CJBB105 – 7
Využívání korpusů

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Využívání korpusů

- **stále ve vývoji, možnosti využití se neustále rozšiřují**
  - vyvíjí se metodologie budování korpusů
  - vyvíjí se metodologie vytěžování korpusů (vytěžování = získávání informací o jazyce)
  - vyvíjí se technologie a aplikace spojené s korpusy
  - vznikají nové typy korpusů
  - stále jsou poměrně málo prozkoumané mluvené korpusy
- **i miliardové korpusy stále poskytují pouze vzorek užívání jazyka**
  - jazykové jevy mohou existovat i mimo korpusy
  - korpusy jsou deskriptivní (popisují, jak se jazyk užívá, a to převážně v psané podobě)

Využívání korpusů

Jazyková data jsou potřeba hlavně:

- **v lingvistice**
  - synchronní i diachronní studie
  - frekvenční studie
  - jazyky v kontrastu (paralelní a srovnatelné korpusy)
  - lexikografie
- **v NLP** (Natural Language Processing, počítačové zpracování přirozeného jazyka)
- okrajově mimo tyto dvě oblasti

Využívání korpusů

- analýzy jsou založeny na důsledném **využívání jazykových dat** pro popis jazyka
  - tento přístup je možný až díky počítačům (hardwaru i softwaru)
- v korpusové lingvistice je klíčový **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 způsobily zvrat v lingvistice
  - poskytují velké množství reálných jazykových dat
  - rozvíjí exaktní přístup k jazyku
- velké korpusy jsou pro výzkum dostatečně reprezentativním vzorkem jazyka
  - výskyty jevů a jejich frekvence nejsou náhoda
  - platí i pro miliardové webové korpusy, které nejsou vyvážené
- **kvantitativní analýza**
  - důležitý je počet výskytů (typické a okrajové jevy, variabilita jazyka)
  - nutná lingvistická interpretace výsledků (co zjištěná čísla vypovídají o jazyce)
- **kvalitativní analýza**
  - nezávisí na počtu výskytů (i málo frekventované jevy jsou důležité, např. hapax legomena a výzkum jazykové periferie)

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

Přístupy k využívání korpusů

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

Kvantitativní a kvalitativní analýza

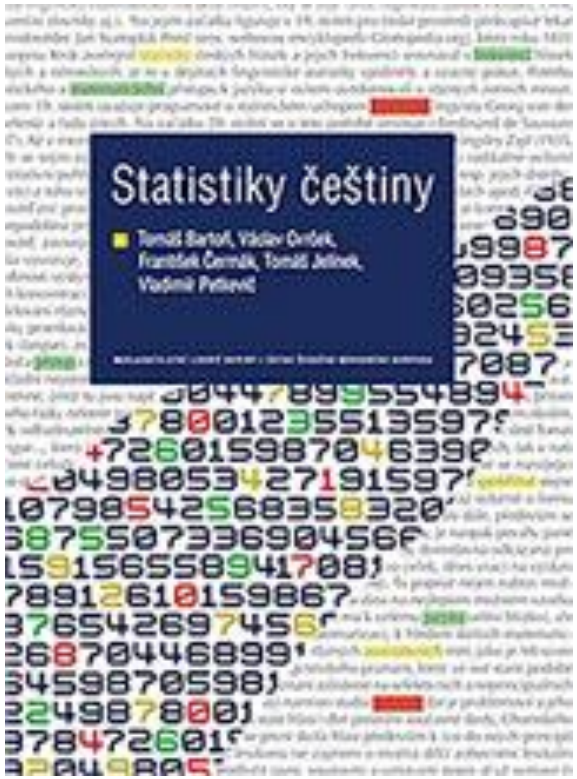
- využívá **frekvenci**, **statistiku**, **pravděpodobnost**
- využívá **řetězec/token** a jeho **kombinace**
  - výzkum kolokací, valence
- **bigramy, trigramy, n-gramy** (shluky slov vyskytujících se vedle sebe v kontextu)
  - výzkum kolokací, termínů apod.
- využívá vložené lingvistické informace – **morfologické značky**
- je vždy důležitá následná **interpretace** výsledků (co zjištěné údaje znamenají)

