Sémantická analýza

Počítačové nástroje pro češtinu
Jaro 2020

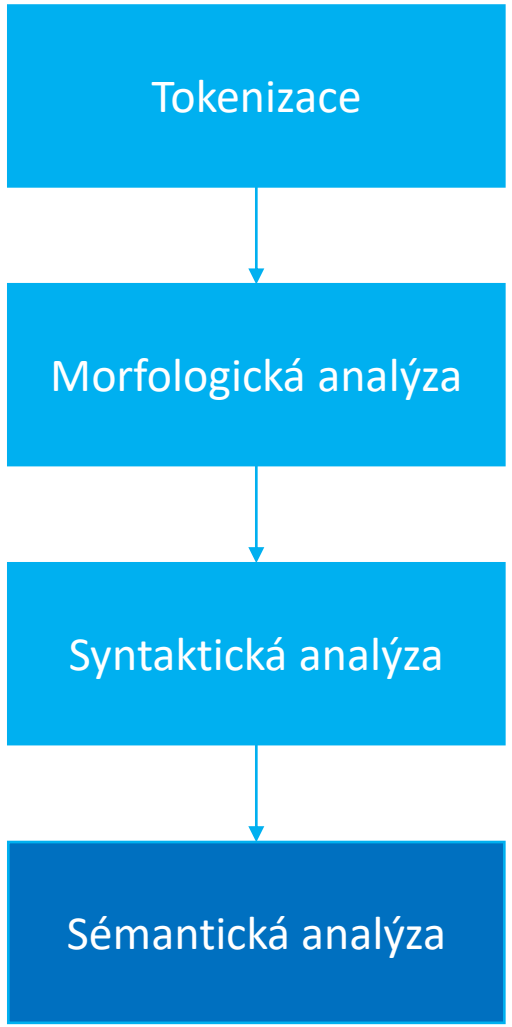
Markéta Masopustová
masopustova@phil.muni.cz

Termíny

- Sémantika – nauka o významu;
- Hyponyma – podřazená slova;
- Hyperonyma – nadřazená slova;
- Kohyponyma – významově shodná slova;
- Meronyma – označuje část celku;
- Synset – synonymická řada.

Sémantická analýza

- Snaží se o formální popis významu – rozdělit slova do skupin a dát jim nějakou nálepku.
- Snaží se o zobecnění světa.
- Měla by být jazykově nezávislá.
- Počátky můžeme najít v ontologiích (v inf. explicitní a formalizovaný popis určité problematiky).
- Většinou se jedná spíše o nějaký slovník, který uchovává znalosti z určité problematiky.
- V ČR se tím zabývají především v rámci CZPJ FI MU a na ÚFAL MFF CUNI.



