CJBB85 Počítačové nástroje pro češtinu – úvod

Mgr. Dana Hlaváčková, Ph.D.

hlavacko@phil.muni.cz

Ústav českého jazyka FF MU

Joštova 13, budova M

Co je to počítačová lingvistika?

- obor mezi lingvistikou a informatikou
- výsledkem jsou denně používané **aplikace**
  - korektor překlepů
  - korektor gramatiky
  - vyhledávání na webu
  - prediktivní psaní
  - překladače jazyků
- **detailní analýza jazyka a jeho formální popis**

Počítačové zpracování přirozeného jazyka – češtiny

- přirozený jazyk x počítačové zpracování
- jak funguje přirozený jazyk?
- jak funguje počítač?
- **formální popis** jazyka
- **algoritmus** – návod, postup při řešení daného problému
- pravidelnost v jazyce (cca 80 %) – algoritmický popis

Počítačové zpracování češtiny – pár zásad

- proč to chceme? (cíl, účel, uživatel)
- jak toho dosáhneme? (efektivita)
- maximum automatizace – minimum ruční práce (při vytváření i používání)
- zpracování velkého objemu dat
- univerzálnost (široká množina vstupů, spojování více nástrojů do jednoho)
- nezávislost na jednotlivých lingvistických teoriích
- při zpracování i používání je nutná PŘESNOST („ono to nefunguje“ ☹)

Počítačové zpracování češtiny

- urychlení a zefektivnění práce lingvisty
- ověřování existujicích teorií
- objevení nového jazykového jevu, zákonitosti
- co a jak mohu použít
- co mohu a nemohu od nástroje očekávat
- autorská práva a přístupy k nástrojům
  - veřejně dostupné (dostupné na MU)
  - hromadný přístup (společné heslo)
  - vlastní přístup (registrace)

Mezioborová spolupráce

- informatika – lingvistika („společný jazyk“)
- počítačová lingvistika (matematická, komputační), jazykové inženýrství, počítačové zpracování přirozeného jazyka
- Natural Language Processing (NLP)

Hlavní oblasti (uživatelský přístup)

- syntéza a analýza řeči
- počítačová lexikografie
- formální analýza jazyka (morfologická, slovotvorná, syntaktická, sémantická, textová)
- korpusová lingvistika
- dialogové systémy, umělá inteligence

Obsah kurzu

- **počítačová lexikografie** – DEBDict, DEBWrite, lexikální databáze, Vokabulář webový a další
- **jazykové korpusy** – KonText, Sketch Engine
- **morfologická analýza** – Ajka, Majka, Morče, MorphoDiTa (atributivní a poziční systém)
- **derivační** rozhraní – Deriv, Morfio a další
- **syntaktická** analýza – Synt, Set, PDT (stromové banky)
- **sémantická** analýza – WordNet, FrameNet, VerbNet
- **valenční databáze** – Vallex, VerbaLex
- **rozpoznávání a syntéza řeči**
- seminární práce

Příbuzná pracoviště

- Centrum zpracování přirozeného jazyka FI MU Brno – http://nlp.fi.muni.cz/
- Ústav formální a aplikované lingvistiky MFF UK Praha – http://ufal.mff.cuni.cz
- Ústav teoretické a komputační lingvistiky FF UK Praha – http://utkl.ff.cuni.cz
- Ústav Českého národního korpusu FF UK Praha – http://www.korpus.cz
- Ústav pro jazyk český AV ČR – http://www.ujc.cas.cz

Příbuzná pracoviště

- Fakulta informačních technologií VUT Brno – http://www.fit.vutbr.cz
- Katedra informatiky a výpočetní techniky – http://www.kiv.zcu.cz, Katedra kybernetiky http://www.kky.zcu.cz FAV ZCU Plzeň
- Ústav informačních technologií a elektroniky FM TU Liberec – http://www.fm.tul.cz
- Slovenský národný korpus, JÚĽŠ SAV Bratislava – http://korpus.juls.savba.sk/

Bonus

- Internetová jazyková příručka
  http://prirucka.ujc.cas.cz

- WebMetaTrans

http://metatrans.fi.muni.cz