Language Resources

Water of life for natural language processing

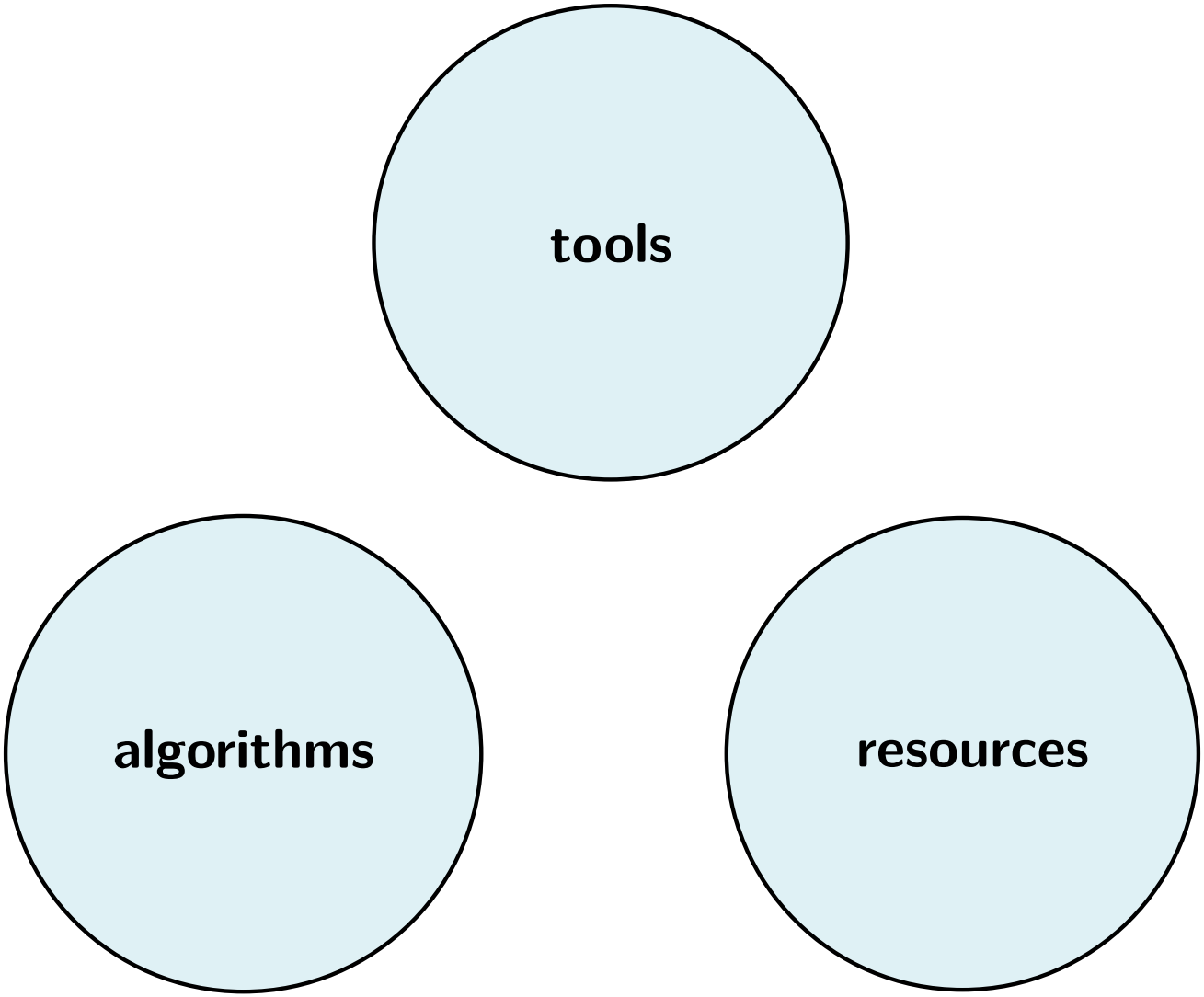
Zuzana Nevěřilová, Adam Rambousek

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

xpopelk@fi.muni.cz, xrambous@fi.muni.cz

Workshop of Natural Language Processing Centre
May 28, 2013

Natural Language Processing



Language resources

similar to dictionaries but more general

- knowledge about language
- knowledge about the world

cat

pronunciation: **kaet**

part of speech: **noun**

definition 1: a small, common mammal with four legs and a long tail. People often keep cats as pets.
Cats purr when they are happy.

