Jazykové korpusy

Počítačové nástroje pro češtinu
Jaro 2020

Adriana Válková

valkova@phil.muni.cz

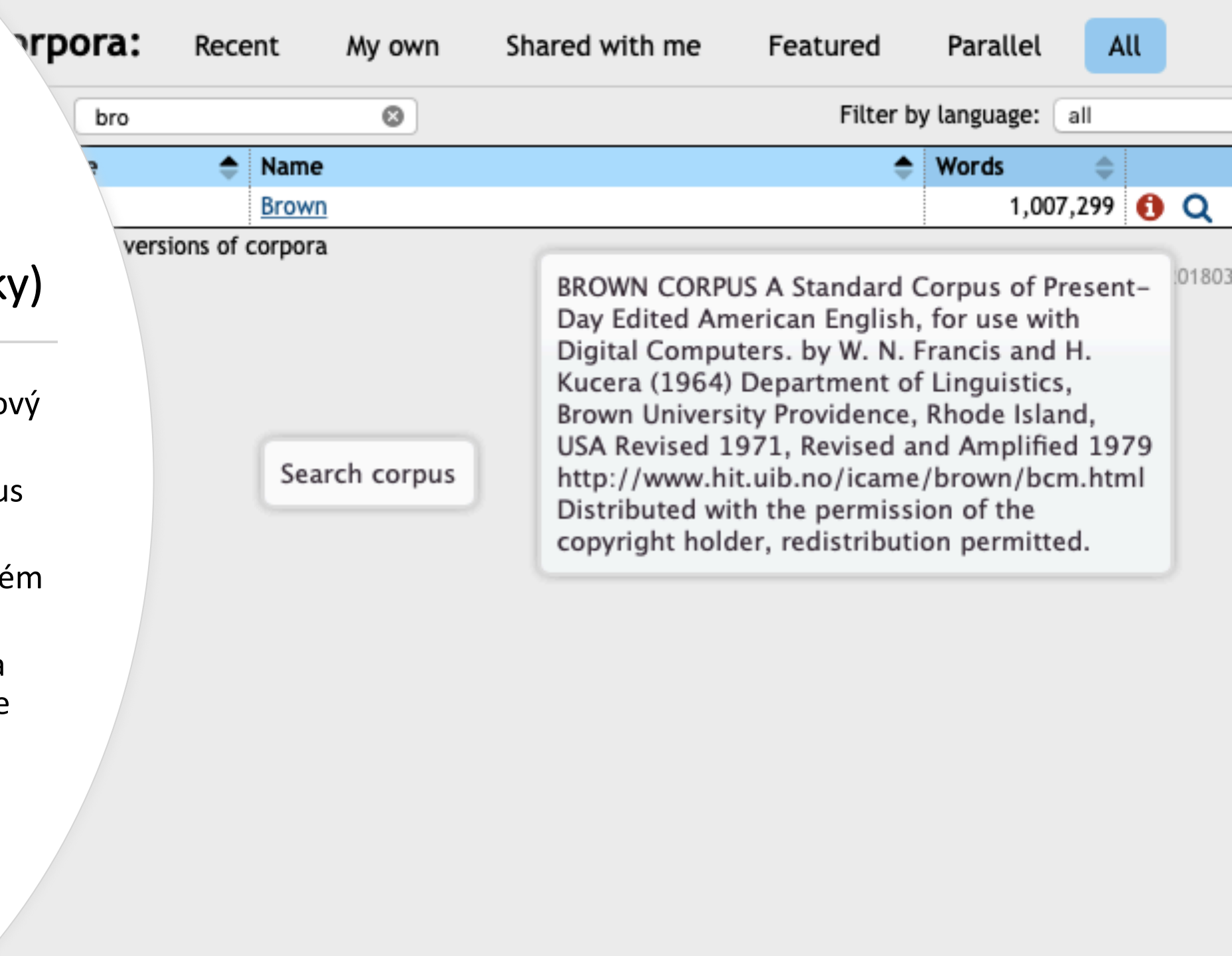
Co je jazykový korpus?

