Wordnet - a multilingual semantic lexicon

University of Ljubljana
Faculty of Arts
Department of Translation

Jožef Stefan
International
Postgraduate School

16th November 2011

Darja Fišer

Outline

1. What is wordnet & Why it's good for
2. 3 approaches to wordnet development
3. Automatic extension of wordnet
4. Automatic cleaning of noisy synsets
5. Browsing, editing and visualization of wordnet
6. Conclusions & future plans

1. Background & Motivation (What is wordnet & Why it's good for)

What are semantic lexicons

- computer databases of human knowledge about our words & worlds
- explicit structural, semantic & relational information
- vocabulary is organized according to the meaning (*tree > birch, car ~ automobile*)
- more structured than dictionaries but less formal than ontologies

Wordnet

Noun

- S: (n) **bug#1** (general term for any insect or similar creeping or crawling invertebrate)
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) insect#1 (small air-breathing arthropod)
  - *derivationally related form*
- S: (n) **bug#2**, glitch#1 (a fault or defect in a computer program, system, or machine)
- S: (n) **bug#3** (a small hidden microphone; for listening secretly)
- S: (n) hemipterous insect#1, **bug#4**, hemipteran#1, hemipteron#1 (insects with sucking mouthparts and forewings thickened and leathery at the base; usually show incomplete metamorphosis)
- S: (n) microbe#1, **bug#5**, germ#3 (a minute life form (especially a disease-causing bacterium); the term is not in technical use)

Verb

- S: (v) tease#1, badger#1, pester#1, **bug#1**, beleaguer#1 (annoy persistently) *"The children teased the boy because of his stammer"*
- S: (v) wiretap#1, tap#5, intercept#2, **bug#2** (tap a telephone or telegraph wire to get information) *"The FBI was tapping the phone line of the suspected spy"; "Is this hotel room bugged?"*

Wordnet

Noun

synset

gloss

- S: (n) **bug#1** (general term for any insect or similar creeping or crawling invertebrate)
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) insect#1 (small air-breathing arthropod)
  - *derivationally related form*

