

Slovene Lexical Database & Slovene Sketch Grammar

Simon Krek

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

Jožef Stefan Institute, Ljubljana, Slovenia



Overview

- project
- lexical database
- data extraction from corpora
- English entry
- discussion & future
- -----
- Slovene Sketch Grammar



Framework

- The operation is partly financed by the European Union,
 - the European Social Fund, and
 - the Ministry of Education and Sport of the Republic of Slovenia.
- The operation is being carried out within the operational programme
 - Human Resources Development for
 - the period 2007–2013, developmental priorities:
 - improvement of the quality and efficiency of educational and training systems 2007–2013.

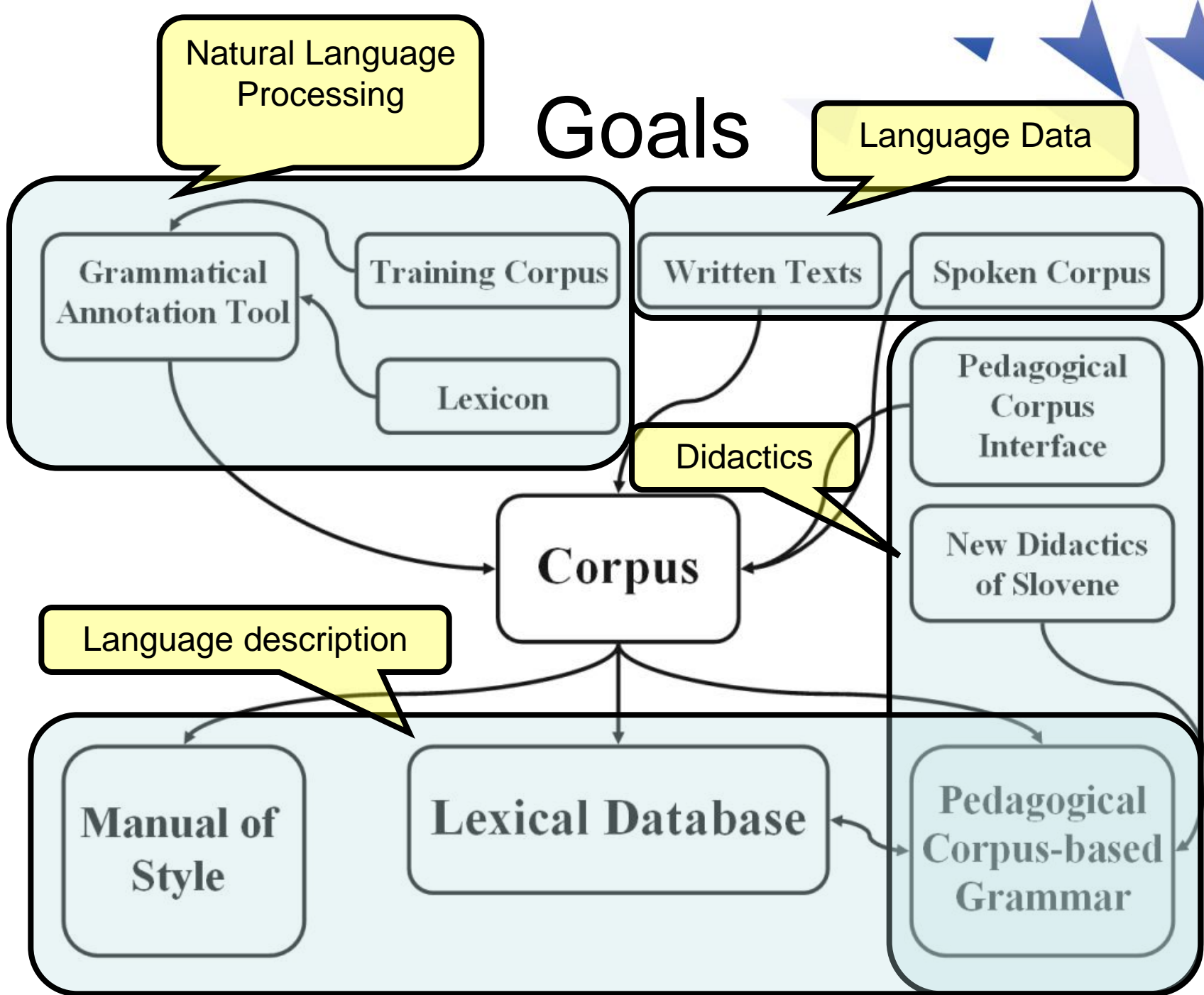


Project

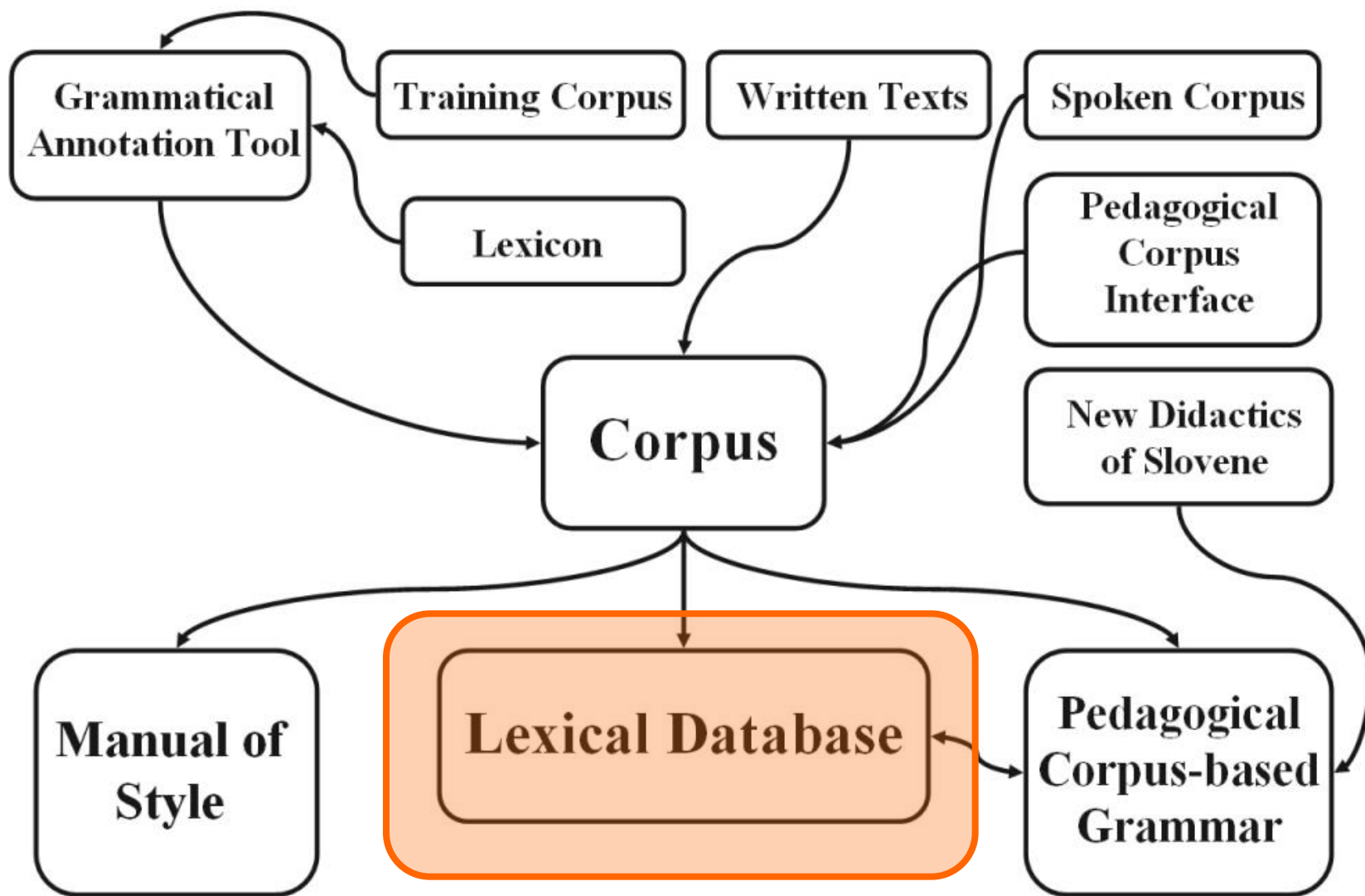
- “Communication in Slovene”
- Web site: <http://www.slovenscina.eu>
- Leading partner: Amebis, d. o. o., Kamnik
- Duration: June 2008 - December 2013
- Total value: 3.2 million Euro
- Project consortium:
 - [Amebis, d.o.o., Kamnik](#)
 - [Jozef Stefan Institute](#)
 - [University of Ljubljana](#)
 - [Scientific Research Centre of the Slovenian Academy of Sciences and Arts](#)
 - [Trojina, Institute for Applied Slovene Studies](#)



Goals



Goals



Timeline

- June – October '08: preparation
- November '08 - June '09: specifications
- June 2010: 1/3
- June 2011: 1/3
- June 2012: 1/3
- Number of lexical units: minimum 2,500



Legal aspects

- Creative Commons

- Attribution
- Share Alike
- Noncommercial



- Availability

- On-line
- Data set

- Owner: Ministry of Education and Sports

Lexical Database

- Dictionary of Linguistics (<http://www.sil.org/>)
 - an organized description of the lexemes of a language
 - a lexeme is the minimal unit of language which has a semantic interpretation and embodies a distinct cultural concept
- Patrick Hanks (<http://www.slovenscina.eu/>)
 - a summary of corpus evidence for each word in the language
 - the focus is typically on syntagmatics and collocations
 - also on lemmatization, morphology, and meaning
 - a primary resource for many applications
 - dictionary writing (monolingual, bilingual)
 - course-book writing
 - education and error correction
 - natural language processing and artificial intelligence
 - codifying the relative importance of each sense of a word



