Základy využití korpusu v praxi **cjbb75**

**Středa: 9.10-10.50 G13**

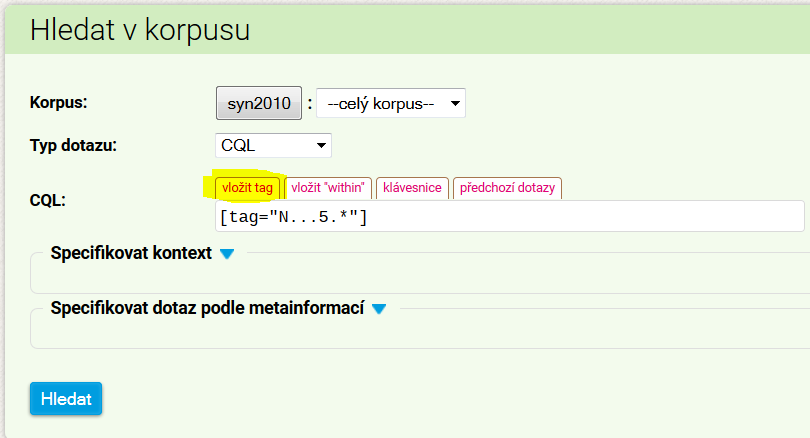
**6. 4. 2016**

**Jak lze v korpusech hledat doklady pro výzkum morfologie?**

**Vyhledávání podle morfologické značky – problém spolehlivosti desambiguace**

**Vyhledejte v korpusu SYN2010 všechny vokativy**

**Jak?**



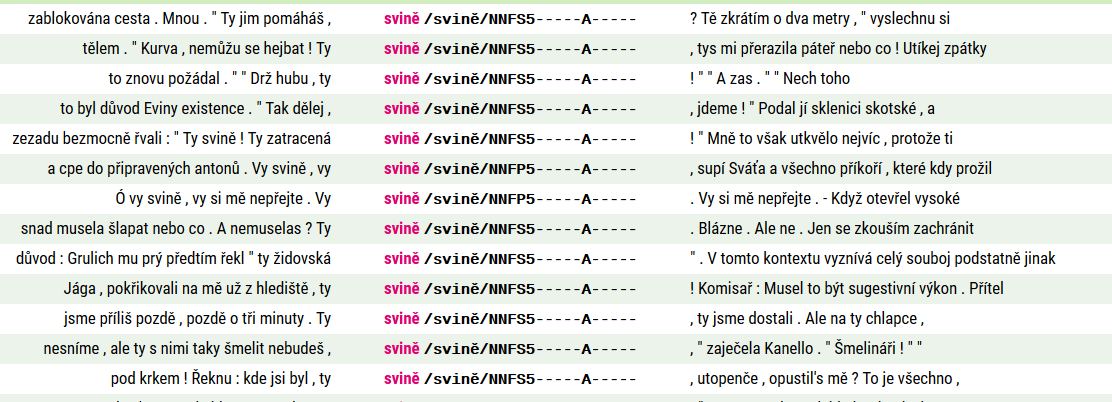
(můžete použít funkci vložit tag).

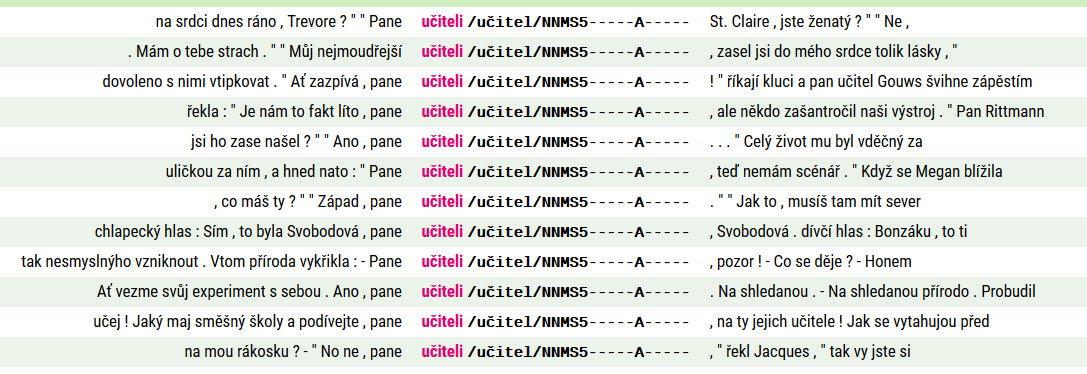


Seznam lc (slovní tvary bez ohledu na velikost písma)



Copak jsou v češtině vokativy jenom u substantiv, která nemají homonymní tvary vokativu?





