Základy využití korpusu v praxi **cjbb75**

**Středa: 9.10-10.50 G13**

**3. 5. 2017:**

**ŘEŠENÍ**

1. Končí-li tvar slova v češtině na ***–ové***, pak jde nutně o tvar životného maskulina. Platí toto tvrzení ? Odpovězte nejdříve, a pak ověřte svoji odpověď v korpusu SYN2015. Součástí odpovědi musí být formulace dotazu, pomocí kterého byla získána data, na jejichž základě odpovídáte.

Neplatí, protože toto zakončení mají pravidelně tvary adjektiv s lemmatem na *.\*ový*. Pokud zadáme dotaz [lc=".\*ové"], zjistíme, přes frekvenci pos, že jsme se nemýlili a že tvary adjektivní flexe, a s nimi i zakončení *–ové* mají i slova tagovaná jako zájmena (*takové*) a číslovky (*třináctimilionový*).

1. Dokažte, že není třeba, abyste věděli, co označuje vyžlucené substantivum (abyste četli H. Pottera) k tomu, abyste byli schopni na základě zadání samého vyřešit následující úkol. **Na základě uvedených vět určete rod a vzor vyžluceného substantiva.**

Mudlové (v originále Muggles) je označení pro lidi, kteří neumí kouzlit a jsou nekouzelnické nebo smíšení krve. Mudlaje tu pro vás.

Maskulinum vzoru *předseda.* Ve druhé věte je tvar nom. sg. a v první nom. pl., obojí lze vyčíst ze syntaktické role, kterou plní. V prvním případě by teoreticky mohlo jít o adjektivum (viz výše), ale vzhledem k tomu, že v závorce je tvar, ktyrý má morfologický příznak angl. pl. substantiv *–s*, lze tuto možnost vyloučit.

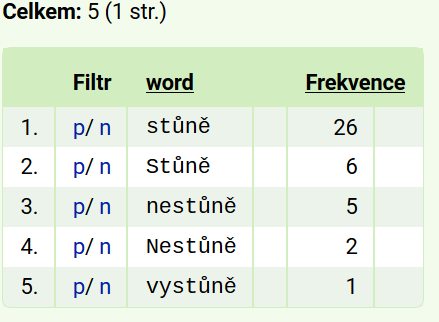
Možnost 2: jde o různá lemmata:



Pak se v první větě jedná o slovo podle *předseda* a ve druhé podle *předseda a/nebo žena.*

1. Hláska *e* má v češtině dvě grafické realizace: *e/ě*. Existuje český slovesný tvar, který končí na grafickou realizaci *ě*? Odpovězte nejdříve, a pak ověřte svoji odpověď v korpusu SYN2015. Součástí odpovědi musí být formulace dotazu, pomocí kterého byla získána data, na jejichž základě odpovídáte.

Pokud jste odpověděli NE, tak je to špatně. Pokud ANO, ale nevím, a pak jste zadali dotaz: [tag="V.\*" & lc=".\*ě"], pak jste patrně zjistili, že jde buď o tvary přechodníků (*nemluvě, nechtě, …*), ale narazili jste i na tvar určitý (*stůně*). N-filtrem jste přechodníky odstranili (N-filtr <0 0> [tag="Ve.\*"] a zjistili jste, že automatická analýza patrně rozpoznává potenciální (nekodifikované tvary) přechodníků přítomných od dokonavých sloves, které ovšem značkuje jako přechodníky minulé. N-filtrem jste přechodníky odstranili (N-filtr <0 0> [tag="Vm.\*"] a zjistili jste, že v korpusu SYN2015 se vyskytuje pouze jeden kořen, od něhož určitý tvar slovesa končící na *–ě.*



Pokud jste na vše přišli bez korpusu, gratuluju.

1. *Místo vašich schůzek sis měl lépe vybírat.* Kolik interpretací má slovní tvar *místo* v uvedené větě? Podívejte se do korpusu SYN2000 a zjistěte, jak jsou desambiguovány interpretace slovního tvaru místo. V případě, že naleznete chybné interpretace, pokuste se navrhnout postum, kterým byste je detekovali.

