Wordnet - a multilingual semantic lexicon

University of Ljubljana
Faculty of Arts
Department of Translation

JCEA Workshop

15th October 2009

Darja Fišer

Outline

1. What is wordnet
2. Why it's good for
3. Wordnet development
4. Analysis of the results
5. Conclusions & future plans

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*)

Wordnet

POS: n ID: ENG20-14311212-n BCS:
Synonyms: vikend:x, konec tedna:x
Definition: a time period usually extending from Friday night through Sunday; more loosely defined as any period of successive days including one and only one Sunday
Domain: time_period
SUMO/MILO: Weekend
-->> [hypernym] obdobje:1, časovno obdobje:1
-->> [holo_part] teden:2
-->> [eng_derivative] _EMPTY_:
<<-- [eng_derivative] _EMPTY_:
<<-- [mero_part] nedelja:1
<<-- [mero_part] sobota:1

STAMP: simonic 2009-05-05 21:44:49 /

- Princeton WordNet
- EuroWordNet
- BalkaNet
- Global Wordnet Association

Wordnet

synset into

POS: n ID: ENG20-14311212-n BCS:
Synonyms: vikend:x, konec tedna:x
Definition: a time period usually extending from Friday night through Sunday; more loosely defined as any period of successive days including one and only one Sunday
Domain: time_period
SUMO/MILO: Weekend
-->> [hypernym] obdobje:1, časovno obdobje:1
-->> [holo_part] teden:2
-->> [eng_derivative] _EMPTY_ :
<<-- [eng_derivative] _EMPTY_ :
<<-- [mero_part] nedelja:1
<<-- [mero_part] sobota:1
STAMP: simonic 2009-05-05 21:44:49 /

literals

gloss

domain & ontology

semantic relations

editor

Synset: milk
Phraset:
Gloss: a white nutritious liquid secreted by mammals and used as food by human beings

1. milk -- (Gastronomy) a white nutritious liquid secreted by mammals and used as food by human beings
   => pasteurized_milk -- (Gastronomy) milk that has been exposed briefly to high temperatures to destroy
   => cows'_milk -- (Gastronomy) milk obtained from dairy cows
   => yak's_milk -- (Gastronomy) the milk of a yak
   => goats'_milk -- (Gastronomy) the milk of a goat
   => acidophilus_milk -- (Gastronomy) milk fermented by bacteria; used to treat gastrointestinal disorders
   => pasturized_milk -- (Gastronomy) subjected to carefully controlled heating to destroy undesirable micr
   => raw_milk -- (Gastronomy) unpasteurized milk
   => scalded_milk -- (Gastronomy) milk heated almost to boiling
   => homogenized_milk -- (Gastronomy) milk
   => certified_milk -- (Gastronomy) from dairi
   => powdered_milk, dry_milk, dried_milk, mi
   => evaporated_milk -- (Gastronomy) milk co
   => condensed_milk -- (Gastronomy) sweeten
   => skim_milk, skimmed_milk -- (Gastronom
   => whole_milk -- (Gastronomy) milk from w
   => low-fat_milk -- (Gastronomy) milk from
   => buttermilk -- (Gastronomy) residue from
   => chocolate_milk -- (Gastronomy) milk flav

Synset: leche
Phraset:
Gloss:

1. leche -- (Gastronomy) *[a white nutritious liquid secreted by mammals and used*
   => leche_pasteurizada -- (Gastronomy) *[milk that has been exposed briefly to hig*
   => leche_de_vaca -- (Gastronomy) *[milk obtained from dairy cows]*
   => leche_de_yac -- (Gastronomy) *[the milk of a yak]*
   => leche_de_cabra -- (Gastronomy) *[the milk of a goat]*
   => *[acidophilus_milk]* -- (Gastronomy) *[milk fermented by bacteria; used to trea*
   => leche_pasteurizada -- (Gastronomy) *[subjected to carefully controlled heating*
   => leche_cruda -- (Gastronomy) *[unpasteurized milk]*
   => leche_hervida -- (Gastronomy) *[milk heated almost to boiling]*
   => leche_homogeneizada -- (Gastronomy) *[milk with the fat particles broken up*
   => leche_certificada -- (Gastronomy) *[from dairies regulated by an authorized m*
   => leche_en_polvo -- (Gastronomy) *[dehydrated milk]*
   => leche_evaporada -- (Gastronomy) *[milk concentrated by evaporation]*
   => leche_condensada -- (Gastronomy) *[sweetened evaporated milk]*
   => leche_descremada, leche_desnatada -- (Gastronomy) *[milk from which the cr*
   => *[whole_milk]* -- (Gastronomy) *[milk from which no constituent (such as fat) h*
   => *[low-fat_milk]* -- (Gastronomy) *[milk from which some of the cream has been*
   => suero_de_la_leche -- (Gastronomy) *[residue from making butter from sour ra*
   => leche_con_chocolate -- (Gastronomy) *[milk flavored with chocolate syrup]*

Why do we need them?

- bridge between language & knowledge
  - semantic normalization (*prst, zemlja*)
  - disambiguation (*prst na roki, rodovitna prst*)

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

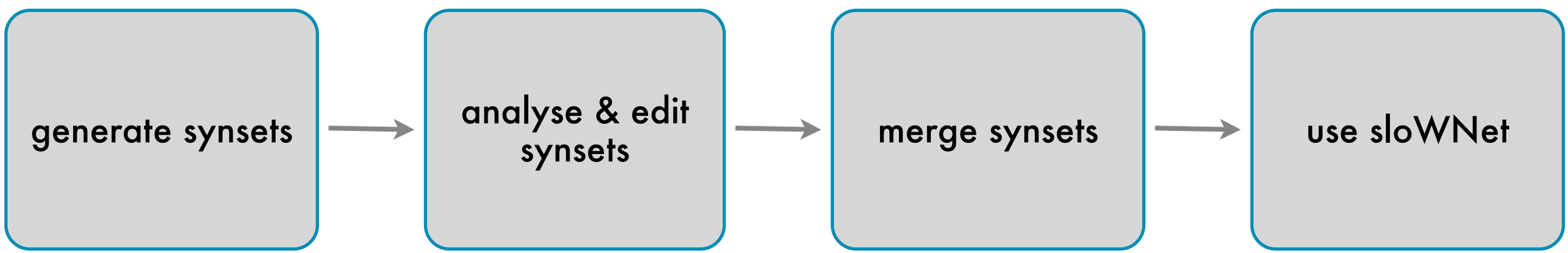
Why automatic construction?

- needs:
  - 1 lexical entry ~30 min
  - lexicon size ~50.000 entries
  - ~25.000 hours / 1000 days

- aims:
  - speed up
  - simplify
  - lower costs
  - recycle

Research goals

- develop methodology & test multilingual approaches
- expand approach
  - translational relation



Dictionary aaproach

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,

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	

Korpusni pristop - primer

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	?

Korpusni pristop - primer

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	?

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.

