***4.4 Složené tvary slovesné a možnosti jejich vyhledávání v korpusu – pravidla slovosledu v češtině***

**4.4.A Motivační úvod**

V korpusech lze vyhledávat výskyty slovních tvarů (word), je možné zvolit vyhledávání podle dalších atributů (lemma, tag, lc, pos atd.). Kromě jednoduchých slovních tvarů můžeme vyhledávat i tvary složené, víceslovné výrazy (fráze). Na tomto místě pojednáme o postupech, které lze použít pro vyhledávání složených slovesných tvarů. Česká slovesná flexe je poměrně chudá na jednoduché slovesné tvary. Kromě indikativu prézentu (futura) aktiva, imperativu a několika sloves tvořících synteticky futurum prefixem *po-* se všechny tvary času, způsobu i rodu vyjadřují analyticky. Chceme-li získat z korpusů podklady pro zkoumání vyjadřování kondicionálu, opisného futura, pasiva, minulého času (1. a 2. osoby), musíme vzít v úvahu: a) z jakých složek (tvary pomocných a významových sloves) se jednotlivé formy času, způsobu a rodu skládají a b) slovosledná pravidla, jimiž se řídí možná pořadí jednotlivých komponent.

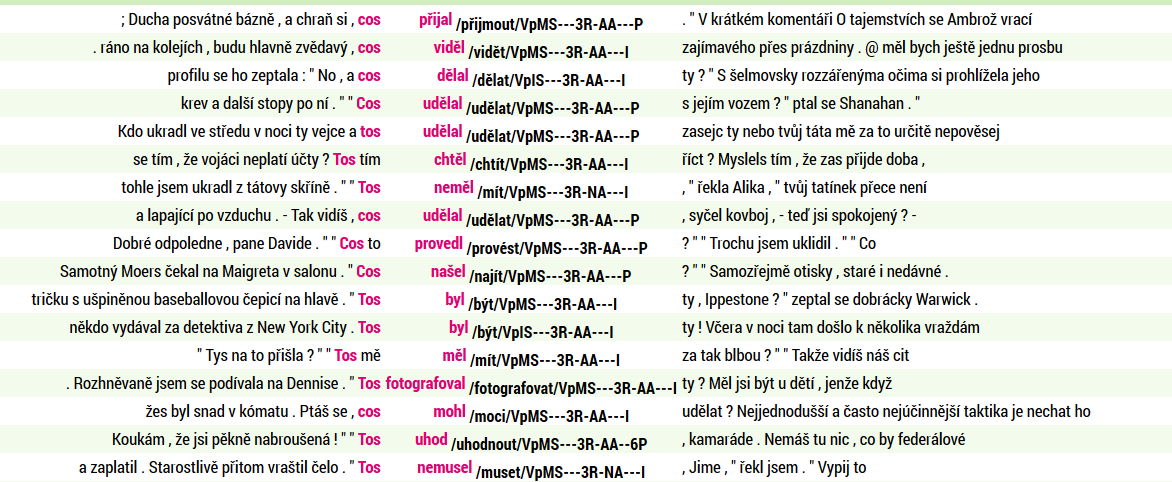
**4.4.B Nastínění problému[[1]](#footnote-1)**

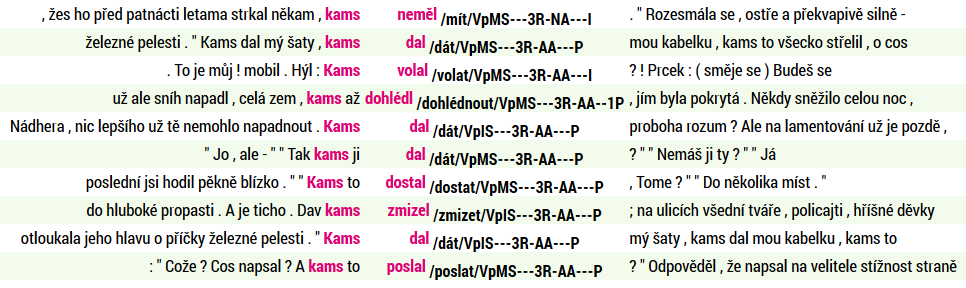
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| singulár | příčestí l-ové | pomocné sloveso | volný morfém *-s* | plurál | příčestí l-ové | pomocné sloveso |
| 1. osoba | přišel | jsem | – | 1. osoba | přišli | jsme |
| 2. osoba | přišel | (jsi/jste) | (-s) | 2. osoba | přišli | jste |
| 3. osoba | přišel | – | – | 3. osoba | přišli | – |