Inspiration

- **International (early):**
 - GENELEX (1990-94)
 - LE-PAROLE (1993-98)
 - SIMPLE (1998-2000)
 - ACQUILEX I, II (1989-1995)
 - CEGLEX (1995-1996)
 - DELIS (1993-1995) ...
- **Individual languages:**
 - [elexico](#) (SP), [ADESSE](#) (SP), [GRIAL](#) (SP), [CLIPS](#) (IT), [CORNETTO](#) (NL), [ALFALEX](#) (FR), [BLF](#) (FR), [STO](#) (DK), [SPRAKBANKEN](#) (S), PRALED (CZ), ...
- **Important for us:**
 - [FrameNet](#)
 - [Corpus Pattern Analysis](#)



Basics

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



Five levels + one

1. Lexical unit

- link to the lexicon in LMF

2. Semantic level

- semantic indicator
- sense frame

3. Syntactic level

- syntactic structure
- syntactic pattern
- syntactic combination

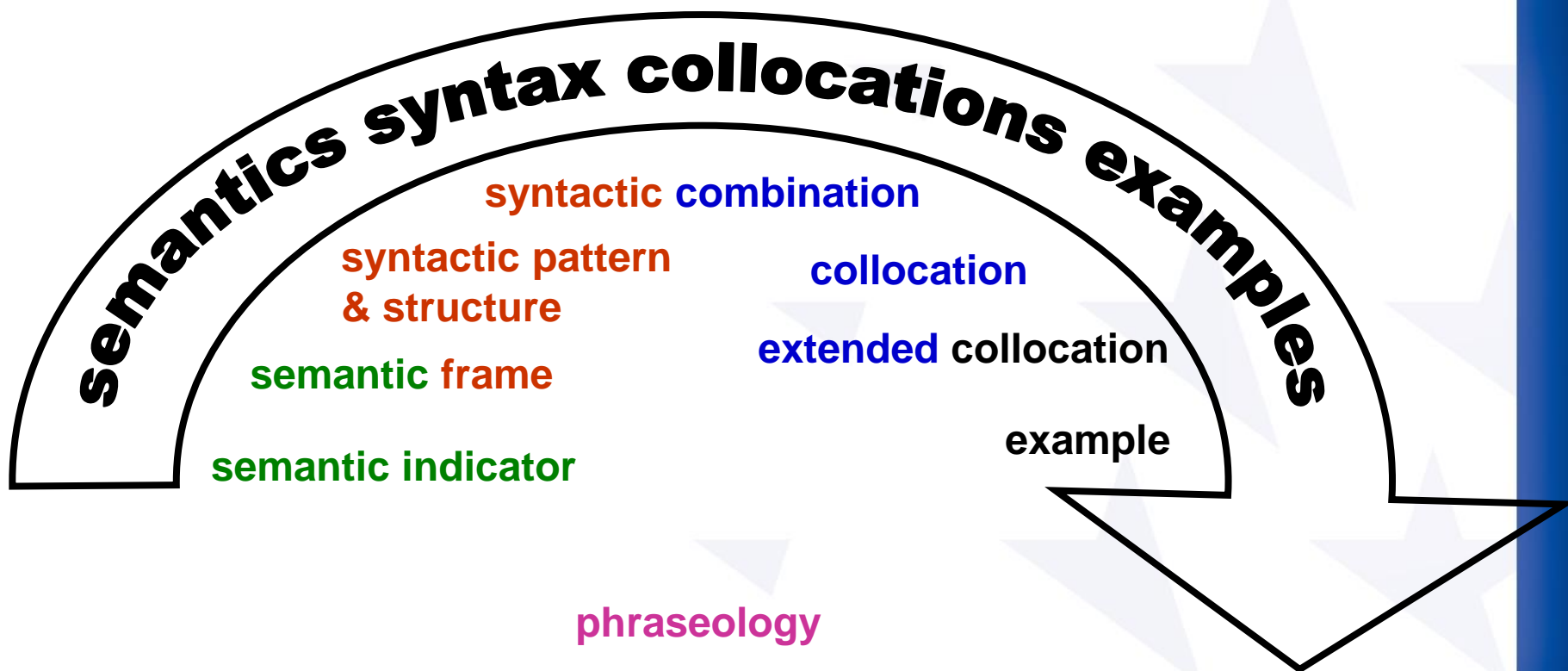
4. Collocation level

- collocation
- extended collocation

5. Corpus examples

6. Phraseology







Level	Type of data	Data
I. Lexical Unit	Headword	to squeeze
	Part-of-speech	verb
II. Semantic Level	Semantic Indicator	<i>press out liquid</i>
	Sense Frame	a PERSON squeezes a CONTAINER with liquid or squeezes LIQUID out of a CONTAINER
III. Syntactic Level	Pattern & Structure (1)	sb squeezes sth -> transitive
	Pattern & Structure (2)	sb squeezes sth out -> PV-out
IV. Collocation Level	Collocation (1)	to squeeze [a lemon, an orange]
	Collocation (2)	to squeeze out [water]
	Extended Collocation	to squeeze [lemon, orange] juice
V. Corpus Example	Example (1)	<i>Squeeze a bit of lemon juice onto the fish.</i>
	Example (2)	<i>Wet the sponge then squeeze out the excess water.</i>
VI. Phraseology	Phraseological Unit	to squeeze a quart into a pint pot
	Semantic Indicator	to attempt to do the impossible
	Corpus Example	<i>There will be little of real substance other than trying to squeeze a quart into a pint pot whilst keeping all of the really contentious areas largely unchanged.</i>

I. Lexical Unit

- link to the lexicon
 - morphosyntactic information
 - Multext-East / JOS tagset
 - corpus frequency
- additional grammatical information
- pronunciation etc.