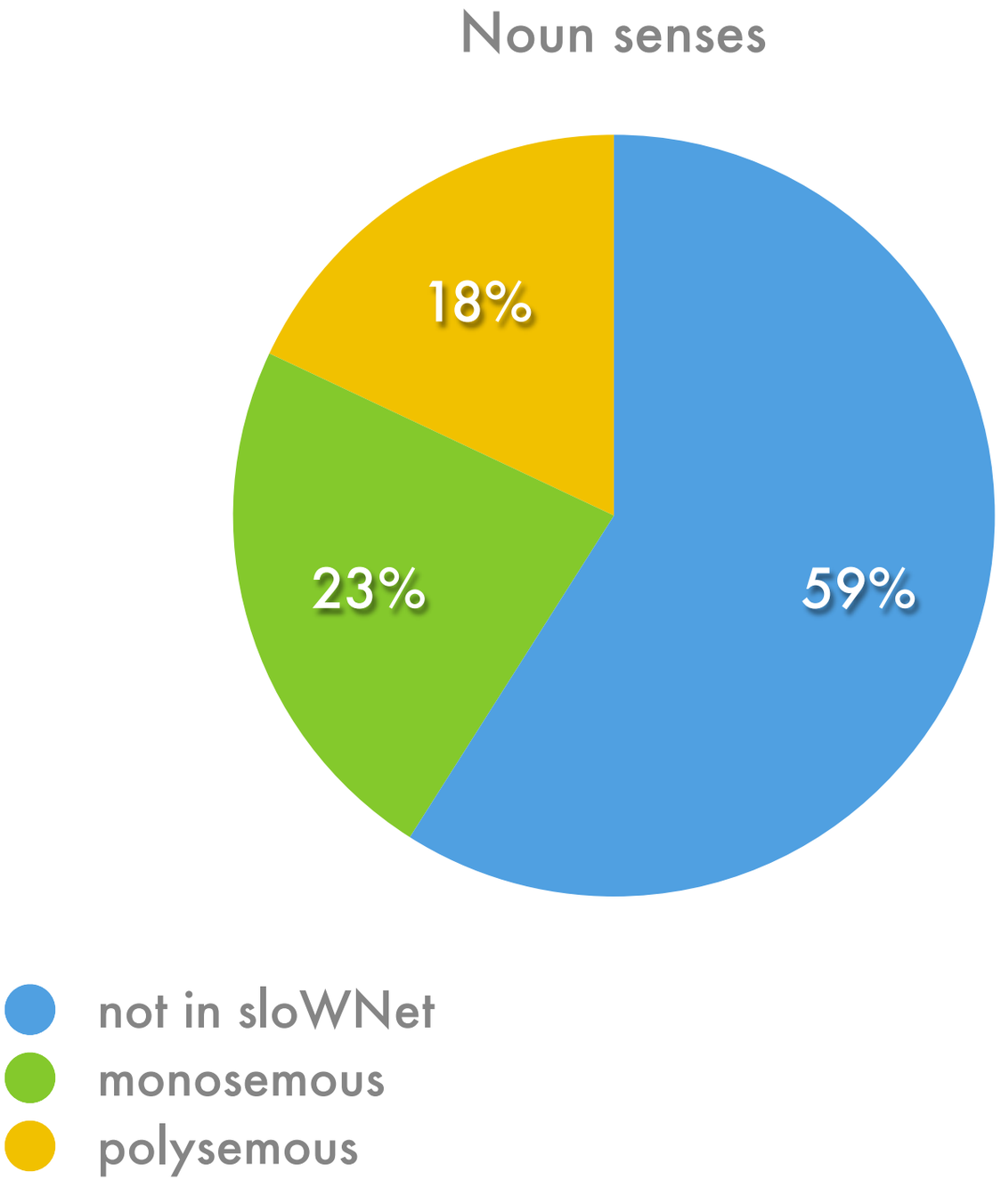
Results

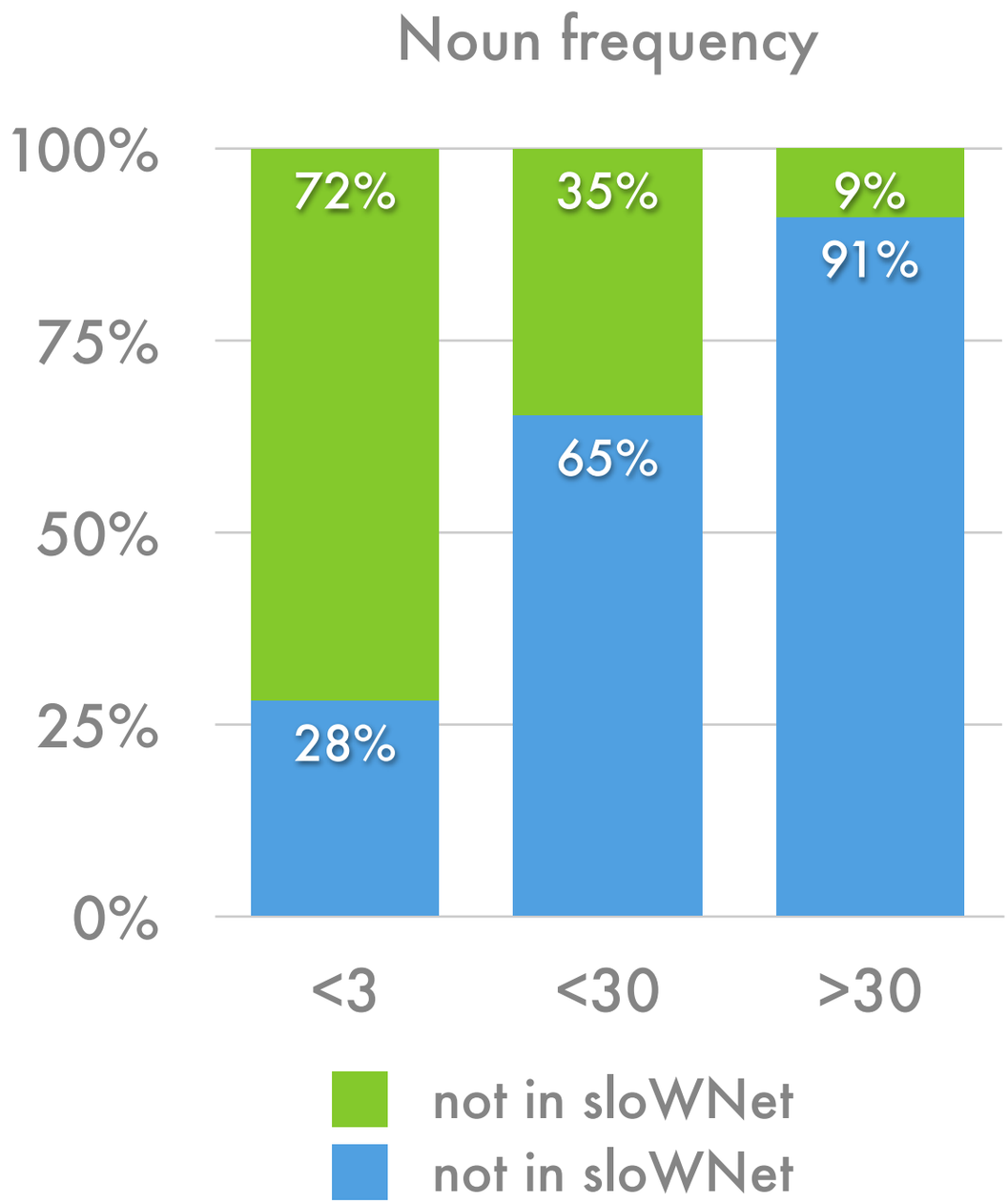
- no. of synsets: 16.886
- no. of literals: 19.582
- % of PWN: 15 %
- % of BCS1 & BCS2: 100 %
- % of nouns: 91%
- % of MWE: 43 %
- 1 literal/synset: 66 %
- synset length: 1,16
- longest synset: 16 literals (*goljufati*)

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





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.

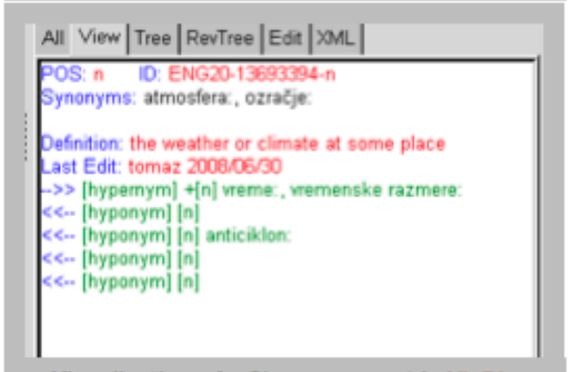
What is in sloWNet?

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:
  - domain-specific terminology (Vintar & Fišer 2008)
- use of sloWNet:
  - semantic annotation of a corpus
  - automatic word-sense disambiguation
  - machine translation

Thank you!