Semantic Lexicons and sloWNet

Darja Fišer
Department of Translation
Faculty of Arts

Presentation Outline

- Computational Lexical Semantics
- Semantic Lexicons
- Wordnet
- Automatic Construction of sloWNet
- Results
- Project Ideas

Computational Lexical Semantics

- in the information age, the volume and importance of electronic document is increasing rapidly, and it is impossible to handle them without automatized support
- CLS helps applications dealing with language understanding
  - machine translation
  - document classification
  - information extraction
  - document summarization

Semantic Lexicons

- bridge the gap between language and knowledge expressed with language
  - semantic normalization (a,c,b = A)
  - semantic disambiguation (xxxaxxx = A1, yyyayyy = A2)
- define the meaning of a word based on its relationship with meanings of other words
  - similarity of meaning corresponds to their distance in the network

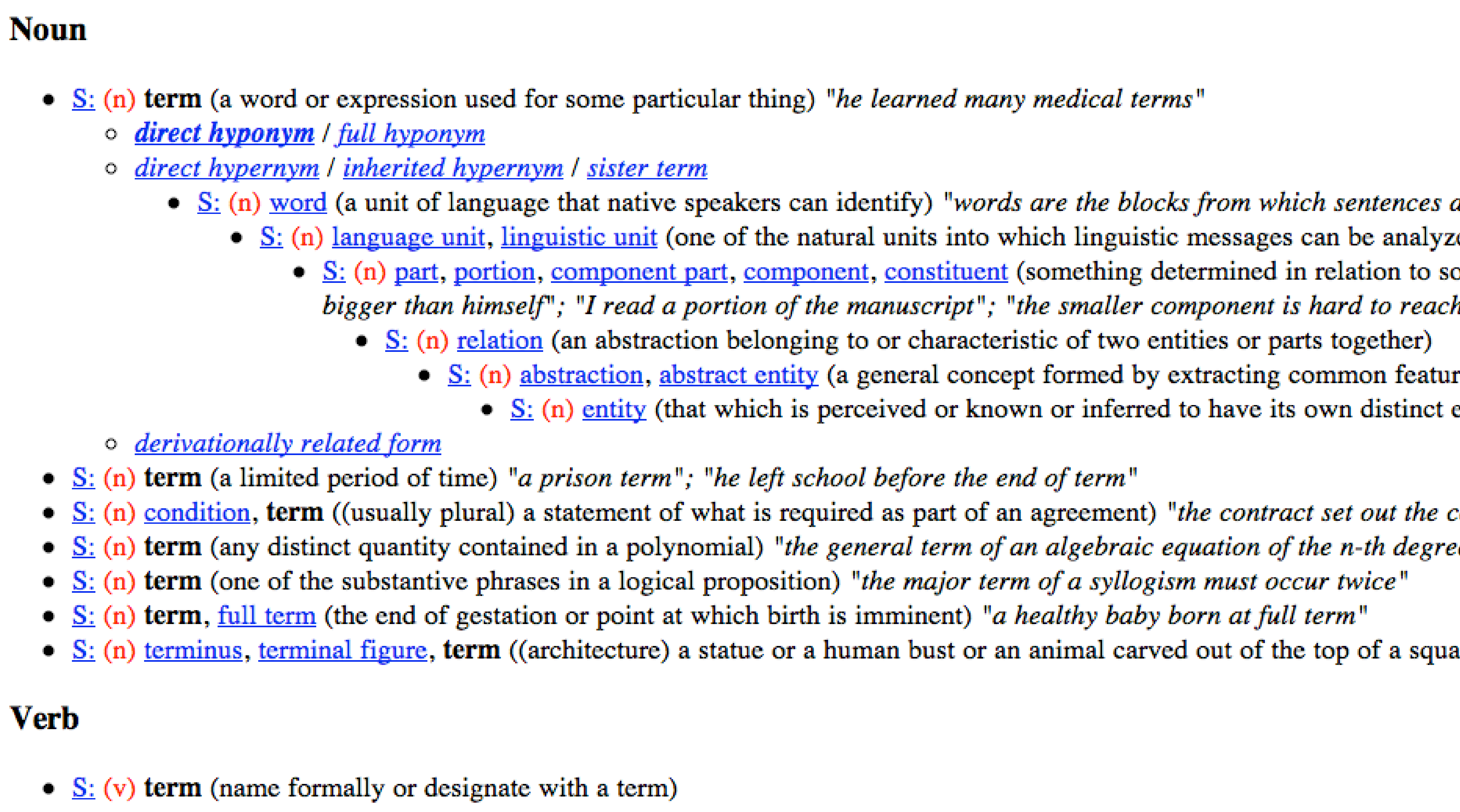
Types of semantic lexicons

- machine-readable dictionaries
  - intended for human use
  - LDOCE
- thesaurus
  - controlled vocabulary organized into a hierarchy
  - Roget
- lexical databases
  - wide coverage of vocabulary, a set of relations
  - FrameNet, WordNet, MindNet
- ontologies & knowledge bases
  - formal representation of world knowledge, language independent
  - Cyc, ConceptNet, HowNet

Why Automatic Construction?

- needs:
  - 1 lexical entry ~30 min
  - lexicon size ~50.000 entries
  - ~25.000 hrs or 1.000 days
- goals:
  - faster
  - easier
  - cheaper
  - recyclable

WordNet



Noun

- S: (n) **term** (a word or expression used for some particular thing) "*he learned many medical terms*"
    - ○ ***direct hyponym*** / *full hyponym*
    - ○ *direct hypernym* / *inherited hypernym* / *sister term*
        - S: (n) word (a unit of language that native speakers can identify) "*words are the blocks from which sentences a*
            - S: (n) language unit, linguistic unit (one of the natural units into which linguistic messages can be analyz
                - S: (n) part, portion, component part, component, constituent (something determined in relation to so *bigger than himself*"; "*I read a portion of the manuscript*"; "*the smaller component is hard to reach*
