

Jazykové korpusy



Počítačové nástroje pro češtinu
Jaro 2020

Adriana Válková
valkova@phil.muni.cz

Co je jazykový korpus?



A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.



Sinclair, J. 2004.

Vznik korpusů (korpusové lingvistiky)

- 2. pol. 20. st., první jazykový korpus – Brown corpus
- 1994 Český národní korpus
- zkoumání jazykového materiálu v jeho přirozeném kontextu
- empiricky podložená data vs. lingvistova introspekce (intuice)

The screenshot shows a web interface for searching corpora. At the top, there are navigation tabs: "Recent", "My own", "Shared with me", "Featured", "Parallel", and "All". A search bar contains the text "bro" and a "Filter by language: all" dropdown. Below the search bar is a table with columns "Name" and "Words". The table lists the "Brown" corpus with 1,007,299 words. A "Search corpus" button is visible. A tooltip or information box provides details about the Brown Corpus: "BROWN CORPUS A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. by W. N. Francis and H. Kucera (1964) Department of Linguistics, Brown University Providence, Rhode Island, USA Revised 1971, Revised and Amplified 1979 http://www.hit.uib.no/icame/brown/bcm.html Distributed with the permission of the copyright holder, redistribution permitted."

Korpusové texty jsou...

- ... autentické
- ... neúplné (autorská práva)
- ... anotované (označkové)

Lidové noviny - Literární noviny	i ilustrace : 1300 citátů se objevuje průběžně ve vlastním	korpusu	hesla a pomáhá tak nejen pochopit kontext , ale též
Mladá fronta DNES	špatná jakost cukrářských výrobků je patrná již pouhými smysly .	Korpusy	zákusků bývají připálené , žlutkové krémy sražené . V
Respekt	tenkrát zcela hypotetické možnosti budoucího sjednocení předpokládal . Například právní	korpus	, jímž se řídí zásady státního života , se nejmenuje
Lidové noviny	za sto let pomůže badatelům enviromentalistům či jiným vědátorům jako	korpus	delikti proti ničitelům životního prostředí . A možná i v
Lidové noviny	češtiny i řady slovníků překladových , terminologických atd . Podobné	korpusy	dnes už existují nejen pro angličtinu , němčinu ap. ,

⏪ 41 / 355 ⏩

◀ se na koncepčního (a už usedlého) sběratele sluší . A řadu těch flaštiček mohu předat jako historickou aqua-štafetu někomu , kdo v tom bude pokračovat . Třeba to za sto let pomůže badatelům enviromentalistům či jiným vědátorům jako **korpus** delikti proti ničitelům životního prostředí . A možná i vznikne nová sběratelská vášeň , jistě levnější než vrcholová filatelistika . Sběratelé vzorků řek , potůčků , jezírek a moří mohou najít zálibení v zatím nezvyklém koníčku . Vznikne fiumenistika . ▶

Corpus, © Lexical Computing Ltd., © NLP Centre FI MU
verze 0.13.3, uses manatee-2.36.7-open-2.158.8
[Ohlásit chybu](#)

Základní terminologie

- v korpusech se slova (vč. interpunkce) označují jako **pozice** (tokeny)
- hledané slovo (zvané **KWIC**) se zobrazuje v tzv. **konkordančním řádku**

1	2	3	4	5	6
Hádej	,	kdo	dnes	přijde	?

6 pozic (tokenů)

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

L - levá pozice, R – pravá pozice

Budování psaných korpusů (tradičních a webových)

Budování tradičních korpusů (např. synchronní ČNK)



Budování webových korpusů (SketchEngine)

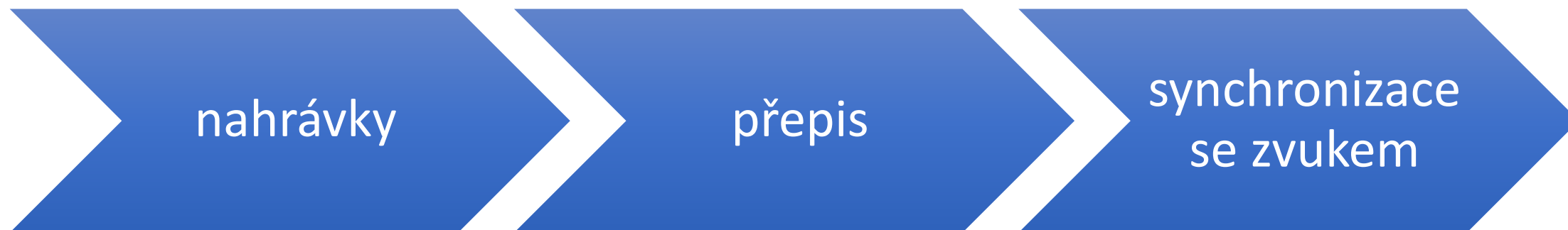
Corpus Architect, WebBootCat – generování textového korpusu z vybraných webových stránek

jusText – odstranění netextového obsahu (obrázky, grafy, tabulky)

Onion – odstranění duplicitních textů

Chared – sjednocení kódování

Budování mluvených korpusů



Vlastnosti korpusů

- anotovaný/označovaný
- reprezentativní
- vyvážený



Proč a kde využívat korpusy?

- zpracování rozsáhlého množství dat za zlomek času
 - typické jevy ve slovní zásobě vs. jevy okrajové
 - ...
- lingvistický výzkum (synchronní/diachronní)
 - výuka cizích jazyků
 - translatologie
 - ...