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)

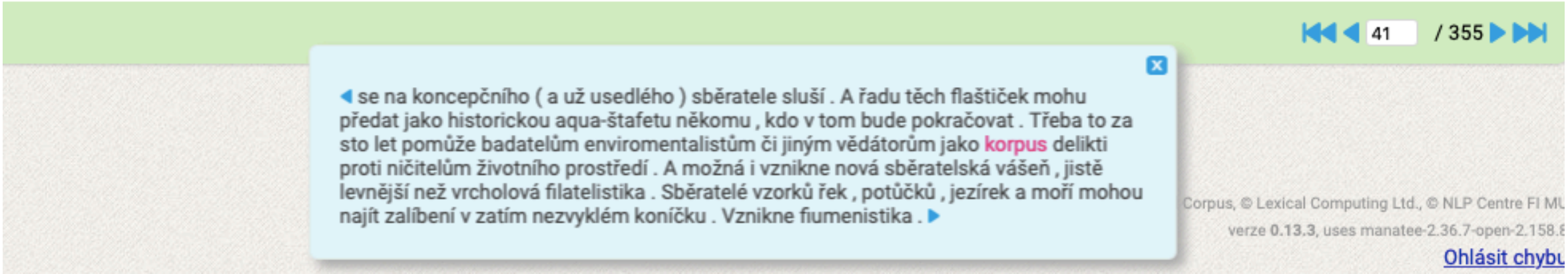
- 2. pol. 20. st., první jazykový korpus – Brown corpus
- 1994 Český národní korpus
- zkoumání jazykového materiálu v jeho přirozeném kontextu
- empiricky podložená data vs. lingvistova introspekce (intuice)



Korpusové texty jsou...

... autentické
... neúplné (autorská práva)
... anotované (označkované)

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 , žloutkové 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. ,



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	KWIC	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/označkovaný**
- **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

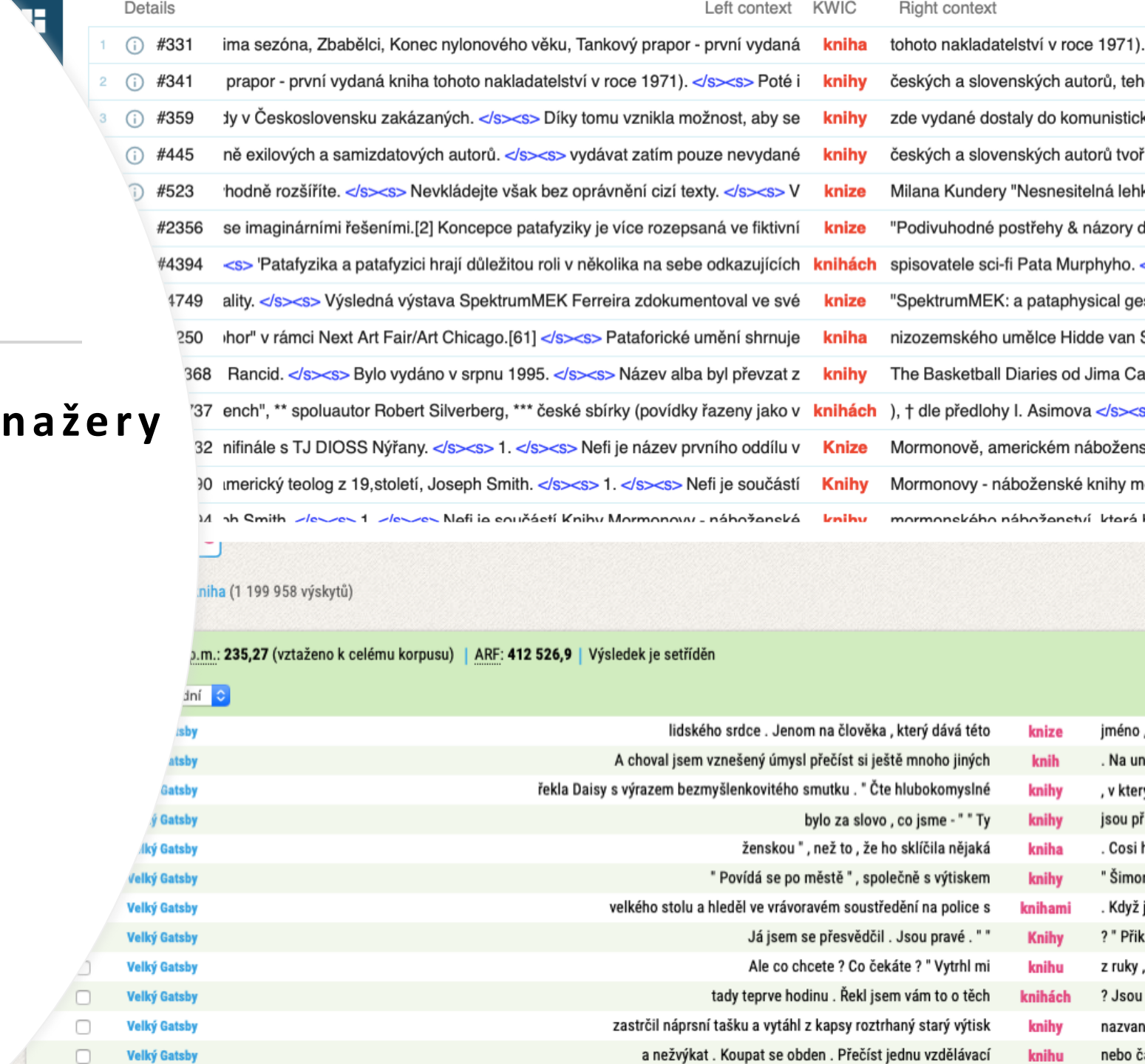
korpus je vždy budován za (s) určitým cílem

