CJBB105 – 7

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Využívání korpusů

- korpusy – zvrat v lingvistice, reálná jazyková data, exaktní přístup k jazyku
- velké korpusy jsou dostatečným vzorkem jazyka – výskyty jevů a jejich frekvence nejsou náhoda
- kvantitativní analýza
  - počet výskytů (typické a okrajové jevy, přechodové oblasti, variabilita jazyka)
  - závisí na velikosti korpusu
  - nutná lingvistická interpretace výsledků
- kvalitativní analýza
  - nezávisí na počtu výskytů (i málo frekventované jevy jsou důležité, hapax legomena a výzkum jazykové periferie)

Čermák, F. Periferie jazyka – Slovník monokolokabilních slov. Praha: NLN, 2014.
Cvrček, V. Kvantitativní analýza kontextu. Praha: NLN, 2013.

Využívání korpusů

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

Využívání korpusů

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

Co v korpusu nenajdeme

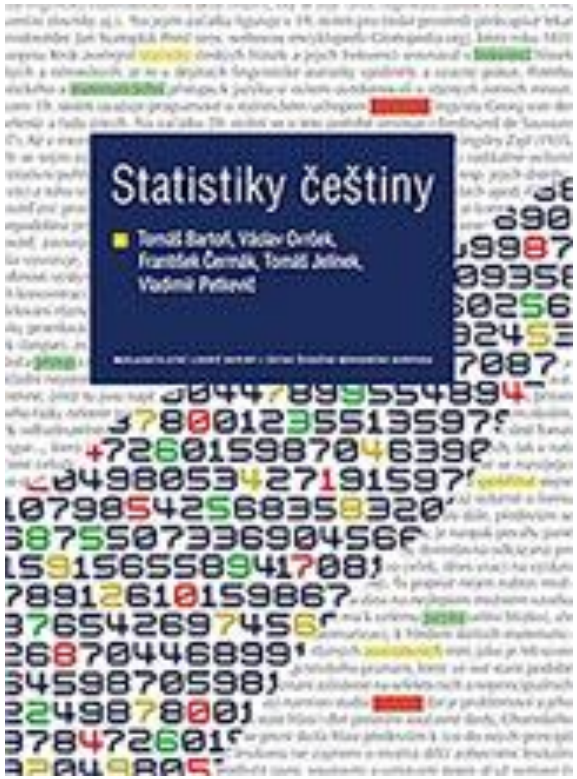
- korpus není encyklopedie
- korpus nejsou pravidla pravopisu
- korpus není výkladový nebo překladový slovník
- korpus neobsahuje všechna česká slova
- jakkoli velký korpus je pouze vzorek jazyka

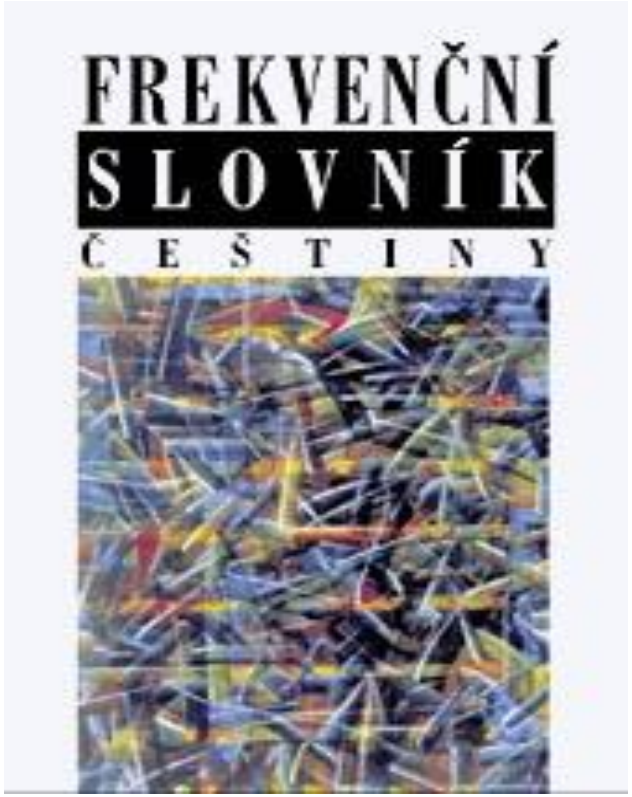
Frekvenční studie

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

Frekvenční studie

- stylistická pozorování – typická slova v určitých typech textů
  - klíčová slova
  - určování sociolingvistických charakteristik
  - projevy emocí
  - určování autorství
- 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
    - atomické značky <orth> <def> <pos>
    - závorkovací značky <entry> <gram> <eg>
- lexikografické stanice – modulární dělení práce

Popis rovin jazyka

- fonetika, fonologie – pokud jsou charakteristiky značkovány (OMK)
- 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 – Machine Learning
  - referenční korpusy
- strojový překlad – Machine Translation
- rozpoznávání a syntéza řeči – Speech Synthesis and Analysis
  - řečové korpusy
- dialogové systémy – Dialogue Systems
- stylometrie, určování autorství – Stylometry, Autorship
- analýza emocí – Sentiment Analysis
- extrakce informací z textu, pojmenované entity – Information Extraction, Named Entity