Povinnou složkou tvaru minulého času indikativu aktiva 1. osoby je v češtině tvar l-ového příčestí slovesa významového a určitý tvar pomocného slovesa *být* (1. osoba sg. nebo pl.).[[2]](#footnote-2) Povinnou složkou tvaru minulého času indikativu aktiva 2. osoby je v češtině tvar l-ového příčestí slovesa významového a určitý tvar pomocného slovesa *být* (2. osoba sg. nebo pl.), kterou lze v případě 2. os. sg. vynechat a nahradit pomocné sloveso *být* volným morfémem *-s*, který se pojí s tvarem l-ového příčestí významového slovesa nebo s některým jiným slovem ve větě. Povinnou složkou tvaru minulého času indikativu aktiva 3. osoby je v češtině tvar l-ového příčestí slovesa významového a absence určitého tvaru pomocného slovesa *být*.[[3]](#footnote-3)

Budeme-li chtít vyhledat tvary minulého času v indikativu aktiva, pak musíme vzít v úvahu specifika značkování v jednotlivých korpusech. V korpusech SYN2000, SYN2005, SYN2006PUB a v korpusu ORWELL mají všechna l-ová příčestí na 8. pozici – osoba hodnotu X (libovolná osoba (1/2/3)). V korpusech SYN2009PUB a SYN2010 jsou uvedeny hodnoty 1, 2 nebo 3. Většinou jsou správně desambiguovány s ohledem na výskyt tvaru pomocného slovesa *být*. Existují ale i chybně desambiguované případy.

Na následujících obrázcích vidíme doklady chybné desambiguace kategorie osoby (8. pozice) ve značce l-ových příčestí, která jsou součástí složeného slovesného tvaru s volným morfémem *-s* spojeným se zájmeny *co/to* a příslovcem *kam*.





**4.4.C Otázky**

Ať už tedy budeme v případě vyhledávání složených slovesných tvarů pracovat s kterýmkoliv korpusem, budeme muset položit dotazy na jednotlivé tvary osoby, (čísla), způsobu a rodu, vždy s ohledem na možná slovosledná schémata.

**4.4.D Formulace dotazu pro získání dat z korpusů**

V případě vyhledávání tvarů minulého času indikativu aktiva se nám jeví jako nejvhodnější vyjít z tvarů pomocného slovesa *být*. Budeme pracovat s korpusem SYN2010.

V menu **Korpus** zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[lemma=”být“ & tag=“VB.....[12]P.\*“].*

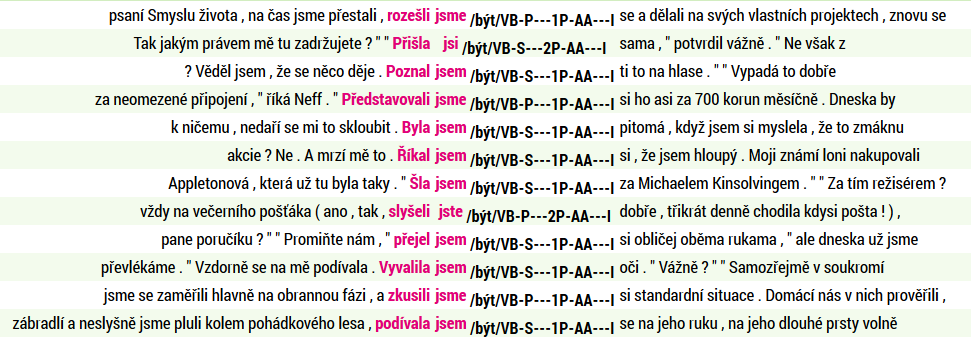
V dalším kroku je třeba zjistit, zda jsou tvary *jsem, jsi, jsme, jste* součástí analytického tvaru minulého času indikativu aktiva, tj. najít v jejich okolí tvar l-ového příčestí a ověřit, že s tvarem pomocného slovesa tvoří hledaný analytický tvar. Zvolíme filtr (pozitivní), rozsah hledání, v němž chceme vyhledávat tvary l-ových příčestí, **Typ dotazu cql** a do dotazovacího řádku zapíšeme dotaz ve formě *[tag=“V[pq].....[12].\*“]*.[[4]](#footnote-4)

Jak rozsáhlý má být zvolený interval?

