KORPUSOVÁ LINGVISTIKA

Mgr. Adriana Válková
415719@mail.muni.cz

Jazykový korpus

- elektronická **databáze** autentických **textů** mluveného nebo psaného jazyka
- „databáze“ – umožňuje vyhledávání jazykových jevů (slov, slovních spojení) v jejich přirozeném kontextu

vylil si z trombónu sliny . Podíval jsem se do	**korpusu**	svého saxofonu a uviděl jsem malou hladinku , jak se
. Byl jsem vždycky rád , když jsem měl v	**korpusu**	hodně mokro . Vyléval jsem to s rozkoší tím větší
ho naklonil na tu stranu , kde nejsou dole na	**korpusu**	klapky , a vylil jsem obsah na podlahu . Hodil
a vylil jsem obsah na podlahu . Hodil jsem do	**korpusu**	závaží čistítka a elegantně jsem převrátil tenora v ruce .
napřažena rovně nad tancující a slunce se třpytilo v jejím	**korpusu**	, nad vlasatýma hlavama se vytahoval a zatahoval Vencův trombón
němu a zpíval jsem a mluvil jsem z jeho pozlaceného	**korpusu**	, že ho přijímám , a že přijímám všechno ,
a tak automatický jako nastupování do paternosteru . Přetlakový kulový	**korpus**	tvořil hlavu celé té křehké , šípovité konstrukce dlouhé přes
těla (ale co říkám ? veškerý pozemský a sublunární	**korpus**	vyhladovělých a žíznivých stolovníků) se proměnila v jedno jediné
boogie Emila Zettnera u piana , pohlédl smutně na lesklý	**korpus**	svého tenora a Zetka se najednou otočil , řekl ,
baletky , ale oči mu neodolatelně přitáhl orchestr . Zlaté	**korpusy**	saxofonové baterie malovaly tam křehký rytmus zeswingované gavoty do půvabných
sexuálním hlasem nástroje a snil o kráse , kterou zlacený	**korpus**	přidá jeho postavě , vztyčené v přitlumeném světle v jakémsi
, Lydie tam sice nechyběla , ale ani hlas zlaceného	**korpusu**	, ani postava čtvrtého saxofonisty ji nelákaly . Neboť hlas

Korpusová lingvistika

- relativně mladý obor (od 2. pol. 20. st.)
- odvětví lingvistiky, která studuje jazyk v jeho přirozeném kontextovém užití
- úzká propojenost s počítačovou (komputační) lingvistikou
- celosvětově převažující metodologie zkoumání jazyka
  - *empiricky podložená data vs. lingvistická intuice a introspekce*

Přednosti korpusů

- jazyková data v přirozeném kontextu
- statistické (frekvenční) zpracování jazykového materiálu
  - *zrychlení a usnadnění lingvistické práce*
  - *na základě toho např. určení typických (centrálních) nebo okrajových jazykových jevů)*
- upřesňuje, opravuje (popř. ruší) některé tvrzení z gramatik
- korpusy odrážejí skutečně užívaný jazyk (jazykový úzus) a jeho variabilitu

Základní terminologie

- **token** (tokenizace) – grafická jednotka oddělené mezerou (nejčastěji slovo)

na opozdilé polské kavaleristy nebo se potuloval po lesích a **vyhledával** ukrytá stáda selských koní . Zapaloval vesnice a odstřeloval polské

- **pozice** – tj. token v korpusu

konkordance	místnosti	.	Byly	z	těžkého	tmavého	**dřeva**	a	zlověstně
pozice	6L	5L	4L	3L	2L	1L	KWIC	1R	2R

- **KWIC** (červeně/růžově vyznačené slovo) – tj. slovo, které hledám
- **konkordance** (konkordanční řádek), ve které se KWIC vyskytuje

