Korpusová lingvistika

PLIN059

Mgr. Dana Hlaváčková, Ph.D.

Korpusová lingvistika

- využívá pro studium jazyka velké soubory elektronických textů
- texty odrážejí a dokládají reálné užívání jazyka
- korpusy jsou deskriptivní (vs. preskriptivní)
- **korpusové manažery** umožňují data prohlížet a třídit a poskytují statistické údaje
- podstatná část počítačové lingvistiky – korpusy poskytují **zdroj jazykových dat**
- studium jazyka založené na jeho **přirozeném kontextovém užívání**
- **metodologický přístup** ke zkoumání jazyka

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.

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
- **morfologické** a **syntaktické** značkování korpusů zvyšuje jejich informační hodnotu

Základní pojmy

- **token, pozice** – 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)
- **atributy** – prvky, které lze hledat v korpusu
- **strukturní značky** – např. hranice dokumentů a vět
- **vertikál** – textový soubor (.vert), ve kterém je text rozdělen na tokeny

<s>
Náměstí
republiky
je
přímo
jejich
skanzenem
<g/>
.
</s>
<s>
Průčelí
je
tvořeno
divadlem
Antonína
Balšánka
<g/>
,
vystavěno
bylo
v
letech
1906
až
1909
<g/>
.
</s>

Typy korpusů

- druh zachycené komunikace
  - psané (written corpora)
  - mluvené (spoken corpora)
- časový záběr
  - diachronní
  - synchronní
- účel
  - všeobecné
  - specializované
- **způsob vytvoření**
  - tradiční
  - webové
- jazyk
  - jednojazyčné
  - paralelní
  - srovnatelné
- 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 užití jazyka v celé šíři
- snaha zachytit variabilitu textů (beletrie, odborné, publicistika)

	SYN2000	SYN2005, SYN2010	SYN2015
publicistika	60 %	33 %	33,33 %
odborná lit.	25 %	27 %	33,33 %
beletrie	15 %	40 %	33,33 %

Tvorba korpusů

- korpusy tradiční a webové (Corpus Architect, WebBootCat)
- sběr dat – sjednocení formátu – externí anotace
- tokenizace (vertikál) – lemmatizace – značkování
- stahování textů (crawler) – webové korpusy
- odstranění netextového obsahu, boilerplate
- odstranění duplicitních textů
- detekce kódování
- mluvené korpusy – nahrávky, přepis, synchronizace textu se zvukem

Korpusové manažery v ČR

- ÚČNK – ČNK – KonText
  - http://kontext.korpus.cz
- FI MU – Sketch Engine
  - https://www.sketchengine.eu/
- Český národní korpus
  - https://www.korpus.cz/