**Gramatika a korpus II plin032**

**Středa: 7.30-9.00 G13**

9. 3. Rozbor dú.

Efektivní vyhledávání homonym slovesných tvarů pomocí počítačových nástrojů I. (Koupil jsem si paštiku, …)

Mějme větu, které se většina Čechů spontánně zasměje:

***Koupil jsem si paštiku. Bylo na ní napsáno: zaječí. A neječí a neječí.***

Proč je tato věta směšná? Inu proto, že se v ní vyskytuje jedno jediné slovo/ jeden jediný slovní tvar, který lze číst dvěma způsoby, tedy tvarové homonymum (více Petkevič 2015).

Mluvčí zvolil jedno z možných čtení a stěžuje si na důsledek své volby.[[1]](#footnote-1)

Formu, která je příznačná pro uvedené příklady jazykové komiky, můžeme gramaticky popsat:

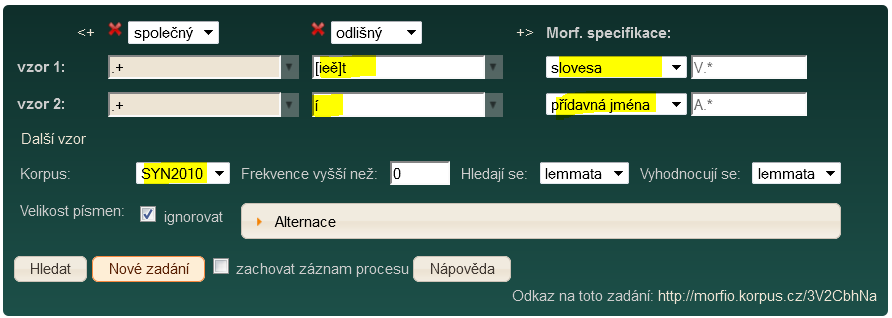
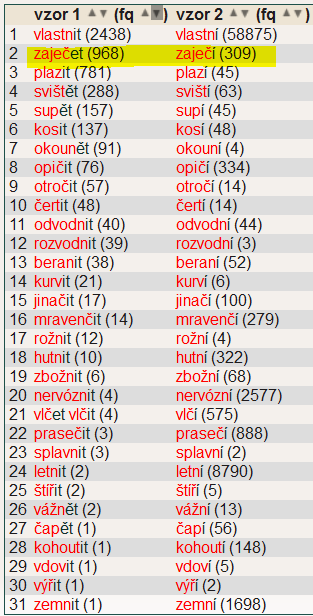
1. *zaječí=* sloveso 3. osoby singuláru indikativu prezentu aktiva slovesa *zaječet*; negované sloveso *(ne)ječet* (=nevydávat výrazný zvuk) příbuzné se slovesem *zaječet* (=vydat výrazný zvuk).
2. *zaječí =* tvar adjektiva měkkého (*zaječí*=příslušející drobnému hlodavci *zajíci*) v několika pádech a rodech;

**Jsou podobné případy pro češtinu typické, nebo jde jen o náhodu, kolik je podobných případů?**

K tomuto účelu můžeme použít jazykový korpus. Jak? Nástroj *Morfio* umožní vyhledat v korpusu o rozsahu 100 milionů slovních tvarů[[2]](#footnote-2) všechny tvary, které lze definovat a) na základě jejich formy a b) na základě jejich gramatických vlastností. Slovní formulace zadání by mohla vypadat následovně:

Chceme vyhledat dvojice (a) sloveso + (b) adjektivum takové, že pro ně platí, že jsou identické v příslušných tvarech. V případě slovesa víme z gramatiky, že (a) slovesa, která ve 3. osobě singuláru indikativu prezentu aktiva končí na *í*, jsou slovesa, která v infinitivu končí na *[ieě]t*. V případě (b) adjektiva víme, že jde o měkké adjektivum, takže slovníkový tvar a řada dalších tvarů končí na *í*.

**Formální zadání při práci s aplikací *Morfio* vypadá následně:**

Výsledkem hledání v korpusu je seznam 31 dvojic. 

