

# TAGSET REDUCTION BASED ON THE FEATURES NEEDED FOR THE SLOVENE SKETCH GRAMMAR

Simon Krek

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

Jožef Stefan Institute, Slovenia

# We'll talk about...

- Slovene morphology
- Tagsets for Slovene
- MTE-JOS tagset
- Slovene corpora and lexical database
- Sketch grammar for Slovene lexical database
- MTE-JOS tagset reduction
- Tagging results with reduced tagsets
- Tagset reduction: conclusions

# Slovene morphology

- The extreme case of adjective:
  - inflectional categories
    - case: 6
    - gender: 3
    - number: 3 (dual!!)
    - definitiveness: 2
      - only nominative & accusative masculine singular
    - degree: 3
  - forms
    - 56 for positive degree
    - 164 for all three degrees

# Adjective "disgusting"

No	Form	Combinations
1	ogabnih	12
2	ogabni	9
3	ogabnima	6
4	ogabna	5
5	ogabnim	5
6	ogabne	4
7	ogabno	4
8	ogabnimi	3
9	ogabnega	2
10	ogabnem	2
11	ogaben	2
12	ogabnemu	2
ALL		56

# 12 combinations for -ih

case	number	gender
genitive	dual	masculine
genitive	dual	feminine
genitive	dual	neutral
genitive	plural	masculine
genitive	plural	feminine
genitive	plural	neutral
locative	dual	masculine
locative	dual	feminine
locative	dual	neutral
locative	plural	masculine
locative	plural	feminine
locative	plural	neutral

# Tagset standardization

- Eagles (*1993-1996*)
- Multext (*1994-1996*)
- Multext-East (*1995-1997*)
  - 1997 - ver. 1 MTE: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, English (7 languages)
  - 2010 - ver. 4 Mondilex: Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, Ukrainian (16 languages)

# From MTE to JOS

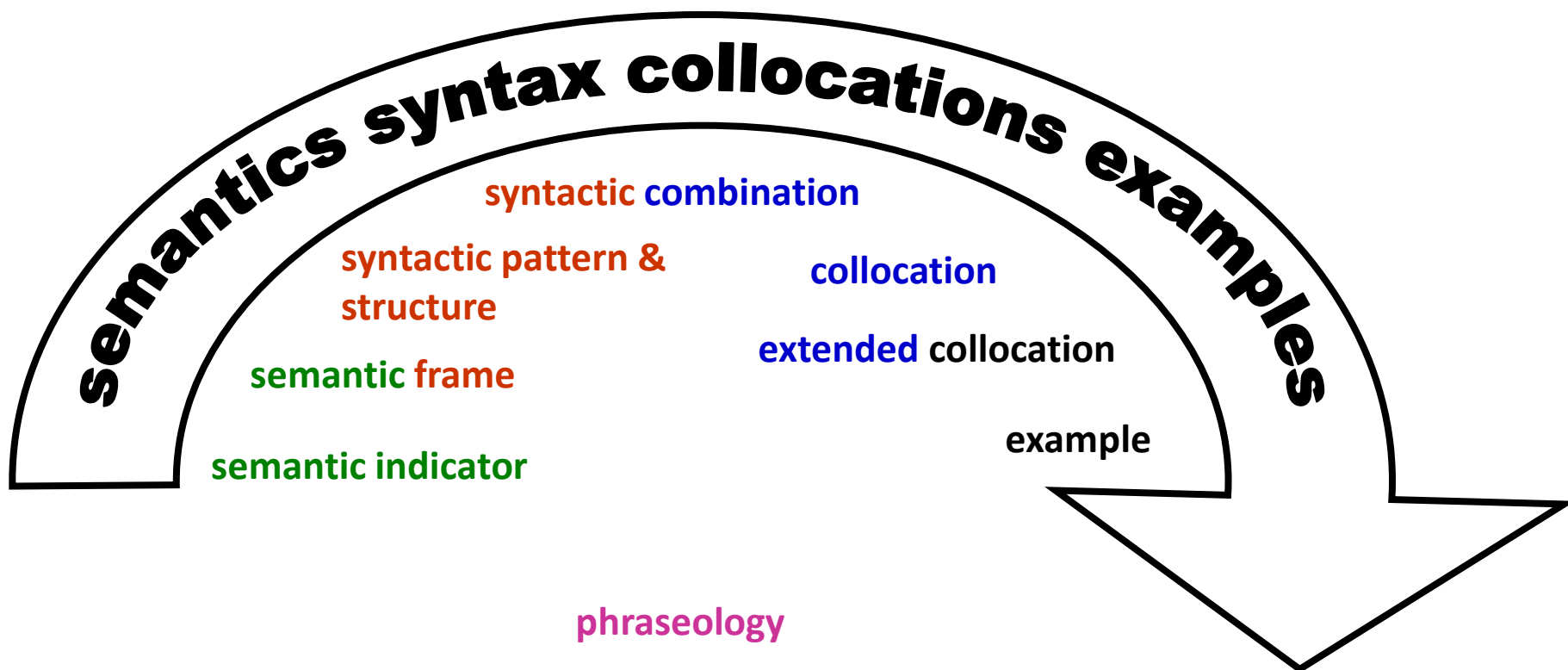
- JOS project (2007-2009)
- MTE principles: positional attributes etc.
- From syntax to morphology
- Reordering of attributes to make tags shorter
- Non-trivial mapping
- 12 categories, 15 attributes, 56 values
- 1.902 tags

