CJBB75 – 1 (G13 8.00-9.30)

K. Osolsobě

Výuka: od 2. 3. každých 14 dní kontaktní, každých 14 dní úkol (viz harmonogram)

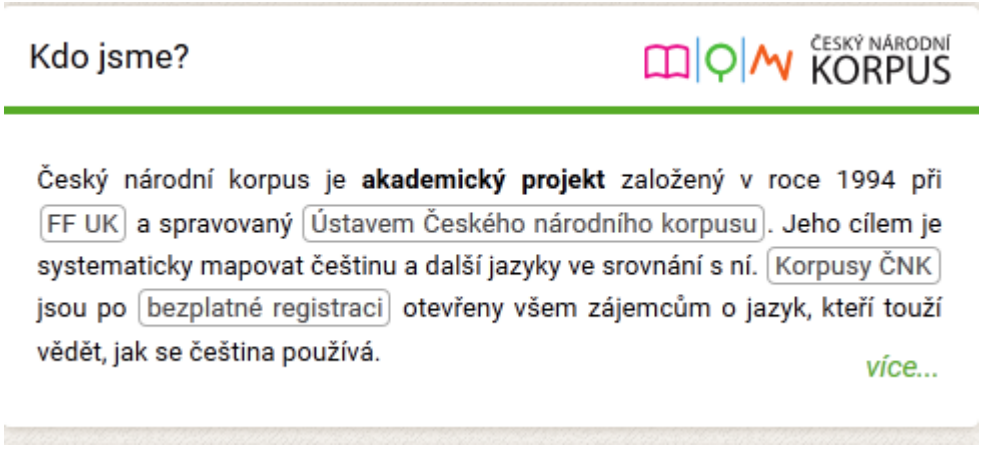
Podmínky ukončení: Průběžné plnění úkolů (5 odevzdaných referátů, test).

Náplň dnešní hodiny

Co je to korpus?

- Soubor textů
- elektronicky uložených a přístupných (korpusové manažery – programy, skrze něž lze ke korpusům přistupovat)
- má stanovený obsah (složený z textů záměrně vybraných dle zveřejněných kritérií)
- má stanovený rozsah/velikost (lze na něm pracovat s frekvenčními/kvantitativně měřitelnými údaji)
- obsahuje standardní anotace (metadata – údaje o každém textu a lingvistické interpretace, anotace jazykových jednotek – vnitřní anotace)

Registrace uživatele pro práci s ČNK (http://ucnk.ff.cuni.cz/)



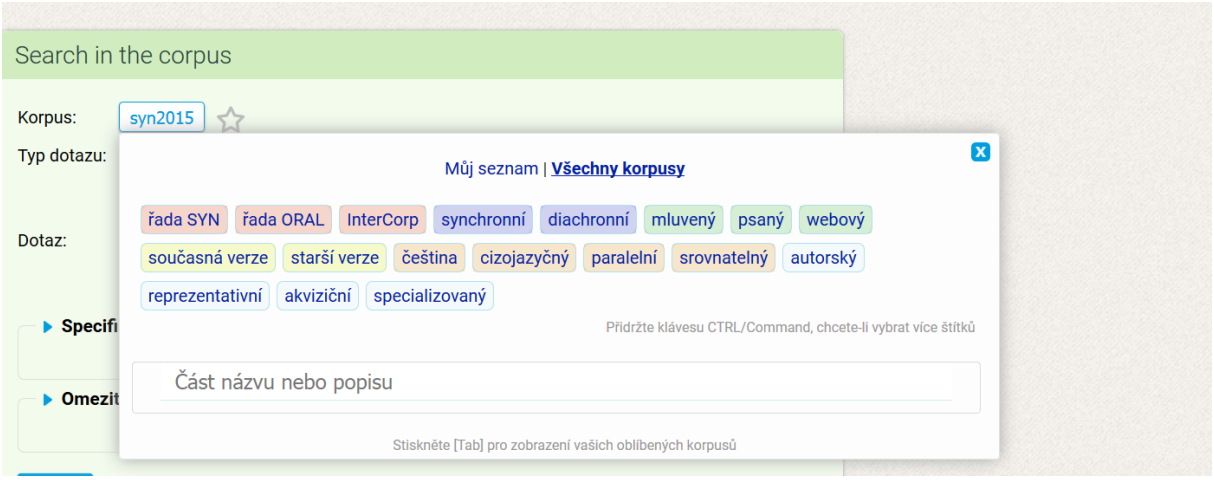
Korpusový manažer

Základy práce s korpusem přes Kontext

KonText

Dotaz

Výběr korpusu



Jaké korpusy jsou k dispozici ?

