for lexicography and NLP)
- **syntactic combinations**
  - more than basic patterns: „pasti za X stopinj"

Syntactic Structures – koža

4 vrhnji del telesa

1.1 pri človeku

- pbz0 SBZ0 [občutljiva, suha, mastna] koža
- SBZ0 sbz2 koža [obraza, telesa, rok, lasišča]
- SBZ0 pod sbz6 koža pod [pazduho, očmi]
- gbz SBZ4 [dražiti, pomirjati, hladiti] kožo

Syntactic Patterns – svitati se

2. dojemati

2.1 nekaj vedeti

če se ČLOVEKU svita o nekem DEJSTVU, potem o tem nekaj ve ali sluti

- komu se svita se o čem
- komu se svita kaj

IV. Collocation Level

- **SEMANTIC FRAME**:

  če se ČLOVEKU svita o nekem DEJSTVU, potem o tem nekaj ve ali sluti

- **SYNTACTIC STRUCTURES AND PATTERNS**:

  NOUN – koža
  pbz0 SBZ0
  SBZ0 sbz2
  SBZ0 pod sbz6
  gbz SBZ4

  VERB – svitati se
  komu se svita se o čem
  komu se svita kaj

If a part of syntactic patterns are collocational, they are shown on the collocation level.

- **COLLOCATIONS**

- [občutljiva, suha, mastna] koža
- koža [obraza, telesa, rok, lasišča]
- koža pod [pazduho, očmi]
- [dražiti, pomirjati, hladiti] kožo

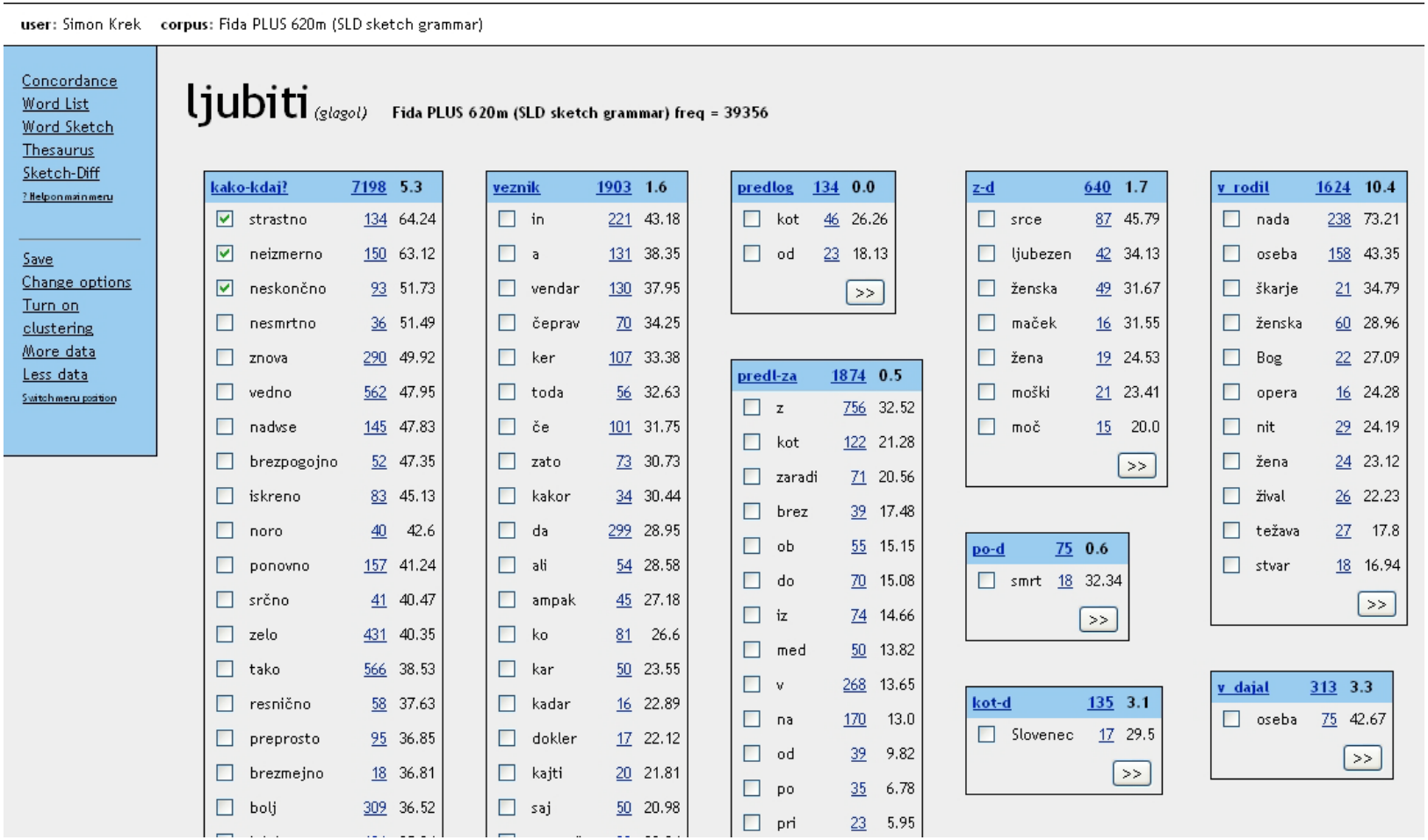
I. Examples

- **COLLOCATIONS**
  - [občutljiva, suha, mastna] koža
  - koža [obraza, telesa, rok, lasišča]
  - koža pod [pazduho, očmi]
  - [dražiti, pomirjati, hladiti] kožo

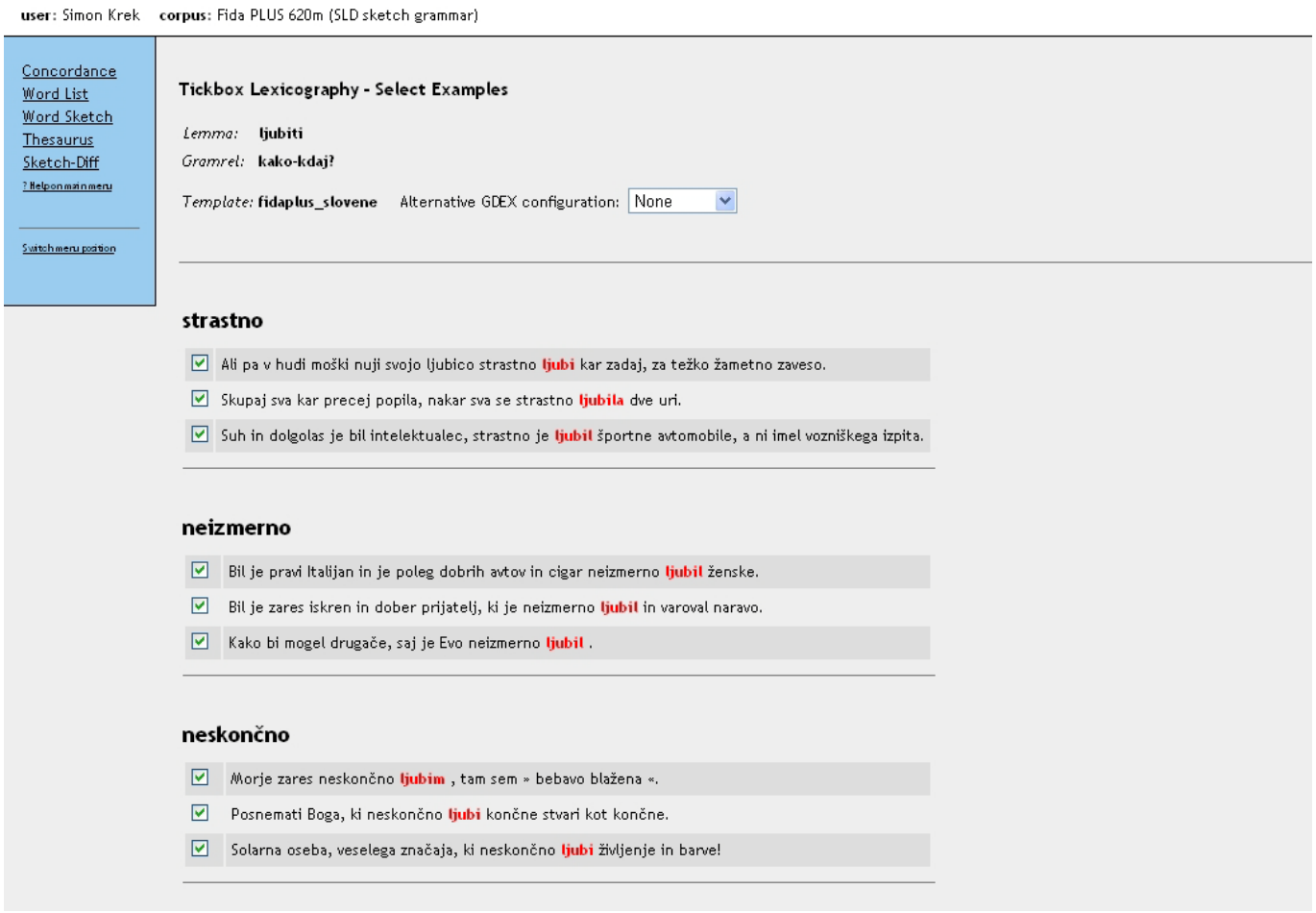
- **EXAMPLES**
- *Tonik je namenjen je **občutljivi koži** in ne vsebuje alkohola.*
- ***Koža rok** postane pozimi občutljivejša.*
- *Opažate na **koži pod očmi** prezgodnja znamenja staranja?*

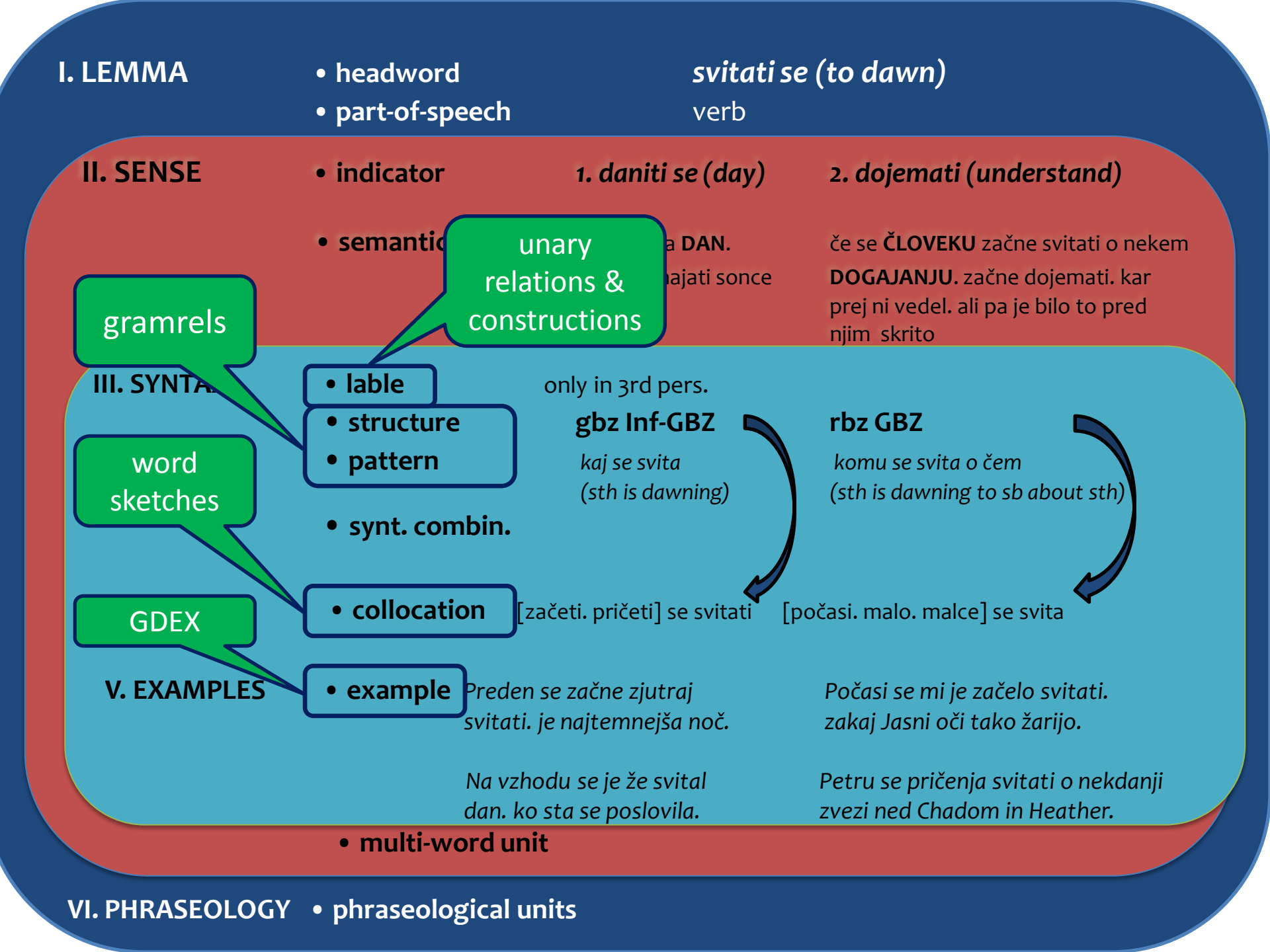
- *Se vam že kaj **svita**, o čem govorim?*
- *Petru pa se pričenja **svitati** o nekdanji zvezi med Chandlerjem in Heather.*
- *Holly je na svojem stolu v klubu Diva zastokala in se prijela za glavo, ko se ji je začelo **svitati**, kaj se bo zgodilo.*

Sketch Engine (word sketch)



Good dictionary examples (GDEX)




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.
GDEX
• collocation
[začeti. pričeti] se svitati
[počasi. malo. malce] se svita
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
- new „directives“
  - *SEPARATEPAGE
  - *CONSTRUCTION
  - *COLLOC

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: *SEPARATEPAGE

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

<struktura>GBZ %s sbz2</struktura>

- *SEPARATEPAGE koga-česa_g2
- *TRINARY

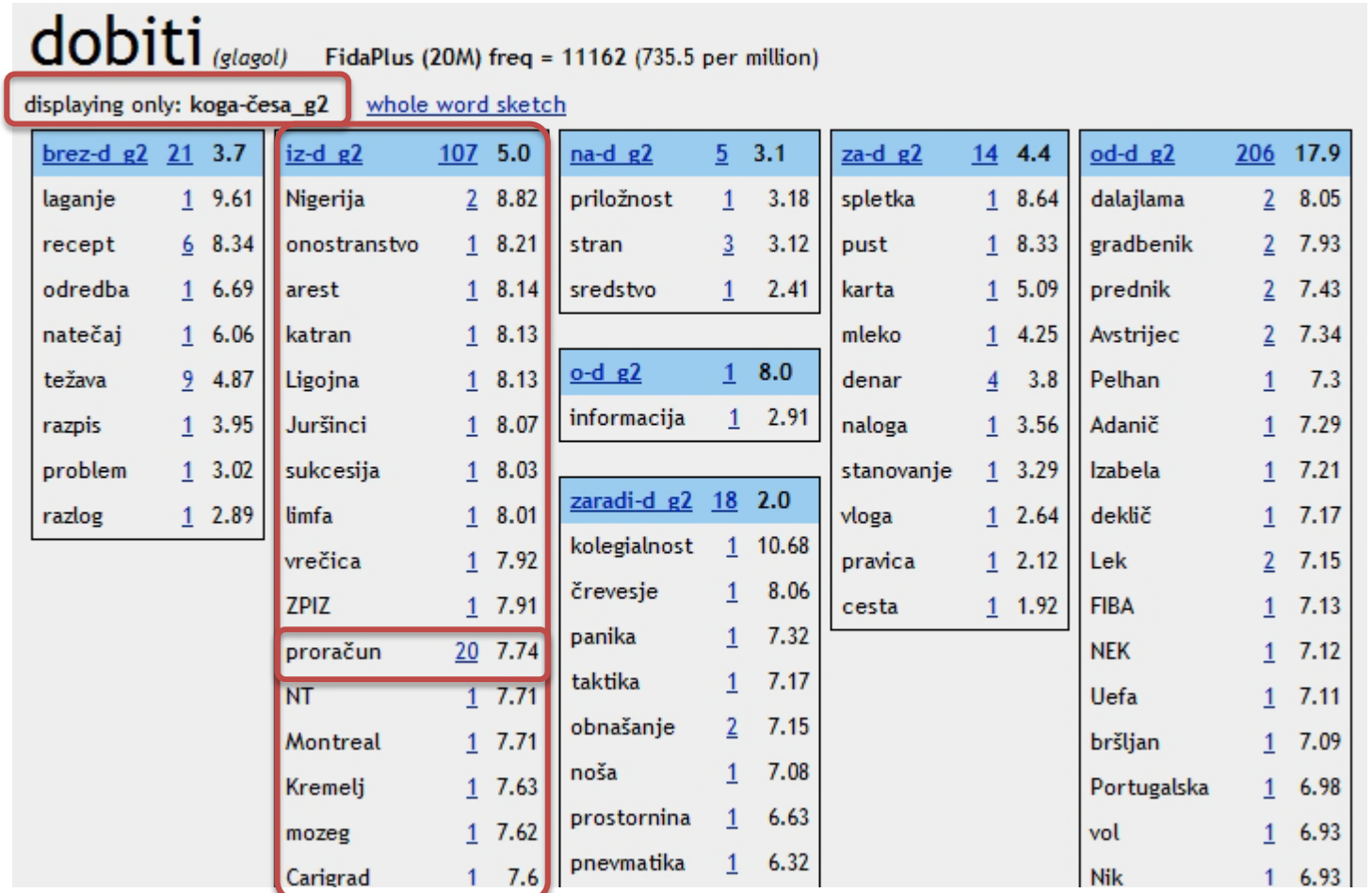
=%s_g2

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

Example: *SEPARATEPAGE



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

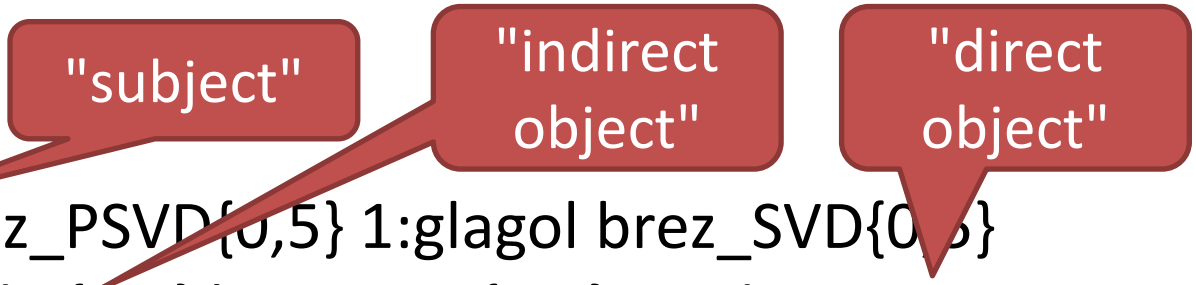
za-d_g2	14	4.4
spletka	1	8.64
pust	1	8.33
karta	1	5.09
mleko	1	4.25
denar	4	3.8
naloga	1	3.56
stanovanje	1	3.29
vloga	1	2.64
pravica	1	2.12
cesta	1	1.92

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

- 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: *CONSTRUCTION



=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

Examples – high precision

poročil z njo. Tako združene	***moči***	**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***	**dale**	čepice - Zadovoljstvo v Luksemburgu - Tudi
sta) Kongresno	***testiranje***	**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

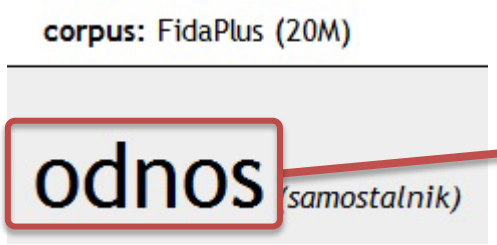
*COLLOC

- For „syntactic combinations“
- Element <zveza> = syntactic combinations
  - "v odnosu do (koga/česa)" (in relation to (sb/sth))
- Mainly nominal headwords
- Under (sub)sense after syntactic structures as a separate category

Example: *COLLOC

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

Example: "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

GDEX – **G**ood **D**ictionary **Ex**amples

- system for evaluation (ranking) of sentences with respect to their suitability to serve as dictionary examples
- sorting sentences so that good examples do not have to be searched for in hundreds of unusable sentences
- initially trained on English, but it did not give good results for other languages

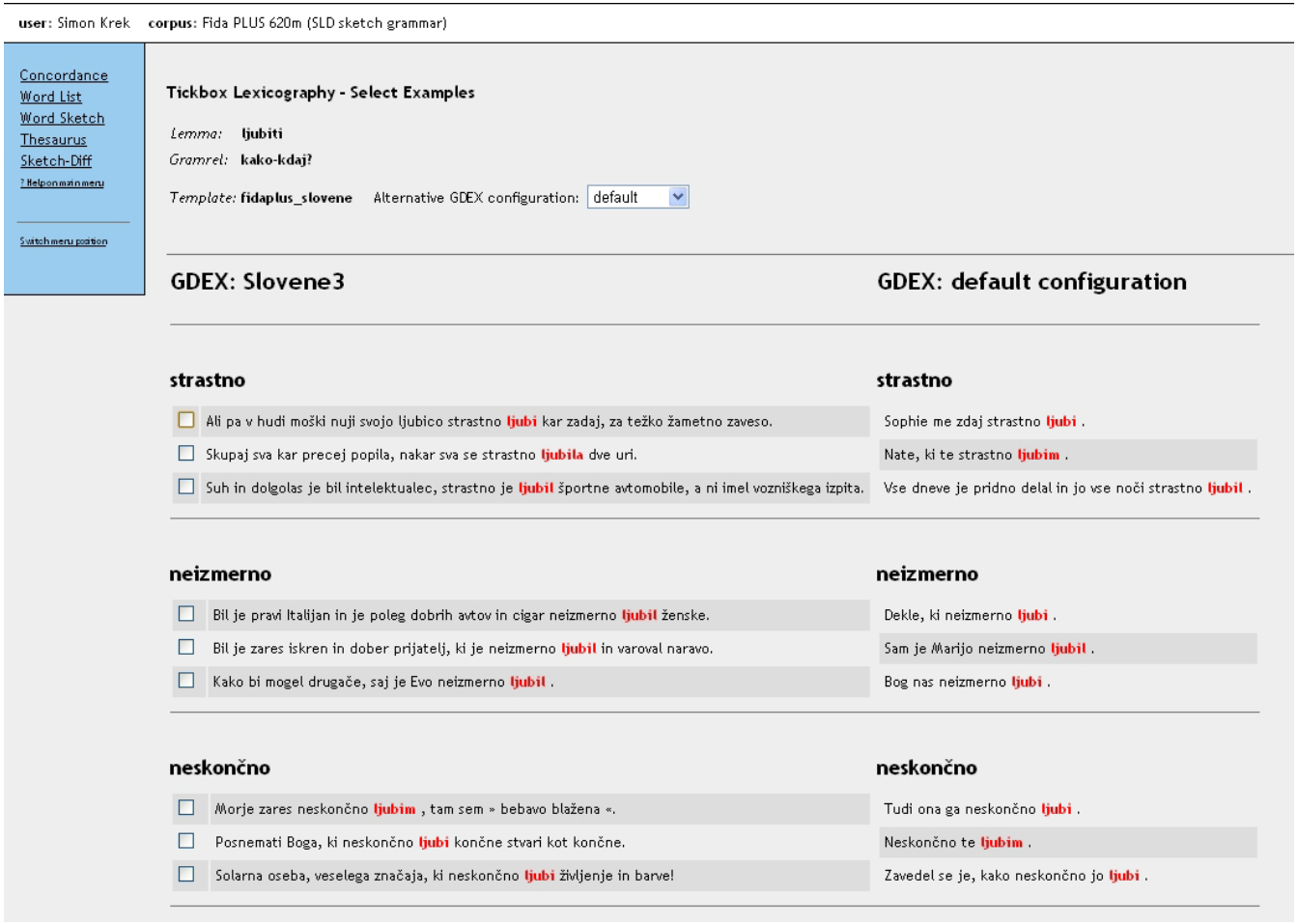
GDEX – configuration

- parameters in a GDEX configuration file
- GDEX Tools web-interface to create and use custom GDEX configurations
- the GDEX evaluation process
  - ranking of out-of-corpus sentences
  - evaluation of TBLex logs
  - cooperation with WEKA

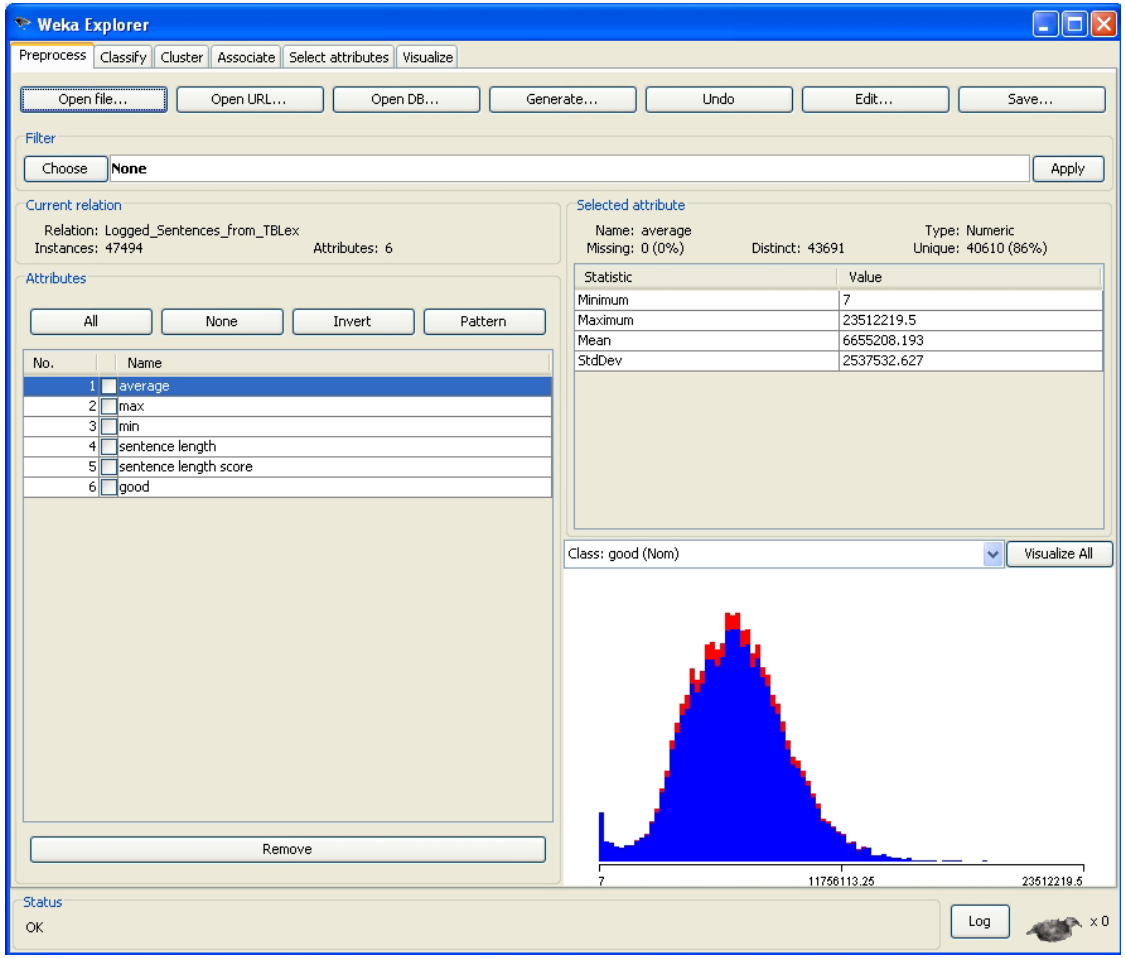
GDEX classifiers

- procedures that quantify measurable features of sentences or tokens
- sentence classifiers: sentence length, keyword position, etc.
- token classifiers: token frequencies, matches to RE, etc.

Evaluation of TBLex logs



Cooperation with WEKA



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 <oznaka>

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

unary to label: "with proper names"

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 (each PoS)
  - 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)

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...

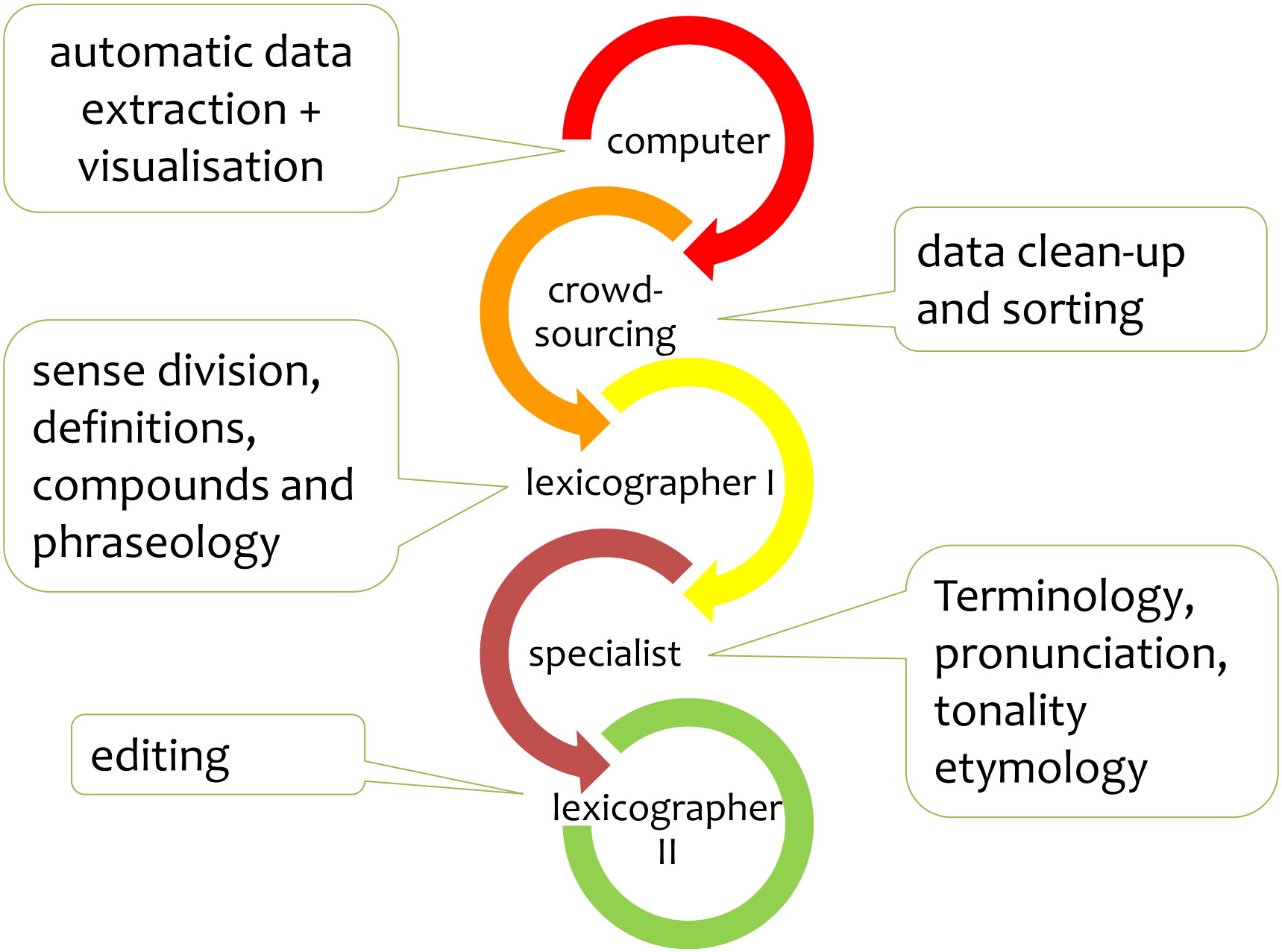
- Minimum number of collocates per relation was increased to 25
- Selection of relevant collocates was ‘left’ to minimum frequency and salience settings
- Number of examples per collocate was reduced to three
- We divided lemmas into frequency groups, and prepared separate settings for each group

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>


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

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.