Bude rozdíl mezi intervalem vpravo a vlevo? Nespouštějme ze zřetele fakt, že tvar pomocného slovesa *být* se chová ve sledovaných případech jako příklonka.

**4.4.E Třídění a pozorování dat získaných z korpusů**

Nejdříve nastavíme rozsah hledání <–1, –1>.



Počet dokladů příslušné slovosledné varianty v korpusu SYN2010 je 269 918. Pokud se v konkordanci vyskytují chyby, pak by k jejich odhalení mohla pomoci následující zkouška. Ponecháme filtr pozitivní, nastavíme rozsah hledání <1,1>,zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”V[pq].....[12].\*“].*

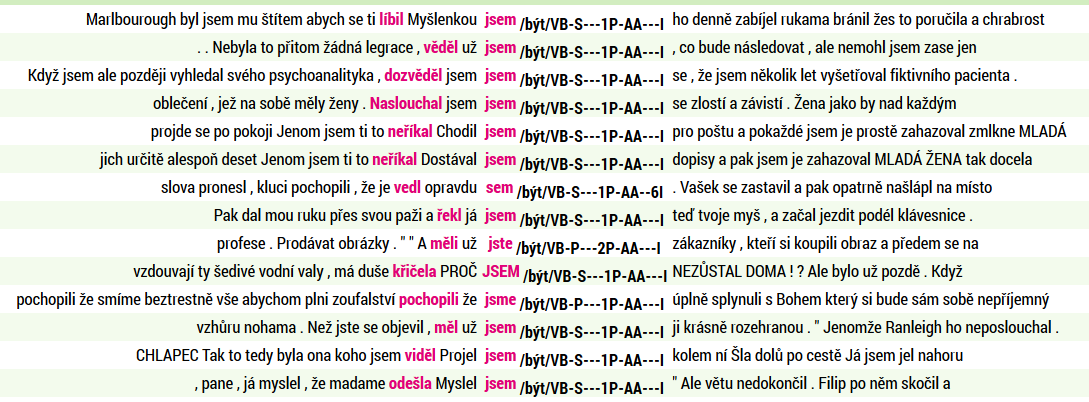
Na obrázku níže vidíme na třetím řádku odspodu chybu (substantivum *Řek* je mylně interpretováno jako l-ové příčestí slovesa *říci*). Na dalších dvou označených řádcích jsou chyby v interpunkci důvodem přegenerovávání. Chyby v interpunkci vidíme i na dalších řádcích, na nich ovšem k přegenerovávání nedochází.



Vrátíme se k původní konkordanci (dva kroky zpět) a nastavíme rozsah hledání <–2,–2>,zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”V[pq].....[12].\*“].*



Pokusíme se odstranit případy, kdy vyhledané řádky neobsahují hledané slovosledné formace. Budeme postupovat následovně: zvolíme filtr (negativní), nastavíme rozsah hledání <–1,–1>,zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”[JZ].\*“]*. Snažíme se odstranit případy, kdy tvar pomocného slovesa a l-ové příčestí netvoří dohromady hledaný složený slovesný tvar. (Správně bychom měli nejprve uvedené případy vybrat pomocí pozitivního filtru, projít všechny řádky a ověřit, zda jde ve všech případech opravdu o případy přegenerovávání, a teprve poté je pomocí negativního filtru odstranit.)