II. Semantic Level

- Semantic Indicators
 - simple EFL-like explanations or synonyms
 - discrimination of senses within the LE
 - self-explanatory in relation to each other
- Semantic Frames
 - FrameNet / Corpus Pattern Analysis
 - simplified, “de-formalized”



Semantic Indicators - squeeze

- hold firmly
 - with your hand
 - with your fingers
 - with your arms
- extract substance
 - press out liquid
 - press out soft matter
- get into limited space
- obtain with difficulty
- just succeed
- persuade
- get maximum
- extort money
- get rid of
- make financial damage
- find time
- push body parts closer



Semantic Frame – extort money

Councils will want to squeeze as much *money* out of *taxpayers* as they can.

Dalglish last night attempted to squeeze *£250,000* out of *Portsmouth* for midfielder Steve Agnew.

a PERSON or an INSTITUTION
squeezes MONEY out of another
PERSON or INSTITUTION



III. Syntactic Level

- “Syntactic Patterns” (Lexicography)
 - sb squeezes sth out of sb
- “Syntactic Structures” (NLP)
 - transitive + |out of|



IV. Collocation Level

- **SEMANTIC FRAME:**

a PERSON or an INSTITUTION squeezes MONEY out of another
PERSON or INSTITUTION

- **SYNTACTIC PATTERNS:**

[sb] squeezes [sth] out of [sb]

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

- **COLLOCATIONS**

to squeeze [money, cash]



Hierarchy vs. direct linking

MONEY	a PERSON or an INSTITUTION squeezes out of another PERSON or INSTITUTION
[sth]	[sb] squeezes out of [sb]
[cash]	to squeeze



Squeeze – sense 4

D extort money

a PERSON or an INSTITUTION squeezes MONEY out of another PERSON or INSTITUTION

Structure: transitive |out of|

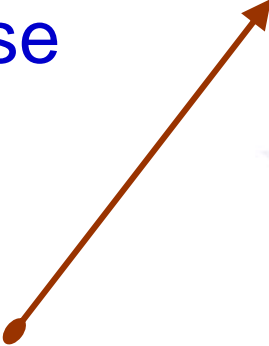
► sb squeezes sth out of sb

■ to squeeze [money, cash]

- *Councils will want to squeeze as much money out of taxpayers as they can.*
- *Dalglish last night attempted to squeeze £250,000 out of Portsmouth for midfielder Steve Agnew.*



DTD

- <!ELEMENT **entry**
 - (**sense**+, **phraseology**?) >
 - <!ELEMENT **sense**
 - (**indicator**,
 - **label***
 - **semantic_frame**,
 - **syntactic_groups**?,
 - **syntactic_combinations**?,
 - **subsense***,
 - **multiword_combinations**?) >
 - <!ELEMENT **Syntactic_groups**
 - (**syntactic_structure**+) >
 - <!ELEMENT **syntactic_structure**
 - (**structure**,
 - **pattern***,
 - **collocation***,
 - **examples**) >
- 



Corpus Data & Authoring Tools

- FidaPLUS: www.fidaplus.net
- Sketch Engine: www.sketchengine.co.uk
 - customized sketch grammar
 - Tickbox Lexicography
 - GDEX
- IDM Dictionary Production System
- custom DTD