Je desambiguace spolehlivá?



Ne tak docela.

**Jak můžeme obejít chyby v desambiguaci?**

Shrňme pozorování:

Seznam obsahoval převážně tvary vokativu, které v češtině nejsou homonymní s jinými pádovými formami. Které to jsou? Vokativ některých substantiv vzorů pán, hrad, muž a všech substantiv vzorů předseda a žena v sg. Všechna ostatní substantiva mají v češtině vokativ homonymní s nominativem (v plurálu). V sg. se homonymie liší u jednotlivých vzorů.

Pokusme se najít chyby v desambiguaci u tvarů vokativu substantiv maskulin životných.

Vzor pán: má vokativ na –e (není homonymní) a vokativ na –u, který se vyskytuje zpravidla u lemmat končících na veláru [kgh] s výjimkou substantiv *Bůh, člověk* a u substantiva *syn*.

Tvar na –u je ovšem homonymní, vyskytuje se totiž nejen ve vokativu, ale může stát také v lokále a dativu sg.

Vyhledejme v korpusu tvary vlku, zobrazme lemma a tag a podívejme se na chyby v desambiguaci





Úkol:

Vyhledejte v korpusu SYN2010 všechny tvary končící na *u* maskulin životných, jejichž lemma končí na k, g, h, ch.

Budeme používat jazyk cql (**C**orpus **Q**uerry **L**anguage). Budeme používat zadání pomocí regulárních výrazů a logických operátorů pro konjunkci.

[tag="N.M.\*" & lc=".\*u" & lemma=".\*[kgh]"]

Dotaz na každou pozici je v jazyce **cql** v hranatých [] závorkách.

Můžeme se ptát na lemma, tag, word, lc, pos, tedy na základní tvar, gramatické vlastnosti –kódované ve značce/tagu, slovní tvar, slovní tvar bez ohledu na to, jakými písmeny je napsán a na slovní druh (1. pozici ve značce – pos – **P**ART **O**F **S**PEECH).

Do hranatých závorek napíšeme typ dotazu, tj. ‚ na jaký typ se ptáme, bez mezery následuje znak ‚=‘ a v úvozovkách "" je slovní tvar/základní tvar/značka atd., na který/kterou se ptáme.

Pokud chceme kombinovat v rámci jednoho dotazu více podmínek, můžeme je spojit znaky pro logické operace (např. konjunkce, disjunkce, negace).

Uvnitř dotazu můžeme také pracovat s disjunkcemi.

Nyní slovně popíšu výše uvedený dotaz:

[tag="N.M.\*" & lc=".\*u" & lemma=".\*[kgh]"]

vyhledej všechna substantiva maskulina : **tag="N.M.\*"**

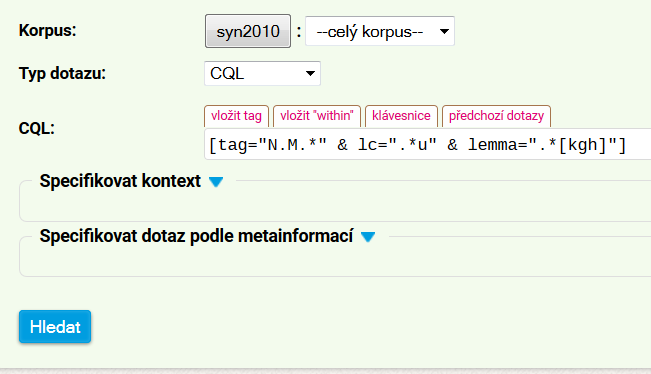
pro něž platí: **&**

že tvary bez ohledu na velikost písma končí na *u*: **lc=".\*u"**

a pro něž platí: **&**

že základní tvar končí na *k,* nebo na *g*, nebo na *h* (tím zahrneme i případy, kdy končí na *ch*): **lemma=".\*[kgh]"**