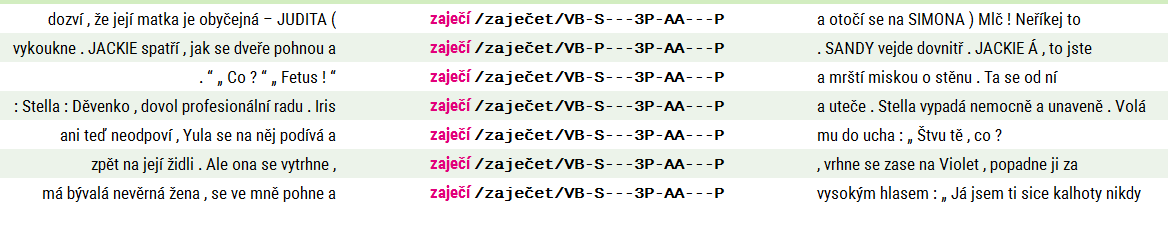
Ke každému členu dvojice můžeme kliknutím zobrazit konkordanční seznam výskytů v kontextech. Seznam je seřazen podle frekvence (čísla v závorkách) slov v prvním sloupci. Frekvence je relevantní pro představu o tom, zda jde o slovo v češtině obvyklé/neobvyklé. I to je zajímavá informace[[3]](#footnote-3). Zamysleme se třeba nad tím, jak bychom interpretovali syntagma ‚*kajmanka supí‘*?

Viděli jsme, že počítačová aplikace *Morfio* užívala k vyhledávání znalosti o české gramatice (znalost slovních druhů). Ty se v korpusech češtiny nevzaly jen tak z ničeho nic. Byly do textů, z nichž je složen korpus, dodány (v procesu **automatické morfologické analýzy**, tedy závisejí na **použitém morfologickém slovníku a použité metodě/metodách disambiguace**).

Podívejme se nyní, jak si s případem dvojice (a) *zaječí/zaječet* a (b) *zaječí/zaječí* poradila automatická morfologická analýza. Na obrázku vidíme konkordanční seznam tvaru *zaječí* spolu s lemmaty a gramatickými tagy. První písmeno v tagu je *A* – adjektivum, nebo *N* – substantivum (název obce) a dále *V* – sloveso. Vyžlucené řádky na prvním seznamu jsou případy, kdy jde o sloveso, ovšem ve značce je chybně uvedeno, že jde o adjektivum a tudíž i lemma je chybně.



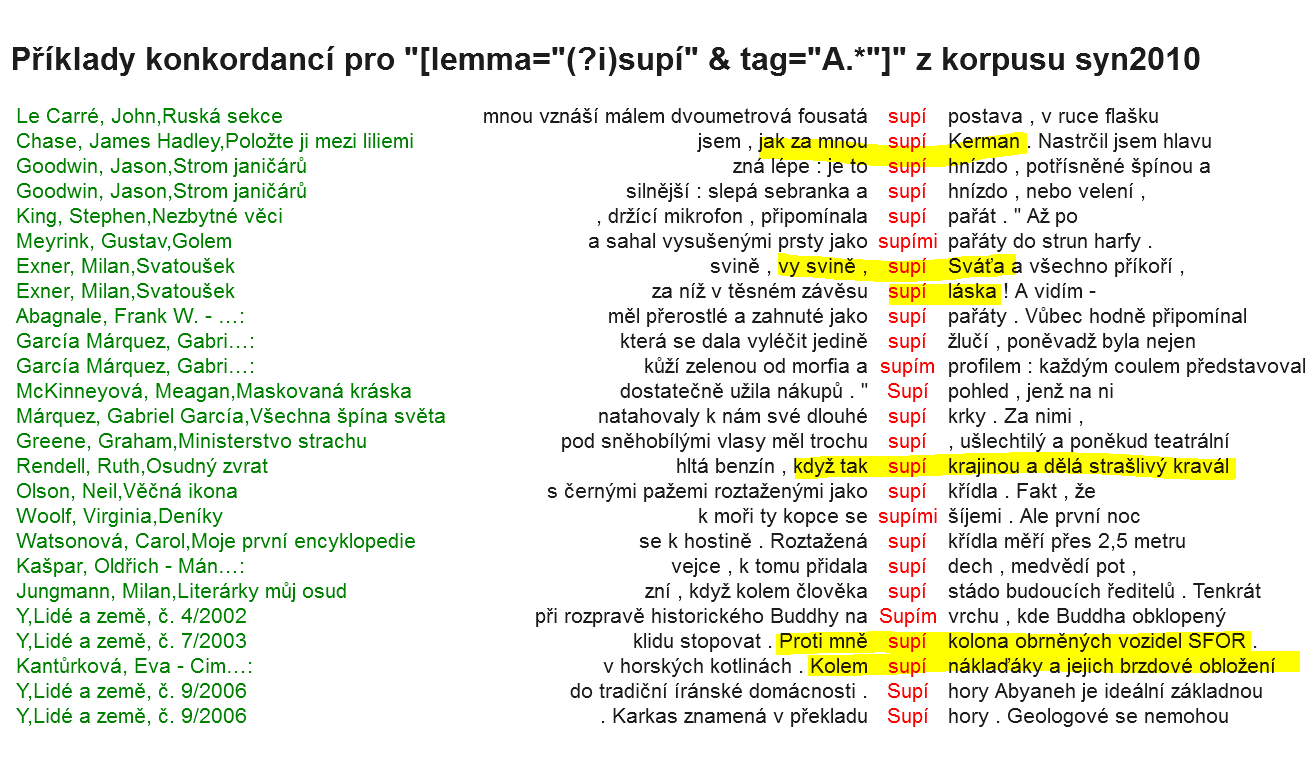
Podívejme se nyní, jak vypadají konkordance případů, které byly označkovány jako slovesa (*V*).