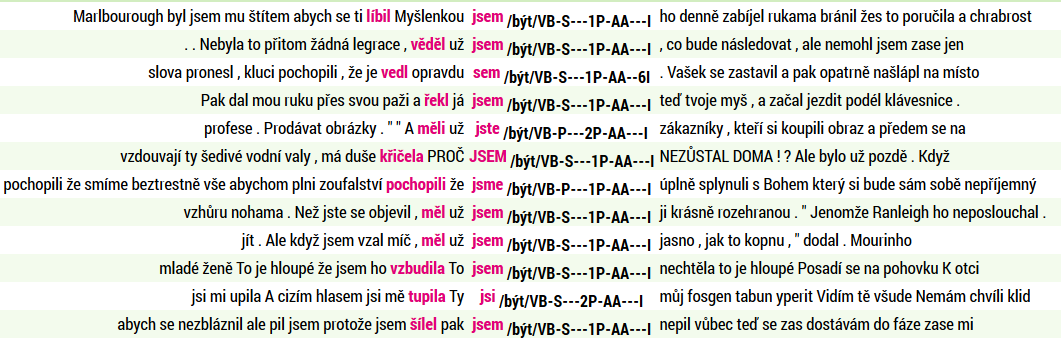


Procházíme-li konkordanci, vidíme, že konkordanční seznam může obsahovat chyby. Je patrné, že kromě příklonek jako *už, my, vy,* které stojí před příklonným tvarem pomocného slovesa *být*, se objevují též řádky s evidentními chybami v morfologickém značkování (*… že je vedl opravdu <sem/být/VB-S---1P—AA—6I> …*), a také chyby v interpunkci zdrojových textů, takže ani pravidlovou disambiguaci není možné opřít o spolehlivá data.

Pokusíme se hromadně odstranit ty případy, kdy vyhledaný tvar l-ového příčestí nepatří k tvaru pomocného slovesa *být*. Budeme postupovat tak, že zvolíme filtr (pozitivní),rozsah hledání <–1,–1>, **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=“V.\*“]*. Z následujícího obrázku je patrné, že se jedná o chyby v originálních textech (především v interpunkci), které nejsou při konverzi textů do formátu korpusu opravovány.

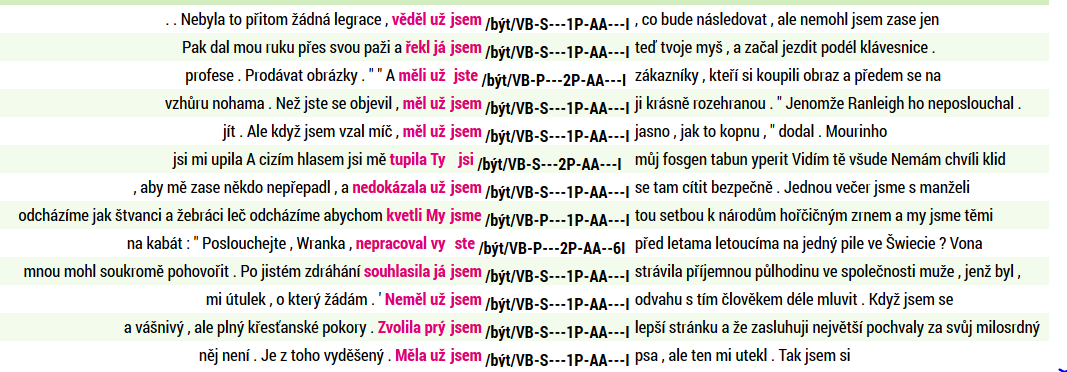


Všechny případy vyloučíme tak, že se vrátíme o krok zpět, zvolíme filtr (negativní), rozsah hledání <–1,–1>, **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=“V.\*“]*.



Ze seznamu zbylých dokladů je patrné, že mezi tvarem l-ového příčestí a tvarem pomocného slovesa *být* mohou stát lemmata osobních zájmen 1. a 2. osoby a slova *už, prý*. Podívejme se a) zda ve všech případech, kdy mezi tvarem l-ového příčestí a tvarem slovesa *být* stojí lemma *já|ty|my|vy|už|prý,* jsou hledané složené tvary minulého času a b) jak vypadají ostatní případy.

1. Budeme postupovat tak, že zvolíme filtr (pozitivní)**,** rozsah hledání <–1,–1>, **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz ve formě *já|ty|my|vy|už|prý*.