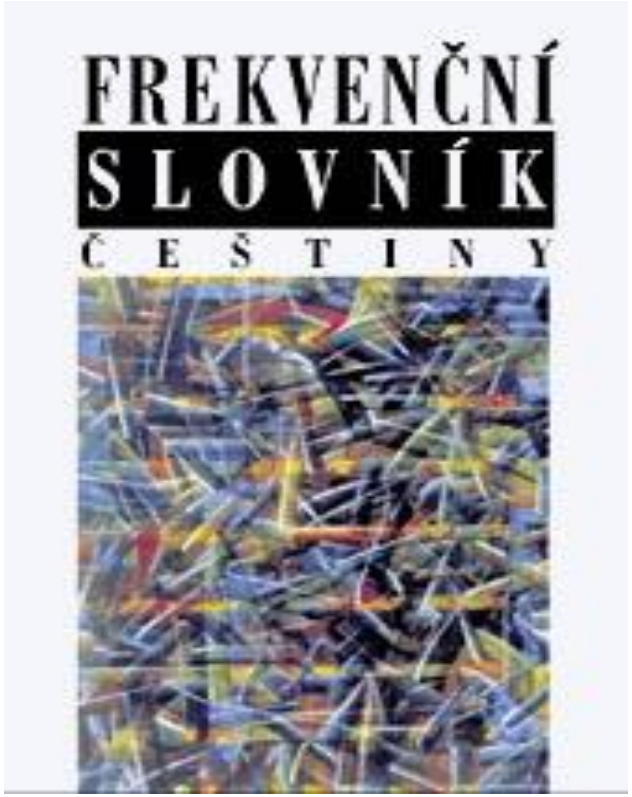
Frekvenční studie

- je možné zkoumat **frekvenci** slov, slovních tvarů, slovních spojení, slovních druhů, slovních segmentů (slabiky, kmeny, sufixy, koncovky), hlásek, znaků (interpunkce)
- vznikají **frekvenční slovníky** (pro češtinu první Těšitelová – 1961); na korpusu založený Frekvenční slovník češtiny – 2004
- výzkum variant (aplikace SyD)
  - např. pravopisné (*filozofie/filosofie*), tvarové (*kopu/kopám*), stylové (*pořád/furt*)
- míra pronikání cizí slovní zásoby, proces počešťování slov, frekvence přejatých slov a jejich různých variant
  - *byznis, byznys, biznis, biznys*

Frekvenční studie

- **stylistická pozorování** – typická slova v určitých typech textů (široké využití)
  - klíčová slova v textu
  - určování sociolingvistických charakteristik (u mluvených korpusů)
  - projevy emocí v jazyce (perspektivní oblast, zajímá komerční subjekty, např. hodnocení zboží v e-shopu)
  - určování autorství a forenzní lingvistika (založeno mj. na stylových rysech typických pro jednoho autora)
- výuka jazyka **pro cizince** (slovníky, slova v kontextech)
- **akvizice** jazyka (korpusy dětského jazyka, výukové korpusy, značkování chyb)
- výzkum **terminologie**
- korpus jako **obraz společnosti** (reálie, společenská situace, která se odráží v jazyce)





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** (lemmat) na základě frekvence v korpusu, hranice min. počtu výskytů (jakou frekvenci musí slovo mít, aby se dostalo do slovníku)
- určování **významů slov** na základě jejich kontextu (např. u homonym, jiný význam – jiný kontext)
- reálné **příklady užití slov** (nemusí se vymýšlet)
  - konkordance (KWIC)
- kolokace, **frazeologismy**, thesaury, Word Sketch
- **metadata**
  - časová datace slovního výskytu, typ textu, autor apod.
- u elektronických slovníků možnost aktualizace dat

Počítačová (korpusová) lexikografie

- **formát slovníku**
  - využívají se značkovací jazyky pro popis struktury slovníkového hesla – vysoká konzistence slovníku (struktura je stejná u všech hesel)
  - starší jazyk SGML (Standard Generalized Markup Language)
  - současný jazyk **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, online zpracování slovníku několika lexikografy, každý má na starost jednu část struktury hesla (definice, gramatika, příklady, frazeologismy atd.), dříve dělení hesel mezi autory podle abecedy

terorismus
[-iz-] (dř. též -ism), -mu m. (z lat.) *způsob vlády vymáhající terorem poslušnost; hrůzovláda, krutovláda, despotismus:* vojenský t.; nesnesitelný t.; demagogie a t.; přen. expr. to je t., nedejte si to líbit

```
<entry>
 <hw>
     <orth>terorismus</orth>
 </hw>
<senses>
  <sense>
    <def>způsob vlády vymáhající terorem poslušnost</def>
    <def>hrůzovláda</def>
    <def>krutovláda</def>
    <def>despotismus</def>
    <eg>vojenský terorismus</eg>
 <eg>nesnesitelný terorismus</eg>
  <eg>demagogie a terorismus</eg>
   <usg type=style>přen.expr.</usg>
        to je terorismus,  nedejte si to líbit
    </eg>
  </sense>
 </senses>
</entry>
```

Ukázka zkráceného zápisu slovníkového hesla v XML

Využití korpusů – popis rovin jazyka

- **fonetika, fonologie** – pokud jsou charakteristiky značkovány (OMK, ORTOFON)
- **morfologie** – tagging, frekvence tagů
- **slovotvorba** – slovotvorné segmenty, derivace, funkční zatížení prefixů/sufixů (Morfio)
- **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, víceslovné jednotky) a jejich značkování
  - *Karel IV., corpus delicti*

Využití korpusů v NLP

- tvorba nových **nástrojů a aplikací** pro výzkum jazyka, minimalizace ručního hledání
- **strojové učení** (systém se učí na vzorovém korpusu)
- **strojový překlad** (založen na strojovém učení)
- **rozpoznávání a syntéza řeči** (přepis řeči na text, převod textu na řeč, strojové učení)
- **dialogové systémy** (komunikace člověka s počítačem, chatboty, asistenti v mobilech, např. Siri)
- určování autorství
- analýza emocí
- **extrakce informací** z textu, pojmenované entity – automatické určování významů částí textu

Další využití korpusů

- toto využití je spíše potenciální, není příliš rozšířené, ale je možné
  - výuka češtiny na ZŠ a SŠ (korpusy SYNEK a LITERA, 2001)
  - výuka češtiny pro cizince (žákovské korpusy)
- literární věda
  - kritici a teoretici
  - autorské korpusy (Karla Čapka, Jana Čepa)
- sociologie (sociolingvistika), psychologie (psycholingvistika) – mluvené korpusy
- neurologie (akvizice jazyka)
- tvůrčí profese
  - spisovatelé, básníci, textaři, žurnalisté, tvůrci reklam
  - mohou se v korpusech inspirovat