([kgh] je v jazyce cql za disjunkce jednotlivých znaků/písmen)



Sledujeme chyby v desambiguaci

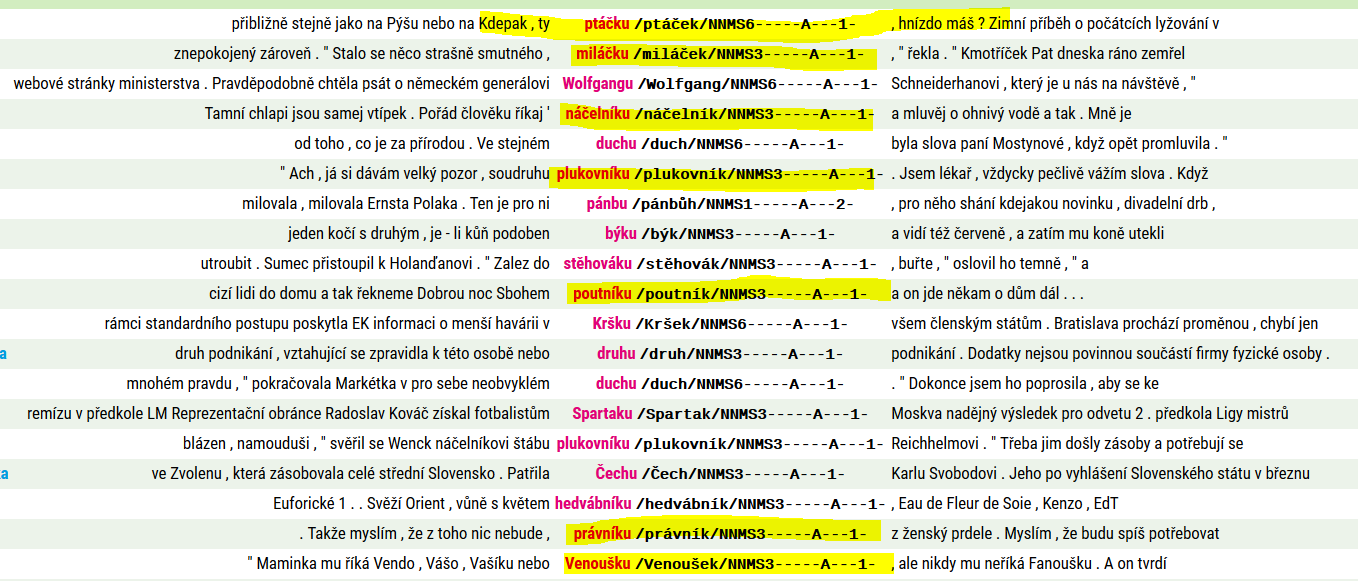


Dotaz vylepšíme, abychom nevyhledali tvary *Bohu, bohu, člověku*, které nemohou být vokativy.

Jak?

Použijeme operátor negace a do dotazu zahrneme podmínku, že nechceme tvary lemmat *bůh* a *člověk*.

**[tag="N.M.\*" & lc=".\*u" & lemma=".\*[kgh]" & lemma!="(bůh)|(člověk)"]**



Všimněme si, kde se nacházejí chyby.

Pokuste se formulovat nějakou vlastnost, jíž se liší (morfosyntakticky) tvar vokativu od tvaru dativu a lokálu.

Např.: před tvarem lokálu/dativu je jméno ve stejném pádě, nebo předložka pojící se se stejným pádem.

Nebo: Tvary vokativu bývají v češtině odděleny čárkou nebo koncovou interpunkcí.

Nebo: před tvarem vokativu se bezprostředně nachází jiný tvar ve vokativu (*pane poručíku, můj miláčku*).

Nebo: Před tvarem vokativu stojí osobní zájmeno *ty* (*ty hlupáku*).

Na základě takovýchto vlastností můžeme „odfiltrovat řádky, které s velkou mírou pravděpodobnosti budou obsahovat tvary se správnou značkou.

Z následujícího seznamu lemmat vybereme řádky, na nichž je tvar *miláčku*.





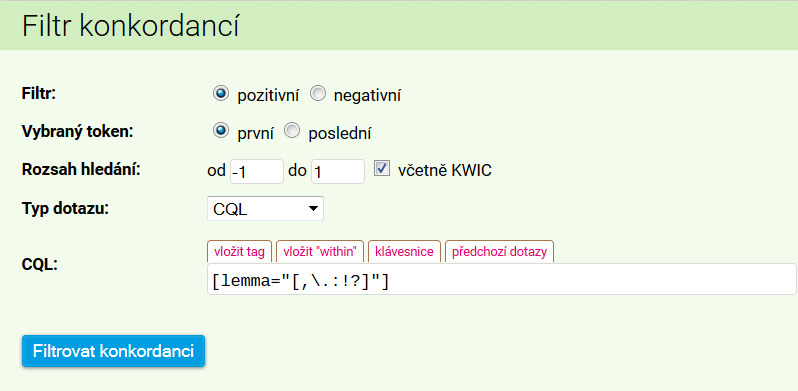
Všimněme si, jak vypadá okolí chybně označkovaných tvarů.

Tvary vokativu bývají v češtině odděleny čárkou, nebo koncovou interpunkcí.

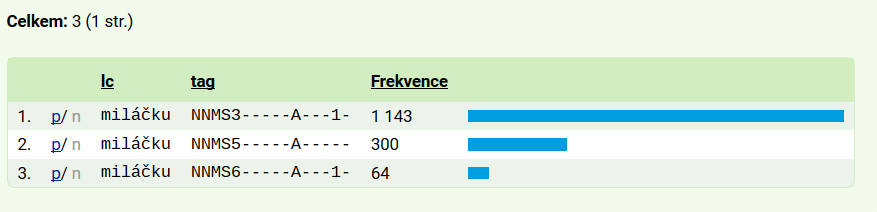
Jak odstraníme řádky, na nichž budou v pozici KWIC tvary oddělené interpunkcí *(<, miláčku, >*  nebo <*, miláčku.>* nebo <*. Miláčku,>* )?

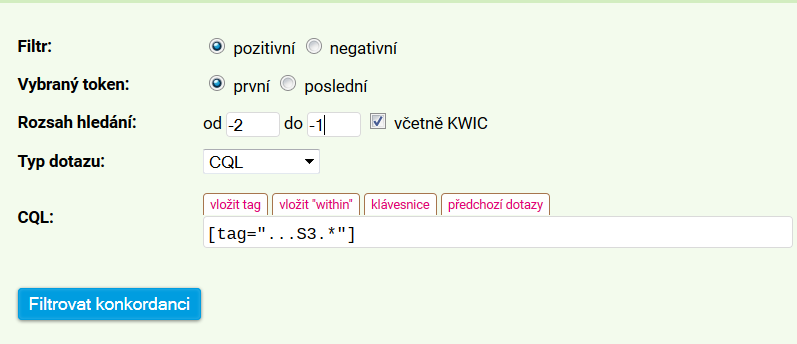
Použijeme pozitivní filtr a zadáme, že chceme vyhledat v intervalu -1,1 příslušné interpunkční znaky:

**[lemma="[,\.:!?]"]**



Získáme seznam 1507 řádků, které patrně budou obsahovat chyby. Pro jistotu je projdeme. Můžeme se také podívat, zda některé tvary nejsou označkovány správně.



Všimněme si, že podíl lokálů je menší, než podíl dativů. A zkontrolujme, zda existují řádky, na nichž by se v levém kontextu vyskytoval tvar v dativu. 