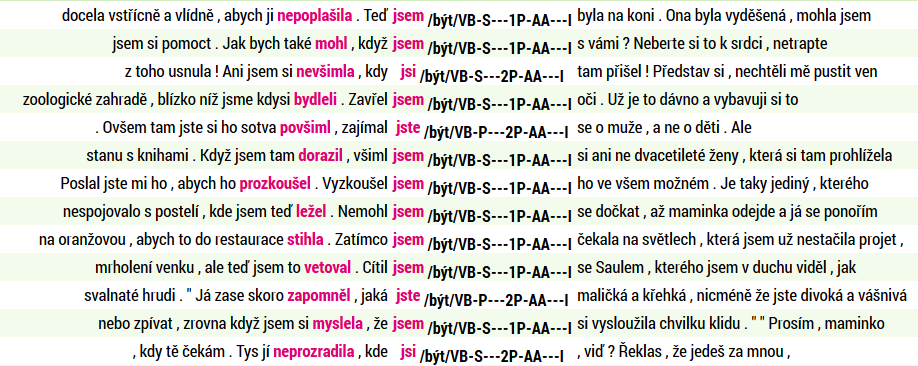
Výsledkem jsou správné doklady analytického minulého času.

1. Vrátíme se zpět a budeme postupovat tak, že zvolíme filtr (negativní)**,** rozsah hledání <–1,–1>, **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz ve formě *já|ty|my|vy|už|prý*.



Výsledkem jsou doklady chybných konkordancí. Chyby jsou mimo jiné způsobeny i chybami v desambiguaci. Např. tvar *že*, před nímž chybí čárka, je interpretován nikoli jako spojka(*J.\**), ale jako (*TT.\** – částice).

Vrátíme se zpět (ke konkordanci tvarů lemmatu *být* označkovaných jako *[lemma=”být” & tag=“VB.....[12].\*“]*), zvolíme filtr (pozitivní), rozsah hledání <–3,–3>, **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=“V[pq].....[12].\*“].*



Zvolíme filtr (pozitivní), rozsah hledání <–2,–1>, **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”[JZ].\*“].*

V zobrazeném konkordančním seznamu je patrné, že vyhledané řádky neobsahují hledané slovosledné varianty. Jedná se o případy jdoucí za hranice větných celků.



Vrátíme se o krok zpět, zvolíme filtr (negativvní), rozsah hledání <–2,–1>, **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”[JZ].\*“].* Pozorujeme, že jde opět především o doklady chyb v interpunkci. Pokusíme se tedy automaticky odstranit tvary, kdy v zadaném intervalu stojí další sloveso.



Zvolíme filtr (negativvní), rozsah hledání <–2,–1>, **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”V.\*“].*

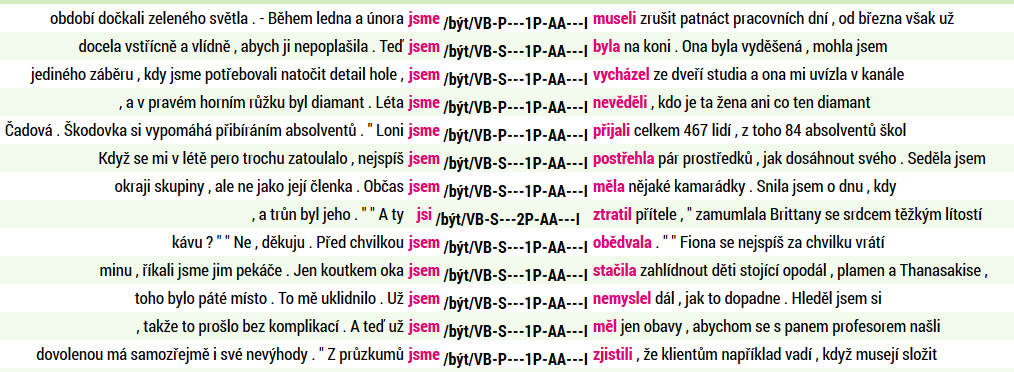


Projdeme ručně seznam zbylých dokladů. Ručním tříděním zjistíme, že všechny nalezené doklady jsou případy přegenerovávání. Důvodem jsou především chyby v interpunkci v originálních textech.

