Základy využití korpusu v praxi **cjbb75**

**Středa: 9.10-10.50 G13**

**4. 4. 2018**

**Korpus jako zdroj dat pro výzkum slovotvorby**

<http://www.ujc.cas.cz/elektronicke-slovniky-a-zdroje/Slovnik-afixu-html>

<http://ufal.mff.cuni.cz/derinet>

<https://nlp.fi.muni.cz/projects/derivancze/>

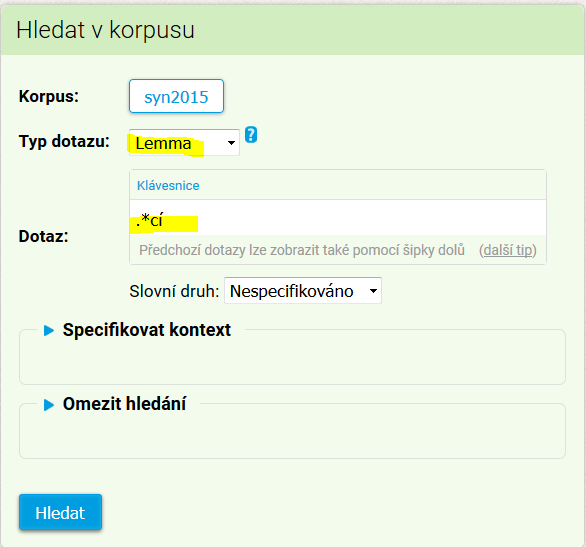
Jak zadat dotaz pro hledání slov, která mají společné slovotvorné vlastnosti

(lingvistické termíny: morfém, afix, kořen, kmen, slovní druh, substantivum, adjektivum, sloveso, …, gramatická kategorie, rod, číslo, pád, osoba …; termíny pro formální zadání: řetězec, token, lemma, morfologická značka/tag, pozice/atribut, hodnota, technické řešení, … )

Adjektiva od sloves typu *koupací*

Slovní formulace – tvrzení, která musí platit, podmínka nutná/dostačující, využití regulárních výrazů a logických operací, využití filtrů, pozorování dat a úprava dotazu.

Chceme vyhledat všechna adjektiva taková, že lemma končí na *cí*.





výsledek je značně přegenerovaný.

Lze dotaz optimalizovat?

Všimněte si:



*obývat – obývací* i *obývající*

*zbývat – zbývající*

Existují dva významy

a) vlastnost vyjadřuje, že něco je určeno k tomu, co označuje základové sloveso: *obývací kuchyň* je nejen k vaření a jedení, ale i k tomu, abychom v ní *pobývali/přebývali*;

b) transpozice významu děje (vlastnost toho, který dělá to, co označuje základové sloveso): zvíře *obývající* oblast/světadíl/ … *=* jde pouze o nominalizaci verba (převod/transpozici z kategorie slovního druhu).

Rozlišujeme dvě slovotvorné třídy:

***-a-cí*** od kmene minulého/infinitivního **(tzv. adjektiva účelová)**

***-aj-í-cí*** od kmene přítomného **(tzv. adjektiva procesuální)**

Jak formulovat dotaz?

Pokusíme se odstranit ta adjektiva, u nichž před sufixem *–cí* předchází kmenotvorný sufix prézentního kmene. Opěrným tvarem je tvar 3. osoby sg. indikativu prézentu aktiva, tedy buď *–ou-*, nebo *–í-* (před *–í-* může ještě předcházet *–j-*/*-uj-/-[eě]j*-/-*aj-*).

Co může předcházet před *–cí* při derivaci od kmene minulého? Mohou nastat případy, kdy je kmen minulý a přítomný homonymní, takže při homonymii sufixu budeme mít možnou homonymii adjektivních derivátů?

Existují v češtině slovesa, jejichž kmen minulý (opěrný tvar je l-ové příčestí) končí na *oul/íl*?

Dokážete si vzpomenout?

Podívejte se do korpusu.

Dotaz je formulouván slovně tak, že **chceme vyhledat všechny tvary slovesa l-ovéhopříčestí maskulina singuláru takové, že končí na *oul* nebo na *íl*.**  V jazyce cql vypadá dotaz takto:

**[lc=".\*(oul|íl)" & tag="Vp[MI]S.\*"]**

Výsledek je prázdný konkordanční seznam.

Můžeme tudíž použít negativní filtr a odstranit lemmata na .\*(ou|í)cí.

Sledujeme-li frekvenčně utříděný seznam, zjistíme, že se v něm vyskytují i adjektiva, která sice na *cí* končí, ale nejsou tvořena od sloves, a tak nespadají ani do kategorie účelových, ani do kategorie procesuálních adjektiv.

První je adjektivum *domácí* se vztahem k adverbiu *doma*.

Zakončení je *ácí*.

Můžeme si položit otázku, zda existují v češtině adjektiva od sloves, která končí na *ácí*?