FidaPLUS

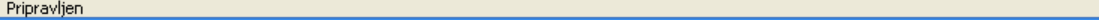
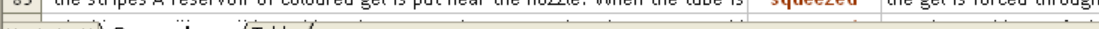
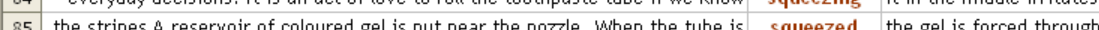
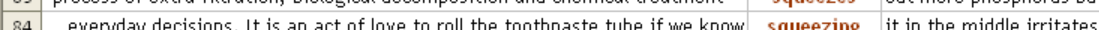
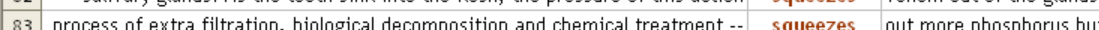
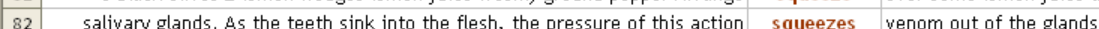
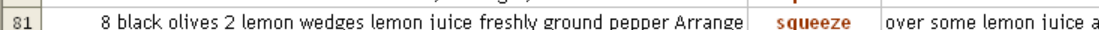
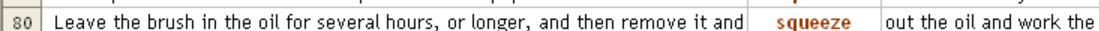
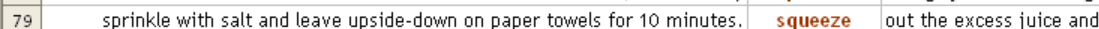
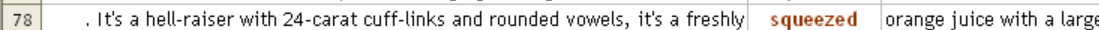
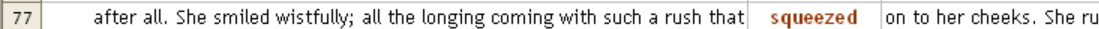
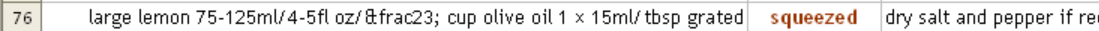
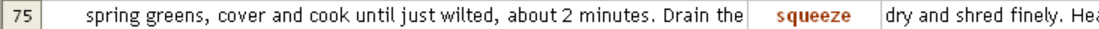
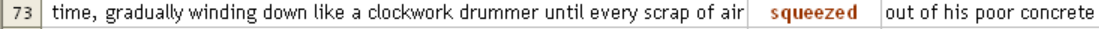
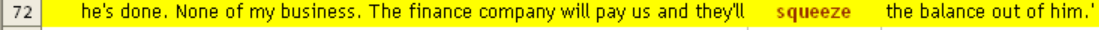
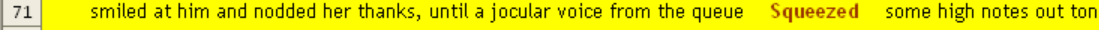
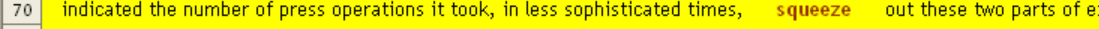
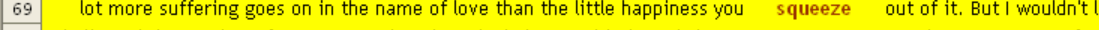
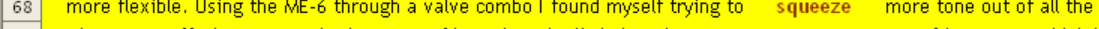
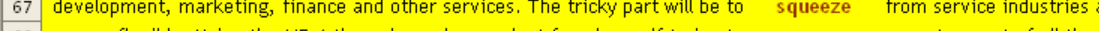
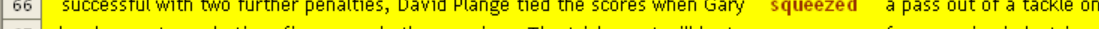
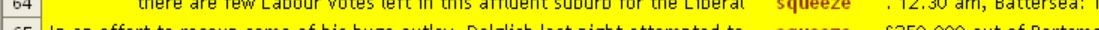
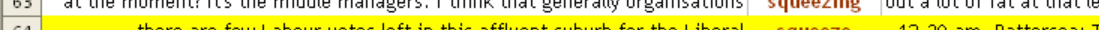
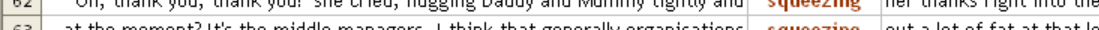
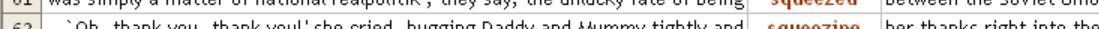
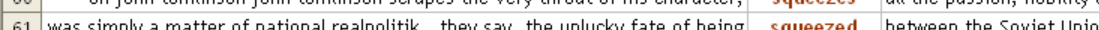
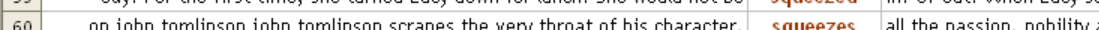
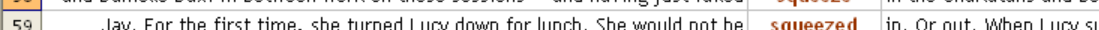
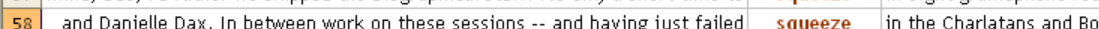
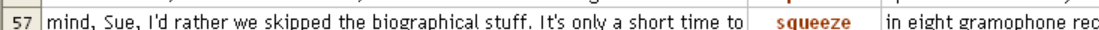
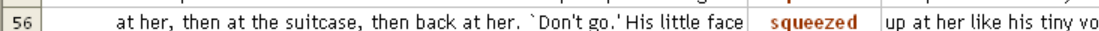
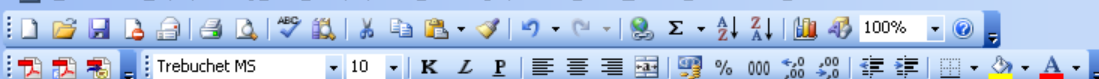
- 621 million tokens
- tagged (85% accuracy)
- text types
 - Literary, scientific, popular science, etc.
- medium
 - Newspapers, magazines, books, internet etc.
- 1990 – 2006 (FIDA 1997-2000)
- available online: <http://www.fidaplus.net/>