Časové hledisko (synchronní / diachronní)

Hledisko textů (psané / mluvené, připravené/spontánní)

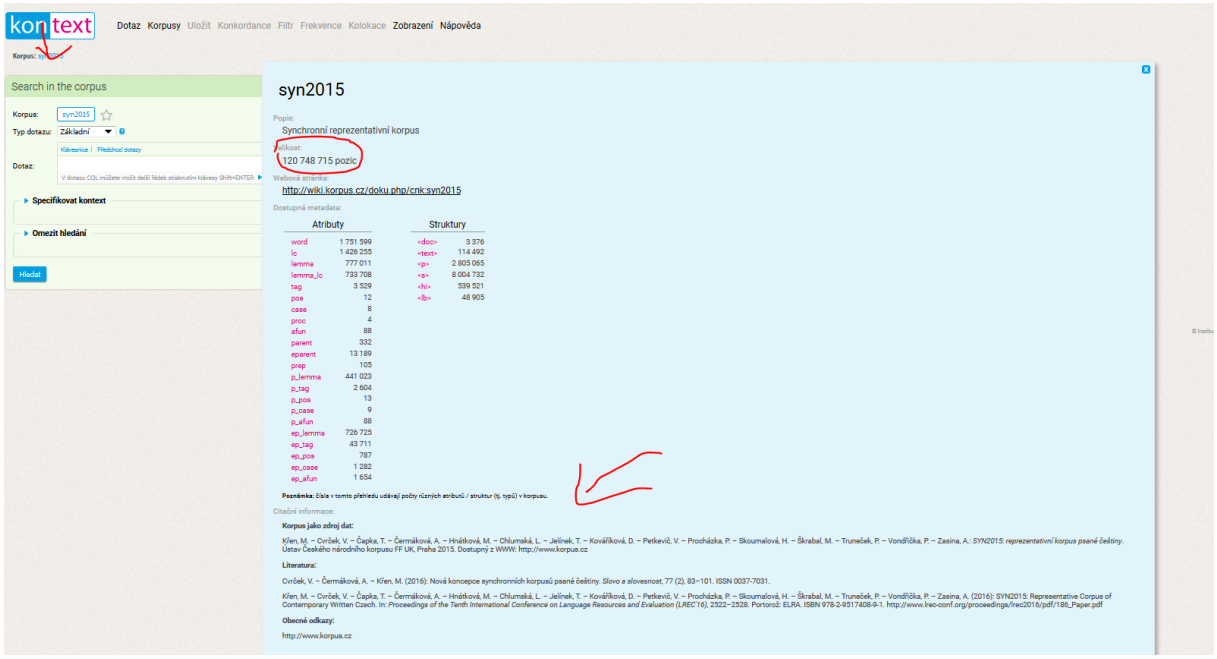
Hledisko žánru (vyvážené žánrově/ žánrově kompaktní – např. korpusy výhradně publicistické, nebo korpus soukromé korespondence, projekt Korpus českého verše).

Hledisko autora (autoři jsou rodilí mluvčí/ autoři se učí jazyk, v němž jsou texty vytvořeny jako tzv. druhý jazyk – learner corpora/žákovské korpusy, autorské korpusy založené na díle/korespondenci význačných osobností).

Hledisko jazyka (jednojazyčné – např. čeština/ vícejazyčné, srovnatelné, paralelní).

Vícejazyčné paralelní korpusy – stejné texty – originál+překlad – zarovnaní/alignment = jednotky, které si odpovídají, jsou propojeny / srovnatelné korpusy – různojazyčné i stejného jazyka vybudované stejným způsobem, mající stejné složení).

Jak čteme informace o zvoleném korpusu?