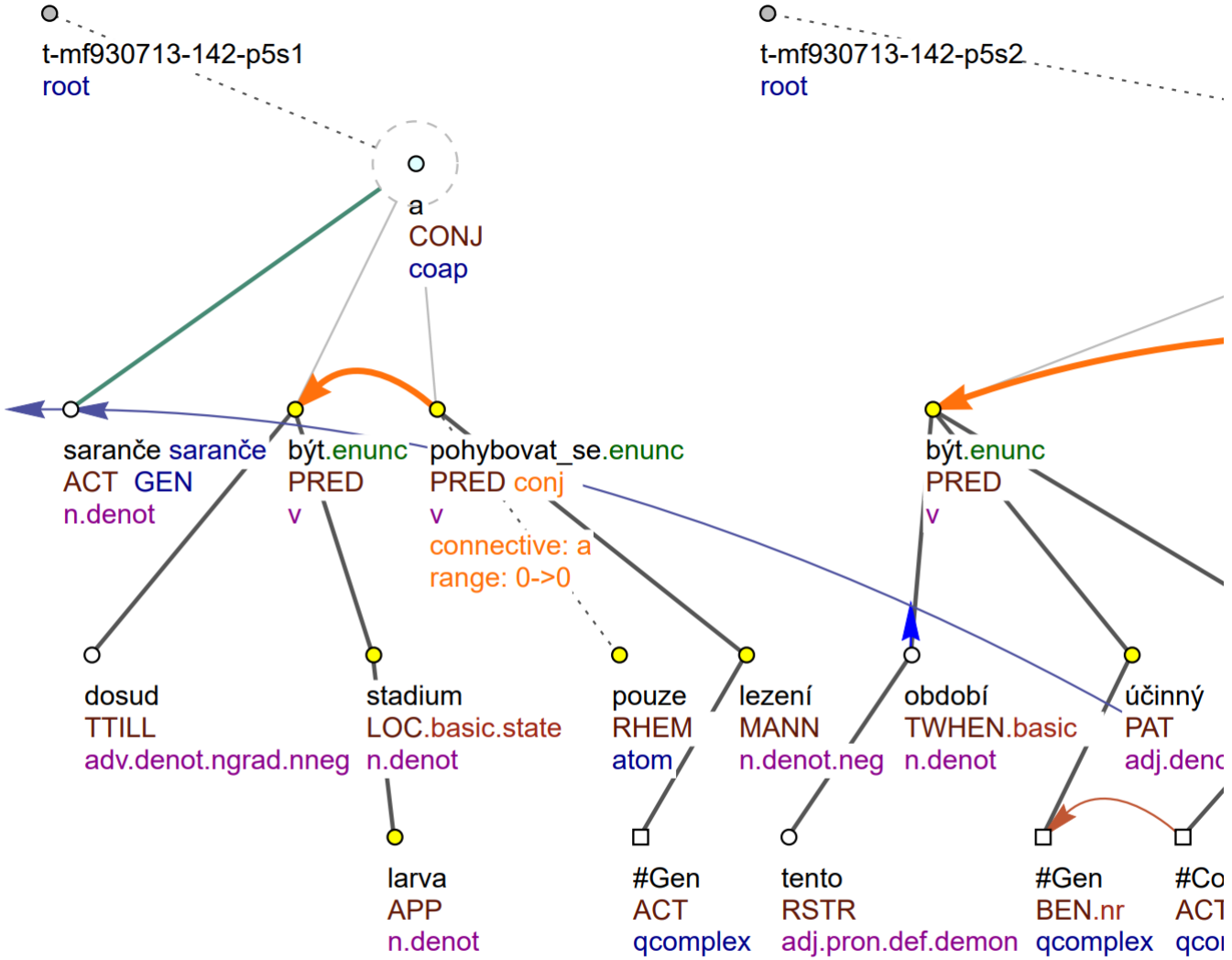
NER

- Rozpoznání pojmenovaných entit (Named Entity Recognition).
- Cílem je najít předem definované kategorie v nestrukturovaném textu.
- Poměrně složité, musí se předem definovat, co je jmenná entita.
- Jmenná entita může být např. jméno, město, datum, značka, …
- https://nlp.fi.muni.cz/projekty/ner/v2/
- http://ufal.mff.cuni.cz/cnec/cnec2.0

Pan Ing. Jan Novák z obce Chýně navrhl, aby se od 15. září 2014 městská rada (dále jen MR) scházela 4 km od památné lípy v centru obce. Svůj návrh podpořil citací § 6 zákona 121/2001 Sb. a svérázným tanečkem.

V rámci ÚFALu

- PDT:
  - Prague Dependency Treebank;
  - České texty doplněné o morfologickou a syntaktickou informaci.
  - Vyznačený není význam, ale o významová roli ve větě (agens, patiens, …).
- SEANCe – projekt ke značkování sentimentu (emocí v textu).

