

# KORPUSOVÁ LINGVISTIKA

Dana Hlaváčková

# JAZYKOVÝ KORPUS

Rozsáhlý soubor elektronicky uložených jazykových dat, obvykle označovaný, 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

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

# DVA PŘÍSTUPY KE ZKOUMÁNÍ JAZYKA

- **raná „korpusová“ lingvistika** – „korpusový“ přístup k jazykovému materiálu, dostatečně velký soubor přirozeně se vyskytujících jazykových dat (konec 19. st. – 50. léta 20. st.)
- **předěl (50. léta 20. st.)** – N. Chomsky a generativní lingvistika
- od 2. pol. 20. st.  
**empirický přístup, observace x intuice a introspekce**

*Ch. Fillmore:*

*„I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore ... [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way. My conclusion is that the two types of linguists need one another.“*

# KORPUSOVÁ LINGVISTIKA V ČR

- **Lexikální archiv ÚJČ**, od r. 1911, 12 mil. ručně psaných lístků
- 1988 **Iniciativní skupina pro přípravu počítačových korpusů, textů a slovníků**  
(sdružení lingvistů, matematiků a programátorů)
- 1991 **Počítačový fond češtiny** – projekt lexikografického počítačového korpusu a tezauru češtiny (Čermák, Sgall, Pala, Hajič, Hajičová, Králík, Schmiedtová, Kučera, Benko)
- 1994 založení **ÚČNK**

# 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