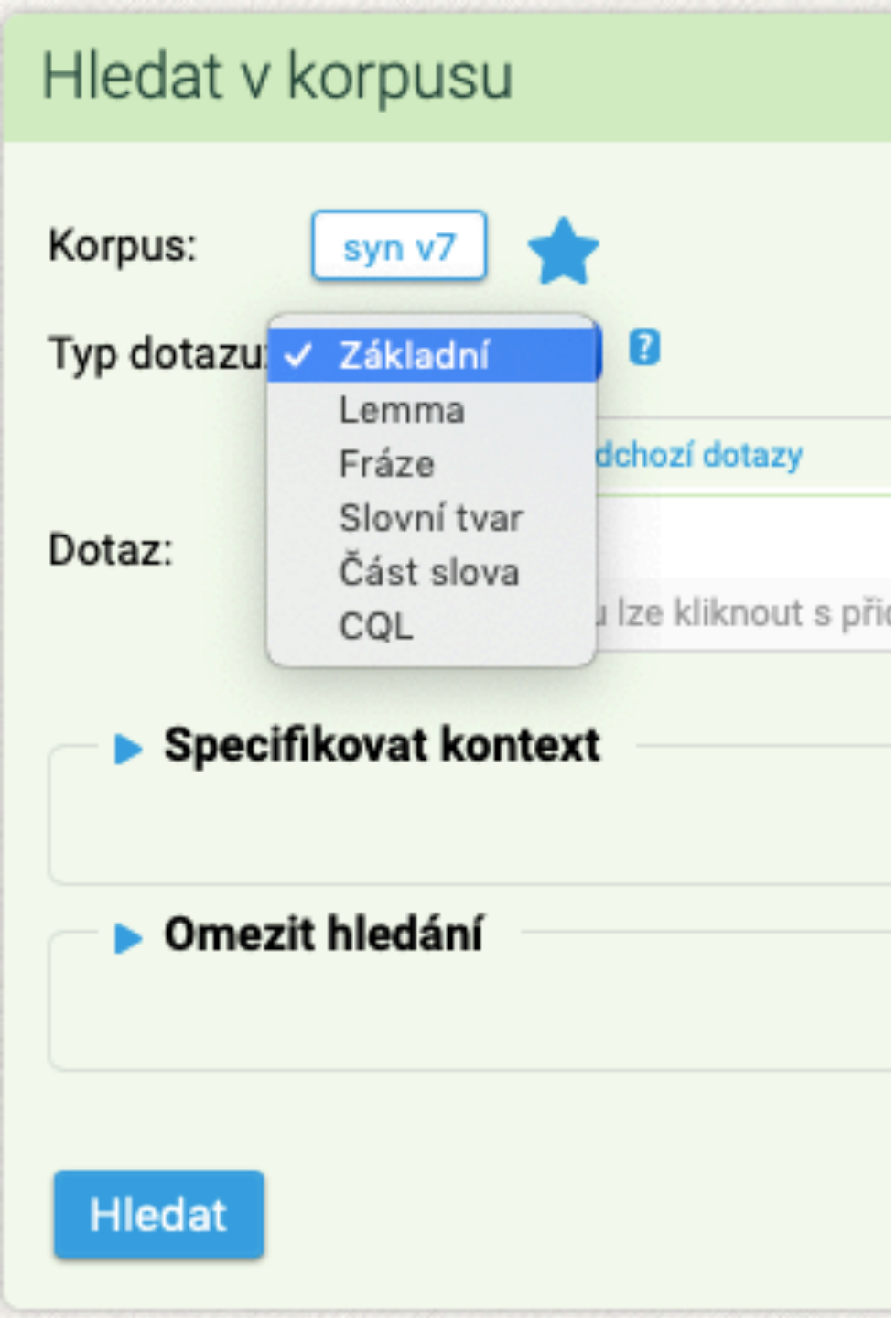
1. jednojazyčný – vícejazyčný (InterCorp)
2. obecný – specializovaný (např. korespondence K. H. Borovského)
3. psaný – mluvený (ORAL)
4. synchronní (SYN) – diachronní (Diakorp – texty ze 14.–20. st.)
5. referenční – nereferenční
6. označkovaný (typ značek?) – neoznačkovaný

Práce s korpusy…

…na webu skrze **korpusové manažery**

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





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)