Zjistíme, že se chyby nevyskytují, nicméně konstatujeme, že jako sloveso je tvar označkován pouze 7krát, přičemž v 25 výskytech na předešlém seznamu jsme našli 6 chyb. Je tedy vidět, že disambiguace homonymií sledovaného typu je neradostná.[[4]](#footnote-4)

V případě homonym bývá dodání jednoznačné interpretace „tvrdým oříškem“ a mnohdy „kamenem úrazu“.

**Sledujme problémy značkování plynoucí z prokazatelné homonymie a všímejme si možností vylepšení.**

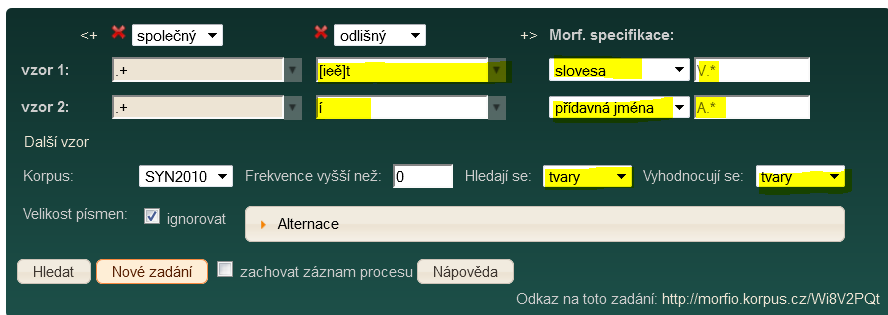


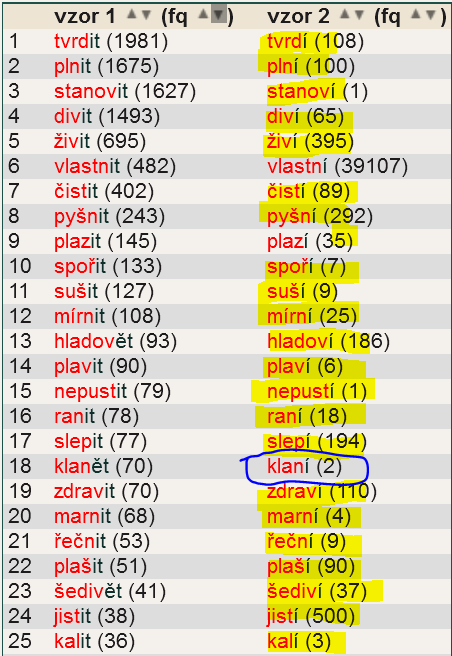
**Zamysleme se, zda existují/mohou existovat v češtině další dvojice tohoto typu. Pokud ano, proč jsme je nenašli a jak můžeme na základě pozorování korpusových dat ověřit správnost/nesprávnost našich předpokladů?**