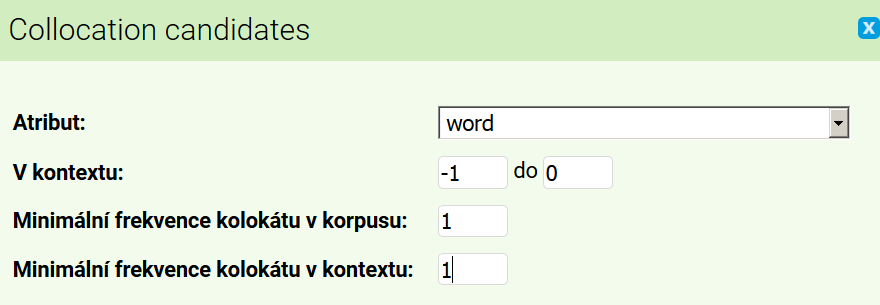
Podíváme se např. na případy označkované jako [lc="místo" & tag="R.\*"] R (předložka). Vidíme chyby. Např. případy jako : Naše družstvo nemělo motivaci hrát ve finále, chtělo se poprat o třetí **místo/R.\*** s Trutnovem. Je jasné, že posloupnost dvou předložek je patrně něco, co by se v jazyce nemělo vyskytovat. Použijeme P-filtr <1 1> [tag="R.\*"] a podíváme se na nalezené doklady a všimneme si, že naše intuice není správná, neboť se objevují i správně označkované doklady jako: **Místo/R.\* do** sněhobílého magnetu míří často střelci do tmavých sítí, různě vzorovaných v klubových barvách a podobně. Podobně lze vyloučit posloupnost „předložka+sloveso“, tedy P-filtr <1 1> [tag="V.\*"] . Výsledkem je nalezení chyb v desambiguaci.



1. Názvy zlomků (substantiva tvořená ze základů číslovek základních sufixem *–ina*, např. *pětina, desetina, dvanáctina, padesátina, …*) a názvy jubileí (*padesátiny, třicetiny, šestnáctiny, …*) mají část tvarů společných (plurálové tvary), což způsobuje nepříjemnosti, chceme-li mít oba významy odlišeny na rovině automatické morfologické analýzy. V korpusu SYN2015 vyhledejte relevantní lemmata a pozorujte chyby v desambiguaci. Navrhněte postup, jak nalézt případy, které by bylo možné zjednoznačnit na základě blízkého kontextu.

Dotaz, jímž je možné vyhledat deriváty na -*ina* od číslovkových základů (bez ohledu na to, zda jsou značkovány jako číslovky, nebo jako substantiva): [lc=".\*(((deset)|(((jeden)|(dva)|(tři)|(čtr)|(pat)|(šest)|(sedm)|(osm)|(devate))(náct))|(((dva)|(tři)|(čtyři))(c[áe]t))|(((pa)|(še)|(sedm)|(osm)|(deva))(desát))|(set)|(tisíc)|(mili[oó]nt))in)|(((deset)|(((jeden)|(dva)|(tři)|(čtr)|(pat)|(šest)|(sedm)|(osm)|(devate))(náct))|(((dva)|(tři)|(čtyři))(c[áe]t))|(((pa)|(še)|(sedm)|(osm)|(deva))(desát))|(set)|(tisíc)|(mili[oó]nt))in((y)|(ám)|(ách)|(ami)))"]

Podíváme-li se na seznam lemmat, tak uvidíme mezi frekventovanými lemmaty lemma **osmnáctina** a **šestnáctina**. Podívejme se na kolokační kandidáty (k, jeho, oslavit, sladký), které se vyskytují v obou případech tam, kde jde o chybnou interpretaci (číslovka dílová Cr).



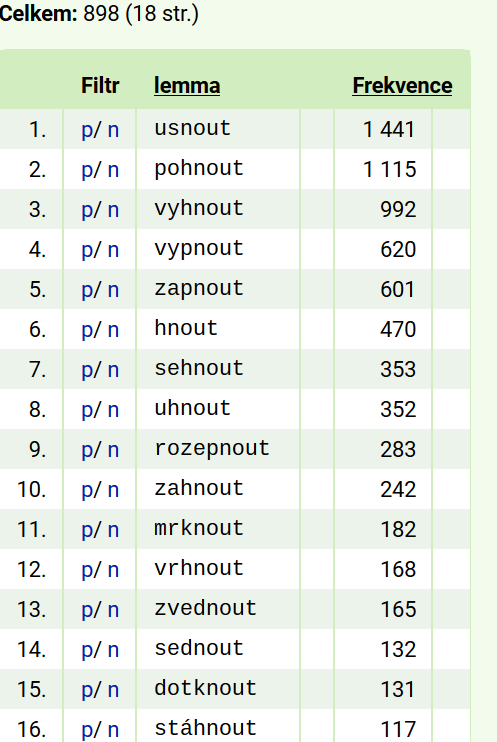
1. V české flexi platí pravidlo, které tvrdí, že substantiva skloňovaná podle vzoru *růže* mají v genitivu pluláru buď koncovku *–í* (*růži*), nebo kocovku *–0* (*vesnic*). Dokázali byste odpověď na následující otázku, kterou by vám položil někdo, kdo se chce naučit česky (cizinec) a která by zněla: Znamená to, že se u každého českého slova, které je ženského rodu a končí v nominativu sg. na *–e*, musím naučit, jak vypadá jeho tvar v genitivu plurálu, nebo mi můžete ještě pomoci nějakým dalším rozšířením pravidla uvedeného výše? Odpovězte nejdříve, a pak ověřte svoji odpověď v korpusu SYN2015. Součástí odpovědi musí být formulace dotazu, pomocí kterého byla získána data, na jejichž základě odpovídáte.