Stručné info. vč. citování

Proč je třeba citovat korpusy

Jak číst informace o velikosti korpusu:

Termíny: viz http://wiki.korpus.cz/doku.php

http://wiki.korpus.cz/doku.php/pojmy:prehled_pojmu

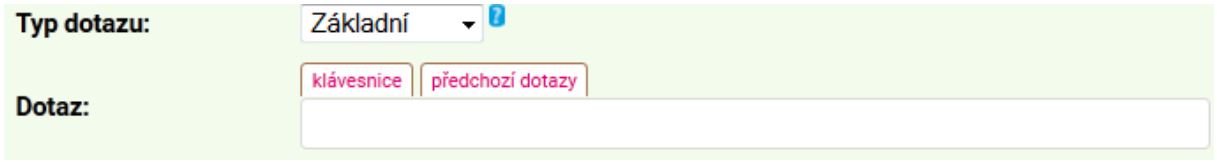
pozice

tokenizace

lemmatizace

desambiguace

Vyhledávání:



Slovní tvar/slovo/word

Fráze

Lemma

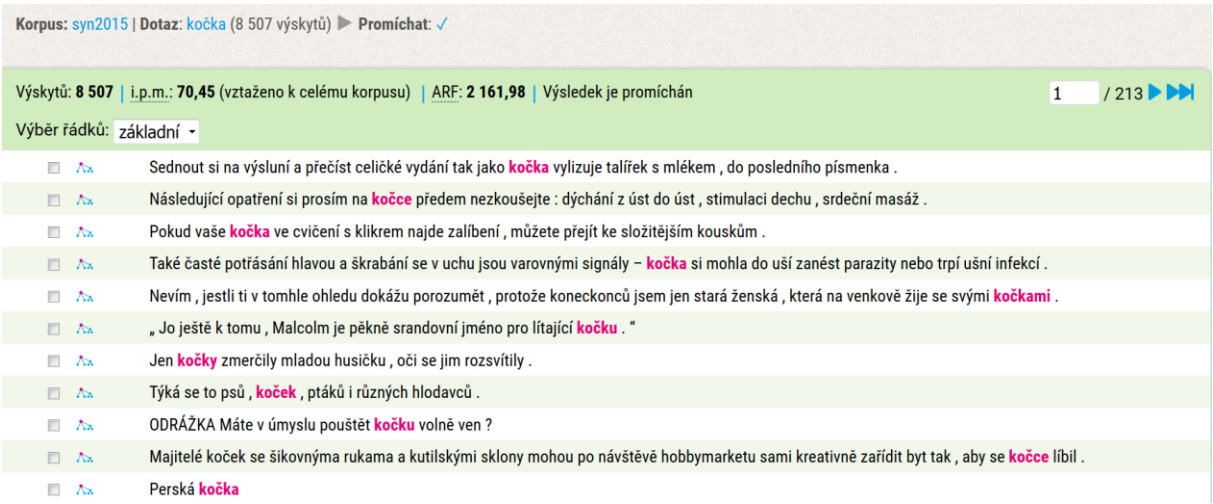
CQL

Regulární výrazy (http://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy)

