**CJBB75 – 18. 2. 2015**

**K. Osolsobě**

**Výuka: každých 14 dní kontaktní**

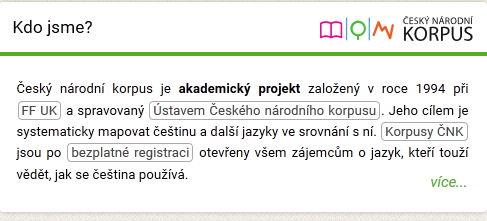
**Podmínky ukončení: Průběžné plnění úkolů (2 studijní ůkoly, 4 písemné práce odevzdané, test).**

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)

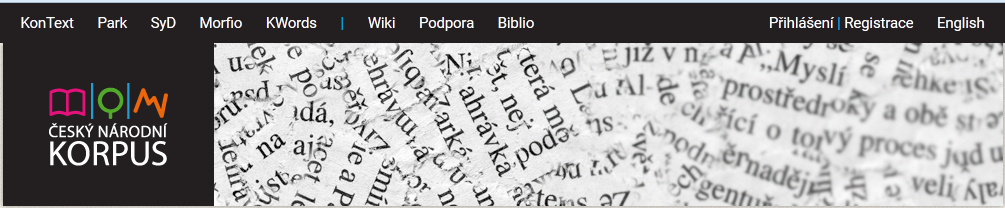
**Registrace uživatele pro práci s ČNK (**[**http://ucnk.ff.cuni.cz/**](http://ucnk.ff.cuni.cz/)**)**



**Korpusový manažer**

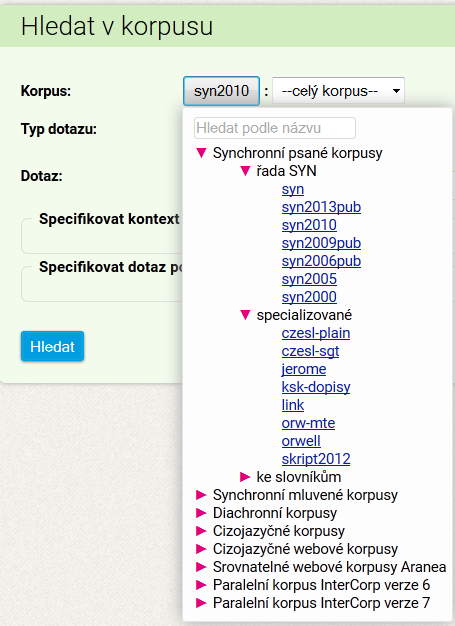
**Základy práce s korpusem přes Kontext**

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).

Hledisko autora (autoři jsou rodilí mluvčí/ autoři se učí jazyk, v němž jsou texty vytvořeny jako tzv. druhý jazyk – learner corpora/žákovské korpusy).

Hledisko jazyka (jednojazyčné – např. čeština/ vícejazyčné).

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 číst informace o velikosti korpusu:**

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

http://wiki.korpus.cz/doku.php/pojmy:prehled\_pojmu

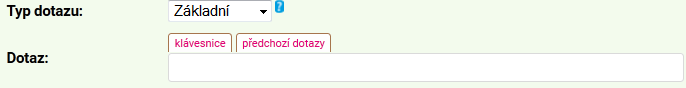
pozice

tokenizace

lemmatizace

desambiguace

Vyhledávání:



Slovní tvar/slovo/word

Fráze

Lemma

CQL

**Regulární výrazy (**[**http://wiki.korpus.cz/doku.php/pojmy:regularni\_vyrazy**](http://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy)**)**

**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 a) slovo, které neznáme?
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!!)**