Odpověď zní? Nikoliv, existuje omezený seznam slov a otevřený seznam slov definovatelných formálně. Jak oba seznamy získat? Dotazem na korpus: [lemma=".\*[eě]" & tag="N.FP2.\*" & lc!=".\*í"]. Jde o slova tvořená sufixem *–ice* (těch je v seznamu nejvíce). Odstraníme je pomocí n-filtru [lemma=".\*ice"]. Dále vidíme substantiva na *–yně*. Odstraníme je pomocí n-filtru [lemma=".\*yně"]. Zbývá 11 lemmat českých substantiv, která jsou frekventovaná a u nichž je třeba si zapamatovat výjimku (nulovou koncovku v gen. pl.). Ostatní můžeme popsat formálně (lemma/nom. sg nončí na *–ice/-yně*).

1. Dokázali byste odpověď na následující otázku, kterou by vám položil někdo, kdo se chce naučit česky (cizinec) a která by zněla: Učil jsem se, jak se tvoří tvar l-ového příčestí od sloves, který musím znát, abych uměl vytvořit větu v minulém čase a v podmiňovacím způsobu. Všiml jsem si, že někdo říka, že ***počítač vypl*** a někdo, že ***počítač vypnul.*** Jak poznám, když se učím nové sloveso, které má v infinitivu zakončení na *–nout*, jak má vypadat tvar příčestí? Odpovězte nejdříve, a pak ověřte svoji odpověď v korpusu SYN2015. Součástí odpovědi musí být formulace dotazu, pomocí kterého byla získána data, na jejichž základě odpovídáte.

Pokud jste z paměti nevydolovali nějaké poučky o třídách a vzorech, tak můžete rovnou začít hledeat odpověď v korpusech a porovnat data vzhledaná dotazem [lemma=".\*nout" & tag="Vp.\*" & lc=".\*(nul)|(nul[aoiy])"] a dotazem [lemma=".\*nout" & tag="Vp.\*" & lc!=".\*(nul)|(nul[aoiy])"]. Možná se vám při pozorování prvního seznamu něco o třídách a vzorech vzbaví. Ne-li, pak se podívejte, zda nenajdete nějaký společný rys, a to především tehdy, když je vyloučeno, aby tvar měl i variantu bez *–nu-*. Mužeme si všimnout, že varianty bez *–nu-* nemohou mít slovesa, u nichž před *–nu-* předchází samohláska (nebo slabičná sonora). Odfiltrujeme je takto. N-filtr <0 0> [lc=".\*[aáeěéiíoóůuyýrlm]((nul)|(nul[aoiy]))"].

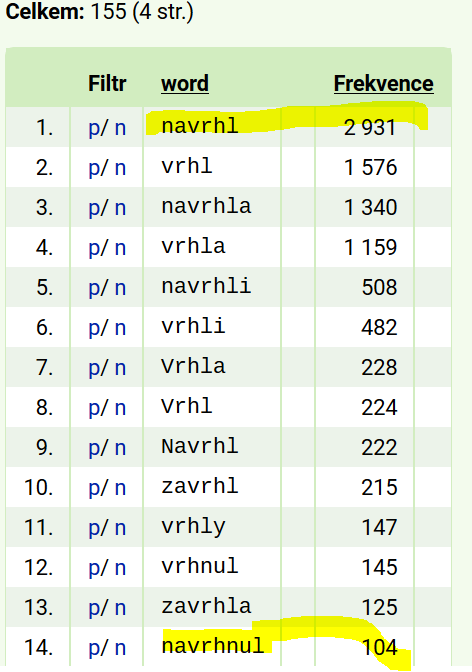
Porovnejme seznam lemmat, k nimž jsme našli tvary s *–nu-*



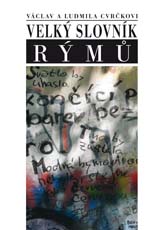
a bez *–nu-*



všimněme si sloves **vrhnout** (1. seznam) a **navrhnout** (2 seznam) a vyhledejme tvary v korpusu takto: [lemma=".\*vrhnout" & tag="Vp.\*"].