Můžeme si na ně a) zkusit vzpomenout, b) vyhledat pomocí pozitivního filtru všechna adjektiva na *.\*ácí* a prohlédnout si je, c) zkontrolovat, zda existují slovesné tvary l-ového příčestí v singuláru maskulina na *.\*ál* a pokud ano, tak jak se od takových tvarů tvoří adjektiva s významem účelu.

Odpověď na a): ?;

b) nalezneme pouze lemma *domácí* a kompozita se druhým členem *domácí;*

c) [lc=".\*ál" & tag="Vp[MI]S.\*"] vyhledáme 126 lemmat, přičemž vidíme, že *á* se při tvoření účelového adjektiva buď krátí (*hrát → hrací, přát → přací, sát → sací,* ), nebo se střídá s *oj* (*stát → stojací*);

Dále vidíme adjektiva jako *zvířecí* a *kuřecí*, jde o jinou slovotvornou třídu (adjektiv tvořených ze substantiv označující vztah přivlastnění/podobnosti základovému substantivu), ale jde také o jiný slovotvorný typ, a sice o skupinu adjektiv tvořených příponou *–í*/ konverzí ke vzoru jarní, přičemž základem je o skupiny substantiv skloňovaných podle vzoru *kuře* kmen rozšířený o kmenotvorné *–[eě]t-*, které alternuje s *–[eě]c-*. V případě, že se základové substantivum neskloňuje podle vzoru *kuře*, pak se příslušné adjektivum tvoří pouze příponou *–í/*konverzí ke vzoru *jarní* (*psí, vlčí, mroží, kozí, myší, …*).

Podívejme se, zda tato adjektiva mohou mít homonymní zakončení s adjektivy tvořenými ze sloves.

Vyhledáme pomocí pozitivního filtru adjektiva, která končí na *[ěe]cí.*

Vidíme, že stejné zakončení mají I adjektiva jako *obráběcí, prováděcí, předváděcí, secí, sklízecí, dobíjecí, secí, …*





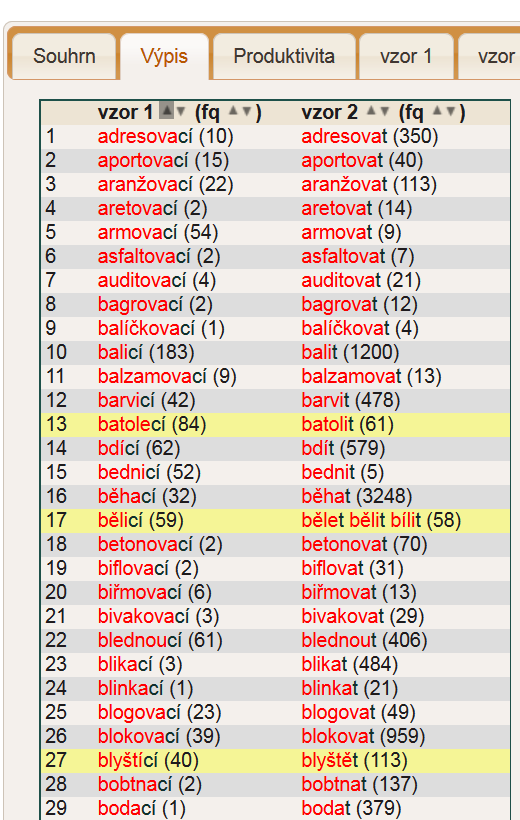
Jak odstranit přegenerované doklady typu *domácí, zvířecí, kuřecí,* ...?

Ruční analýza dat.

Další nástroje pro hledání slovotvorných zákonitostí.

Vyhledávání dvojic slov podle zadání společné a odlišné časti pomocí nástroje Morfio.





Jak vypadají výsledky korpusových analýz v lexikografickém díle zaměřeném na popis slovotvorných prostředků:

<http://www.slovnikafixu.cz/heslar/-c%C3%AD>

11. 4. Dú: Jak lze v korpusu vyhledat substantiva tvořená příponou *–č*, která označují osoby vykonávající činnost označenou základovým slovesem: *řidič, nosič, sběrač, očišťovač, hráč, …*

Pokuste se:

a) formulovat dotaz (v přirozeném jazyce a v jazyce cql);

b) pozorovat korpusová data vyhledaná na základě dotazu a dotaz specifikovat;

c) popsat, jaká pozorování lze využít při reformulaci dotazu/používání filtrů;

d) popsat možnosti využití nástroje Morfio k vyhledání dvojic sloveso/substantivum označující člověka, který dělá to, co označuje základové sloveso typ *nosit/nosič*;

e) podívejte se na heslo *–č* ve Slovníku afixů užívaných v češtině a porovnejte svá pozorování s tím, co se uvádí v hesle. Narazili jste v korpusových datech na něco, co by bylo v rozporu s tím, co se uvádí v textu hesla?

f) Jak byste vysvětlili slova jako *držič, sedič, tiskač, snoubič, strojič*?