Analysis

- analyze a random sample of concordances
- assign (provisional) sense to each concordance
- go to the word sketch
- make the sense/subsense structure
- **PROVE IT!** through example-collocation-pattern





Trebuchet MS 10 K L P

E58 manage to find time

B

C

D

E

55 much as 40 per cent of the cost of a house. The developers' profit margin is

56 at her, then at the suitcase, then back at her. 'Don't go.' His little face

57 mind, Sue, I'd rather we skipped the biographical stuff. It's only a short time to

58 and Danielle Dax. In between work on these sessions -- and having just failed

59 Jay. For the first time, she turned Lucy down for lunch. She would not be

60 on John Tomlinson John Tomlinson scrapes the very throat of his character,

61 was simply a matter of national realpolitik, they say, the unlucky fate of being

62 'Oh, thank you, thank you!' she cried, hugging Daddy and Mummy tightly and

63 at the moment? It's the middle managers. I think that generally organisations

64 there are few Labour votes left in this affluent suburb for the Liberal

65 In an effort to recoup some of his huge outlay, Dalglish last night attempted to

66 successful with two further penalties, David Plange tied the scores when Gary

67 development, marketing, finance and other services. The tricky part will be to

68 more flexible. Using the ME-6 through a valve combo I found myself trying to

69 lot more suffering goes on in the name of love than the little happiness you

70 indicated the number of press operations it took, in less sophisticated times,

71 smiled at him and nodded her thanks, until a jocular voice from the queue

72 he's done. None of my business. The finance company will pay us and they'll

73 time, gradually winding down like a clockwork drummer until every scrap of air

74 vital nutrients. Chop and mix together two or three pieces of fresh fruit.

75 spring greens, cover and cook until just wilted, about 2 minutes. Drain the

76 large lemon 75-125ml/4-5fl oz/8frac23; cup olive oil 1 x 15ml/1tbsp grated

77 after all. She smiled wistfully; all the longing coming with such a rush that

78 . It's a hell-raiser with 24-carat cuff-links and rounded vowels, it's a freshly

79 sprinkle with salt and leave upside-down on paper towels for 10 minutes.

80 Leave the brush in the oil for several hours, or longer, and then remove it and

81 8 black olives 2 lemon wedges lemon juice freshly ground pepper Arrange

82 salivary glands. As the teeth sink into the flesh, the pressure of this action

83 process of extra filtration, biological decomposition and chemical treatment --

84 everyday decisions. It is an act of love to roll the toothpaste tube if we know

85 the stripes A reservoir of coloured gel is put near the nozzle. When the tube is

squeezed

squeezed

squeeze

squeeze

squeezed

squeezes

squeezed

squeezing

squeezing

squeeze

squeeze

squeezed

squeeze

squeeze

squeeze

squeeze

Squeezed

squeeze

squeezed

Squeeze

squeeze

squeezed

squeezed

squeezed

squeezed

squeeze

squeeze

squeeze

squeezes

squeezes

squeezing

squeezed

to a point where they cannot (should they, uncharacteristically, be so minded)

up at her like his tiny voice. 'I got to,' she said through her teeth. Opening

in eight gramophone records, isn't it? Besides, I'd like to hear what your reaction

in the Charlatans and Bob Geldof -- he's been releasing his own dance records under

in. Or out. When Lucy suggested a quick drink, Jay was cut to the quick and

all the passion, nobility and madness out of him (boris) SUNDAY TIMES physically

between the Soviet Union and the Axis alliance. As for the Romanian Holocaust,

her thanks right into them. 'I think perhaps the bike will make up for the

out a lot of fat at that level simply because they're finding more efficient ways

. 12.30 am, Battersea: The Tories' gain in Battersea in 1987 was one of the most

£250,000 out of Portsmouth for midfielder Steve Agnew. But Rovers blocked

a pass out of a tackle on the line 12 minutes from the end. Rallying: Biasion

from service industries alone the entire 4% of annual productivity increases that

more tone out of all the distortion settings, something I don't ever remember

out of it. But I wouldn't like to dwell on it. Perhaps you could lighten up a little

out these two parts of extraction. Subsequent writers, however, have copied

some high notes out tonight, Lemon. Juicy. Juicy. Only Lemon left in the country

the balance out of him.' 'For heaven's sake!' Such a scandalous-sounding

out of his poor concrete lungs. Eyes bulging from a purple, blue-lipped face, he

a little fresh orange juice over the fruit -- you can almost taste the energy!