konkordance, KWIC

Zobrazení

KWIC/Věta

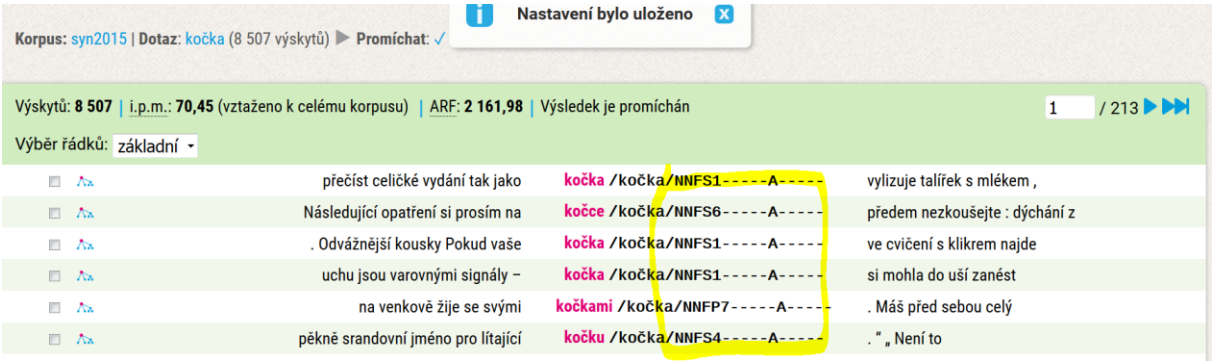


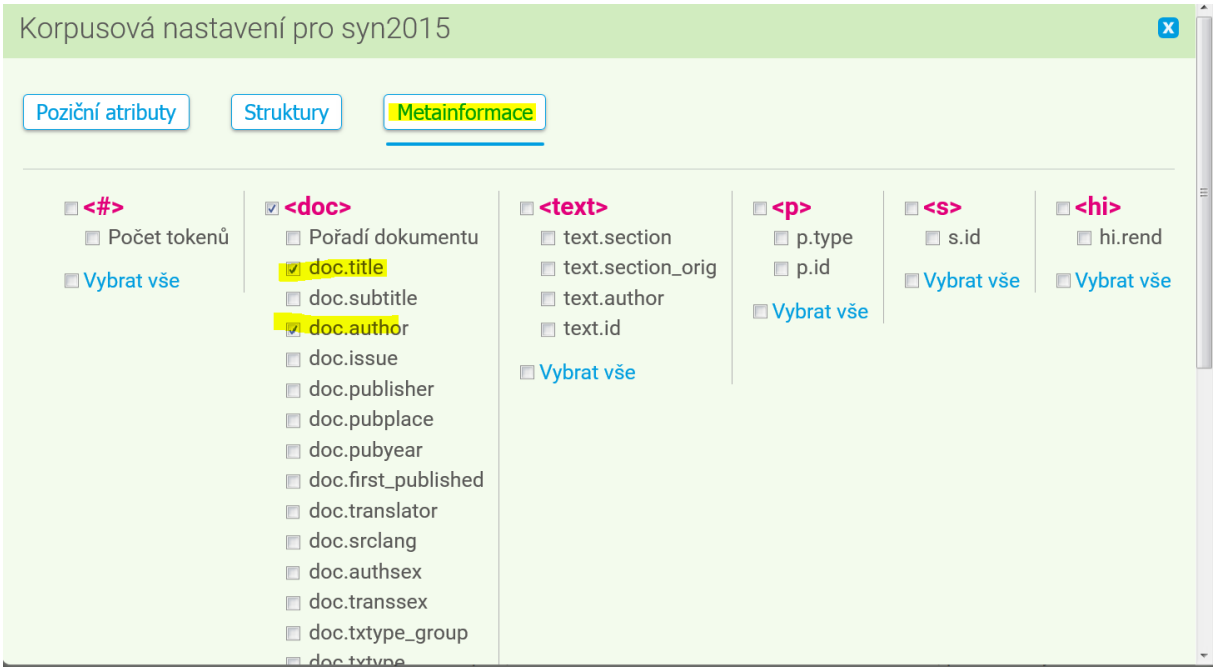
Korpusová nastavení

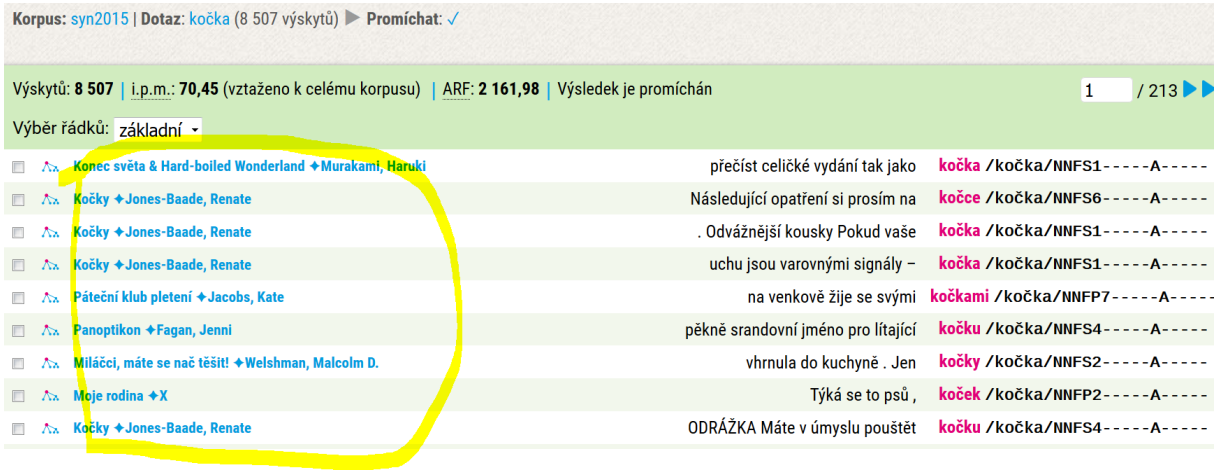
(Lemma, POS – **p**art **o**f **s**peech)


