



KORPUSOVÁ LINGVISTIKA

Mgr. Adriana Válková
415719@mail.muni.cz



Jazykový korpus

- elektronická **data**báze 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 mokró . 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 jakémisi
, 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	<u>KWIC</u>	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/označkový** (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

Hledat v korpusu

Korpus:

syn v7



Typ dotazu

✓ Základní



Lemma

Fráze

Slovní tvar

Část slova

CQL

[dchozí dotazy](#)

Dotaz:

[u lze kliknout s př](#)

► **Specifikovat kontext**

► **Omezit hledání**

Hledat

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. *označovaný (typ značek?) nebo neoznačovaný*

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>)

The screenshot shows the CONCORDANCE search interface. At the top, it displays 'Czech Web 2017 (csTen7)' and a search bar with the query 'lemma="knihy"'. Below the search bar, there are several search results listed in a table. Each row contains a result ID, a snippet of text with highlighted search terms, and the source of the text. The search terms 'knihy' and 'knih' are highlighted in red in the original image. The results include various sources such as 'Tankový prapor', 'Československu zakázaných', 'Patafyzika a patafyzičtí hrají', 'Světová výstava SpektrumMEK', 'Next Art Fair/Art Chicago', 'Rancid', 'Robert Silverberg', 'T.J. DiOSS Njřany', 'Joseph Smith', and 'Joseph Smithem'.

The screenshot shows the KonText search interface. At the top, it displays 'Dotaz Korpusy Uložit Konkordance Filtr Frekvence Kolokace Zobrazení Nápvěda'. Below the search bar, there are several search results listed in a table. Each row contains a result ID, a snippet of text with highlighted search terms, and the source of the text. The search terms 'knihy' and 'knih' are highlighted in red in the original image. The results include various sources such as 'lidského srdce', 'A choval jsem vzrušený úmysl přelstít si ještě mnoho jiných', 'lekl a Daley s výrazem bezmyšlenkovitého smutku', 'bylo za slovo, co jsme - "Ty', 'ženskou", než to, že ho skličila nějaká', '"Povídá se po městě", společně s výtiskem', 'velkého stolu a ležel ve vzrušeném soustředění na police s', 'Já jsem se přesvědčil, jsou pravé. ...', 'Ale co chcete? Co čekáte? Vytáhni mi', 'tady špatně hodinu. Řekni jsem vám to a těch', 'zastříhli nápravní tašku a vytáhli z kapsy rozřihaný starý výtisk', 'a nevyšel. Koupat se obden. Přičíst jednu vzdušnici', 'Během měsíce našel v Gatsbuiho knihovně, jak obdivoval jeho

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).