

# Sémantická analýza

PLIN059

Mgr. Dana Hlaváčková, Ph.D.

# Sémantická analýza

- **sémantika** (významy slov, slovních spojení)
- **sémantické vztahy** (hyperonymie – hyponymie)
- někdy navazuje na analýzu morfologickou a syntaktickou
  - min. určení slovních druhů (entity a abstrakce – substantiva, děje – slovesa, vlastnosti – adjektiva)
- **snaha o formální popis významů** (jazykově nezávislý)
  
- ontologie
- **sémantické sítě**
- pojmenované entity (information retrieval)
- analýza sentimentu

# Ontologie

- **ontologie** – významové struktury, skládají se z tzv. konceptů
  - mělká (shallow)
  - strukturovaná, hierarchická
  - **vrcholová** (top ontology, upper ontology)
  - **doménová** (znalostní obor, terminologie, taxonomie)

# Sémantické sítě

- FrameNet
  - <https://framenet.icsi.berkeley.edu/>
- VerbNet
  - <https://verbs.colorado.edu/~mpalmer/projects/verbnet.html>
- WordNet
  - <http://wordnet.princeton.edu>

# WordNet – motivace

- **G. A. Miller** (*Princeton University*) – psycholog a psycholinguista, psycholexikologie
- uspořádání významů v lidské lexikální paměti
- hierarchie
- experimenty – asociační testy, schopnost zpracovávat anaforické výrazy

# WordNet – struktura

- model lexikální paměti, sémantická síť
- **synset** – synonymická řada (blízká synonyma), **literál** + číslo významu
- substantiva, adjektiva, verba, adverbia
- **hierarchická struktura** – hyperonyma, hyponyma, kohyponyma
- substantiva – tematické hierarchie
- verba – vztah vyplývání
- další **sémantické vztahy** – antonyma, holonyma, meronyma, domény ontologií SUMO, MILO
- derivační vztahy

# EuroWordNet

- **Base Concepts** – jádro slovní zásoby (cca 1000 synsetů)
- **Top-Ontology** – 63 konceptů
  - entity 1. řádu = objekty
  - entity 2. řádu = stavy a procesy
  - entity 3. řádu = abstraktní pojmy (množina)
- **Interlingual Index** – číslo, které propojuje významy v jednotlivých wordnetech

# Odkazy

- Global WordNet Association
  - <http://www.globalwordnet.org>
- nástroje, prohlížeče
  - VisDic, DEBVisDic (doplněk Firefoxu), DEBVisDic 2 (webové rozhraní)
  - Responsivity Aware Wordnet viewer (RAW)
  - <https://deb.fi.muni.cz/raw-viewer/rawviewer.html>



# Pojmenované entity

- named entity
- hledání předem definovaných kategorií v nestrukturovaném textu
- „sémantické nálepky“ ke slovům, slovním spojením, číslicím a znakům
- Named Entity Recognition (NER, CZPJ FI MU)
  - <https://nlp.fi.muni.cz/projekty/ner/v2/>
- Czech Named Entity Corpus (CNEC, ÚFAL MFF UK)
  - <http://ufal.mff.cuni.cz/cnec/cnec2.0>