

CJBB105 – 7

Využívání korpusů

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Využívání korpusů

Jazyková data:

- využití v lingvistice
- využití v NLP

Využívání korpusů

- analýzy založeny na důsledném **využívání jazykových dat** pro popis jazyka
 - díky počítačům (hardwaru i softwaru)
- klíčový je **mimořádný rozsah dat** a jejich přístupnost, jazykový materiál je:
 - odrazem skutečného užívání jazyka
 - aktuální (v daném časovém období)
 - objektivní (vyváženost, reprezentativnost)
 - dostatečný (velikost)
 - lehce přístupný (korpusové manažery)

Využívání korpusů

- korpusy – zvrát v lingvistice, velké množství reálných jazykových dat, rozvoj exaktního přístupu k jazyku
- velké korpusy jsou dostatečným reprezentativním vzorkem jazyka – výskyty jevů a jejich frekvence nejsou náhoda
- **kvantitativní analýza**
 - počet výskytů (typické a okrajové jevy, přechodové oblasti, variabilita jazyka, tři pásma frekvenčního seznamu)
 - nutná lingvistická interpretace výsledků
- **kvalitativní analýza**
 - nezávisí na počtu výskytů (i málo frekventované jevy jsou důležité, hapax legomena a výzkum jazykové periferie)

Využívání korpusů

- **corpus-based** (korpusem ověřovaný) přístup
 - ověřování stávajících teorií (založených na introspekci a několika příkladech)
 - od hypotézy k dokladům
 - *např. doložení existence variantních koncovek, posouzení jejich frekvence*
- **corpus-driven** (korpusem řízený) přístup
 - průzkum korpusového materiálu, tvorba nových teorií (úprava stávajících)
 - od dokladu k hypotéze
 - *např. výzkum aktuálních kolokací*

Co v korpusu nenajdeme

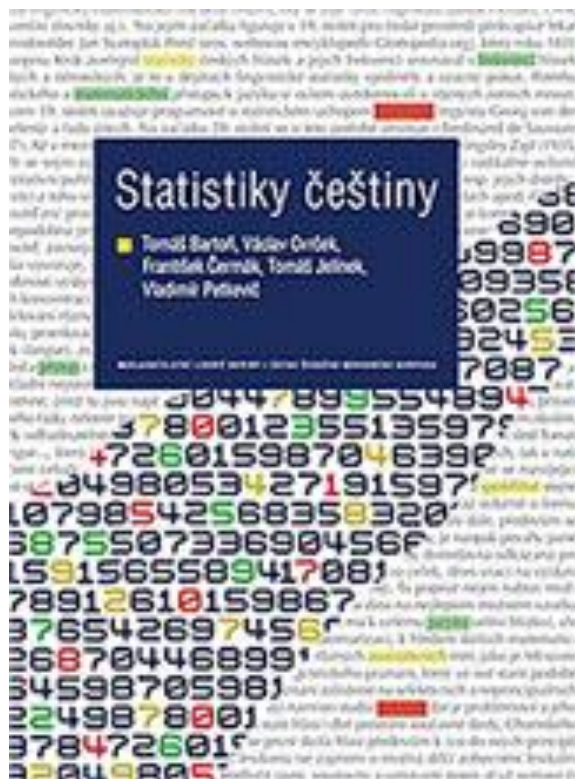
- korpus není preskriptivní, ale deskriptivní
- korpus nejsou pravidla pravopisu
- korpus není encyklopedie
- korpus není výkladový nebo překladový slovník
- korpus neobsahuje všechna česká slova
- jakkoli velký korpus je pouze vzorek jazyka

Frekvenční studie

- frekvence slov, slovních tvarů, slovních spojení, slovních druhů, slovních segmentů (slabiky, kmeny, sufixy, koncovky), hlásek, znaků (interpunkce)
- frekvenční slovníky (Těšitelová – 1961; **FSČ – 2004**)
- výzkum variant (SyD)
 - např. pravopisné (*filozofie/filosofie*), tvarové (*kopu/kopám*), stylové (*pořád/furt*)
- míra pronikání cizí slovní zásoby, proces počešťování slov
 - *byznis, byznys, biznis, biznys*

Frekvenční studie

- stylistická pozorování – typická slova v určitých typech textů (široké využití)
 - klíčová slova
 - určování sociolingvistických charakteristik
 - projevy emocí
 - určování autorství
 - forenzní lingvistika
- výuka jazyka pro cizince (slova v kontextech)
- akvizice jazyka (korpusy dětského jazyka, výukové korpusy, značkování chyb)
- výzkum terminologie
- korpus jako obraz společnosti (reálie, společenská situace)



Bartoň, T. a kol. Statistiky češtiny. Praha: NLN, 2009.

Čermák, F. - Křen, M. (eds.) Frekvenční slovník češtiny. Praha: NLN, 2004.

Počítačová (korpusová) lexikografie

- od počátků je vznik korpusů spojen s tvorbou slovníků a gramatik
- výběr slovníkových hesel, lemmata, hranice min. počtu výskytů
- významy slov na základě jejich kontextu
- reálné příklady užití
 - konkordance (KWIC)
- souvýskyty, kolokace, frazeologismy, thesaury, Word Sketch
- metadata
 - časová datace slovního výskytu, typ textu, autor
- možnost aktualizace
- na slovnících budovaná kontrola překlepů, gramatiky a stylu

Počítačová (korpusová) lexikografie

- formát slovníku – značkovací jazyky – popis struktury slovníkového hesla – konzistence slovníku
 - SGML (Standard Generalized Markup Language)
 - XML (eXtensible Markup Language)
 - DTD (Document Type Definition) – definice atributů textu
 - počáteční a ukončovací značky, např. <orth> <def> <pos> <gram> <eg>
- lexikografické stanice – modulární dělení práce

Popis rovin jazyka

- **fonetika, fonologie** – pokud jsou charakteristiky značkovány (OMK, nově ORAL)
- **morfologie** – tagging, frekvence tagů
- **slovotvorba** – slovotvorné segmenty, derivace, funkční zatížení prefixů/sufixů (Morfio, Deriv)
- **syntax** – syntaktická analýza, nominální a verbální fráze, koreferenční vztahy, aktuální větné členění
- **sémantika** – odvození významu na základě kontextu
- **vývoj jazyka** (diachronní korpusy)
- **multiword expression (MWE)**
 - *Karel IV., corpus delicti*

Využití korpusů v NLP

- tvorba nových nástrojů, minimalizace ručního hledání
- strojové učení
- strojový překlad
- rozpoznávání a syntéza řeči
- dialogové systémy
- určování autorství
- analýza emocí
- extrakce informací z textu, pojmenované entity