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

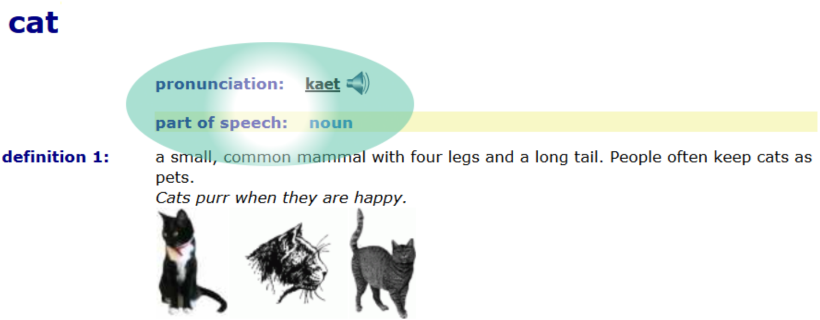
xpopelk@fi.muni.cz, xrambous@fi.muni.cz

Workshop of Natural Language Processing Centre
May 16, 2014

Language resources

similar to dictionaries but more general

- knowledge about language
- knowledge about the world



cat

pronunciation: kaet

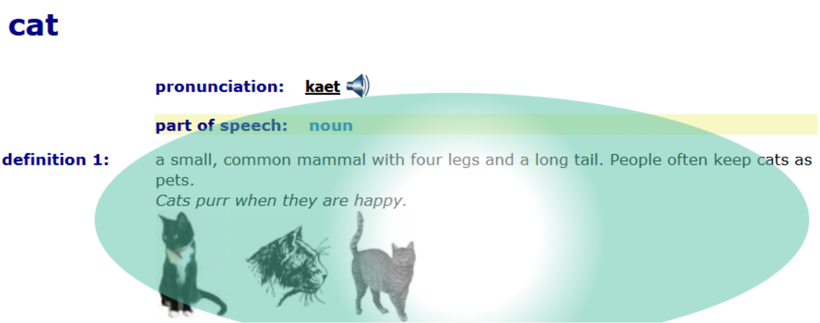
part of speech: noun

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

similar to dictionaries but more general

- knowledge about language
- knowledge about the world



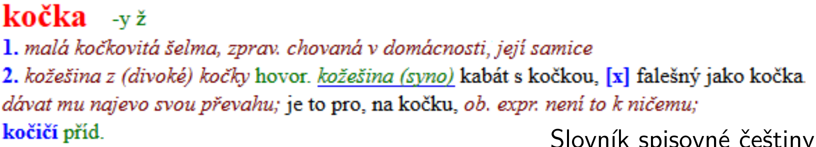
cat

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*

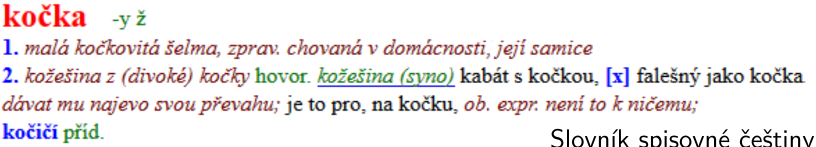
2. *kožešina z (divoké) kočky* hovor. *kožešina (syno)* kabát s kočkou, **[x]** falešný jako kočka *dávat mu najevo svou převahu;* je to pro, na kočku, *ob. expr. není to k ničemu;*

kočičí příd.

Slovník spisovné češtiny

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

Language resources



kočka -y ž

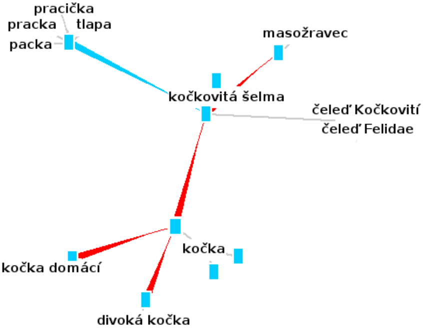
1. *malá kočkovitá šelma, zprav. chovaná v domácnosti, její samice*

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.

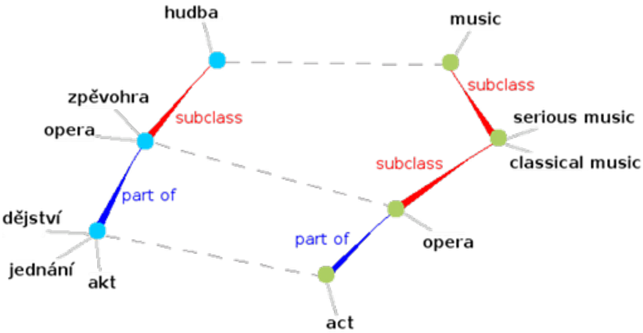
Slovník spisovné češtiny

- 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

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 (The Sketch Engine)
- 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

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 (The Sketch Engine)
- 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

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 (The Sketch Engine)
- 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

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 (The Sketch Engine)
- 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

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

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

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

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

Conclusions

- language resources:

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

Thank you