                    - S: (n) relation (an abstraction belonging to or characteristic of two entities or parts together)
                        - S: (n) abstraction, abstract entity (a general concept formed by extracting common featur
                            - S: (n) entity (that which is perceived or known or inferred to have its own distinct e
    - ○ *derivationally related form*
- S: (n) **term** (a limited period of time) "*a prison term*"; "*he left school before the end of term*"
- S: (n) condition, **term** ((usually plural) a statement of what is required as part of an agreement) "*the contract set out the c*
- S: (n) **term** (any distinct quantity contained in a polynomial) "*the general term of an algebraic equation of the n-th degre*
- S: (n) **term** (one of the substantive phrases in a logical proposition) "*the major term of a syllogism must occur twice*"
- S: (n) **term**, full term (the end of gestation or point at which birth is imminent) "*a healthy baby born at full term*"
- S: (n) terminus, terminal figure, **term** ((architecture) a statue or a human bust or an animal carved out of the top of a squa

Verb

- S: (v) **term** (name formally or designate with a term)

Dictionary Approach - Example

- Serbian wordnet:
  - **konac, kraj, svršetak, završetak**
- Serbian-Slovene dictionary:
  - **konac**: izid, iztek, konec, končanje, kraj, krajnik, obrobje, nit, sklep, sukanec, zaključek, zatrep
- sloWNet:
  - **izid, iztek, konec, končanje, kraj, sklep, zaključek**

Dictionary Approach: Overview

- Erjavec&Fišer 2006
- Approach: Serbian synsets were translated with a bilingual dictionary & the results were validated by hand
- Result: a set of manually validated basic concepts which exist in wordnets for many other languages
- Problems: no disambiguation for polysemous literals was performed -> many errors lead to a lot of manual editing

Corpus Approach - Example

EN		CS		RO		BG		SI	
beseda	id	beseda	id	beseda	id	beseda	id	beseda	id
party	01	strana	01	partid	01	партия	01	stranka	01
party	02	večírek	02	petrecere	02	забава	02	zabava	02
army	03	armáda	03	armată	03	армия	03	armada	03
army	03	armáda	03	armată	03	армия	03	vojska	03