Z výše uvedeného je patrné, že od <–3,–3> pozice vlevo od tvaru pomocného slovesa *být* již nenacházíme tvary l-ových příčestí, které by s tvarem pomocného slovesa *být* tvořily analytický slovesný tvar minulého času. Lze dobře vidět, že v naprosté většině případů jde o konkordance s chybami v interpunkci (*...* ***zmizela*** *kancelář kde <jsem> pracoval ...*), dále překlepy a narazili jsme i na jedno z úskalí tokenizace (*...* ***Zavolala*** *D. Řekla <jsem> ...*), kdy jeden token fyzicky zastupuje dva (tečku za iniciálovou zkratkou i tečku jakožto interpunkční znak) a je ovšem interpretován pouze jako první z možností (úplná desambiguace je při automatické morfologické analýze požadována, není ale vždy bezchybná[[5]](#footnote-5)). Chyby se ovšem vyskytují i v případech, kdy l-ové příčestí stojí na pozici <–2,–2> od tvaru pomocného slovesa *být*, ať už jsou příčinou chyby v lemmatizaci a značkování (*... často* ***zamířil*** *právě <sem/být/VB.\*> ...*) nebo v interpunkci (*... jedna věta co mi* ***ujela*** *napsala <jsem> jí ujetou větu ...*), překlepy, u nichž nelze dokázat, zda jde o vynechanou interpunkci, nebo o nezáměrné opakování téhož slova (překlep: *...* ***Bála*** *bála <jsem> se, že ....*).

Potenciální chyby lze vybrat tak, že hledáme tvary, které nemohou být příklonkami (např. l-ová příčestí), vztažná zájmena (*kdo, co*)a příslovce (*kde, jak, proč, kolikrát*), před nimiž chybí čárka, slova jako *právě, nyní, opravdu, zrovna, pořád, …*, za nimiž následuje tvar *sem* mylně lemmatizovaný jako tvar slovesa *být*.

Podívejme se nyní na pravý kontext pomocného slovesa *být*. Postup bude analogický. Vrátíme se zpět (ke konkordanci tvarů 1. a 2. osoby slovesa *být*). Zvolíme filtr (pozitivní), rozsah hledání <1,1>, **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”V[pq].....[12].\*“].*



Možné chyby jsou opět případy, kdy i v intervalu <–1,–1> stojí l-ové příčestí (viz výše). Na obrázku níže vidíme, že je možno najít případy přegenerovávání.



Vrátíme se zpět (ke konkordanci tvarů 1. a 2. osoby slovesa *být*). Zvolíme filtr (pozitivní), rozsah hledání <2,2> , **Typ dotazu tag** a do dotazovacího řádku napíšeme dotaz ve formě *V[pq].....[12].\*.*



Zvolíme filtr (pozitivní), nastavíme rozsah hledání <1,1>,zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”[JZ].\*“]*. Snažíme se odstranit případy, kdy zvolený interval překračuje hranice větných celků. Pokud chceme mít přesné výsledky, musíme provést vždy ruční kontrolu řádků, které jsou kandidáty na chybné případy (proto je nejdříve vyhledáme pomocí pozitivního filtru). Teprve poté můžeme pomocí negativního filtru příslušné řádky odstranit.



Dále se snažíme odstranit případy, kdy tvar pomocného slovesa a l-ové příčestí netvoří dohromady hledaný složený slovesný tvar. Pokoušíme se najít možné doklady přegenerovávání analogické těm, které jsme pozorovali v případě levého kontextu.

Vrátíme se zpět, zvolíme filtr (pozitivní), nastavíme rozsah hledání <1,1>,zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”V[pq]-----[12].\*“]*.



Na obrázku vidíme, že kromě případů plusquamperfekta se objevují i hledané případy přegenerovávání.

Pokusíme se je automaticky vybrat: zvolíme filtr (pozitivní), nastavíme rozsah hledání <1,1>,zvolíme **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz ve formě *být|bývat*.



Kromě posledního řádku, na němž je chybně desambiguován tvar *žvanil* (*… vždycky jsem byl <žvanil/žvanit/V.\*> …*), jde ve všech případech o plusquamperfektum.

Vrátíme se zpět ke konkordanci tvarů 1. a 2. osoby slovesa *být*. Zvolíme filtr (pozitivní), rozsah hledání <3,3>, **Typ dotazu tag** a do dotazovacího řádku napíšeme dotaz ve formě *V[pq].....[12].\*.*



Zvolíme filtr (pozitivní), nastavíme rozsah hledání <1,2>,zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=“[JZ].\*“]*. Snažíme se odstranit případy, kdy zvolený interval překračuje hranice větných celků. Pokud chceme mít přesné výsledky, musíme provést vždy ruční kontrolu řádků, které jsou kandidáty na chybné případy (proto je nejdříve vyhledáme pomocí pozitivního filtru). Teprve poté můžeme pomocí negativního filtru příslušné řádky odstranit. Na následujícím obrázku je část konkordancí získaných výše uvedeným postupem. Na první pohled je patrné, že zahrnuje jak hledané doklady přegenerovávání, tak případy opačné.



Pokud bychom mechanicky odstranili řádky uvedeným způsobem, tak bychom mohli, jak se říká „vylít s vaničkou i dítě“. Podívejme se na následující obrázek (např. lemma *však* je spojka *J.\**, nesignalizuje ovšem hranici větného celku).



Dále se snažíme odstranit případy, kdy tvar pomocného slovesa a l-ové příčestí netvoří dohromady hledaný složený slovesný tvar. Pokoušíme se najít možné doklady přegenerovávání analogické těm, které jsme pozorovali v případě levého kontextu.

Vrátíme se zpět, zvolíme filtr (pozitivní), nastavíme rozsah hledání <1,2>,zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=”V[pq]-----[12].\*“]*.



Tímto způsobem můžeme pokračovat dále. Procházení velkého počtu řádků je časově náročné a nutné je ovšem především tehdy, pokud chceme mít k dispozici nezkreslená, statisticky relevantní data.

Proto si práci opět rozdělíme. Zvolíme filtr (pozitivní), nastavíme rozsah hledání <1,2>,zvolíme **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz ve formě *být|bývat*.



Dále zvolíme filtr (negativní), nastavíme rozsah hledání <1,2>,zvolíme **Typ dotazu tag** a do dotazovacího řádku napíšeme dotaz ve formě *[tag=“[ZJ].\*“]*. Na následujícím obrázku vidíme, že kromě případů plusquamperfekta (např. 2., 3., 8., 9., 11., 12., 13. řádek) se objevují i hledané případy přegenerovávání.



**4.4.F Formulace závěrů**

Ukázali jsme návod postupů pro vyhledávání analytických tvarů (1. a 2. osoby) minulého času indikativu aktiva. V další fázi lze např. zkoumat frekvenci slovosledných schémat. Na první pohled je patrné, že nejfrekventovanější jsou případy, kdy oba tvary stojí těsně vedle sebe, přičemž varianta *„přišel jsem“* je asi o třetinu frekventovanější, než varianta *„jsem přišel“*. Omezen je výskyt variant *„přišel <.\*> jsem“*. Dále vlevo od pomocného tvaru slovesa už l-ové příčestí tvořící s pomocným slovesem složený tvar nenajdeme. Pokud jde o výskyt l-ového příčestí na dalších pozicích vpravo od pomocného slovesa (varianty *„jsem <.\*> přišel“*, *„jsem <.\*> <.\*> přišel“*), tak lze konstatovat, že frekvence postupně klesá.

**4.4.G Formulace dalších otázek vyplynuvších ze zkoumání daného jevu**

Pozorujte chyby (většinou v interpunkci), které brání efektivnímu vyhledávání případů, kdy tvar l-ového příčestí významového slovesa stojí na pozici <–2,–2> vlevo od pomocného slovesa *být*. Repertoár jednotek, které mohou stát na prvním místě vlevo od tvaru pomocného slovesa *být*, stojí-li tvar l-ového příčestí významového slovesa rovněž vlevo, je omezený. Projděte znovu konkordance v ostatních korpusech řady SYN a pokuste se najít další slova, která by tento repertoár mohla obohatit.

V jaké vzdálenosti vpravo jste našli poslední případ l-ového příčestí, které tvoří s pomocným slovesem *být* analytický tvar minulého času? Podívejte se, jak vypadá situace na pozici <12,12> vpravo od tvaru pomocného slovesa *být.* Je patrné, že podíl doložených dokladů klesá, nicméně i v takto rozšířeném kontextu stále lze nalézt další doklady.