Všimněme si, že tento aspekt je pro desambiguaci nerelevantní. Jména v dativu nerozvíjejí tvar substantiva v pozici KWIC (*svému miláčku*), ani s ním nejsou koordinovány (*Rosťovi, miláčku svému, Vlastíkovi, drahoušku mému, ...*).

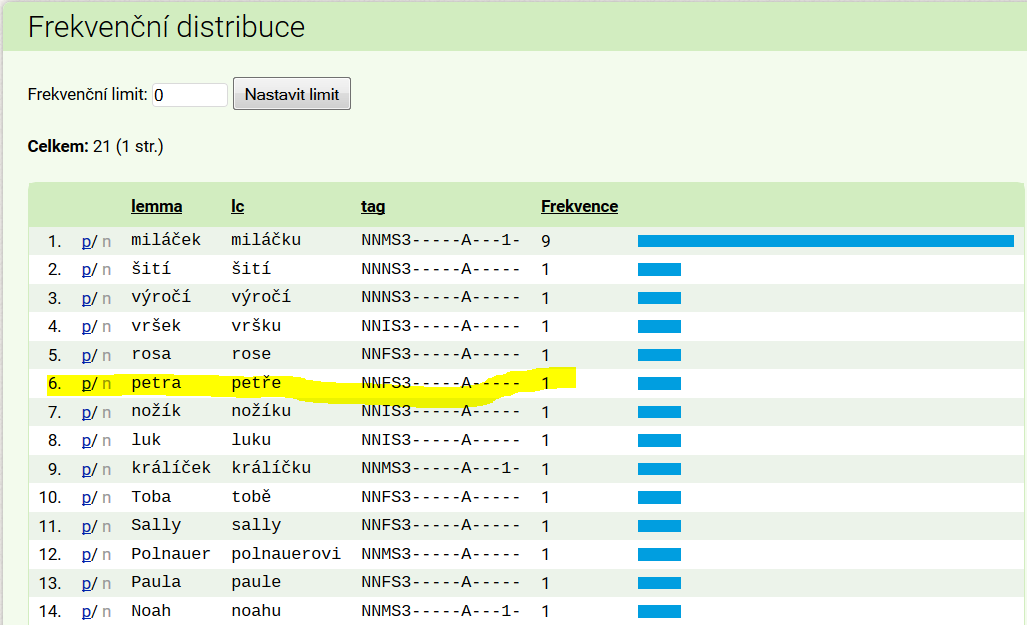
Nicméně můžeme se pokusit ještě více specifikovat dotaz.

Vrátíme se zpět a zadání zpřesníme:





Všimněme si, jak chyby v disambiguaci vzájemně souvisejí:



Vrátíme se a odfiltrujeme všech 1507 řádků z 1557 a 50 řádků uložíme, projdeme a označíme chyby v lemmatizaci.

Správně tagované jsou pouze 3 doklady, ve všech ostatních případech jde o chybně otagované vokativy.



[**https://kontext.korpus.cz/view?q=~D4I02CI0;fromp=1;corpname=omezeni%2Fsyn2010&viewmode=kwic&pagesize=40&attrs=word%2Clemma%2Ctag&attr\_allpos=kw&ctxattrs=word&structs=p%2Cg%2Cerr%2Ccorr&refs=doc%2C%3Dopus.nazev**](https://kontext.korpus.cz/view?q=~D4I02CI0;fromp=1;corpname=omezeni%2Fsyn2010&viewmode=kwic&pagesize=40&attrs=word%2Clemma%2Ctag&attr_allpos=kw&ctxattrs=word&structs=p%2Cg%2Cerr%2Ccorr&refs=doc%2C%3Dopus.nazev)

**ÚKOL na 13. 4. 2015**

**Ze seznamu**



**Dle abecedy si každý student vybere jeden tvar (začneme s tvarem *duchu* a budeme pokračovat). Odstraní případy mezi interpunkcí a zbytek konkordančních řádků projde a vyznačí chyby v lemmatizaci (uloží seznam konkordančních řádků a vyznačí je). Seznam s vyznačenými chybami odevzdá do 18. 4. 2016.**