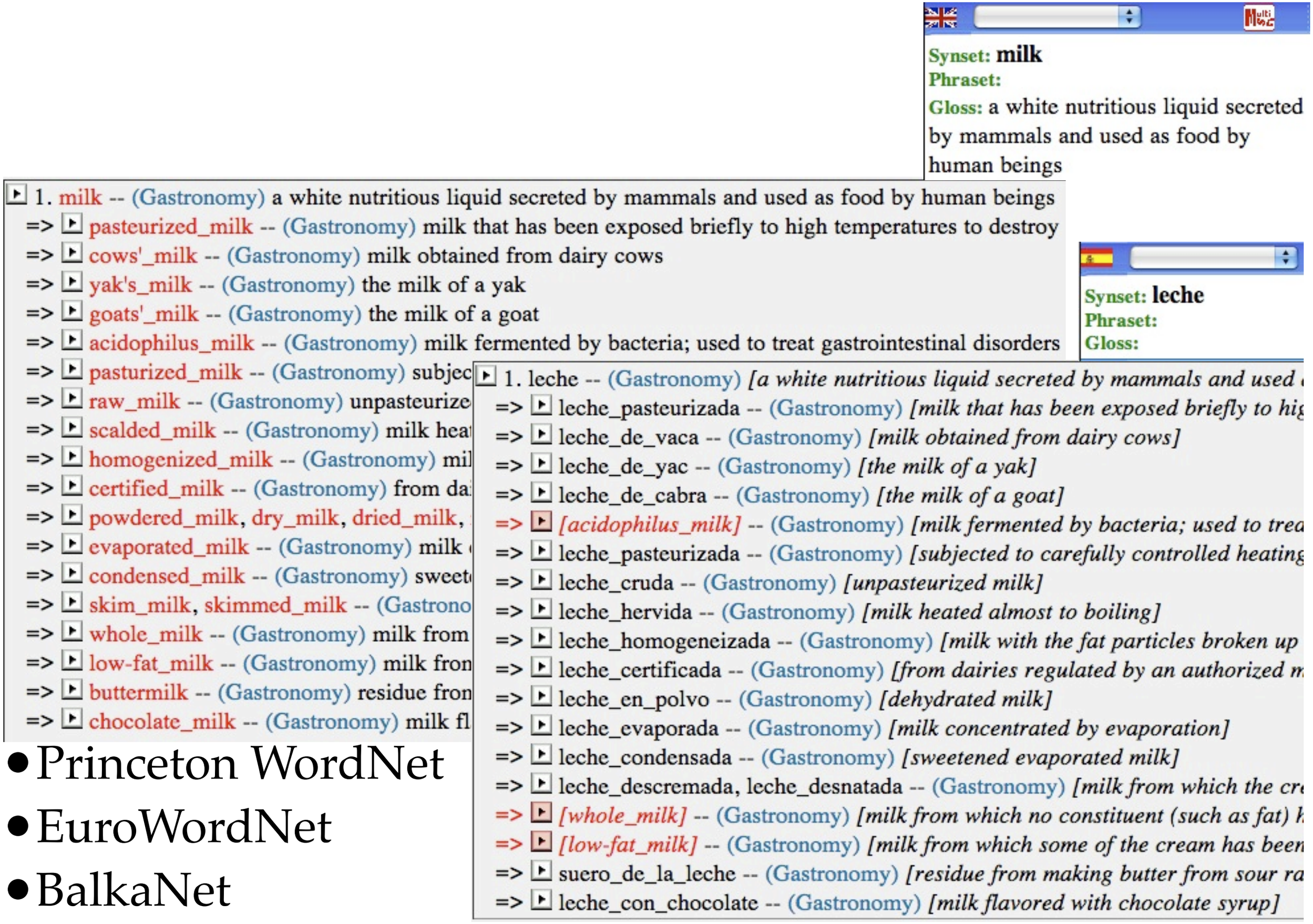
semantic relations

- S: (n) **bug#2**, glitch#1 (a fault or defect in a computer program, system, or machine)
- S: (n) **bug#3** (a small hidd one; for listening secretly)
- S: (n) hemipterous insect#1, **bug#4**, hemipteran#1, hemipteron#1 (insects with sucking mouthparts and forewings thickened and leathery at the base; usually show incomplete metamorphosis)
- S: (n) microbe#1, **bug#5**, germ#3 (a minute life form (especially a disease-causing bacterium); the term is not in technical use)

literals

Verb

- S: (v) tease#1, badger#1, pester#1, **bug#1**, beleaguer#1 (annoy persistently) *"The children teased the boy because of his stammer"*
- S: (v) wiretap#1, tap#5, intercept#2, **bug#2** (tap a telephone or telegraph wire to get information) *"The FBI was tapping the phone line of the suspected spy"; "Is this hotel room bugged?"*



- Princeton WordNet
- EuroWordNet
- BalkaNet
- Global Wordnet Association

Why do we need them?

- bridge between language & knowledge
  - semantic normalization (*bug, pester*)
  - disambiguation (*bug_insect, bug_defect*)

- HLT applications:
  - search engines
  - machine translation
  - document classification
  - information extraction
  - text summarisaton

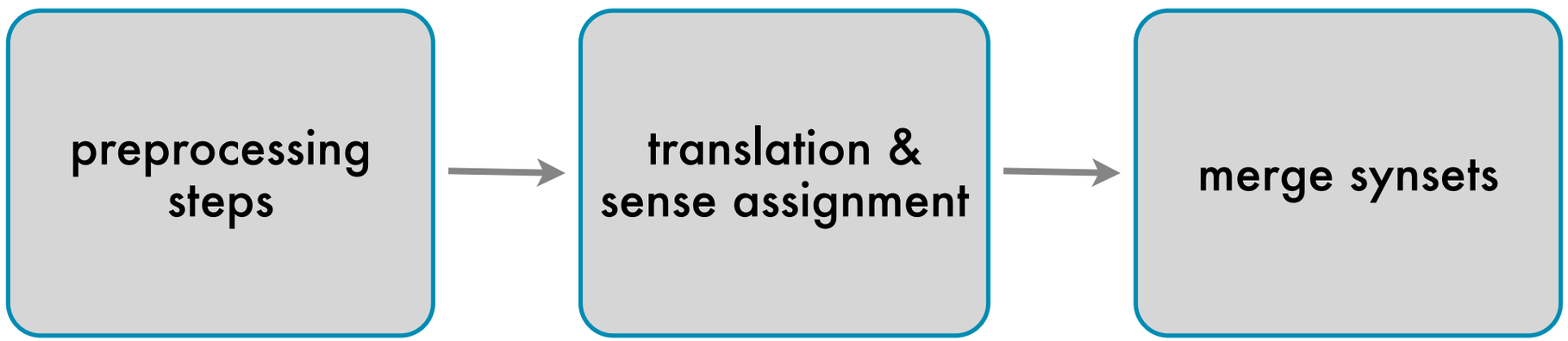
Why automatic construction?

- needs:
  - 1 lexical entry ~30 min
  - lexicon size ~100,000 entries
  - ~50,000 hours / ~2,000 days / ~6 years
- aims:
  - speed up
  - simplify
  - lower costs
  - recycle

2. Wordnet development (3 approaches)

Research goals

- develop methodology & test 3 different multilingual approaches
- **expand** vs. merge approach
  - translational relation
  - parallel wordnets



1. Dictionary approach

- Rigau idr. (1998)
  - bilingual dictionary
  - obvious choice
  - rich vocabulary
  - ready-made translations
  - different sense inventory
  - mapping to wn synsets not trivial

1. Dictionary approach

I. **bug** [bʌg] SAMOST *fam*	
***1. bug** (insect):*	
bugs *pl*	**mrčes** *m*
bed **bug**	**stenica**
***2. bug** MED:*	
bug	**bacil** *m*
***3. bug** COMPUT:*	
bug (fault)	**hrošč** *m*
***4. bug** (listening device):*	
bug	**prisluškovalni aparat** *m*
***5. bug** (enthusiasm):*	
bug	**mrzlica** *ž*
to catch the travel **bug**	**imeti potovalno mrzlico**

Noun

- S: (n) **bug#1** (general term for any insect or similar creeping c invertebrate)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) insect#1 (small air-breathing arthropod)
  - *derivationally related form*
- S: (n) **bug#2**, glitch#1 (a fault or defect in a computer prograr machine)
- S: (n) **bug#3** (a small hidden microphone; for listening secretl
- S: (n) hemipterous insect#1, **bug#4**, hemipteran#1, hemipterc with sucking mouthparts and forewings thickened and leather usually show incomplete metamorphosis)
- S: (n) microbe#1, **bug#5**, germ#3 (a minute life form (especia causing bacterium); the term is not in technical use)

Verb

- S: (v) tease#1, badger#1, pester#1, **bug#1**, beleaguer#1 (anno *"The children teased the boy because of his stammer"*
- S: (v) wiretap#1, tap#5, intercept#2, **bug#2** (tap a telephone c wire to get information) *"The FBI was tapping the phone line c suspected spy"; "Is this hotel room bugged?"*

1. Dictionary approach

wordnet A → dictionary A-B → wordnet B

Serbian WN

konac, kraj, svršetak, završetak

→

Srp-Slo dict.

konac: izid, iztek, konec, končanje, kraj, ~~krajnik~~, ~~obrobje~~, ~~nit~~, sklep, ~~sukanec~~, zaključek, ~~zatrep~~

→

Slovene WN

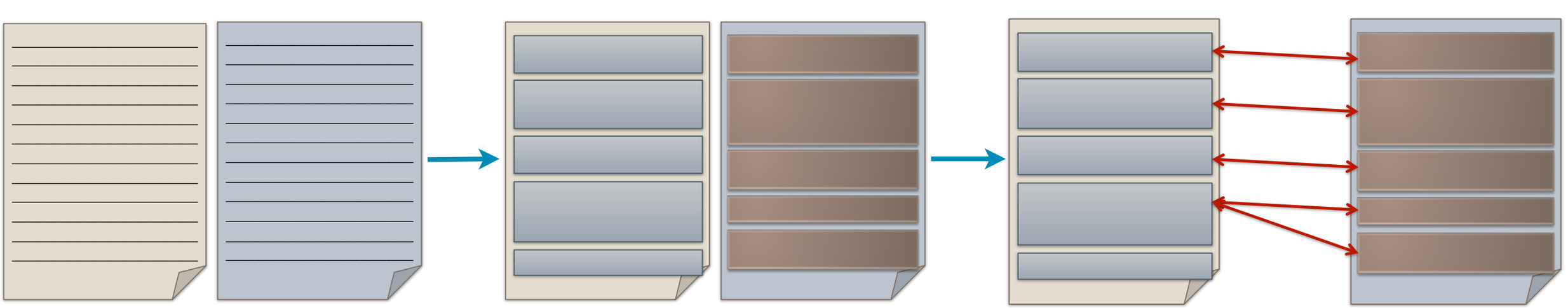
izid, iztek, konec, končanje, kraj, sklep,

2. Corpus approach

- Diab (2004)
  - multilingual parallel corpora
  - existing wordnets for several languages
  - automatic sense-assignment to polysemous words
  - more pre-processing
  - limited to single-word literals

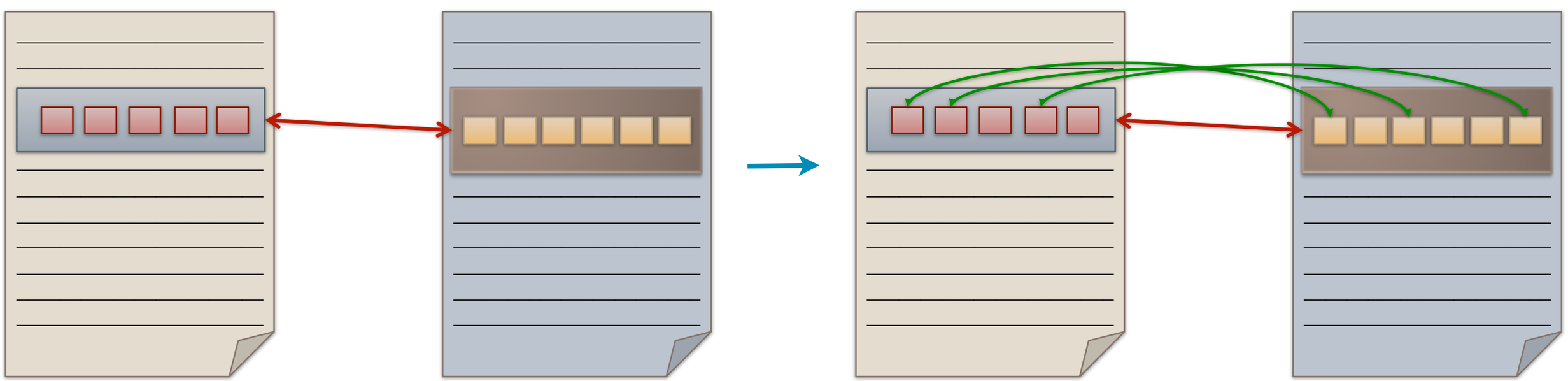
2. Corpus approach

- sentence alignment



2. Corpus approach

- POS-tagging
- lemmatization
- word alignment
- lexicon extraction



2. Corpus approach

EN		CS		RO		BG		SI	
word		word		word		word		word	
party		strana		partid		партия		stranka	
party		večírek		petrecere		забава		zabava	
army		armáda		armată		армия		armada	
army		armáda		armată		армия		vojska	

2. Corpus approach

EN		CS		RO		BG		SI	
word	wn	word	wn	word	wn	word	wn	word	wn
party	01 11 22	strana	01	partid	01 27 57	партия	01 23	stranka	?
party	02 17 50	večírek	02 09	petrecere	02	забава	02 15 20	zabava	?
army	03 16	armáda	03 99 55	armată	03 10	армия	03	armada	?
army	03 33 66	armáda	03 29	armată	03 29	армия	03	vojska	?

2. Corpus approach

EN		CS		RO		BG		SI	
word	wn	word	wn	word	wn	word	wn	word	wn
party	01 11 22	strana	01	partid	01 27 57	партия	01 23	stranka	?
party	02 17 50	večírek	02 09	petrecere	02	забава	02 15 20	zabava	?
army	03 16	armáda	03 99 55	armată	03 10	армия	03	armada	?
army	03 33 66	armáda	03 29	armată	03 29	армия	03	vojska	?

2. Encyclopedic approach

- Navigli & Ponzetto (2010)
  - extensive & multilingual publicly available resource
  - multi-word terms
  - domain-specific terms
  - small size of Slovene Wikipedia (64,000 articles << 2.5 million articles in English)
  - fake monosemy

3. Encyclopedic approach

- Lietuvių
- Magyar
- Nederlands
- 日本語
- Polski
- Português
- Русский
- Simple English
- Slovenščina
- ไทย
- Suomi
- Svenska

Crop rotation

From Wikipedia, the free encyclopedia

Please expand this article with text translate

A↔あ After translating, {{Translated|nl|Vruchtwisseling}} must b

Translation instructions · Translate via Google

"Fallow" redirects here. For other uses, see Fallow (disambiguation).

Crop rotation or **Crop sequencing** is the practice of growing a series of

Kolobarjenje

Iz Wikipedije, proste enciklopedije

Kolobarjenje (tudi **kolobar**) je metoda, pri kateri se vrtnine različnih talnih škodljivcev, ki napadajo točno določene vrst bolezni.

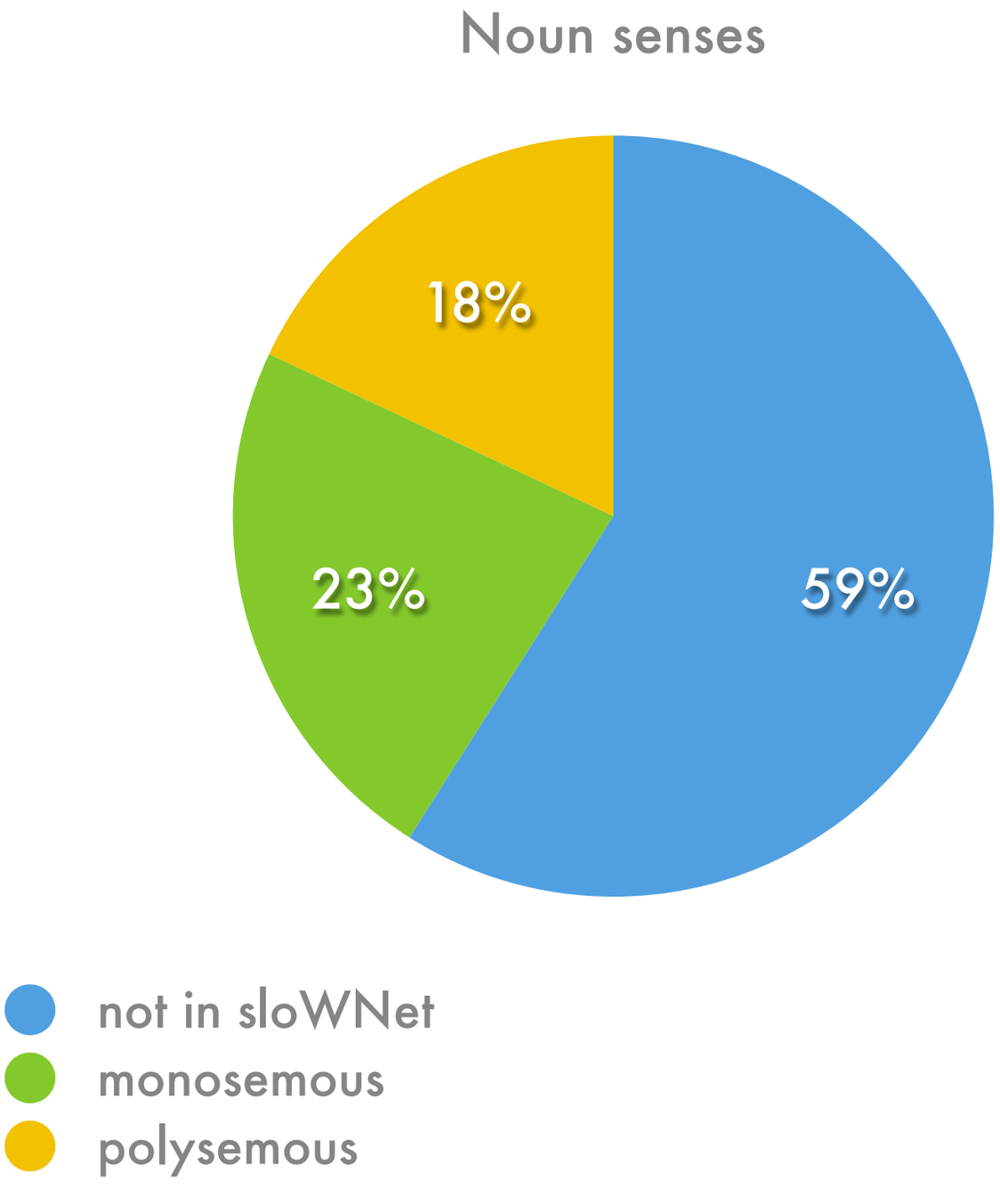
Results

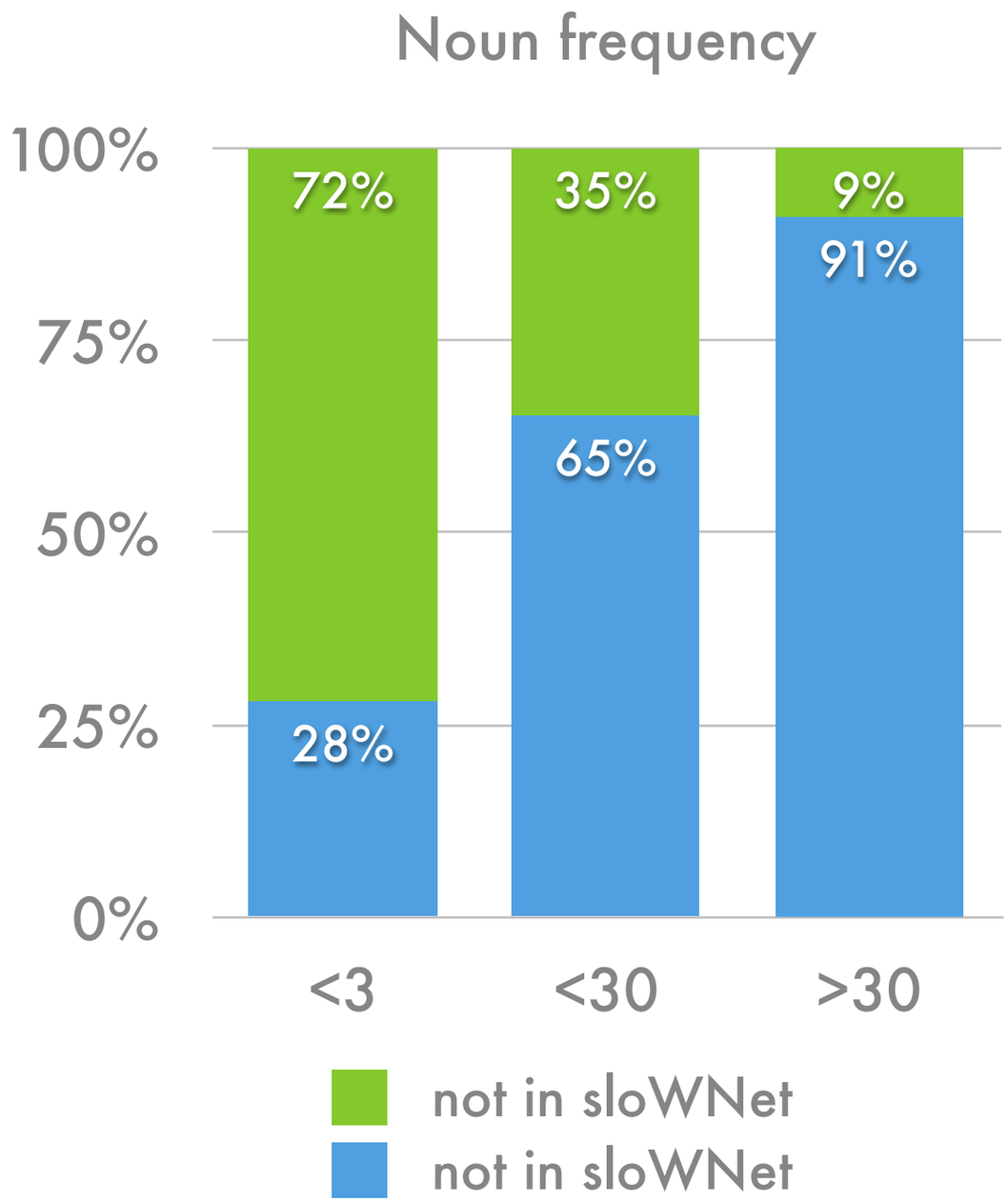
- no. of synsets: 16.886
- no. of literals: 19.582
- % of PWN: 15 %
- % of BCS1 & BCS2: 100 %
- % of nouns: 91%
- % of MWE: 43 %
- 1 literal/synset: 66 %
- synset length: 1,16
- longest synset: 16 literals (*goljufati*)

Analysis

- domains:
  - factotum 25 % (dictionary & corpus)
  - zoology 17 % (wiki)
  - botany 13 % (wiki)
  - biology 7 % (wiki)
  - (agriculture ~330 synsets)
- semantic relations:
  - hypernymy 46 %, 91 % for nouns
  - complete chains 46 %
  - longest chain 16 nodes (*telica*)

Vocabulary coverage





3. Wordnet extension (with Benoît Sagot, INRIA, France)

Motivation & aproach

- current wordnet reliable but small
- available resources not used to their full potential
- idea:
  - extract all possible translation equivalents
  - use current wordnet as “copper standard”
  - train a Maximum Entropy classifier
  - add (synset,literal) pairs that are above threshold

Classifier

- ~ 300,000 translation equivalents extracted
- features used for the classifier:
  - semantic proximity (wn vs. corpus - SemanticVectors)
  - number of sources yielding the same (synset,literal) pair
  - level of polysemy
  - number of tokens in literal
- threshold was set empirically to 0.1

Evaluation of the results

- ~ 68,000 (literal, synset) pairs were added to wn
- 63,010 (63%) new (literal, synset) pairs were added
- 25,102 synsets that were empty before now have at least 1 literal
- sloWNet 3.0: 82,721 (literal,synset) pairs & 42,919 synets
- manual evaluation of 400 (literal, synset) pairs above the threshold: 64% accuracy
- automatic evaluation against goldstandard: 85% accuracy

3. Cleaning noisy synsets (work in progress)

Motivation & aproach

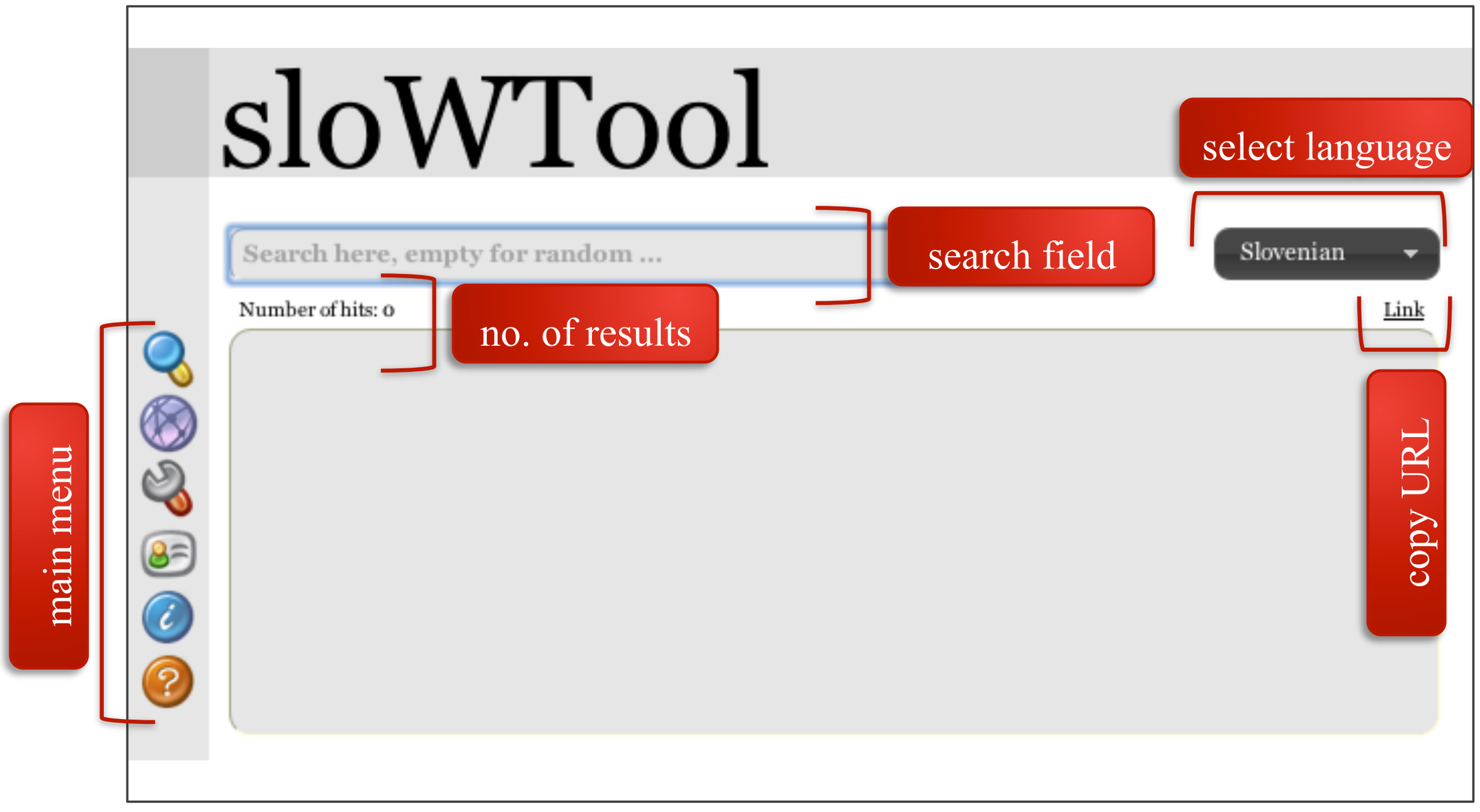
- current wordnet large but noisy
- biggest errors are due to poor wsd (*organ_body, organ_instrument*)
- idea: rank a (noisy) list of synonym candidates with distributional methods for detecting semantic similarity between words
- hypothesis: lexemes tend to co-occur in corpora with other semantically related lexemes, as made explicit by relations between synsets in a wordnet

4. sloWTool
(browsing, editing & visualization)

Motivation

- available tools:
  - Princeton WordNet: Princeton WordNet Browser
  - EuroWordNet: Polaris & Periscope
  - BalkaNet: DEBVisDic
- ideal tool:
  - freely available, platform-independent, on-line
  - all-in-one tool (browsing, editing & visualization)
  - support for standardized formats (e.g. LMF)
  - support for multilingual scenarios
  - easy integration of third-party resources (e.g. domains, coarse-grained sense clusters, images)

Browsing

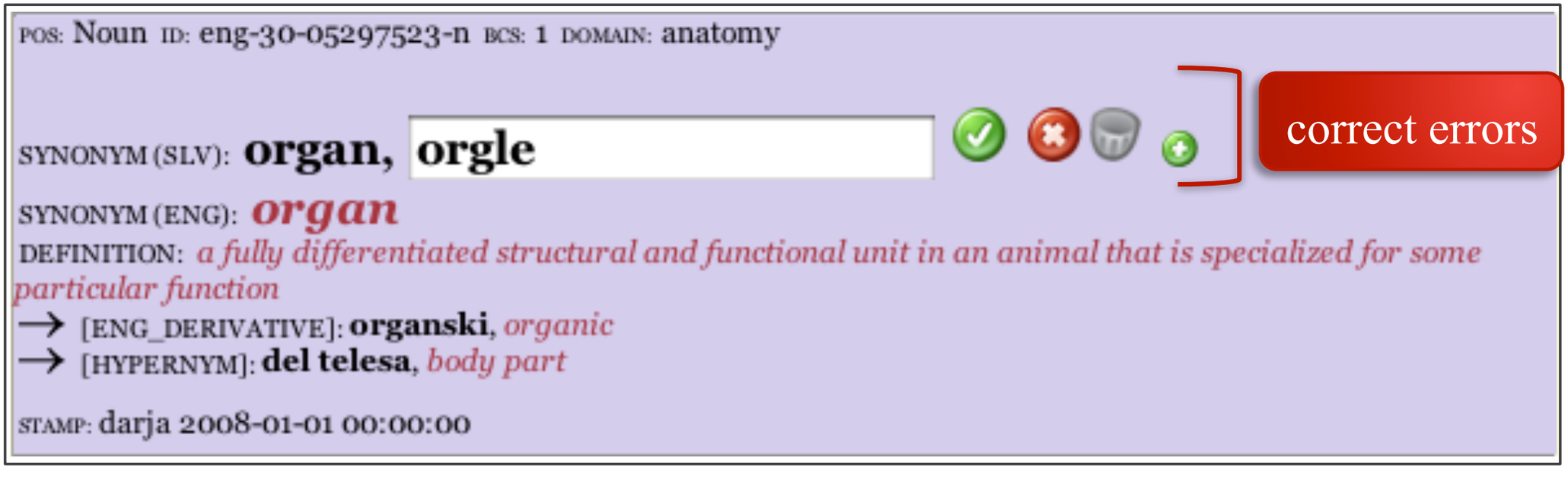

sloWTool
select language
Search here, empty for random ...
search field
Slovenian
Number of hits: 0
no. of results
Link
main menu
copy URL

Browsing


prst
Slovenian
Number of hits: 9
Link
POS: Noun ID: eng-30-05566097-n BCS: 2 DOMAIN: anatomy CLUSTERID: finger3
meta info
SYNONYM (SLV): daktil, prst
SYNONYM (ENG): dactyl, digit
synonyms
DEFINITION: a finger or toe in human beings or corresponding body part in other vertebrates
definition
→ [ENG_DERIVATIVE]: digitalen, digital
→ [HOLO_PART]: vretenčar, craniate, vertebrate
→ [HYPERNYM]: ud, del, član, ud, okončina, končina, ekstremiteta, member, appendage, extremity
semantic relations
POS: Noun ID: eng-30-05559908-n BCS: 2 DOMAIN: anatomy CLUSTERID: life10
SYNONYM (SLV): član, del, ekstremiteta, končina, okončina, ud, ud
SYNONYM (ENG): appendage, extremity, member
DEFINITION: an external body part that projects from the
USAGE: it is important to keep the extremities warm
example
→ [HYPERNYM]: zunanji del telesa, external body part
STAMP: darja 2008-01-01 00:00:00
STAMP: darja 2008-01-01 00:00:00
edit info