Language resources

kočka -y ž

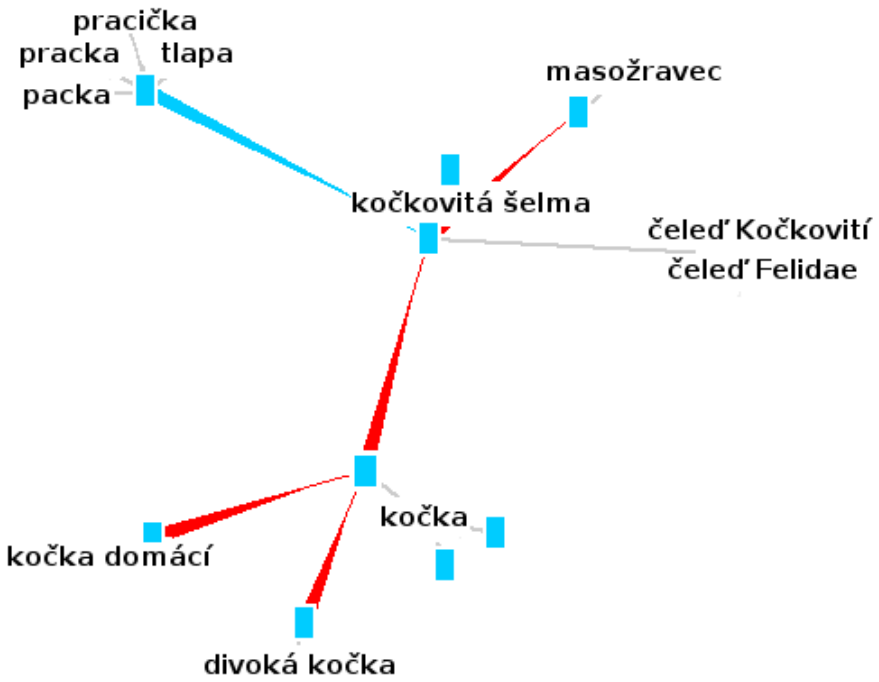
1. *malá kočkovitá šelma, zprav. chovaná v domácnosti, její samice*

2. *kožešina z (divoké) kočky* hovor. *kožešina (syno)* kabát s kočkou, **[x]** falešný jako kočka *dávat mu najevo svou převahu;* je to pro, na kočku, *ob. expr. není to k ničemu;*

kočičí příd.