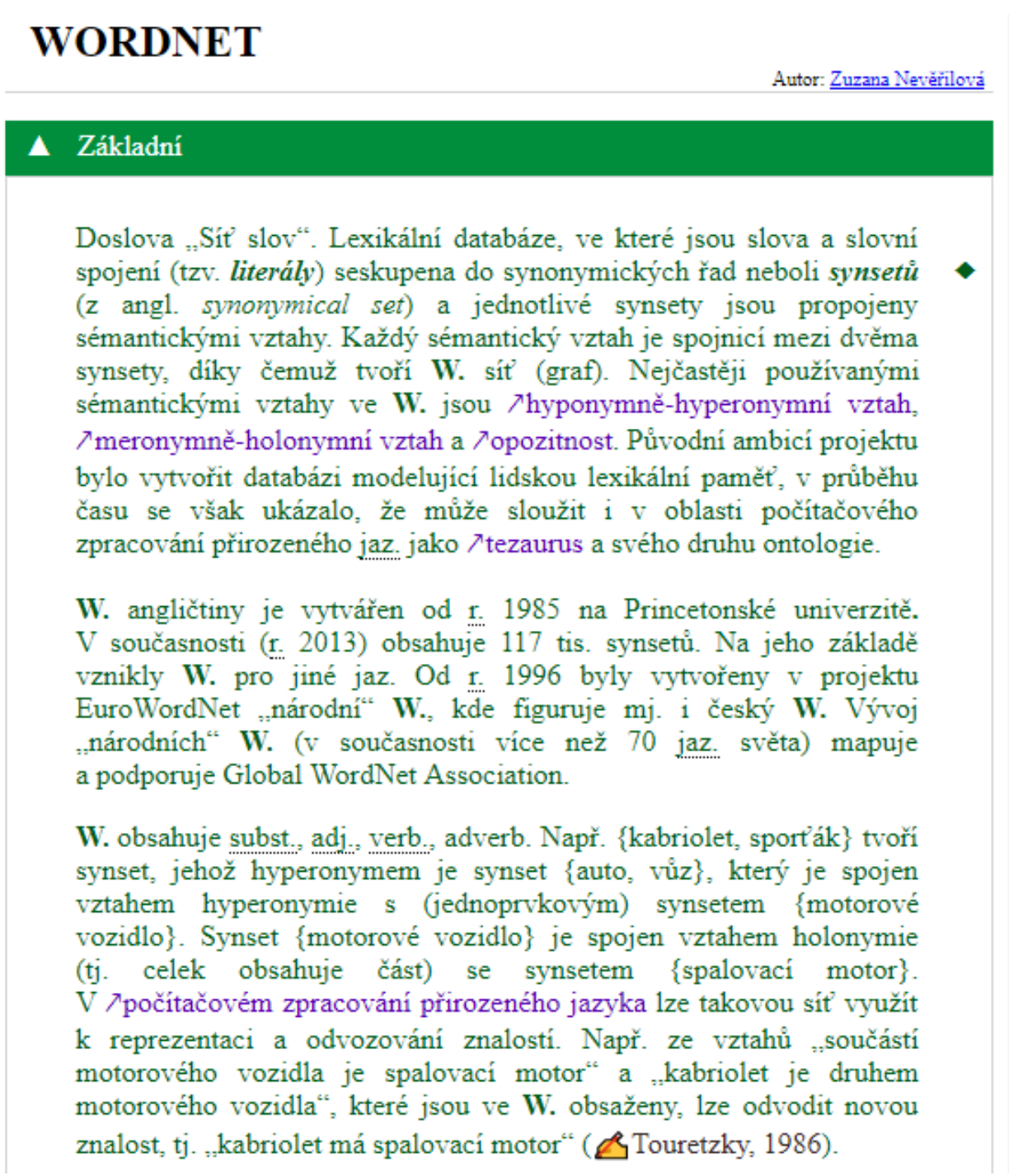


Sémantické sítě (anglické)

- FrameNet
  - lexikální síť čitelná pro člověka i stroj
  - https://framenet.icsi.berkeley.edu/fndrupal/
- VerbNet
  - slovník sloves
  - https://verbs.colorado.edu/~mpalmer/projects/verbnet.html
- ConceptNet
  - sémantická síť vytvořená pro lepší porozumění významu pro stroj
  - http://conceptnet.io/
  - obrázek z ConceptNetu ->

WordNet

- G. A. Miller (*Princeton University*) – psycholog, psycholingvista, psycholexikolog.
- Základním je Princeton WordNet (1985), postupně vytvářeny národní Wordnety.
- http://wordnet.princeton.edu
- http://globalwordnet.org/resources/wordnets-in-the-world/
- podrobnosti viz NESČ a obrázek ->