(Nástroj SyD).

1. V rámci publikací ÚČNK vyšla i tato publikace: Cvrček, V., Cvrčková Porkertová, L.: Velký slovník rýmů. Nakladatelství Lidové noviny, Praha 2011. Její anotace zní: Příručka rýmů usnadní psaní jakýchkoli rýmovaných textů, což je úkol, před kterým **stojí nejen začínající básníci, ale také autoři písňových textů, reklamních sloganů a mnozí další. Slovník zároveň může sloužit jako doplňková pomůcka při výuce literární teorie na základních a středních školách. Svým rozsahem i způsobem zpracování se jedná o unikátní projekt, který na současném knižním trhu nemá obdoby (posledním takovým pokusem byl Puchmajerův Rýmovník z roku 1824).

Nemáte slovník po ruce. Zaexperimentujte se svojí jazykovou vynalézavostí a s korpusem. Najděte (po paměti) rýmy slov *běhna*, *loutna, pouť*. Vyhledejte v korpusu synv4 doklady, které mohou být rýmem slova/tvaru a porovnejte svoji jazykovou vynalézavost s tím, co lze najít v korpusu. Použili jste jako pomůcku k vyhledávání v paměti nějaký rýmovaný text, který umíte zpaměti? Pokud ano, uvědomte si, nakolik byl autor textu v otázce rýmu vynelázavý (rýmové slovo má v korpusu velkou – malou – žádnou frekvenci).

vyhledáme slovní tvary .\*[eě]hn[aá], .\*ou[td]n[aá], [lc="(.ou[dtďť])|(..ou[dtďť])"]

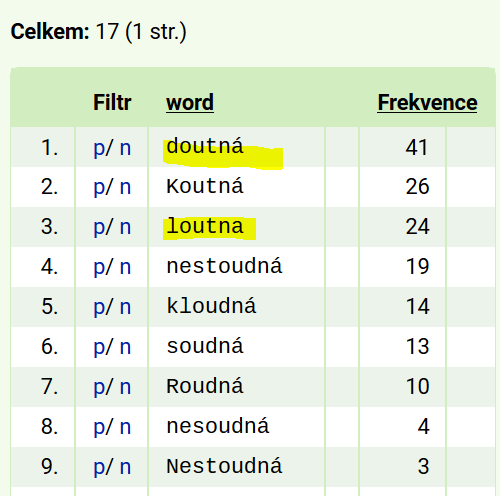


Hnusím si, nákazy se střehna

vojandu, byť se křížem žehná,

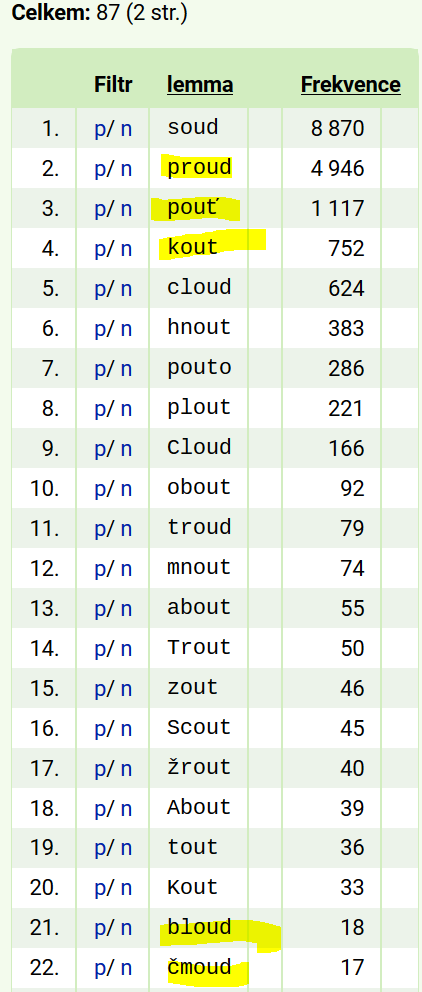
dráždí mne stará vlezlá běhna,

hoch obdivující svá stehna.