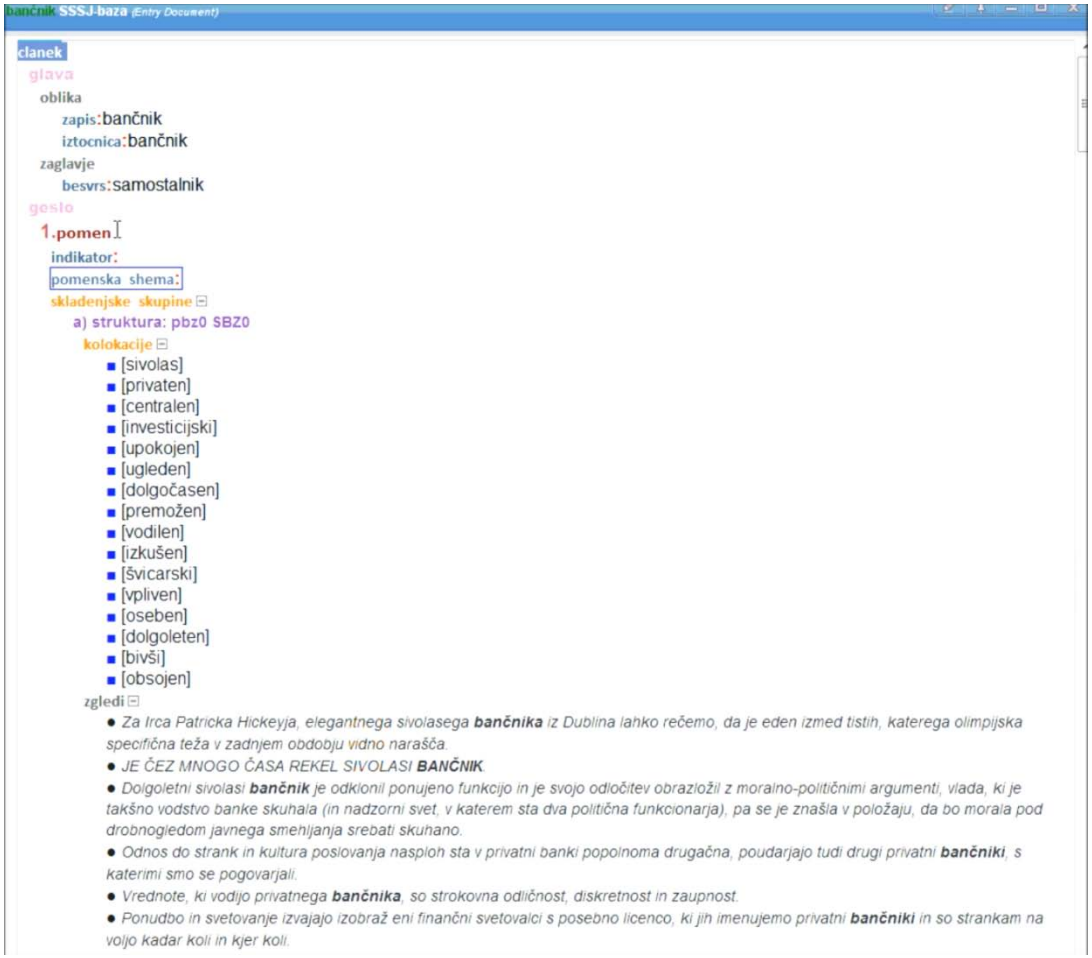
DA | NE | Ne vem

30%

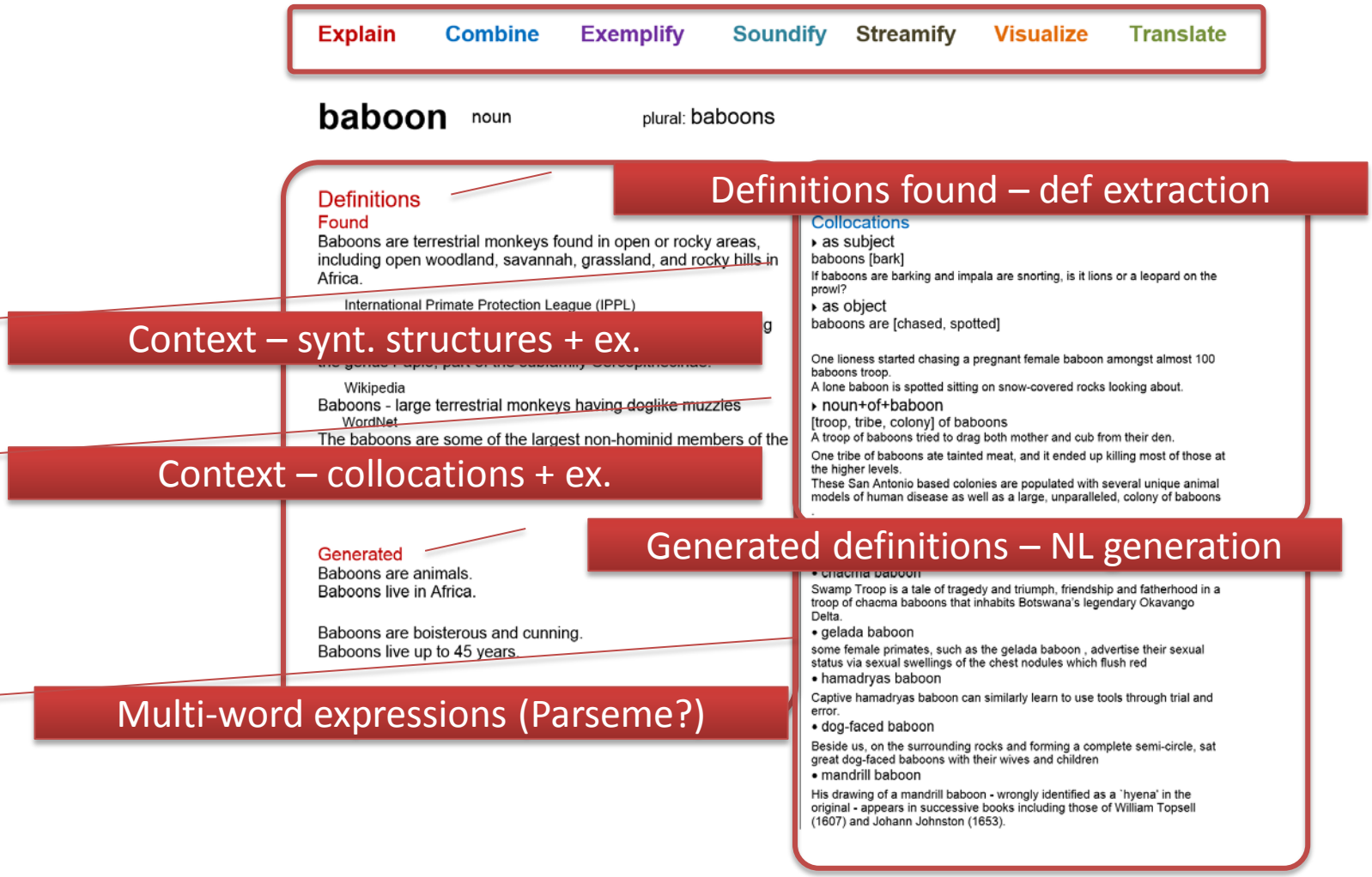
Work left for lexicographers

- 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.)

Lexicographer I.



ACDC


Explain
Combine
Exemplify
Soundify
Streamify
Visualize
Translate
baboon noun
plural: baboons
Definitions
Found
Baboons are terrestrial monkeys found in open or rocky areas, including open woodland, savannah, grassland, and rocky hills in Africa.
International Primate Protection League (IPPL)
Wikipedia
Baboons - large terrestrial monkeys having doglike muzzles
WordNet
The baboons are some of the largest non-hominid members of the
Generated
Baboons are animals.
Baboons live in Africa.
Baboons are boisterous and cunning.
Baboons live up to 45 years.
Collocations
as subject
baboons [bark]
If baboons are barking and impala are snorting, is it lions or a leopard on the prowl?
as object
baboons are [chased, spotted]
One lioness started chasing a pregnant female baboon amongst almost 100 baboons troop.
A lone baboon is spotted sitting on snow-covered rocks looking about.
noun+of+baboon
[troop, tribe, colony] of baboons
A troop of baboons tried to drag both mother and cub from their den.
One tribe of baboons ate tainted meat, and it ended up killing most of those at the higher levels.
These San Antonio based colonies are populated with several unique animal models of human disease as well as a large, unparalleled, colony of baboons
Swamp Troop is a tale of tragedy and triumph, friendship and fatherhood in a troop of chacma baboons that inhabits Botswana's legendary Okavango Delta.
• gelada baboon
some female primates, such as the gelada baboon , advertise their sexual status via sexual swellings of the chest nodules which flush red
• hamadryas baboon
Captive hamadryas baboon can similarly learn to use tools through trial and error.
• dog-faced baboon
Beside us, on the surrounding rocks and forming a complete semi-circle, sat great dog-faced baboons with their wives and children
• mandrill baboon
His drawing of a mandrill baboon - wrongly identified as a `hyena' in the original - appears in successive books including those of William Topsell (1607) and Johann Johnston (1653).
Definitions found – def extraction
Context – synt. structures + ex.
Context – collocations + ex.
Generated definitions – NL generation
Multi-word expressions (Parseme?)