syn1 (party1) : **stranka**

syn2 (party2): **zabava**

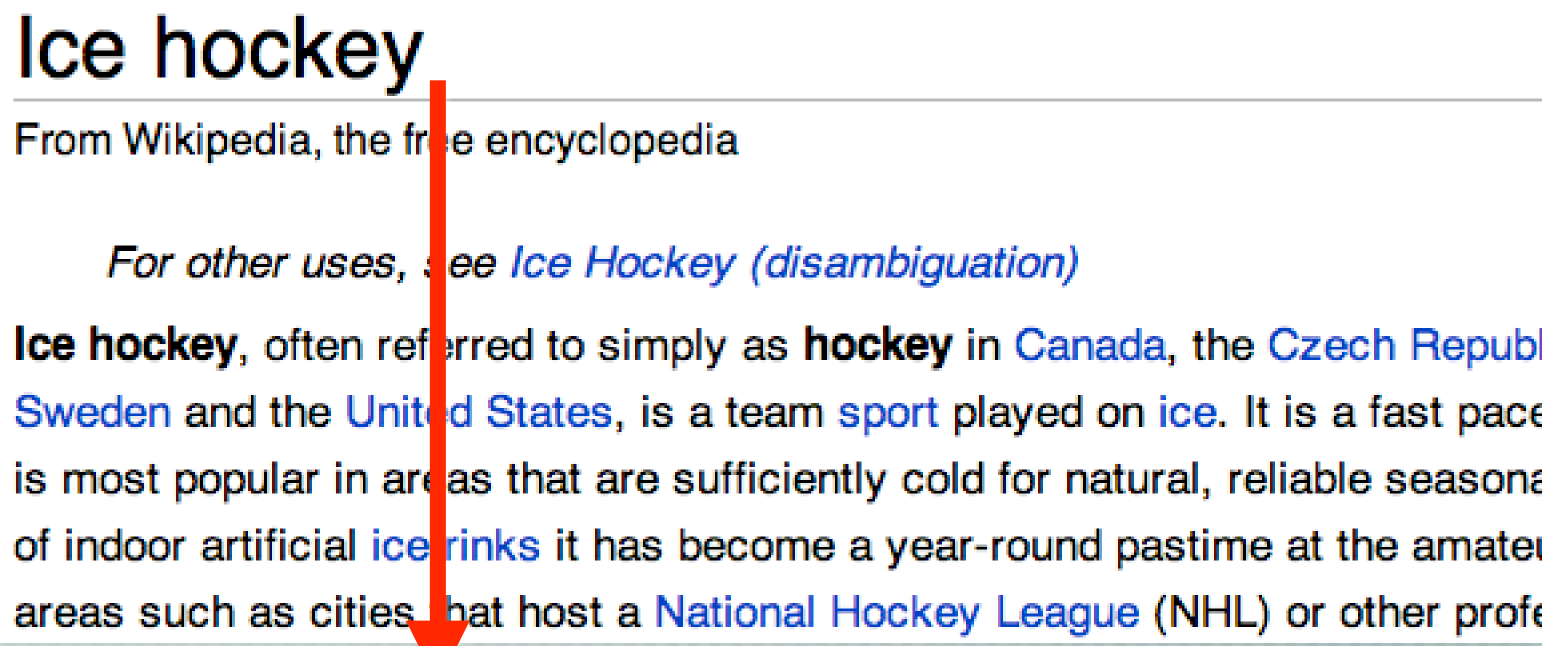
syn3 (army): **armada, vojska**

Corpus Approach - Overview

- Fišer 2007
- Approach: a multilingual parallel corpus was word-aligned, a multilingual lexicon was extracted & disambiguated with the existing wordnets for these languages
- Result: base concepts were extended with other polysemous words
- Problems: word-alignment limited to single-word literals, different sizes of the existing wordnets, holes in the network

Encyclopedic Approach - Example

- English wordnet: **ice hockey**
- Wnglish Wikipedia:



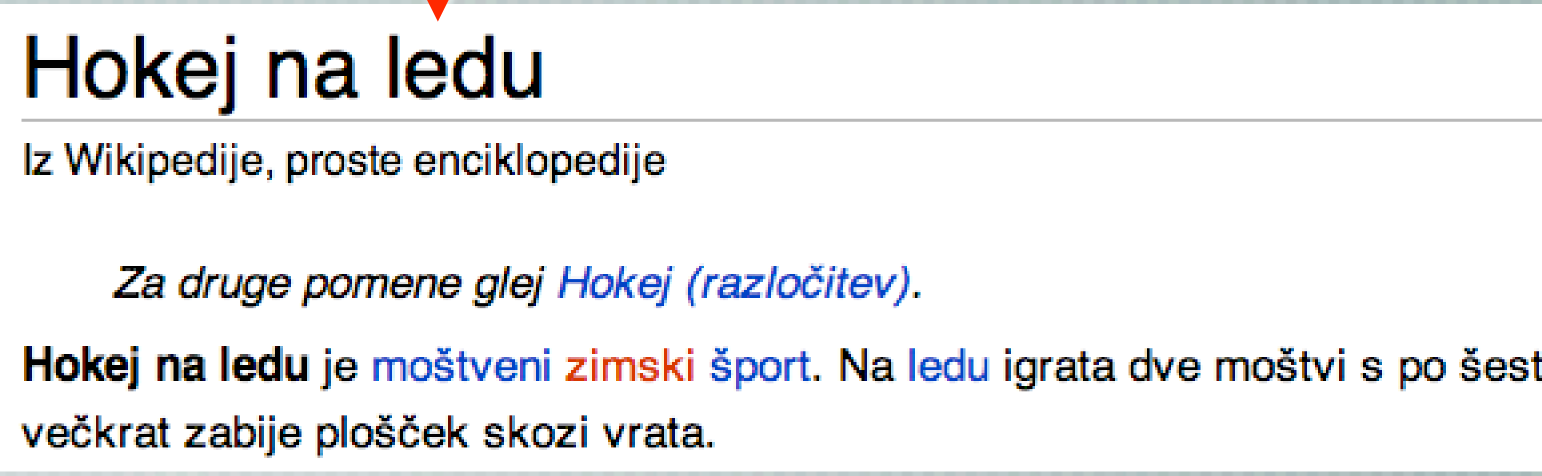
Ice hockey

From Wikipedia, the free encyclopedia

For other uses, see Ice Hockey (disambiguation)

Ice hockey, often referred to simply as **hockey** in Canada, the Czech Republ
Sweden and the United States, is a team sport played on ice. It is a fast pace
is most popular in areas that are sufficiently cold for natural, reliable seasona
of indoor artificial ice rinks it has become a year-round pastime at the amateu
areas such as cities that host a National Hockey League (NHL) or other profe

- Slovene Wikipedia:



Hokej na ledu

Iz Wikipedije, proste enciklopedije

Za druge pomene glej Hokej (razločitev).

Hokej na ledu je moštveni zimski šport. Na ledu igrata dve moštvi s po šest
večkrat zabije plošček skozi vrata.

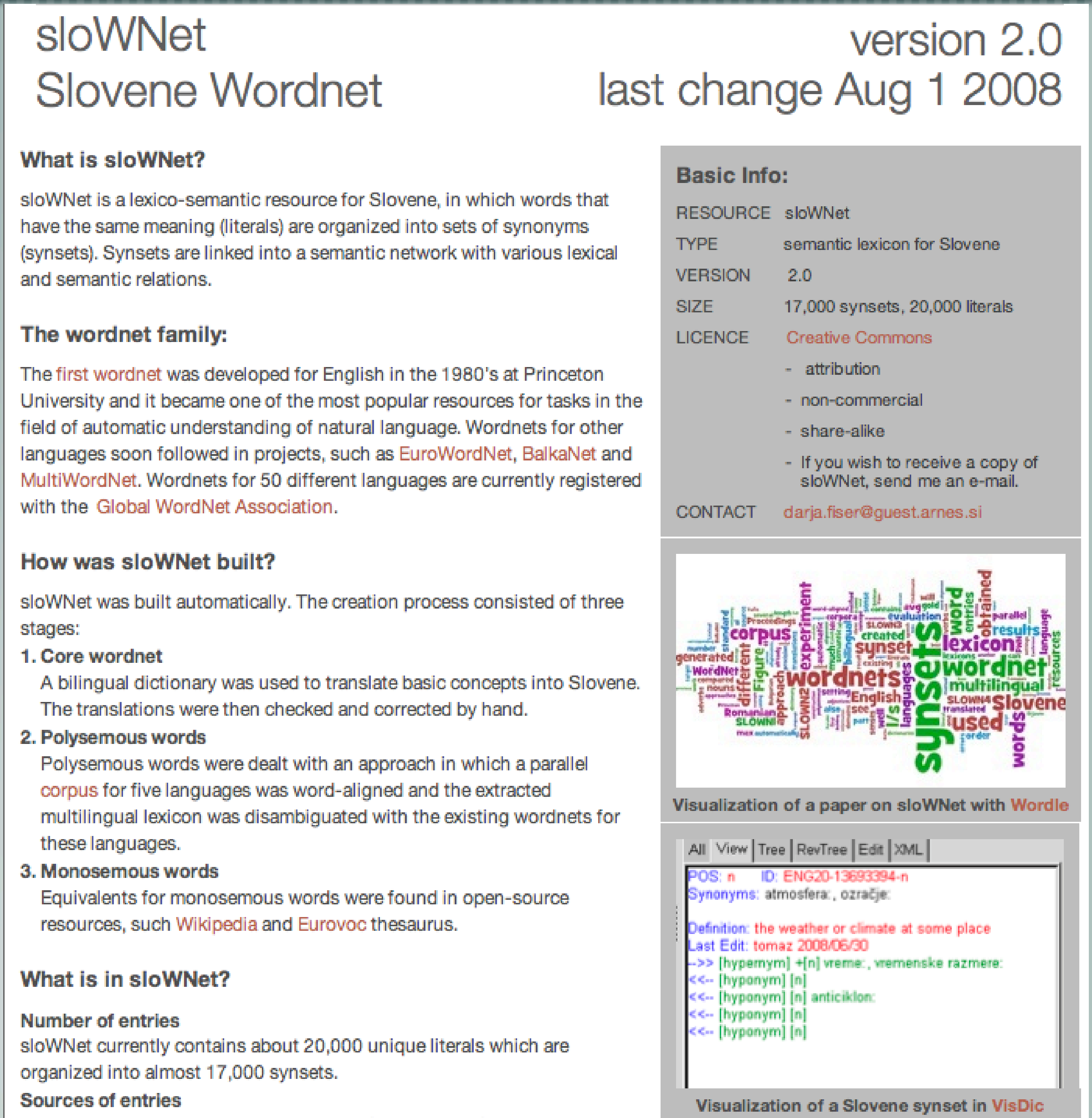
- Slovene wordnet: **hokej na ledu**

Encyclopedic approach - Overview

- Fišer&Sagot 2008
- Approach: monosemous literals were translated with Wikipedia& EuroVoc
- Result: a lot of scecific concepts were added, also many multi-word expressions
- Problem: small size of Slovene Wikipedia (64.000 articles in Slovene vs. 2.5 mio articles in English)

SloWNet 2.0

- 17.000 concepts
- 20.000 entries
- general & specific vocabulary
- mostly nouns
- xml format
- freely available for research



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

```
POS: n      ID: ENG20-13693394-n
Synonyms: atmosfera: , ozračje:

Definition: the weather or climate at some place
Last Edit: tomaz 2008/06/30
-->> [hypernym] +[n] vreme: , vremenske razmere:
<<-- [hyponym] [n]
<<-- [hyponym] [n] anticiklon:
<<-- [hyponym] [n]
<<-- [hyponym] [n]
```

Visualization of a Slovene synset in VisDic

Discussion

- Achievement: the automatic approach recycled existing resources to create a wordnet which is aligned with other wordnets (mono- & multilingual applications)
- Problems: quality of the generated synsets depends on the quality of the resources used, automatically generated synsets contain errors, gaps in the network
- Plans: ensure good coverage of Slovene vocabulary, use sloWNet in applications

On-going work

- student project: annotation of a corpus with wordnet senses
  - take a word in context & try to assing one of its meanings from wordnet
  - can be used as training data for future wsd research
- bilateral project: using wordnet in a machine translation system (Eng-Slo, Eng-Hun)
  - can wordnet help in better wsd and lexical choice in MT

Project ideas

- wordnet browser & visualization tool
- automatic assignment of synset reliability scores
- automatic detection of errors in synsets
- automatic filling of the gaps in the network
- extension of a particular domain
- extension of the wiki approach to polysemous words (wsd)
- using sloWNet in an application

- Thanks for your attention
- Questions, comments & suggestions welcome
- Cooperation in refining, extending & using sloWNet welcome

- http://lojze.lugos.si/~darja/slownet.html