CJBB105 – 6
Morfologické značkování

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Morfologické značkování

- **token/word – lemma – tag**
- tokenizace – rozdělení na pozice
- lemmatizace
  - přiřazení základního slovního tvaru (jednoslovné)
  - subst. – nom.sg., adj. – nom. sg. masc., verb. – infinitiv
  - význam spojený s tvarem – *nerv/nervy, na holičkách*
- tagging
  - přiřazení morfologické značky (všechny interpretace tvaru)
- desambiguace – zjednoznačnění lemmat a tagů na základě kontextu
- tagger – nástroj, který provádí morfologickou analýzu a desambiguaci

Morfologické značkování

- na úrovni **slovních druhů**
  - PoS tagging (angličtina)
  - neohebné slovní druhy (spojky, částice, citoslovce)
    - adverbia – značena negace a stupeň
- **kompletní**
  - všechny morfologické kategorie (slovanské/flektivní jazyky)
    - ohebné slovní druhy
  - nutné pro další stupně automatického zpracování jazyka a navazující aplikace

Morfologické značky

- transparentní – tagset
  - jednoznačná interpretace značky
- zachycují morfologické charakteristiky
  - křížení se sémantickými vlastnostmi (např. druhy zájmen a adverbií)
- nezávislé na lingvistických teoriích
  - orientované na uživatele a současně strojově čitelné
- podoba – kód sestavený z písmen a čísel
  - **kočka/kočka/NNFS1-----A-----**
  - **kočka /kočka/k1gFnSc1**
  - **kot** [kot:subst:sg:nom:m2]
  - **cat /NN/cat**
  - **Katze /N.Reg.Nom.Sg.Fem/Katze**

Homonymie

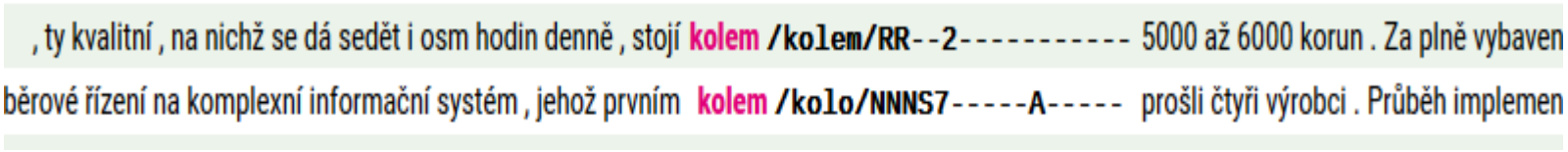
- významová – rozdíl v morfologických kategoriích
  - *koruna, sladit (uvést v soulad/činit sladkým – vid)*
- tvarová – nejfrekventovanější
  - *jarní (rod, číslo, pád)*
- slovnědruhová
  - *jak (adverbium, spojka, částice)*
- kombinovaná
  - *ženu (subst., f, ak., sg./verb., 1. os., sg.)*
- *Sním je místo něho. Praštil se sluchátkem.*
  - *http://nlp.fi.muni.cz/projekty/wwwajka*

Metody automatického značkování

- morfologické značkování včetně desambiguace
- závisí na velikosti a kvalitě morfologického slovníku
- Stochastické (statistické, pravděpodobnostní)
  - strojové učení (referenční data)
- Pravidlové
  - pravidla stanovená lingvisty nebo vyvozená z textu
  - pozitivní i negativní
- Hybridní
  - kombinace obou přístupů, nejúspěšnější
- neznámé tvary – **guesser** – odhadne možné lemma a tag
- úspěšnost taggerů – přesnost (precision) a pokrytí (recall)

Morfologická analýza v ČR

- ÚČNK Praha – Český národní korpus, manažer KonText
  - Ústav formální a aplikované lingvistiky MFF UK
  - Ústav teoretické a komputační lingvistiky FF UK
- **poziční systém**
  - značka se skládá z 16 pozic, každá vyjadřuje jednu morfologickou charakteristiku
  - 2 rezervní, 1 stylová, 1 smíšená
- analyzátor stochastický, pravidlový, hybridní



Morfologická analýza v ČR

- FF + FI MU Brno, manažer Sketch Engine
  - Centrum zpracování přirozeného jazyka FI MU
  - Ústav českého jazyka FF MU
- **atributivní systém**
  - atribut – morfologická kategorie obecně (c = pád)
  - hodnota – morfologická kategorie konkrétně (1–7)
- analyzátor pravidlový, hybridní (+ guesser)

ový výsledek nebyl ovlivněn . Druhým kolem /kolo/k1gNnSc7 prezidentských voleb se Rusko ve středu

luma požaduje šetření korupce kliky kolem /kolem/k7c2 Gračova Jako člověka po uši zapleteného

Syntaktická analýza

- v korpusu SYN2015 (zveřejněn 2016)
- zobrazení závislostních vztahů mezi slovy ve větě – závislostní strom
- vychází z **PDT** (Prague Dependency Treebank) z ÚFAL MFF UK
  - manuálně označkovaná data
  - východiskem je syntax Vl. Šmilauera
  - syntaktický analyzátor (parser) – úspěšnost cca 80 %
- zobrazení v KonTextu, možnosti vyhledávání podle syntaktických atributů (*parent, afun*)

Práce s morfologickými značkami

- uživatel – lingvista
- znalost tagsetu a principu analýzy
- vyhledávání v korpusu podle morfologických charakteristik
- kontrola správnosti značkování
- jazykové a frekvenční studie
- schopnost interpretace značky a nalezených informací
- projekt NovaMorf