Počítačové nástroje pro češtinu

Mgr. Dana Hlaváčková, Ph.D.

hlavack@fi.muni.cz

Centrum zpracování přirozeného jazyka

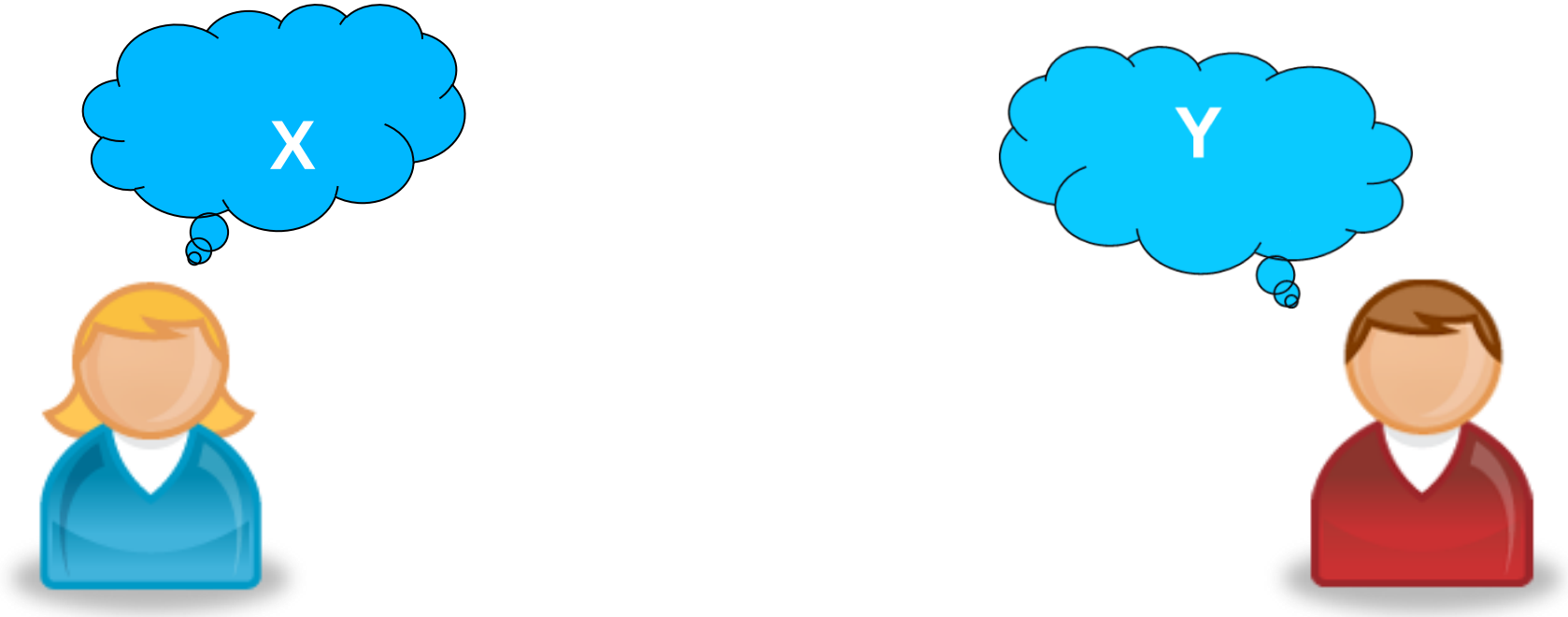
budova B, 2. patro, prostřední laboratoř (B206)

Fakulta informatiky MU

Formality

- kód CJBB85 (dříve Lingvistický software)
- 3 kredity, ukončení zápočtem
- praktický seminář, teoretický úvod
- aktivní účast v semináři, 1 neomluvená absence, jinak omluvenky v ISu
- seminární práce (praktická)

Komunikace



ZDROJ
(vysílač, mluvčí)

KOMUNIKAČNÍ KANÁL
(signál, kód)

PŘÍJEMCE
(adresát, posluchač)

Počítačové zpracování češtiny

- přirozený jazyk x počítačové zpracování
- jak funguje přirozený jazyk?
- jak funguje počítač?
- **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
- zpracování velkého objemu dat
- univerzálnost (široká množina vstupů)
- nezávislost na jednotlivých lingvistických teoriích
- PŘESNOST („ono to nefunguje“ ☹)

Počítačové zpracování češtiny

- urychlení a zefektivnění práce lingvisty
- zpracování velkého množství jazykových dat
- ověřování existujicích teorií
- objevení nového jazykového jevu, zákonitosti
- co a jak mohu použít
- co mohu a nemohu od nástroje očekávat
- autorská práva a přístupy

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)
- korpusová lingvistika
- dialogové systémy, umělá inteligence

Obsah kurzu

- počítačová lexikografie, prohlížeč a editor slovníků – DebDict a další
- rozpoznávání a syntéza řeči
- jazykové korpusy – Bonito2, Word Sketches
- morfologická analýza – AJKA
- derivační rozhraní – Deriv
- syntaktická analýza – KLARA, SYNT, PDTB
- sémantická analýza – WordNet, Ontologie
- valenční databáze – Vallex, VerbaLex
- 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 FAV ZCU Plzeň – http://www.kiv.zcu.cz, Katedra kybernetiky http://www.kky.zcu.cz
- Ústav informačních technologií a elektroniky FM TU Liberec – http://www.fm.tul.cz

Bonus

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

- Web MetaTrans

http://metatrans.fi.muni.cz