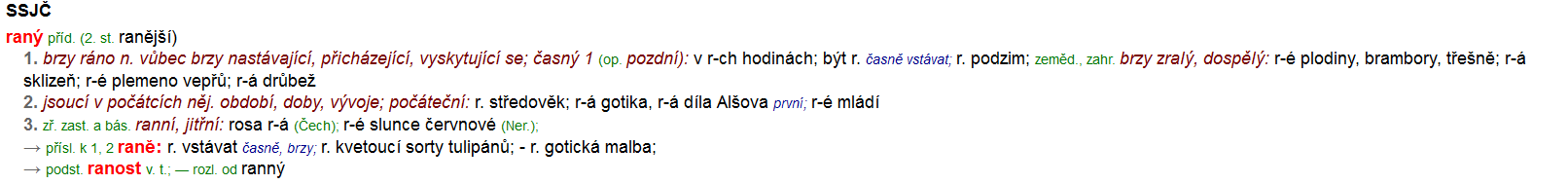
**Lingvisticky** vzato je hledané slovesné lemma (.\*[ieě]t) **lemmatem produktivní slovesné třídy** (**IV. tř. vzory *prosit, trpět, sázet***). Dále nám **lingvistická intuice** napoví, že produktivita třídy se mimo jiné opírá o tvoření **sloves od adjektiv s významem** **činit/stávat se takovým jako motivující adjektivum** (viz dvojice ***vážnět/vážní*** bez ohledu na doklady, které nejsou pro naši intuici nijak průkazné).

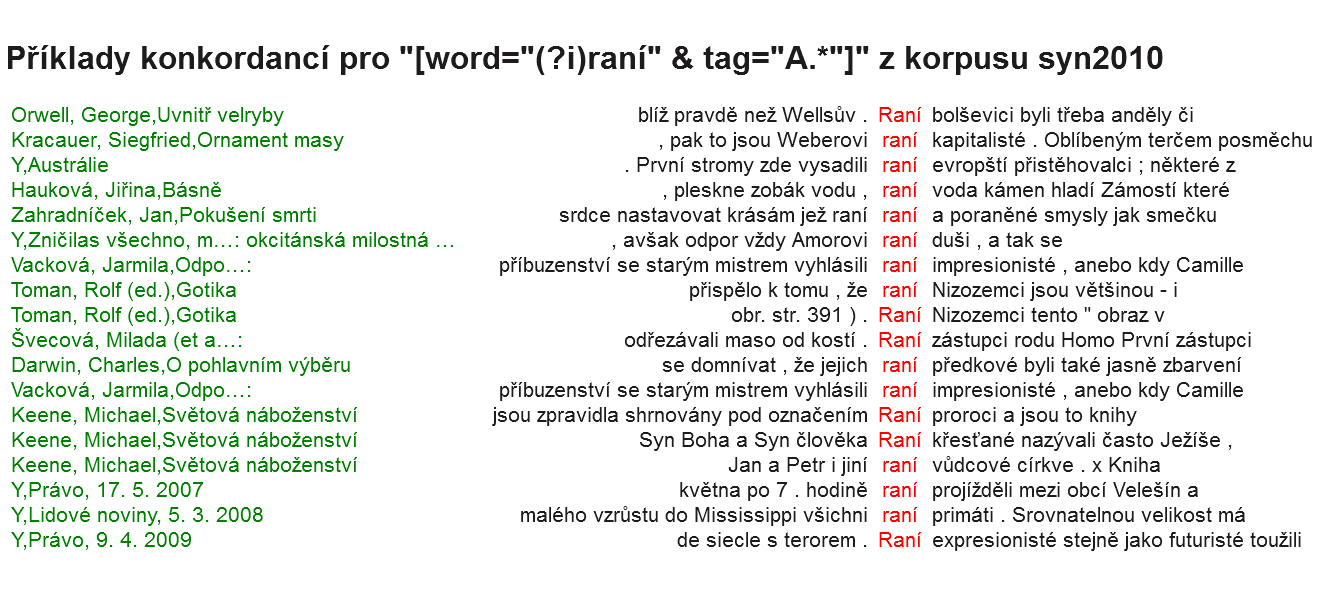
**Lingvisticky vzato** nám ovšem pozorování této dvojice napoví, že bylo **chybou vyhodnocovat pouze lemmata**. Proč? Inu proto, že **homonymní tvar nemusí mít pouze adjektiva měkká** (jako je ***zaječí***), ale i **tvrdá**, u nichž je homonymní **nominativ plurálu maskulin životných** (tedy ***vážní*** jakožto tvar adjektiva ***vážný***).