Večerem dům zvolna doutná,

k tanci tiše hraje loutna …



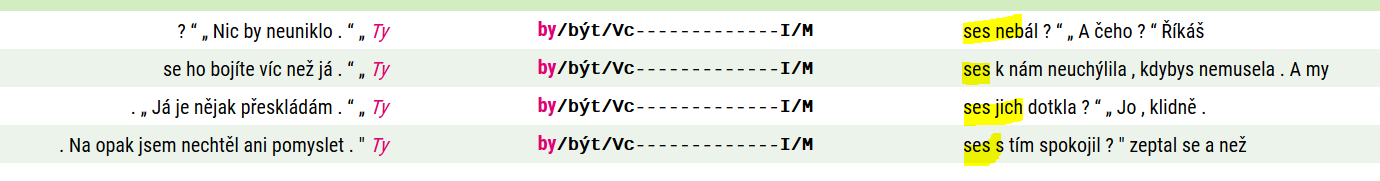
Přijela pouť a lidí proud   
zaplnil na návsi kdejakej kout   
přijela pouť hudba a čmoud   
marcipán má zas chuť náramnou   
  
Přijela pouť no a já bloud   
srdce tvý perníkem chci odemknout   
jestli jseš sám přijeď hned k nám   
protože právě dnes pouť u nás začíná

1. Tabulkové paradigma tvarů *by*, které v češtině slouží ke tvoření kondicionálu bývá v českých mluvnicích uvedeno následovně:

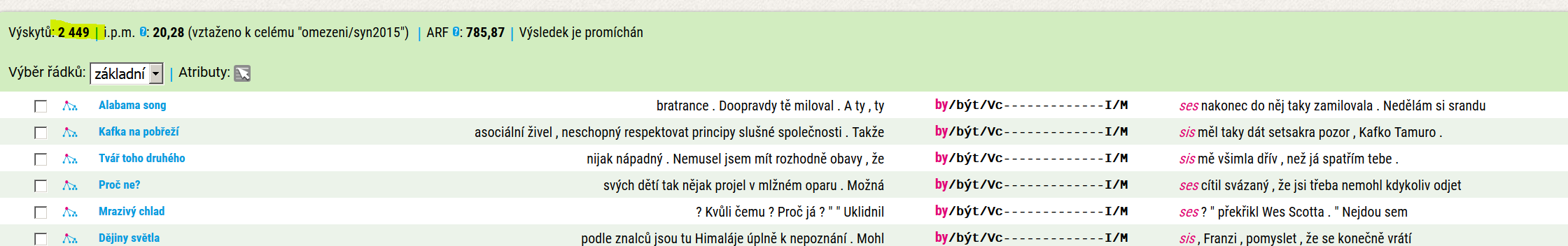
|  |  |  |  |
| --- | --- | --- | --- |
| osoba | singulár | osoba | plurál |
| 1. | ***bych*** | 1. | ***bychom*** |
| 2. | ***bys*** | 2. | ***byste*** |
| 3. | ***by*** | 3 | ***by*** |

Bez komentáře k tabulce, by tudíž následující tvrzení: „**Tvar *by* může být v češtině součástí tvaru kondicionálu 2. osoby singuláru**“ nemělo platit. Platí, nebo neplatí? Pokuste se odpovědět na základě introspekce. Pokuste si ověřit svoji odpověď v korpusu SYN2015.

Pokud si na základě intrspekce nevíme rady, nezoufejme. Vyhledáme v korpusu tvar *by* a všimneme si, že nemá označkováno ani číslo, ani osobu. Pokud by osoba byla nutně 3., pak by tak měla být označkována. Asi nebude. Pokud Vám to ještě nedošlo, zkuste v kontextu <-1 -1> najít P-filtrem zájmeno ty [tag="PP.\*" & lc="ty"].



A máme pravidlo: **Pokud je významové sloveso zvratné, pak při tvoření kondicionálu ve druhé osobě sg. se užívá tvar *by* nikoli *bys* a *–s* se přimyká k zvratné částici/zájmenu *se/si,* tedy *ses/sis*.**

Ještě si to ověřme. Vraťe se o krok zpět a vyhledejte v kontextu <1 1> P-filtrem [lc="s[ei] s"].

A zkusme také najít doklad bys s[ei]s a bys s[ei]. Je to správně česky?



A tohle sem nepatří: Nešla **bys se** mnou do kina.

1. I když je letec **více DiCapriův než Scorseseho** film, je zřejmé, že pro Scorseseho, vždy fascinovaného filmem i zuřivými hrdiny, to byla mnohem více než pouhá zakázka.

Podívejte se na větu z korpusu synv4 Pokuste se najít podobné doklady a zamyslete se nad tím, zda jde o analytické stupňování subjektově posesivních adjektiv.

Kandidáty můžeme najít jako fráze, na které se ptáme:

[lc="(víc.\*)|(méně)|(míň)"][tag="AU.\*"][lc="než"], vhodné doklady si označíme:

