

Korpusová lingvistika – 4

Budování korpusů

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Budování korpusů

- specifikace cíle a účelu korpusu
- specifikace cílové skupiny uživatelů
- specifikace typu korpusu
- výběr a sběr jazykového materiálu
- autorská práva
- zpracování textů – vertikál, kódování
- vnitřní a vnější značkování
- hardwarové a softwarové vybavení
- personální zajištění a finanční podpora

Budování korpusů podle jejich typu

- tradiční synchronní psané korpusy
- webové synchronní psané korpusy
- specializované korpusy
- mluvené korpusy

Budování tradičních korpusů

- Český národní korpus (korpus.cz)
- výběr textů – vyváženost a reprezentativnost korpusu
- dohody s poskytovateli textů (autorská práva)
- texty v elektronické podobě
 - odstranění netextového obsahu (obrázky, grafy, tabulky)
 - sjednocení kódování
 - záznam metatextových informací (autor, dílo, rok vydání)
 - vertikál – tokenizace
 - atributy – **poziční** (*word, lemma, tag*) a **strukturní** (*opus, doc, s*)
 - lemmatizace a morfologické značkování

Budování webových korpusů

- Centrum zpracování přirozeného jazyka FI MU, czTenTen12
- **Sketch Engine** (ske.fi.muni.cz)
- autorská práva – veřejně dostupné texty na Internetu
- web crawler **SpiderLing** – stahování textů
- **jusText** – odstranění boilerplate (netextového obsahu) z webové stránky
 - vybírá text obsahující celé věty
- **onion** (ONE Instance ONLY) – odstranění duplikátů
- **chared** (character encoding) – sjednocení kódování, pro řadu jazyků
- **Corpus Architect** – tvorba korpusů
 - nahrávání textů v elektronické podobě uživatelem
- **WebBootCaT** – texty z webu
 - seed words (klíčová slova)
 - URLs (adresy webových stránek)

Budování mluvených korpusů – nahrávka

- specifikace typu zaznamenané promluvy
 - monolog – dialog
 - formální – neformální (poloformální)
 - připravená – nepřipravená
- specifikace délky nahrávky
- specifikace mluvčích a sociolingvistických kategorií (pohlaví, věk, vzdělání, teritoriální zařazení), vyváženost korpusu
- autorská práva (prohlášení nahrávajícího) a anonymizace citlivých údajů
- pořízení kvalitní digitální nahrávky (diktafony), příp. úprava nahrávky (oříznutí)

Budování mluvených korpusů – přepis

- korpusy řady ORAL – spontánní dialogy
- přepis nahrávek podle stanovených pravidel
- nástroj ELAN, dříve Transcriber
- **segmentace** přepisu a synchronizace segmentů
- **ortografický přepis**, fonetický přepis
- **pauzová interpunkce**
- hezitační a jiné zvuky, přeřeknutí, smích, citoslovce, nesrozumitelné úseky, neverbální zvuky
- **simultánní úseky**

Pražský mluvený korpus

- první korpus mluvené češtiny, **675** tis. slov
- autentická mluvená čeština, tematicky nesespecializovaná
- z městského prostředí Prahy a jejího okolí
- neformální dialogy, poloformální řízený rozhovor (dotazník)
- magnetofonové nahrávky (**304**), přepis do MS Word
- z let **1988–1996**, odrážejí jazyk jak konce předchozího společenského období, tak začátek nového
- pravidla přepisu ortografická, pro obecnou češtinu
- větná interpunkce

Brněnský mluvený korpus

- první korpus mluvené češtiny z oblasti Moravy, **490** tis. slov
- běžně mluvený jazyk z městského prostředí Brna
- **250** anonymních magnetofonových nahrávek z let **1994–1999**, **294** mluvčích
- prolíná se středomoravský interdialekt s obecnou češtinou
- v oblasti slovní zásoby zbytky někdejšího soužití brněnské češtiny s německým jazykem a vliv brněnského slangu (hantecu)
- neformální a poloformální dialogy
- v pravidlech přepisu zohledněna specifika brněnské mluvy
- pauzová interpunkce

Korpusy řady ORAL

- **ORAL2006** – mluvená čeština z celé oblasti českých nářečí
- 221 nahrávek z let 2002–2006
- pouze neformální dialogy, přátelský vztah mezi mluvčími
- 111,5 hodin, 1 000 798 slov od 754 mluvčích
- **ORAL2008** – plně vyvážený v základních sociolingvistických kategoriích (pohlaví, věk, vzdělání, oblast pobytu v dětství)
- 297 nahrávek z let 2002–2007
- výhradně neformální situace
- 115 hodin, 1 000 097 slov od 995 mluvčích

Korpusy řady ORAL

- **ORAL2013** – nahrávky pořízeny v Čechách, na Moravě i ve Slezsku
- 835 nahrávek z let 2008–2011
- 2 785 189 textových slov, tj. celkem 3 285 508 pozic
- 2 544 mluvčích, z toho 1 297 unikátních
- délka téměř 300 hodin