# Slovene corpora

- FIDA (2000)
  - 100 million
  - MTE tagset, rule-based tagger
- FidaPLUS (2006)
  - 620 million
  - MTE tagset (2006) – rule-based tagger (Sketch Engine)
  - JOS tagset (2010), rule-based & statistical + metatagger
- Gigafida (2011)
  - 1.1 billion
  - JOS tagset (2011) – statistical tagger (Sketch Engine – before summer?)



# Slovene lexical database



## 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

- restriction

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.

- 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
- new directives
  - \*SEPARATEPAGE
  - \*CONSTRUCTION
- together with the switch to Gigafida?
- before summer 2011
- work on constructions
- info from the new dependency parser

# Tagset reduction – ver. 15

- 12 categories: 64 tags
- verb
  - **type**: main, auxilliary
  - **form**: infinitive, supine, participle, present, future, conditional, imperative
  - **person**: 1st, 2nd, 3rd
  - **negation**: yes, no
- noun
  - **type**: common, proper
  - **case**: nominative, genitive, dative, accusative, locative, instrumental
- adjective
  - **type**: general, participle
  - **case**: nominative, genitive, dative, accusative, locative, instrumental
- pronoun
  - **type**: personal, possessive, demonstrative, relative, reflexive, general, interrogative, indefinite, negative

# Tagset reduction – ver. 16

- 10 categories: 154 tags
- number (3/2): noun, verb, adjective
- degree (3): adjective, adverb
- type: numeral (3), adverb (3), conjunction (2)
- case (6): preposition
- number: adjective = no dual -> more than 1

# Statistical tagger – three tagsets

- 500.000 word training corpus

	JOS	JOS-L1	JOS-L2	JOS/L1	JOS/L2	L1/L2
Accuracy – known words	92.20	94.95	94.60	2.75	2.40	0.35
Accuracy – unknown words	64.02	73.17	72.46	9.14	8.43	0.71
Accuracy	89.66	92.99	92.60	3.33	2.94	0.39
Accuracy – known words (ctg)	97.89	98.27	98.33	0.38	0.44	-0.06
Accuracy – unknown words (ctg)	88.17	89.07	89.21	0.89	1.04	-0.14
Accuracy (ctg)	97.01	97.44	97.51	0.43	0.49	-0.07

# Conclusions

- 2.94% = 32,340,000 tokens (1.1 billion corpus)
- 0.43% = 4,730,000 tokens (1.1 billion corpus)
- significant: unknown words
- not significant: categories
- Reduced tagset: automatic extraction of data from the corpus
- Full tagset: manual corpus analysis in Sketch Engine & other concordancers