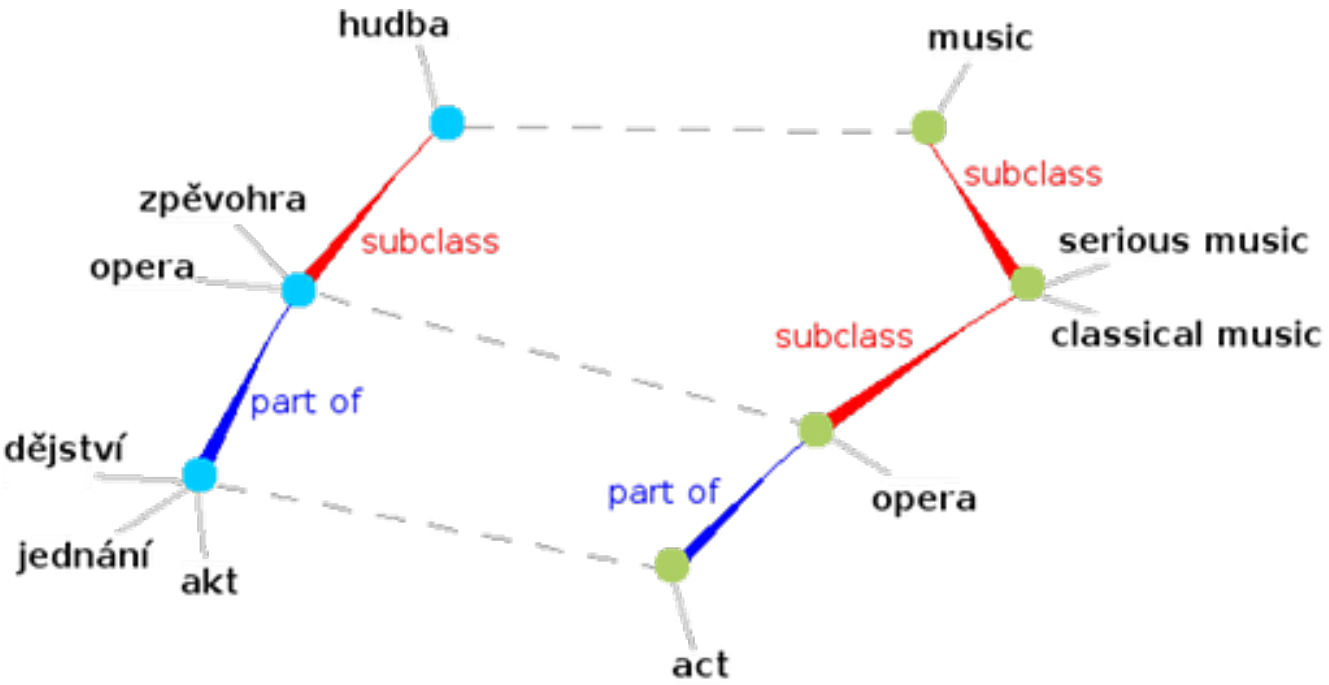
Slovník spisovné češtiny

- intended for humans: multilingual dictionaries, explanatory dictionaries, thesauri, encyclopedias
- intended for computer programs: translation memory, knowledge bases, semantic networks



Czech WordNet

WordNets



- 85,592 words organized in 40,919 synonymical sets
- several relation types: subclass, part-of, translation, synonymy

Synonyms: Dictionary vs. thesaurus

handsome

▶ **adjective** 1. *a handsome, dark-haired young man*

—SYNONYMS **good-looking**, nice-looking, attractive, personable, striking, stunning, fine, well-proportioned, well-formed; [informal] hunky, dishy, gorgeous, drop-dead gorgeous, tasty, fanciable, knockout; [Brit.] [informal] fit; [N. Amer.] [informal] cute; [Austral./NZ]

handsome

Lemma	Score	Frequency
charming	0.438	144228
gorgeous	0.391	268712
elegant	0.386	284702
attractive	0.378	521645
sexy	0.377	317087
lovely	0.362	457754
tall	0.358	392665
good-looking	0.351	19948
stylish	0.34	248842
beautiful	0.338	1654731
intelligent	0.331	255455
cute	0.33	310776

- from the contemporary language
- similarity score
- available for many languages
- for every word used in the language

Selected language resources at NLPC

- 6 dictionaries of Czech language, 512,000 of entries
- synonyms
  - Czech synonyms (K. Pala): 23,000 entries, 56,000 synonyms
  - Czech WordNet: 85,592 words organized in 40,919 synonymical sets
  - automatically generated thesaurus
- translation
  - interconnected wordnets: Czech, English, Dutch, Italian, Spanish, French, Greek, Polish, Romanian, Turkish
- specials
  - contemporary vulgar words (April 2013): 600 words/collocations + rules to detect concealing
  - sign language dictionary with gesture videos

Tools for language resources

Language resources have to be

- built and continuously maintained
- digitalized (OCR to XML)
- connected with other language resources
- shared among computer programs
- readable for humans

Language resource tools: the DEB platform

- platform for dictionary editing and browsing
  - strict client-server architecture
  - basically any XML data
- server
  - server side modules
  - database backend (XML database)
- client
  - lightweight
  - graphical interface
  - web interface
- practically used in 22 international scientific/commercial projects

Conclusions

- language resources:
  - dictionaries
  - corpus-based thesauri
  - semantic networks (WordNet)
- flexible and powerful tool for language resources processing: the DEB platform