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

1. podstatná část počítačové lingvistiky – korpusy poskytují **zdroj jazykových dat**
2. studium jazyka založené na jeho **přirozeném kontextovém užívání**
3. **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

<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>

Základní pojmy

- **token, pozice** – řetězec znaků oddělený z obou stran mezerami
- **tokenizace** – proces rozdělení textu na tokeny
- **vertikál** – textový soubor (.vert), ve kterém je text rozdělen na tokeny
- **strukturní atributy** – informace, které se vztahují k textu (hranice dokumentů, hranice vět, …)
  - korpusová metadata (žánrové dělení textů, autor, název díla, …)
- **poziční atributy** – informace, které se vztahují k jednomu tokenu (word, lemma, tag, …)
- **korpusový prohlížeč**, **korpusový manažer** (Bonito, Bonito2, **Sketch Engine**, **KonText**)

čisté jako z prádelny . Když se oblékáš , strakatá	kočka	vyskočí na umyvadlo a tře se ti hlavou o bok
vrhla na trávník a protahovala si ruce i nohy jako	kočka	. Ležet bez hnutí na lehátku , daleko od svěží
, Regina , Leopard , Jupiter , SmČlandský lev ,	Kočka	, Tygr , Měsíc , Koruna , Klíč , Stockholm
podařilo odplout . Kapitán Tönnes Speck na lodi Kattan (	Kočka	) se o to chtěl pokusit . Se svými dvaadvaceti
závod vyhraje . V červnu 1649 opustila loď Kattan (	Kočka	) Švédsko a zamířila do Severní Ameriky . Loď byla
holubího . Malý mužík pozoroval celých pět minut kabelku jako	kočka	pohybující se klubíčko a pouhá představa jejího obsahu ho hypnotizovala
mnou zacházeli , jako kdybych byl něco , co přitáhla	kočka	z ulice , a jediná osoba , která se ke
mně chovala trochu slušně , byla upírka . " "	Kočka	ne , " řekla lady Sibyla . " Cože ?
ale v pádu přešel do salta nazad . Dopadl jako	kočka	, vrhl se zpět k užaslému Karotkovi a udeřil ho
" Černocha . " " Nabral jsem ho hned před	Kočka	Barem , na Padesátý čtvrtý a Broadwayi , a on
píchaly ho do vědomí . Talibe mhouřil oči jak divoká	kočka	, ale i tak mu bělostné biče světla pronikaly skrz
Jindřišky . Za košili . Přesně tím pohybem , jakým	kočka	zdvíhá koťata za kůži na hřbetě . " Co si
Jsi si jistý , že žádnému z mých sousedů neschází	kočka	nebo pes ? " vyptával se dál . " Znám
vrčení gazíku . Stáli tak , dokud se odněkud neobjevila	kočka	, sedla si mezi ně a začala si olizovat tlapky
dva fosforeskující světelné body , jako by se tam skrývala	kočka	. „ Běž ven a pošli ji do prdele !
vítr hnal před sebou kupu suchého listí , vyděšená černá	kočka	se hbitě protáhla plotem u ředitelova domu . Odstrčil knihu
okna . Futaki je pořád uvnitř . K řediteli šla	kočka	, ještě sem ji tu neviděl , co tu taková
, ještě sem ji tu neviděl , co tu taková	kočka	k čertu hledá ? ! Zřejmě se něčeho lekla ,
. „ Micur ! “ Na kuchyňském stole seděla černá	kočka	a z červeného hrnce vesele chlemtala zbytek paprikáše od oběda
Sem silnější ! “ – blesklo jí hlavou	Kočka	k ní přiběhla a třela se jí o nohy

konkordance, konkordanční řádek, konkordanční seznam

KWIC – key word in context (hledaný výraz v korpusu)

Typy korpusů

- druh zachycené komunikace
  - psané (written corpora)
  - mluvené (spoken corpora)
  - multimodální
- 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

- v závislosti na účelu korpusu (kvantita a kvalita)
- národní korpusy – obraz užívání jazyka
- malý vzorek vzhledem k celku jazyka, nezobrazuje užití jazyka v celé šíři
- snaha zachytit variabilitu textů (beletrie, odborná lit., publicistika)

	SYN2000	SYN2005, SYN2010	SYN2015, SYN2020
publicistika	60 %	33 %	33,33 %
odborná lit.	25 %	27 %	33,33 %
beletrie	15 %	40 %	33,33 %

Tvorba korpusů

- korpusy tradiční a webové
- sběr dat
  - poskytovatelé textů
  - webové korpusy – stahování textů (crawler)
- sjednocení formátu a kódování
- odstranění netextového obsahu (boilerplate)
- odstranění duplicitních textů (webové korpusy)
- interní anotace
- tokenizace (vertikál) – lemmatizace – externí anotace (značkování)
- 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/