

Jak si čeština vede v počítačovém světě

Mgr. Dana Hlaváčková, Ph.D.

Ústav českého jazyka FF MU

hlavacko@phil.muni.cz

Čeština ve světě počítačů

- čeština je pro počítačový svět **dostatečně velký jazyk**
 - 13,2 mil. mluvčích (10,7 mil. v ČR a 2,5 mil. v zahraničí)
 - slovenština 7 mil. mluvčích
- využívání softwaru (a dalších produktů) v češtině
- komunikace s počítačem v češtině
 - lokalizace
 - strojový překlad
- **počítačové zpracování češtiny**
 - počítačové nástroje zpracovávají přirozený jazyk

Přirozený jazyk vs. počítač

- **přirozený jazyk**

- složitý systém – vyvíjí se a mění
- užívaný konkrétním jazykovým společenstvím
- pravidla užívání nejsou definována formálně, ale vyplývají z úzu
- **nejednoznačnost** (homonymie)
- *plod/plot, koncovka -ý, ženu, Praštil se sluchátkem.*

- **počítač**

- schopen pracovat pouze na základě formálních pravidel
- vyžaduje jednoznačnost a přesnost

Počítačová lingvistika

- **vymezení oboru**
 - propojení humanitních a přírodních věd
 - **lingvistika – informatika – matematika**
(statistika, pravděpodobnost)
- **počítačová lingvistka** (*Computational Linguistics*)
 - matematická, kvantitativní, počítačová
- **počítačové zpracování přirozeného jazyka**
(*Natural Language Processing, NLP*)

Hlavní oblasti počítačové lingvistiky

- **formální (algoritmický) popis rovin jazyka**
 - morfologie, slovtvorba, syntax, sémantika (WordNet)
- **korpusová lingvistika** (zdroj jazykových dat)
- **počítačová lexikografie**
- **rozpoznávání a syntéza řeči**
- **strojový překlad**
- **dialogové a otázkové systémy** (chatboty)
- **reprezentace znalostí** (text)

Počítačový model přirozeného jazyka

- počítač rozeznává **akustické signály** a **grafické znaky a jejich řetězce**
- chceme, aby uměl:
 - mluvení, psaní, čtení
 - vedení dialogu, porozumění textu
- **hláska** (foném, difón, trifóny)
- **slovo** a jeho segmentace
 - řetězec znaků ohraničený z obou stran mezerami
 - významy slov (*sladit cukrem, sladit části v celek*)
- spojování slov do větších celků – **věta**
- spojování vět – **text**

Český jazyk (a další slovanské jazyky)

- *obtížný pro počítačové zpracování*
- asimilace (znělosti, artikulace)
- bohatá morfologie
- slovtvorba (alternace hlásek v kmeni)
- volný slovosled
- bohatá synonymie a **homonymie**

S čím se běžně setkáváte

- **korektory** překlepů a gramatiky (*MS Word*)
- **dělení slov** v textových editorech
- **prediktivní psaní** (našeptávače)
- **vyhledávání** na internetu (zapojení lemmatizace, *Seznam.cz*)
- **Internetová jazyková příručka** (<http://prirucka.ujc.cas.cz>)
- **syntéza a rozpoznávání řeči**
- **překladače** (*Google Translate*)
- určování autorství a **plagiátorství** (*IS MU, Odevzdej.cz*)
- programy pro **extrakci informací** z textu (<https://nlp.fi.muni.cz/projekty/topicks>)

Jazykové korektory

- **korektory překlepů** (spellchecker + našeptávač)
 - slovník (slova, příp. slovní spojení)
- **korektor gramatiky**
 - interpunkce (je součástí pravopisu)
 - shoda podmětu s přísudkem, shoda přívlastku
 - styl (opakování slov, hovorové tvary, nevhodné výrazy)
 - časté chyby (**vyjímka, *standartní, *abyjsme*)
 - typografie
- **Řek sem *nu aby my loupil* lísky.**
Petr se včera večer v lese s Karlem ztratil a byly tři hodiny pryč.

Korpusová lingvistika

- **elektronické zdroje jazykových dat**
- **národní korpusy**
 - vyvážené, reprezentativní – Český národní korpus
 - složka diachronní, synchronní, psaná, mluvená
 - paralelní korpusy
- specializované korpusy
- morfologické a syntaktické značkování
- **webové korpusy**
 - miliardy slovních výskytů (čeština 10 mld.)
 - proměnlivost zdroje, neznámé autorství

Korpusová lingvistika

- **ČNK, KonText** (korpusový manažer, <http://www.korpus.cz>)
- analýzy jazyka na všech dostupných rovinách
- využití v literárněvědných výzkumech
- pro studenty – využití při psaní závěrečných prací
- některé nástroje dostupné p na MU
- **Sketch Engine** (korpusový manažer, <https://www.sketchengine.eu/>, UČO, primární heslo)
- možnost budování vlastních korpusů

Hlavní pracoviště

- *Ústav českého jazyka, FF MU Brno*
- *Centrum zpracování přirozeného jazyka, FI MU Brno*
- *Ústav formální a aplikované lingvistiky, MFF UK Praha*
- *Ústav teoretické a počítačové lingvistiky, FF UK Praha*
- *Ústav Českého národního korpusu, FF UK Praha*
- *Fakulta informačních technologií, VÚT Brno*
- *Fakulta aplikovaných věd, ZČU Plzeň*
- *Ústav informačních technologií a elektroniky,*
TU Liberec

Studium

- ***Počítačová lingvistika***, FF MU Brno (bc., mgr.)
- ***Zpracování přirozeného jazyka (bc.), Umělá inteligence a zpracování dat (mgr.)***, FI MU
- ***Matematická lingvistika***, MFF UK Praha (mgr., dr.)
- ***Matematická lingvistika***, FF UK (dr.)
- Saarland University, Germany
- Brandeis University, USA
- Saint Petersburg State University, Russia

Čeština v počítačovém světě ... si vede dobře

- množství korpusů klasických i webových
- morfologické a syntaktické analyzátory
- aplikace pro formální popis derivace
- systémy pro rozpoznávání a syntézu řeči
- manažer Sketch Engine pro cizojazyčné korpusy
- celosvětová sémantická síť WordNet
- spolupráce se zahraničními firmami (Appen, What3Words)
- slabší pozice v oblastech, kde je potřeba obrovské množství dat (např. strojový překlad)

Děkuji za pozornost