Korpus: syn2015 | Dotaz: kočka (8 507 výskytů) ▶ Promíchat: ✓
Výskytů: 8 507 | i.p.m.: 70,45 (vztaženo k celému korpusu) | ARF: 2 161,98 | Výsledek je promíchán
1 / 213
Výběr řádků: základní
přečíst celičké vydání tak jako kočka /kočka/N vylizuje talířek s mlékem ,
Následující opatření si prosím na kočce /kočka/N předem nezkoušejte : dýchání z
. Odvážnější kousky Pokud vaše kočka /kočka/N ve cvičení s klikrem najde
uchu jsou varovnými signály – kočka /kočka/N si mohla do uší zanést
na venkově žije se svými kočkami /kočka/N . Máš před sebou celý
pěkně srandovní jméno pro létající kočku /kočka/N . “ „ Není to
vhrnula do kuchyně . Jen kočky /kočka/N zmerčily mladou husičku , oči
Týká se to psů , koček /kočka/N , ptáků i různých hlodavců
ODRÁŽKA Máte v úmyslu pouštět kočku /kočka/N volně ven ? S prvním
být tak , aby se kočce /kočka/N líbil . Ale i pokud

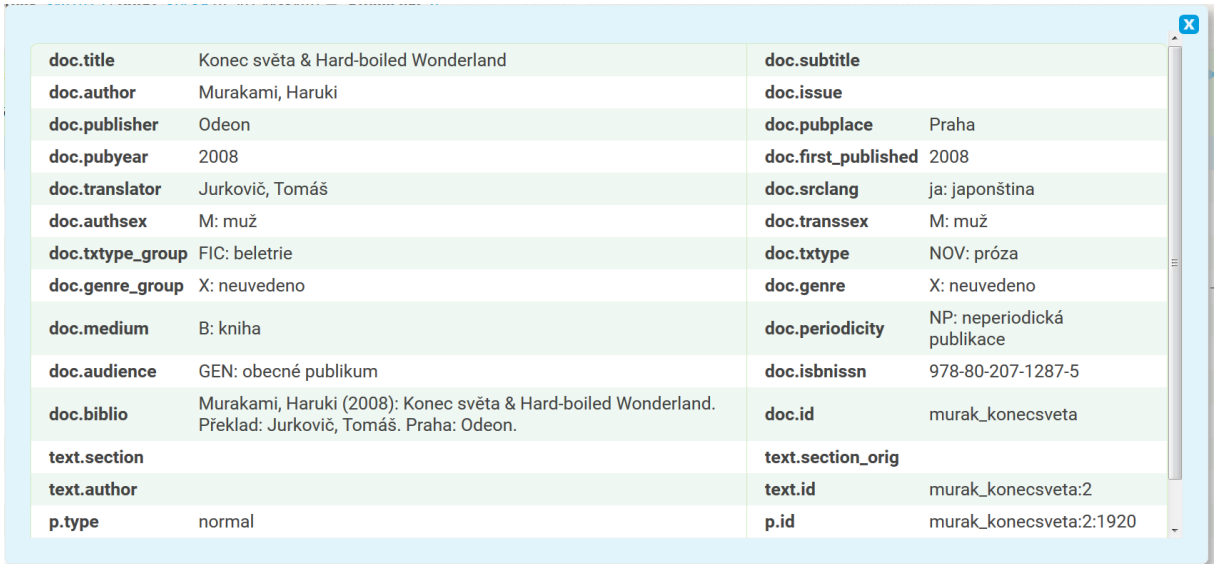
(Lemma, tag)