WORDNET

Autor: Zuzana Nevěřilová

▲ Základní

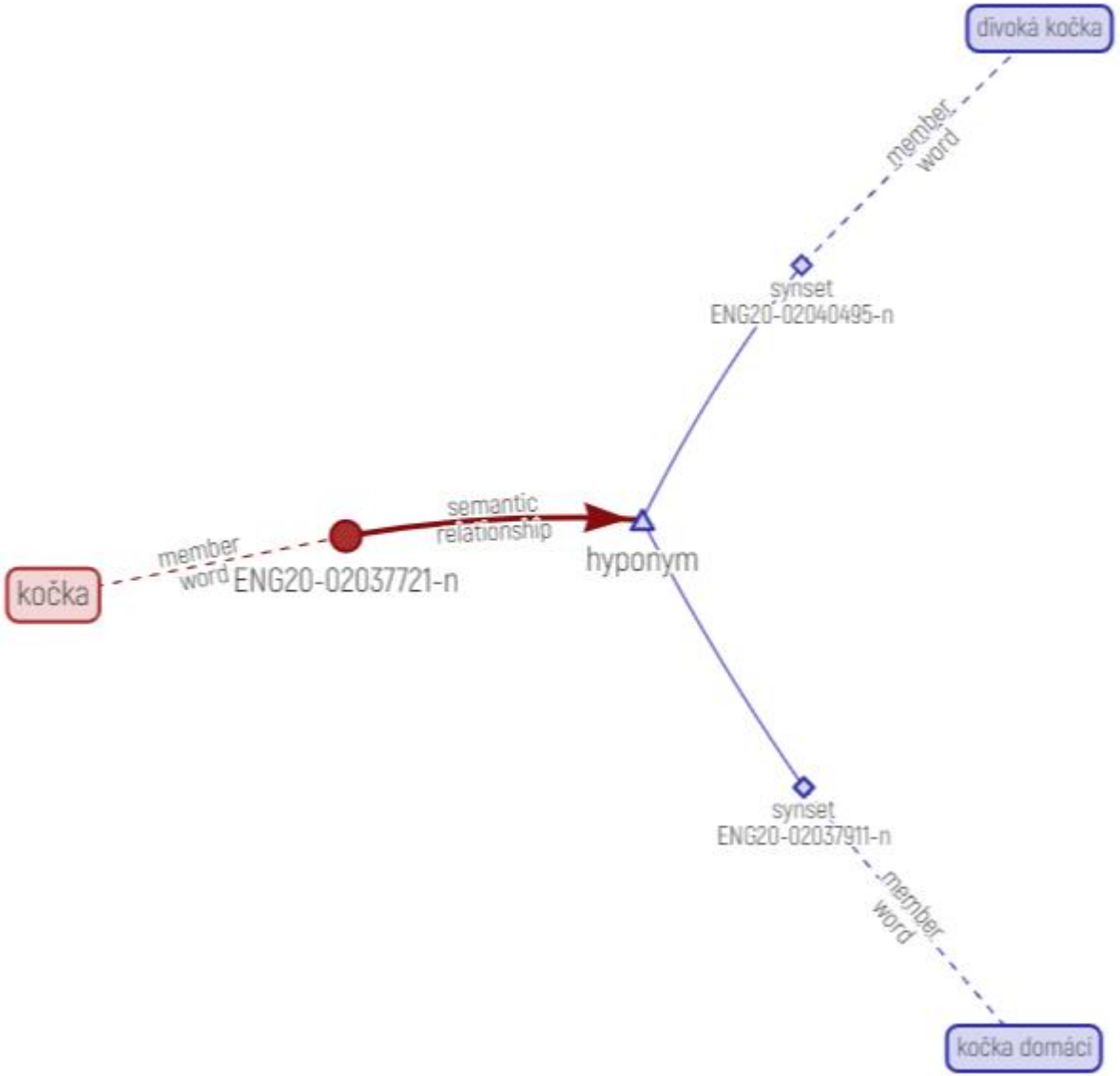
Doslova „Síť slov“. Lexikální databáze, ve které jsou slova a slovní spojení (tzv. ***literály***) seskupena do synonymických řad neboli ***synsetů*** (z angl. *synonymical set*) a jednotlivé synsety jsou propojeny sémantickými vztahy. Každý sémantický vztah je spojnicí mezi dvěma synsety, díky čemuž tvoří **W.** síť (graf). Nejčastěji používanými sémantickými vztahy ve **W.** jsou ↗hyponymně-hyperonymní vztah, ↗meronymně-holonymní vztah a ↗opozitnost. Původní ambicí projektu bylo vytvořit databázi modelující lidskou lexikální paměť, v průběhu času se však ukázalo, že může sloužit i v oblasti počítačového zpracování přirozeného jaz. jako ↗tezaurus a svého druhu ontologie.

W. angličtiny je vytvářen od r. 1985 na Princetonské univerzitě. V současnosti (r. 2013) obsahuje 117 tis. synsetů. Na jeho základě vznikly **W.** pro jiné jaz. Od r. 1996 byly vytvořeny v projektu EuroWordNet „národní“ **W.**, kde figuruje mj. i český **W.** Vývoj „národních“ **W.** (v současnosti více než 70 jaz. světa) mapuje a podporuje Global WordNet Association.

W. obsahuje subst., adj., verb., adverb. Např. {kabriolet, sporťák} tvoří synset, jehož hyperonymem je synset {auto, vůz}, který je spojen vztahem hyperonymie s (jednoprvkovým) synsetem {motorové vozidlo}. Synset {motorové vozidlo} je spojen vztahem holonymie (tj. celek obsahuje část) se synsetem {spalovací motor}. V ↗počítačovém zpracování přirozeného jazyka lze takovou síť využít k reprezentaci a odvozování znalostí. Např. ze vztahů „součástí motorového vozidla je spalovací motor“ a „kabriolet je druhem motorového vozidla“, které jsou ve **W.** obsaženy, lze odvodit novou znalost, tj. „kabriolet má spalovací motor“ (Touretzky, 1986).

WordNet prakticky

- Přístup: demo/demo; read/read.
- DebVisDic 2:
  - https://deb.fi.muni.cz/proj_debvisdic-cs.php
- RAW viewer:
  - https://deb.fi.muni.cz/raw-viewer/rawviewer.html



Děkuji za pozornost.