dry and shred finely. Heat the oil over medium heat. Mince the remaining garlic

dry salt and pepper if required 1. Soak the bread in water for about 10min -- this

on to her cheeks. She rubbed them away with an angry fist. She was a fool to let

orange juice with a large shot of vodka -- and like a Mickey Finn, the beauty of

out the excess juice and scoop away the pulp. Place the pulp in the bowl with the

out the oil and work the brush backwards and forwards over plain brown paper or

over some lemon juice and add freshly ground pepper. Garnish with lemon wedges

venom out of the glands and down the hollow tubes of the paired fangs. It spreads

out more phosphorus but is not proven to reduce large amounts of nitrogen. And

it in the middle irritates our partner. Little considerations such as replacing

the gel is forced through tiny holes, into the paste. Laughter Lines 'Did

limit money

make grimace

manage to find time

manage to find time

manage to find time

metaph.

metaph.

metaph.

metaph.

obtain with difficulty

obtain with difficulty

obtain with difficulty

obtain with difficulty

obtain with difficulty

obtain with difficulty

obtain with difficulty

obtain with difficulty

press out gas

press out liquid

press out liquid

press out liquid

press out liquid

press out liquid

press out liquid

press out liquid

press out liquid

press out particles

press out soft substance

press out soft substance

Concordance Table

TBL-GDEX: 'rana' / 'wound'

rana Fida PLUS 620m freq = 20319

<u>a</u>	<u>modifier</u>	<u>8068</u>	<u>2.1</u>	<u>prec</u>	<u>zaradi</u>	<u>589</u>	<u>31.0</u>	<u>is</u>	<u>obj2</u>	<u>1304</u>	<u>7.2</u>
<input checked="" type="checkbox"/>	vboden	<u>353</u>	87.13	<input type="checkbox"/>	biti	<u>320</u>	34.67	<input type="checkbox"/>	zacetiti	<u>222</u>	73.67
<input checked="" type="checkbox"/>	strelen	<u>583</u>	84.18	<input type="checkbox"/>	umreti	<u>78</u>	33.99	<input type="checkbox"/>	celiti	<u>156</u>	66.48
<input checked="" type="checkbox"/>	zacetjen	<u>162</u>	77.56	>>				<input type="checkbox"/>	lizati	<u>49</u>	46.58
<input type="checkbox"/>	nezacetjen	<u>147</u>	77.51					<input type="checkbox"/>	odpirati	<u>90</u>	37.97
<input checked="" type="checkbox"/>	hud	<u>963</u>	62.27	<u>prec</u> <u>kljub</u>		<u>36</u>	16.3	<input type="checkbox"/>	oskrbeti	<u>41</u>	36.2
<input checked="" type="checkbox"/>	krvaveč	<u>99</u>	60.55	<input type="checkbox"/>	biti	<u>25</u>	20.73	<input type="checkbox"/>	zadati	<u>30</u>	31.39
<input checked="" type="checkbox"/>	rakav	<u>168</u>	57.22	>>				<input type="checkbox"/>	odpreti	<u>56</u>	20.61
<input type="checkbox"/>	gnojen	<u>94</u>	53.44					<input type="checkbox"/>	zdraviti	<u>23</u>	18.99
<input type="checkbox"/>	globok	<u>315</u>	52.69					<input type="checkbox"/>	povzročiti	<u>26</u>	16.66
<input type="checkbox"/>	razpočen	<u>34</u>	52.11					<input type="checkbox"/>	dobiti	<u>39</u>	11.56
<input type="checkbox"/>	ugrizen	<u>27</u>	51.71					<input type="checkbox"/>	imeti	<u>56</u>	6.64
<input type="checkbox"/>	rakast	<u>81</u>	49.3	>>				>>			

Tickbox
Lexicography

Tickbox Lexicography - Select Examples

Lemma: rana

Gramrel: a_modifier

Template: vanilla

vboden

- ☐ V kliničnem centru so sprejeli 160 ljudi , med njimi tudi človeka z vbodnimi **ranami** , j
- ☐ Osebe je oskrbelo še človeka z vbodnimi **ranami** ter nesrečnika z opeklinami .
- ☐ - Janko Makoter je v nočnih urah 26. septembra 1999 v lastni spalnici umrl zaradi vbo
- ☒ Oskrbeli človeka z vbodnimi **ranami**
- ☐ Oskrbeli so vbodno **rano**

strelen

- ☐ Oskrbeli so tudi osebo s strelno **rano** .
- ☐ » Strelna **rana** , « si je rekel , » ne laserska .
- ☐ Uporabil je pištolo Makarov , kalibra 7,65 milimetra , strelna **rana** skozi glavo in mož
- ☒ Urgenco je obiskal poškodovanec s strelnimi **ranami** in drugi z vbodnimi ranami , dva
- ☐ Po njihovih podatkih ob truplu , ki je imelo dve strelni **rani** na zatilju , ni bilo nobenih c

zaceljen

- ☐ Moje telo je žarelo iz teh točk kot znova oživiljeni človek , ki uživa v zaceljenih **ranah**
- ☒ Nekatere bo začelo trgati tudi po starih , zaceljenih **ranah** .