Struktura anotovaného korpusu

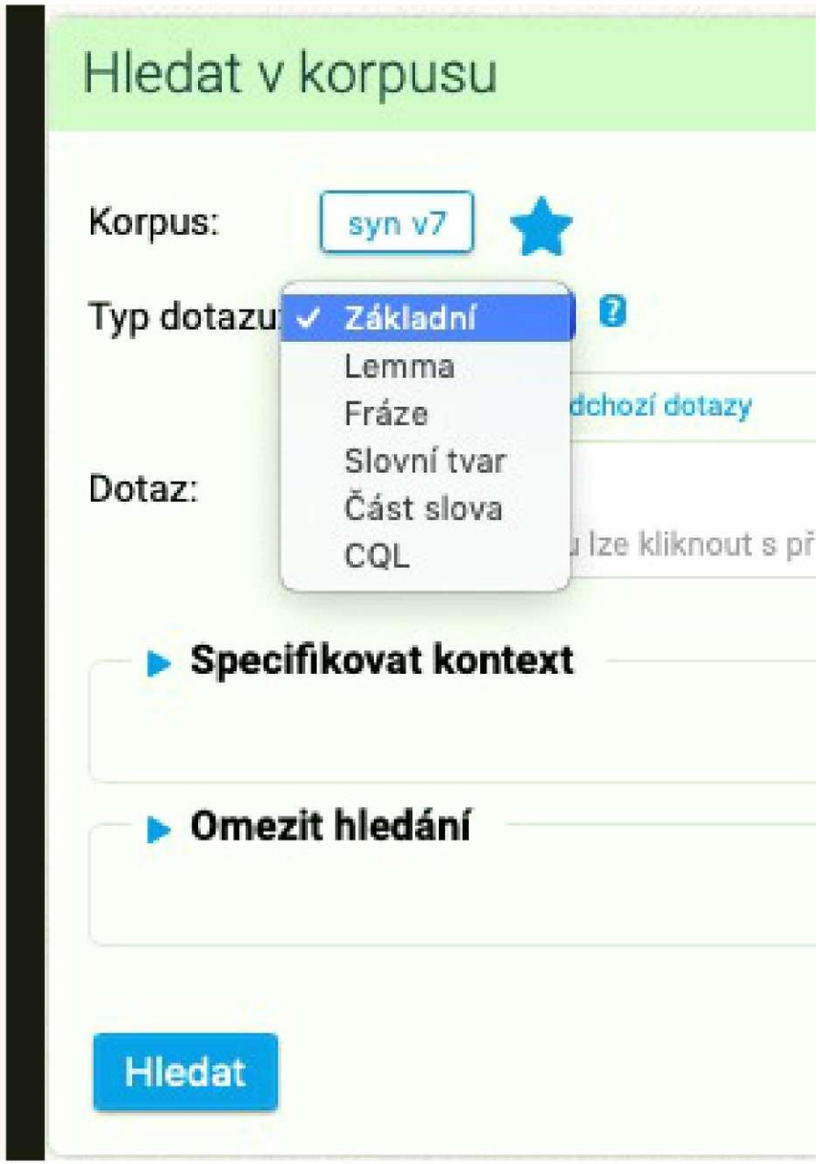
- **strukturní atributy**
  - *struktura dokumentu (hranice vět, odstavců atd.) a informace o nich (autor textu, rok vydání, žánr atd.)*
- **poziční atributy**
  - *(morfologické) informace k jednotlivým pozicím*
  - *morfologická značka = tag*
  - *tagset = souhrn morfologických značek*
  - *lemma – slovníkový tvar hesla*
  - *lemmatizace – přiřazení lemmatu*

Vlastnosti korpusu

- **anotovaný/označkovaný** (tokenizace a lemmatizace)
- **reprezentativní** (obsahuje všechny variety jazyka)
- **vyvážený** (odpovídá-li poměr jazykových variet jejich poměru v jazykovém úzu)

Co a jak vyhledávat?

- **poziční atributy**
  - *word (základní)*
  - *lemma*
- dotazovací jazyk **CQL** (Corpus Query Language)
  - *práce s regulárními výrazy*
  - *složitější typy dotazů*
  - *poziční atribut tag*
- korpusový manažer



Mgr. Adriana Válková, 415719@mail.muni.cz

Typy korpusů

- **korpus je vždy budován za určitým cílem**
  1. *jednojazyčný nebo vícejazyčný (InterCorp)*
  2. *obecný nebo specializovaný (např. korespondence K. H. Borovského)*
  3. *psaný mluvený (ORAL)*
  4. *synchronní (SYN) nebo diachronní (Diakorp – texty ze 14.–20. st.)*
  5. *referenční nebo nereferenční*
  6. *označkovaný (typ značek?) nebo neoznačkovaný*

Tvorba korpusů

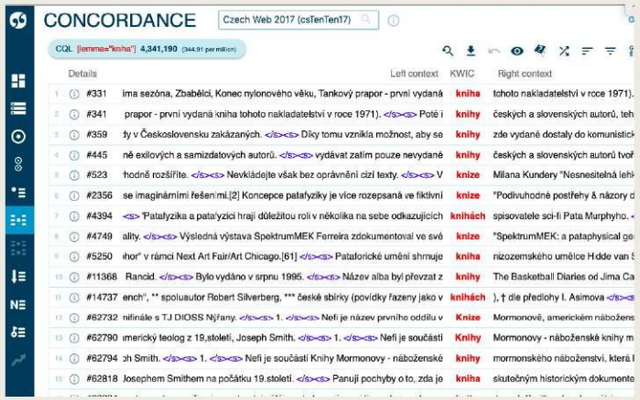
korpusy tradiční a webové

- 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

Kde vyhledávat?

- na webu pomocí **korpusových manažerů**
  - *Praha ÚČNK – KonText* (http://kontext.korpus.cz)
  - *Brno FI MU – SketchEngine* (*https://app.sketchengine.eu)*



úkol: v korpusech hledám…

- frekvenci slov
- kontext určitých slov
- informace o typech textu (v nichž se slova vyskytují)
- rozsah užití přejatých slov, jejich pravopisná podoba
- jazykové varianty
  - pravopisné (*realismus/realizmus*)
  - morfologické ( )
  - lexikální (alespoň/aspoň).