# KORPUSOVÁ LINGVISTIKA

Dana Hlaváčková

# JAZYKOVÝ KORPUS

Rozsáhlý soubor elektronicky uložených jazykových dat, obvykle označkovaný, organizovaný se zřetelem k využití pro určitý cíl, vůči němuž je také považován za reprezentativní.

Čermák, F. Jazykový korpus: Prostředek a zdroj poznání.

In *Studie z korpusové lingvistiky.* Praha: Karolinum, 2000, s. 15-38.

# KORPUSOVÁ LINGVISTIKA

- podstatná část počítačové lingvistiky – korpusy poskytují zdroj jazykových dat
- studium jazyka založené na jeho přirozeném kontextovém užívání
- metodologický přístup ke zkoumání jazyka

# PŘEDNOSTI KORPUSŮ

- velký rozsah s možností dalšího rozšiřování
- jazyková data v přirozené kontextové podobě
- převaha typických jazykových jevů nad okrajovými
- reprezentativní korpus je schopen zachytit variabilitu jazyka
- zrychlení a usnadnění lingvistické práce

# ZÁKLADNÍ POJMY

- **textové slovo, pozice, token** – řetězec znaků oddělený z obou stran mezerami
- **tokenizace** – proces rozdělení textu na tokeny
- **korpusový prohlížeč**, **korpusový manažer** (Bonito, Bonito2, Sketch Engine, KonText)
- **konkordance**, konkordanční řádek, konkordanční seznam
- **KWIC** – key word in context (hledaný výraz v korpusu)

# ZÁKLADNÍ POJMY

- **atributy** – prvky, které lze hledat v korpusu (word, lemma, tag, lc, pos)
- **lemma** – základní slovní tvar
- **lemmatizace** – přiřazení základního slovního tvaru
- **strukturní značky** – hranice dokumentů a vět
- **regulární výrazy** – speciální znaky umožňující efektivní hledání v korpusu

# TYPY KORPUSŮ

- druh zachycené komunikace – psané (written corpora)
  – mluvené (spoken corpora)
- časový záběr – diachronní
  – synchronní
- účel – všeobecné
  – specializované
- jazyk – jednojazyčné
  – paralelní
- možnost rozšíření – uzavřené (referenční)
  – otevřené (nereferenční)
- značkování – tagging (POS tagging, morfologie)
  – parsing (syntax, treebank)
  – alignment (párování)

# REPREZENTATIVNOST KORPUSŮ

Relativní

- v závislosti na účelu korpusu (kvantita x kvalita)
- malý vzorek vzhledem k celku jazyka
- nezobrazuje reálné užití jazyka
- snaha zachytit variabilitu textů

**SYN2000**

denní tisk / 60 %

naučná literatura / 25 %

krásná literatura / 15 %

**SYN2005, SYN2010**

publicistika / 33 %

odborná literatura / 27 %

beletrie / 40 %

# TVORBA KORPUSŮ

- sběr dat – sjednocení formátu – externí anotace
- tokenizace (vertikál) – lemmatizace – značkování
- Corpus Architect, WebBootCat
- jusText – odstranění netextového obsahu, boilerplate
- Onion – odstranění duplicitních textů
- Chared – detekce kódování

- mluvené korpusy – nahrávky, přepis, synchronizace textu se zvukem