User : simon.krek

Project SLOLEKS_A
Headword pasti
Senses 5
Words 0 Characters 0



FILE EDIT CONFIG TOOLS HELP

pasti *glagol*

1 premikati se od zgoraj proti navzdol v 3. osebi

PREDMET zaradi sile težnosti pade KAM

a) Struktura: sbz1 GBZ rbz/na sbz4/v sbz4

- ▶ kaj pade kam
- ▶ kaj pade na kaj
- ▶ kaj pade v kaj

- [bomba, granata, drevo] pade
- pasti na [dno, Zemljo, zemljo, tla]
- pasti v [prepad, globino, morje]

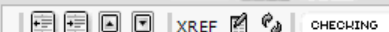
- Kaj pa, če helikopter **pade** dol?
- Se še spominjate dne, ko je **padla** bomba čisto blizu vaše gostilne.
- Svetleča granata je **padla** kakih petdeset metrov stran.
- Pri odcepu za Kostanjevico je **padlo** na dolenjo drevo.
- Mnoge živali in rastline poginejo in **padejo** na vodno dno.
- Vsako leto **pade** na Zemljo iz vesolja obilica kamenin.
- Debele kostanjeve ježice so **padle** na zemljo.
- V stanju breztežnosti nič ne **pade** na tla.
- Dve gondoli sta **padli** v globok prepad.
- Propeler je razsekalo in helikopter je kot kamen **padel** v globino.
- Po izjavah posadke naj bi bomba **padla** v morje daleč proč od letalonosilke.

skladenjske zveze:

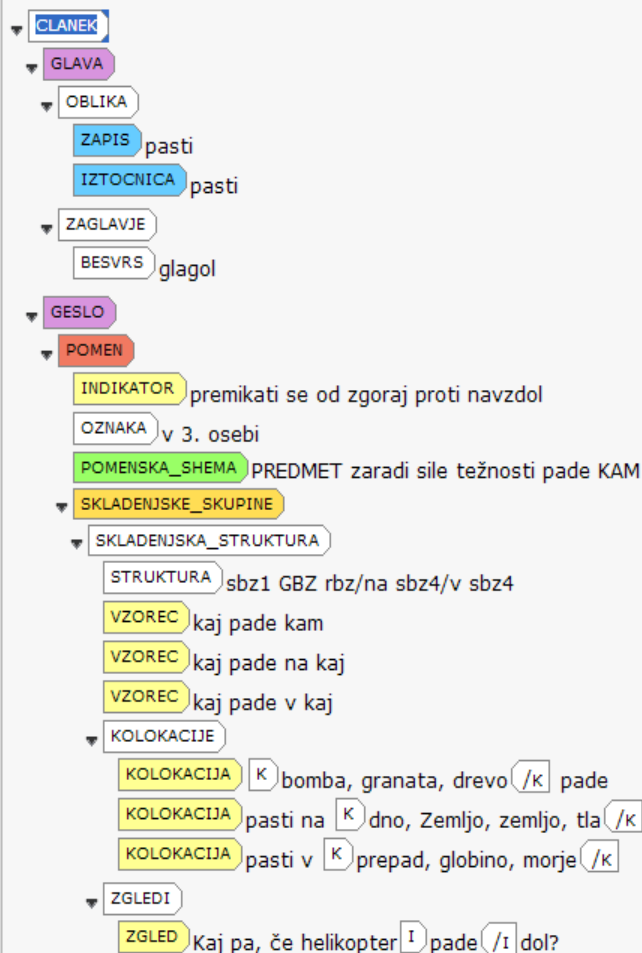
pasti pod kotom

pasti pod kotom [x] stopinj

- Kovinski meteorit s premerom 40 metrov, ki pade pod kotom več kot 30 stopinj, bi tla še dosegel, preden bi se povsem stalil oziroma izhlapel.
- Pred 50 milijoni let je na Luno padel pod blagim kotom težak asteroid.



<clanek>



squeeze

- Sketch Engine – ukWaC
- (N)ODE
- MEDAL
- LDOCE
- Cobuild



Discussion

- system of sense distribution
 - sense/subsense (two levels?)
- closed “list” of syntactic patterns
 - sketch grammar?
- variation in syntactic combinations
- “extended” extended collocations
- multiword units & phraseology



Future

- relation to SloWNet – Slovene WordNet
- comparison with FrameNet frames
- LMF compatibility checkup
- semantic role analysis & consolidation
- complete automation of pattern-collocation-example extraction
- crude automatic WSD



SLDB Sketch Grammar

- 32 relations
- compatibility with SLDB “syntactic structures”
- naming the relations is not easy
- mostly based on part-of-speech info



<http://www.slovenscina.eu/>

Thank you!

simon.krek@ijs.si