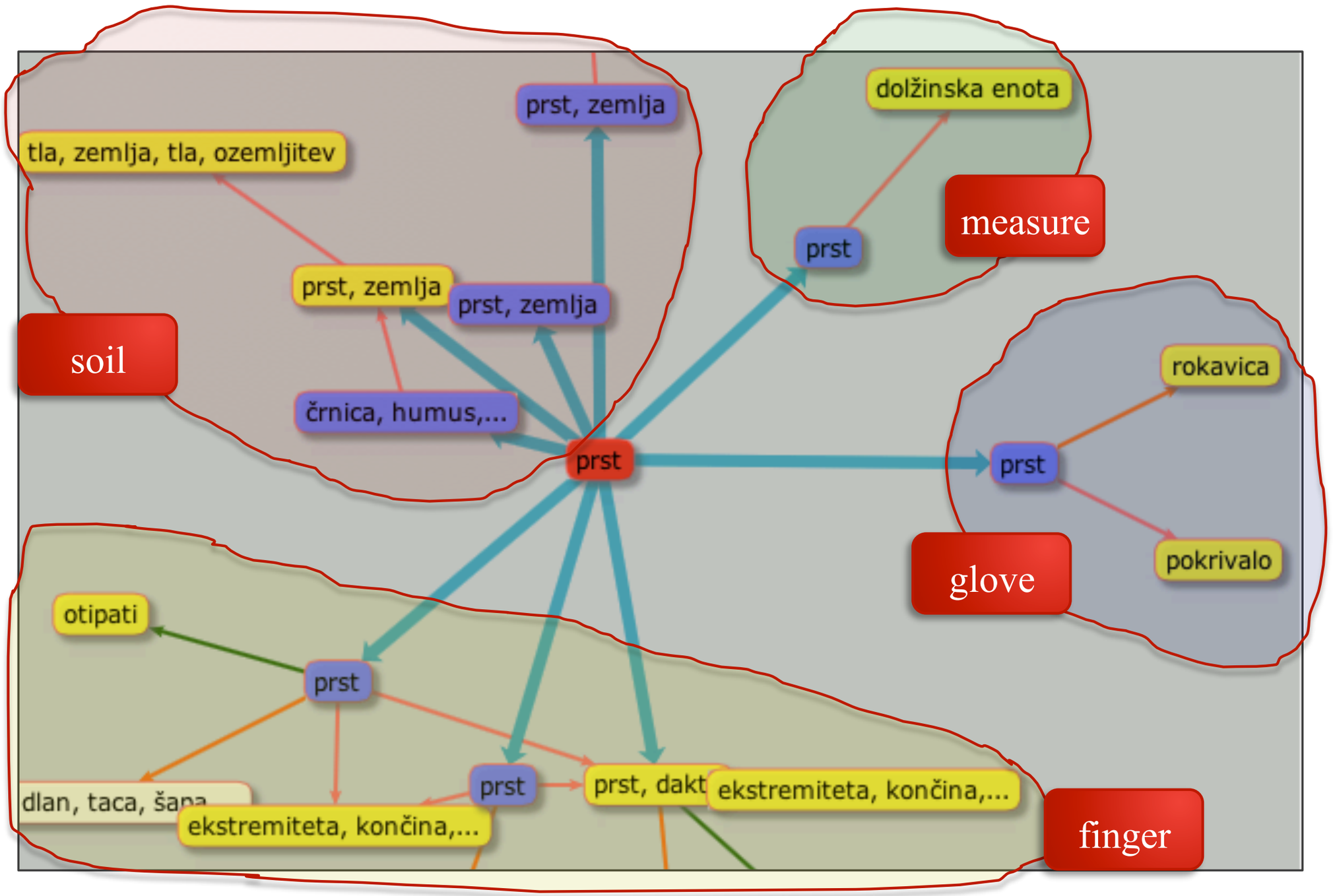
Editing

register & login



correct errors

Visualization



5. Conclusions
(and more future work)

nl.ijs.si/slownet

- XML format
- CC licence
- viewing & editing in DEBVisDic

sloWNet
Slovene Wordnet

version 2.0
last change Aug 1 2008

What is sloWNet?

sloWNet is a lexico-semantic resource for Slovene, in which words that have the same meaning (literals) are organized into sets of synonyms (synsets). Synsets are linked into a semantic network with various lexical and semantic relations.

The wordnet family:

The first wordnet was developed for English in the 1980's at Princeton University and it became one of the most popular resources for tasks in the field of automatic understanding of natural language. Wordnets for other languages soon followed in projects, such as EuroWordNet, BalkaNet and MultiWordNet. Wordnets for 50 different languages are currently registered with the Global WordNet Association.

How was sloWNet built?

sloWNet was built automatically. The creation process consisted of three stages:

1. **Core wordnet**
   A bilingual dictionary was used to translate basic concepts into Slovene. The translations were then checked and corrected by hand.
2. **Polysemous words**
   Polysemous words were dealt with an approach in which a parallel corpus for five languages was word-aligned and the extracted multilingual lexicon was disambiguated with the existing wordnets for these languages.
3. **Monosemous words**
   Equivalents for monosemous words were found in open-source resources, such Wikipedia and Eurovoc thesaurus.

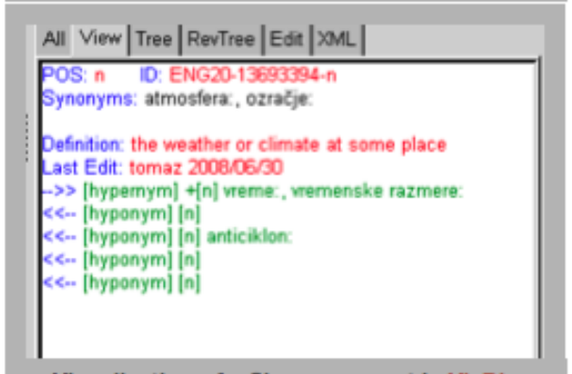
What is in sloWNet?

Number of entries
sloWNet currently contains about 20,000 unique literals which are organized into almost 17,000 synsets.
Sources of entries

Basic Info:

RESOURCE	sloWNet
TYPE	semantic lexicon for Slovene
VERSION	2.0
SIZE	17,000 synsets, 20,000 literals
LICENCE	Creative Commons - attribution - non-commercial - share-alike - If you wish to receive a copy of sloWNet, send me an e-mail.
CONTACT	darja.fiser@guest.arnes.si

Visualization of a paper on sloWNet with Wordle



Visualization of a Slovene synset in VisDic

Conclusions

- advantages of the model:
  - faster & easier construction
  - modularity
  - language-independentend (WOLF, Fišer & Sagot 2008)
- disadvantages of the model
  - fine-grained senses
  - inherited inconsistencies
  - English-centered

Future plans

- further development of sloWNet:
  - cleaning & refinement
  - add domain-specific terminology
  - verify the sense inventory in a monolingual reference corpus
- use of sloWNet:
  - semantic annotation of a corpus (on-going)
  - automatic word-sense disambiguation
  - use sloWNet in NLP tasks & applications

Thank you!