Typy korpusů

korpus je vždy budován za (s) určitým cílem

1. jednojazyčný – vícejazyčný (InterCorp)
2. obecný – specializovaný (např. korespondence K. H. Borovského)
3. psaný – mluvený (ORAL)
4. synchronní (SYN) – diachronní (Diakorp – texty ze 14.–20. st.)
5. referenční – nereferenční
6. označovaný (typ značek?) – neoznačovaný

Práce s korpusy...

...na webu skrze korpusové manažery

- Praha ÚČNK – KonText (<http://kontext.korpus.cz>)
- Brno FI MU – SketchEngine (<https://app.sketchengine.eu>)

Details	Left context	KWIC	Right context
1	i #331	ima sezóna, Zbabělci, Konec nylonového věku, Tankový prapor - první vydaná	knihá tohoto nakladatelství v roce 1971).
2	i #341	prapor - první vydaná kniha tohoto nakladatelství v roce 1971). </s><s> Poté i	knihy českých a slovenských autorů, teh
3	i #359	ly v Československu zakázaných. </s><s> Díky tomu vznikla možnost, aby se	knihy zde vydané dostaly do komunistick
	i #445	ně exilových a samizdatových autorů. </s><s> vydávat zatím pouze nevydané	knihy českých a slovenských autorů tvoř
	i #523	hodně rozšíříte. </s><s> Nevkládejte však bez oprávnění cizí texty. </s><s> V	knize Milana Kundery "Nesnesitelná lehk
	#2356	se imaginárními řešeními.[2] Koncepce patafyziky je více rozepsaná ve fiktivní	knize "Podivuhodné postřehy & názory d
	#4394	<s> 'Patafyzika a patafyzici hrají důležitou roli v několika na sebe odkazujících	knihách spisovatele sci-fi Pata Murphyho. <
	4749	ality. </s><s> Výsledná výstava SpektrumMEK Ferreira zdokumentoval ve své	knize "SpektrumMEK: a pataphysical ge
	250	ihor" v rámci Next Art Fair/Art Chicago.[61] </s><s> Pataforické umění shrnuje	knihá nizozemského umělce Hidde van S
	368	Rancid. </s><s> Bylo vydáno v srpnu 1995. </s><s> Název alba byl převzat z	knihy The Basketball Diaries od Jima Ca
	37	ench", ** spoluautor Robert Silverberg, *** české sbírky (povídky řazeny jako v	knihách), † dle předlohy I. Asimova </s><s>
	32	nifinále s TJ DIOSS Nýřany. </s><s> 1. </s><s> Nefi je název prvního oddílu v	Knize Mormonově, americkém nábožens
	90	imerický teolog z 19.století, Joseph Smith. </s><s> 1. </s><s> Nefi je součástí	Knihy Mormonovy - náboženské knihy m
	4	h Smith </s><s> 1. </s><s> Nefi je součástí Knihy Mormonovy - nábožensk	knihu mormonského náboženství, která
		niha (1 199 958 výskytů)	
		p.m.: 235,27 (vztaženo k celému korpusu) ARF: 412 526,9 Výsledek je setříděn	
	dní		
	sbysy	lidského srdce . Jenom na člověka , který dává této	knize jméno ,
	atsby	A choval jsem vznešený úmysl přečíst si ještě mnoho jiných	knih . Na un
	Gatsby	řekla Daisy s výrazem bezmyšlenkovitého smutku . Čte hlubokomyslné	knihy , v kter
	y Gatsby	bylo za slovo , co jsme - " " Ty	knihy jsou př
	lký Gatsby	ženskou " , než to , že ho sklíčila nějaká	knihá . Cosi h
	Velký Gatsby	" Povídá se po městě " , společně s výtiskem	knihy " Šimon
	Velký Gatsby	velkého stolu a hleděl ve vrávoravém soustředění na police s	knihami . Když j
	Velký Gatsby	Já jsem se přesvědčil . Jsou pravé . " "	Knihy ? " Přík
	Velký Gatsby	Ale co chcete ? Co čekáte ? " Vytrhl mi	knihu z ruky ,
	Velký Gatsby	tady teprve hodinu . Řekl jsem vám to o těch	knihách ? Jsou
	Velký Gatsby	zastrčil náprsní tašku a vytáhl z kapsy rozrhaný starý výtisk	knihy nazvan
	Velký Gatsby	a nežvýkat . Koupat se obden . Přečíst jednu vzdělávací	knihu nebo č

Hledat v korpusu

Korpus:

syn v7



Typ dotazu

✓ Základní



Lemma

Fráze

Slovní tvar

Část slova

CQL

Dotaz:

dchozí dotazy

u lze kliknout s přík

► Specifikovat kontext

► Omezit hledání

Hledat

Jak vyhledávat v korpusech?

Jednodušší typ dotazů (lemma, fráze, slovní tvar, část slova)

příklad: existuje slovo XY?

Složitější typ dotazů

- **CQL** (Corpus Query Language)
 - práce s regulárními výrazy
 - zadání pozičních atributů (word (základní), lemma, tag)

příklad: jaká jsou nejfrekventovanější substantiva středního rodu končící v 1. pádě singuláru koncovkou -o?

Úkol: v korpusech hledám...

- ... je používanější slovo X, nebo Y?
- ... v jakém kontextu se dané slovo používá?
- ... jaký význam dané slovo má?
- ... v jakých typech textů se slovo nejčastěji používá?
- ... jak se slovo X nejčastěji překládá do angličtiny (ruštiny...)?
- ... jaká slova nejčastěji daný literát používal?
- ... která jazyková varianta se častěji používá?
 - varianty pravopisné (realismus/realizmus)
 - varianty morfologické (kope/kopá)
 - varianty lexikální (alespoň/aspoň)