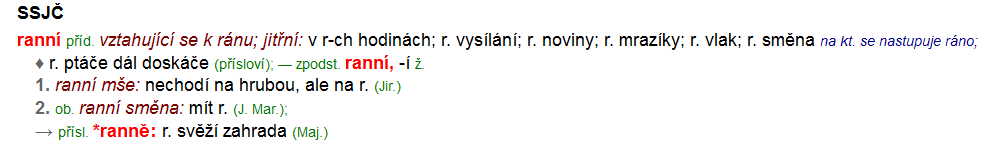
V aplikaci ***Morfio*** budeme tedy vyhodnocovat tvary a získáme 119 dvojic, přičemž vidíme, že tvrdých adjektiv je mezi nimi podstatné množství.

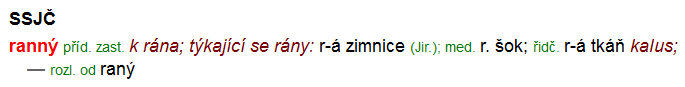




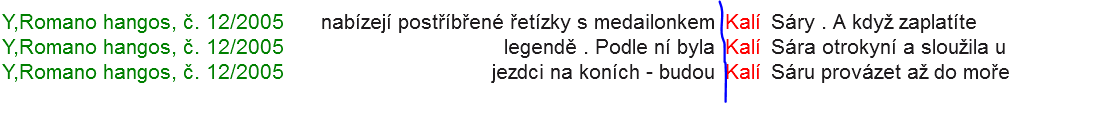




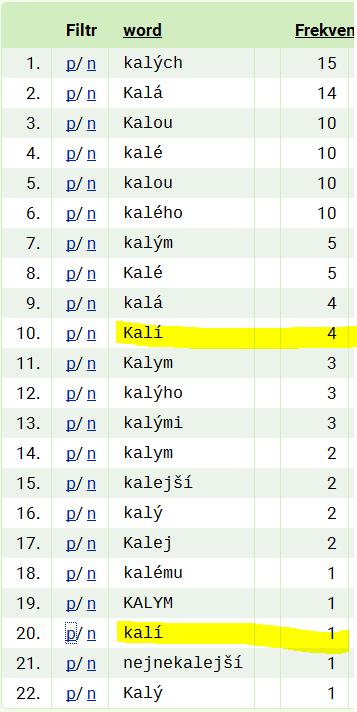


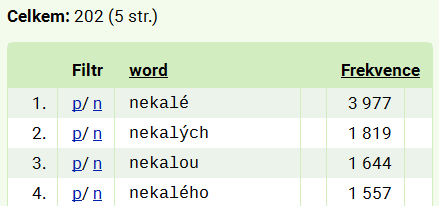


**Chyba: *klaný,* ale i *kalý.***



**Argumenty proti uvádění adjektiv typu *kalý* ve slovníku.**





**16. 3. Dú: Analýza zadaných dvojic homonymních tvarů z hlediska disambiguace.**

1. Mluvčí byl zklamán ve svém očekávání a my se mu smějeme. Jak to? Větě: *„Koupil jsem si robot, bylo na něm napsáno: zaječí. A neječí a neječí.*“ se ovšem nesmějeme. Proč? [↑](#footnote-ref-1)
2. Budeme pracovat s korpusem SYN2010 (http://wiki.korpus.cz/doku.php/cnk:syn2010). [↑](#footnote-ref-2)
3. Slova s frekvencí 1 jsou velmi neobvyklá, o to mohou být pro poznání jazyka zajímavější. Např. sledujeme-li kontext slovesa *výřit*: *… jak se to milostně zubí! S jakou by rozkoší zatkly mi asi do těla zuby. Hahaha to jsi ty drahoušku milý, a máš tak samoten* ***vířit*** *Počkej mne trochu začneme spolu chvíli. To budou na nás* ***výřit****! …*, zjistíme, že jde o slovní hříčku, neboť v bezprostředním kontextu se nachází sloveso *vířit =* ‚dělat *vír*, rychle se krouživě pohybovat‘, zatímco sloveso *výřit* není překlep, ale okazionálně utvořené sloveso napodobující kolokviální sloveso *vejrat* = ‚dívat se udiveně jako *výr*‘. Jde o homofonii (v češtině slova lišící se *i/y* a *í/ý* znějí stejně), která je rovněž častým zdrojem českého humoru. [↑](#footnote-ref-3)
4. Výše uvedený seznam vygenerovaný aplikací *Morfio* je také omezen chybami v disambiguaci, protože pracuje na základě zadání, v němž je užito morfologické značkování (informace o slovních druzích hledaných lemmat). [↑](#footnote-ref-4)