Nastavení bylo uloženo
Korpus: syn2015 | Dotaz: kočka (8 507 výskytů) ▶ Promíchat: ✓
Výskytů: 8 507 | i.p.m.: 70,45 (vztaženo k celému korpusu) | ARF: 2 161,98 | Výsledek je promíchán
1 / 213
Výběr řádků: základní
přečíst celičké vydání tak jako kočka /kočka/NNFS1-----A----- vylizuje talířek s mlékem ,
Následující opatření si prosím na kočce /kočka/NNFS6-----A----- předem nezkoušejte : dýchání z
. Odvážnější kousky Pokud vaše kočka /kočka/NNFS1-----A----- ve cvičení s klikrem najde
uchu jsou varovnými signály – kočka /kočka/NNFS1-----A----- si mohla do uší zanést
na venkově žije se svými kočkami /kočka/NNFP7-----A----- . Máš před sebou celý
pěkně srandovní jméno pro létající kočku /kočka/NNFS4-----A----- . “ „ Není to


Korpusová nastavení pro syn2015
Poziční atributy
Struktury
Metainformace
<#>
Počet tokenů
Vybrat vše
<doc>
Pořadí dokumentu
doc.title
doc.subtitle
doc.author
doc.issue
doc.publisher
doc.pubplace
doc.pubyear
doc.first_published
doc.translator
doc.srclang
doc.authsex
doc.transsex
doc.txtype_group
<text>
text.section
text.section_orig
text.author
text.id
Vybrat vše
<p>
p.type
p.id
Vybrat vše
<s>
s.id
Vybrat vše
<hi>
hi.rend
Vybrat vše



Kompletní info o zdrojovém textu:



doc.title	Konec světa & Hard-boiled Wonderland	**doc.subtitle**	
doc.author	Murakami, Haruki	**doc.issue**	
doc.publisher	Odeon	**doc.pubplace**	Praha
doc.pubyear	2008	**doc.first_published**	2008
doc.translator	Jurkovič, Tomáš	**doc.srclang**	ja: japonština
doc.authsex	M: muž	**doc.transsex**	M: muž
doc.txtype_group	FIC: beletrie	**doc.txtype**	NOV: próza
doc.genre_group	X: neuvedeno	**doc.genre**	X: neuvedeno
doc.medium	B: kniha	**doc.periodicity**	NP: neperiodická publikace
doc.audience	GEN: obecné publikum	**doc.isbnissn**	978-80-207-1287-5
doc.biblio	Murakami, Haruki (2008): Konec světa & Hard-boiled Wonderland. Překlad: Jurkovič, Tomáš. Praha: Odeon.	**doc.id**	murak_konecsveta
text.section		**text.section_orig**	
text.author		**text.id**	murak_konecsveta:2
p.type	normal	**p.id**	murak_konecsveta:2:1920

Úkol na příště:

Prostudovat www stránky ÚČNK

Umět odpovědět na otázky:

1. Co je to korpusu?
2. Co je to Český národní korpus?
3. Jaké typy korpusů máme k dispozici?
4. Co to znamená, když řeknu, že korpus má 100 milionů slov?
5. Jak komunikujeme s korpusem (jak jej můžeme využívat pro lingvistickou práci)?
6. Jak můžeme vyhledat v korpusu výskyt slova, jak se se zobrazí v korpusu výskyt slova a co můžeme se zobrazenými výskyty dále dělat?
7. Jak můžeme vyhledat v korpusu všechna slova, která mají společnou vlastnost, že jsou tvary jednoho základního tvaru?

8. Jak můžeme v korpusu vyhledat všechny tvary na rovině gramatické abstrakce (třeba podstatné jména rodu ženského ve 3. pádě, nebo slovesa v přítomném čase v první osobě)?

A připravit si otázky, na něž byste rádi znali odpověď (souvisí s korpusy!!)