Základy korpusové lingvistiky – termíny:

Viz též
http://wiki.korpus.cz/doku.php/manual:zobrazeni_dotazu

Jazykový korpus je rozsáhlý soubor **autentických textů** (psaných nebo mluvených) převedený do **elektronické podoby** v jednotném formátu tak, aby v něm bylo možné jednoduše **vyhledávat** jazykové jevy, zejména slova a slovní spojení (kolokace).

Každý text je při vstupu do korpusů zpřístupňovaných rozhraním KonText opatřen anotací = slova v něm jsou tzv. **označkována**.

Každému slovu jsou přiděleny tzv. **poziční atributy** = vlastnosti, podle kterých je můžeme hledat.

Základními pozičními atributy jsou:

- word - slovní tvar (např. *kočkou, modrou, píše*)
- lc - (z angl. *lowercase*) ekvivalent slovního tvaru, který ovšem zanedbává velikost písmen
- lemma - základní (slovníkový) tvar (např. *kočka, modrý, psát*)
- tag - značka zachycující morfologickou (příp. i jinou, obvykle gramatickou) informaci

Příklad:

kočkou – **word/lc**: *kočkou*

lemma: *kočka* (word/lc *kočkou* patří k lemmatu *kočka*)

tag: podstatné jméno, rod ženský, jednotné číslo, 7. pád (NNFS7.*)

(N = podstatné jméno, N – obyčejné, tj. bez další zvláštní charakteristiky, F = femininum, S = singulár, 7 = 7. pád; značka „.*“ znamená, že další gramatické vlastnosti nepotřebujeme)

modrou – **word/lc**: *modrou*

lemma: *modrý* (word/lc *modrou* patří k lemmatu *modrý* – nikoliv *modrá*!!!)

tag: přídavné jméno, rod ženský, jednotné číslo, 7. pád (AAFS7….1A.*) nebo 4. pád (AAFS4….1A.*)

(A = přídavné jméno, A – obyčejné, tj. bez další zvláštní charakteristiky, F = femininum, S = singulár, 7 = 7. pád // 4 = 4. pád; pak 4 nedefinované pozice, 1 = 1. stupeň, A = bez negace, značka „.*“ znamená, že další gramatické vlastnosti nepotřebujeme)

píše – **word/lc**: *píše*

lemma: *psát* (word/lc píše patří k lemmatu *kočka*)

tag: – sloveso, 3. osoba jednotného čísla přítomného času, činný rod, kladný tvar (VB.S…3P.AA.*)

(V = verbum/sloveso, B – tvar přítomného nebo budoucího času, „.“ = pozice není definována, S = singulár, pak 3x nedefinovaná pozice, 3 = 3. osoba, P = prézens, pak nedefinovaná pozice, A = aktivum, A = kladný tvar)

Toto značkování umožní vyhledat právě ten výraz, který potřebujeme, nebo právě ty tvary slov (např. dativ podstatných jmen rodu ženského, všechny přechodníky přítomné apod.

Další termíny:

dotaz

příkazový řádek

konkordance

Jednoduché hledání: kombinace nastavení „atributu“ (word, lemma…) a vepsání slova do příkazového řádku: např. nastavíme „word“ a zadáme „kočkou“.

Dotazy:

základní

lemma

slovní tvar (= word, lc)

část slova

CQL

Další dotazy: základní, fráze, CQL – necháme na později

Další možnosti dotazů:

„základní“:

- zadáme-li „slovníkový tvar“, tj. 1. pád jedn. č. jména nebo infinitiv slovesa, funguje „základní“ jako lemma (většinou)
- zadáme-li jiný tvar (např. *kočkou*, *píše*), funguje „základní“ jako „word“.

V tomto je tedy „nevýhoda“ dotazu „základní“. Výhodou je, že můžeme zadat i slovní spojení (frázi). Pozor, pokud chceme všechny tvary daného spojení,