Explain Combine Exemplify **Soundify** Streamify **Visualize** Translate

baboon noun plural: baboons

Sounds, Graphics and Visuals

Sounds

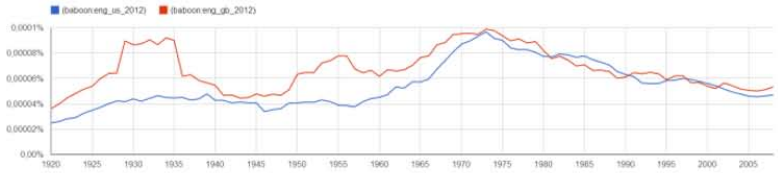
Recorded

baboon ▶

Speech Synthesis

baboon ▶

Graphics



Visuals

Images

http://www.image-net.org/ - Synset: baboon

http://www.image-net.org/ - Synset: chacma, chacma baboon

Videos

www.youtube.com

Explain Combine Exemplify Soundify **Streamify** Visualize Translate

baboon noun plural: baboons

Streaming

Twitter

Sam Kanyi @snkanyi Its a shame how my standards have dipped. I have to settle for those whose spelling is as bad as IR **baboon**'s! Smh

37m **Lesley Marshall** @lesleyrocksface Socotra Island Blue **Baboon**... @devonmassyn is the best boyfriend ever! I can't believe I got her for Christm

1h **____^ XinYee ♥** @dream180793 thanks for talking to me hahaha thanks **baboon ♥**

1h **k y l e †** @KylleJarvis people are arguing on here on about who got the best presents.. Be grateful you **baboon!**

2h **Yaʁa Darwish** ∞ hahahah you know me too well. I literally can't wait, counting down the seconds & milliseconds till you're here **baboon! ♥**

2h **Maximus aurelius** @don_jide How can this **baboon** next to me say the queen would get the D i am officially dead

News Feeds

BABOONS CAN RECOGNIZE WORDS The findings might weigh in on debates about how best to teach children to read...
http://news.discovery.com/animals/baboons-words-120412.html

Penn Researchers Connect Baboon Personalities to Social Success and Health Benefits...
http://www.upenn.edu/pennnews/news/penn-researchers-connect-baboon-personalities-social-success-and-health-benefits

IT PAYS TO BE A NICE BABOON Nice baboons enjoy stronger social circles and...
http://news.discovery.com/animals/baboons-nice-121002.html

Foraging Baboons Are Picky Punters: Baboon Foraging Choices Depend On Their Habitat and Social Status
http://www.sciencedaily.com/releases/2012/09/120913203917.htm

Showdown over baboons
Cape Town - Baboon conservationists, scientists, animal welfare activists and wildlife managers hold a no-holds-barred meeting on Monday...
http://www.iol.co.za/news/south-africa/western-cape/showdown-over-baboons-1.1421348#.UNnwFOQmbh4

Word-detecting baboons are a tough read
Studies differ on whether monkeys learn to recognize letter combinations
http://www.sciencenews.org/view/generic/id/346591/description/Word-detecting_baboons_are_a_tough_read

baboon noun plural: baboons

Translate

Language	Translation(s)
Basque	
Czech	pavián
Estonian	paavian
Slovenian	pavijan

Czech

Tak já ti něco řeknu , ty mechanická **opice** .

Představa lady Nicholsonové jako jeho choti , skutečnost , že její jemná krása byla někdy vydána zvrhlostem tohoto degenerovaného **paviána** , naplnila Doyla niternou zuřivostí .

Dále Slater cituje příklady , kdy se vampýři dělí o krev , kterou obstaral jen jeden z nich , a později si zas vyměňují úlohy , nebo kdy subdominantní **paviání** samec vyvádí takové kousky , že odláká pozornost hlavního samce v tlupě a další samec se mezitím může pářit z jednou ze samic . Poté dodává :

Medvědí ' oplátkový systém ' neprobíhá v podmínkách , kdy jeden oplácí druhému , jako tomu bylo u Slaterových příkladů s vampýry a **paviány** ;

Jejich zjevná radost nebyla tak úplně nepodobná známému štěstí návštěvníků zoologické zahrady , kterým se podaří zastihnout chocholatého právě v době krmení .

English

Just let me tell you something , my brass **baboon** .

At very high doses, a poor gastric tolerability (gastritis, gastric erosions and/or vomiting) of clopidogrel was also reported in rat and **baboon**.

The idea of Lady Nicholson as his spouse , that her handsome refinement had ever been subject to the vicissitudes of this **baboon** 's degeneracy , filled Doyle with moral outrage .

During non clinical studies in rat and **baboon**, the most frequently observed effects were liver changes.

Citing examples of vampire bats sharing blood that only one has obtained , then reversing roles later , and of sub-dominant male **baboons** taking turns distracting the alpha male so the other can mate with one of its females , Slater goes on to say :

At very high doses, a poor gastric tolerability (gastritis, gastric erosions and/or vomiting) of clopidogrel was also reported in rat and **baboon**.

A bear 's reciprocity is n't one-on-one like Slater 's examples of bats and **baboons** ;

During non clinical studies in rat and **baboon**, the most frequently observed effects were liver changes.

Their obvious delight was not entirely unlike the often seen elation of visitors to a zoo on coming upon the crested **baboon** at feeding time .

Slovenian

Pri zelo visokih odmerkih so pri podganah in **pavijanih** poročali tudi o slabem želodčnem prenašanju (gastritis, želodčne erozije in/ ali bruhanje).

V predkliničnih študijah na podganah in **pavijanih**, so bile jetrne spremembe najpogosteje opaženi učinek.

Pri zelo visokih odmerkih so pri podganah in **pavijanih** poročali tudi o slabem želodčnem prenašanju (gastritis, želodčne erozije in/ ali bruhanje).

V predkliničnih študijah na podganah in **pavijanih**, so bile jetrne spremembe najpogosteje opaženi učinek.

Plan

- Slovene Lexical Database
- Extraction of data (Sketch Engine)
    - Sketch Grammar
    - GDEX (Good Dictionary EXamples)
- Workflow / crowdsourcing
- ACDC (Automatically Constructed Dictionary Content)