Pokud bychom skutečně chtěli shromáždit úplný materiál, pak je třeba vzít v úvahu i případy minulého času 2. osoby s volným morfémem *-s*. Volný morfém *-s* tvoří se slovem, k němuž se „přiklání“, grafickou jednotu. Předpokládáme, že se může pojit s řadou slov. Kromě tvarů příčestí, některých spojek, vztažných a tázacích zájmen a zájmenných příslovcí by to mohla být i některá další slova, s výjimkou těch, která končí na sykavky, dále slovesných tvarů prézentních, imperativních, infinitivních a možná i dalších. Značkování grafických slov s připojeným volným morfémem *-s*, která jsou tokenizována jako jedno lemma, není v korpusech řady SYN vždy důsledné (srov. více Osolsobě 2007a). Volný morfém *-s* se často pojí s tvarem l-ového příčestí slovesa významového (*přišels*)*.* V morfologické značce takovýchto l-ových příčestí je na 8. pozici (osoba) uvedeno, že se jedná o 2. osobu (i v korpusech SYN2000, SYN2005, SYN2006PUB).

(Pokud se volný morfém *-s* pojí s tvarem zvratného *se/si*, pak je z řady důvodů snazší začít vyhledávání právě od tohoto tvaru, a to proto, abychom odlišili případy, kdy l-ové příčestí je součástí tvarů minulého času indikativu a kdy je součástí tvarů kondicionálu.)

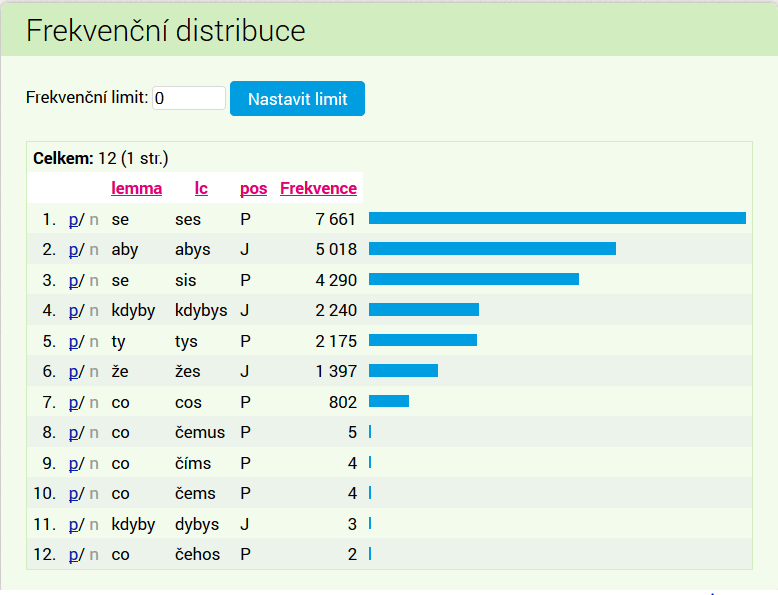
Volný morfém *-s* se dále pojí s tvary některých dalších slov. Pokusme se vyhledat ty případy, kdy se volný morfém *-s* nepojí s tvarem slovesa (l-ovým příčestím, např. *přišels*, nebo poneciálně n-/t-ovým příčestím, např. *bitas*, viz výše), a zároveň lemmatem ve slovníku automatického morfologického analyzátoru byl zvolen tvar bez volného morfému *-s* a zároveň ve značce je na 8. pozici zaznamenáno, že na tvaru je formálně signalizován tvar 2. osoby přísudkového slovesa. Pro tyto tvary platí a) tvar končí na *-s*, b) lemma nekončí na *-s* (podmínka nutná, nikoli postačující, viz tvary jako *nás, vás, pros, bos, ...*), c) jsou označkovány jako tvary 2. osoby (na 8. pozici je *2*, opět podmínka nutná nikoli postačující, viz tvary jako *vás, pros, dones, …*).

Přehled o značkování v korpusu SYN2010 získáme takto. Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme *[lc=“.\*s“ & lemma!=“.\*s“ & lemma!=“vy“ & tag=“*.*......2.\*“ & tag!=”V.\*”]*.



Podívejme se na frekvenční distribuci lemmat, tvarů a slovních druhů.





Připomeňme ovšem, že na lemmatizaci a značkování se nelze spolehnout absolutně. Problematické případy nelze efektivně hledat tak, že budeme hledat kandidáty na slova s připojeným volným morfémem *-s* mezi lemmaty označkovanými jako *X.\** končícími na *s*, protože mezi nimi převažují cizojazyčná slova[[6]](#footnote-6), kterých je velké množství. Vhodnější postup je tedy „namátková kontrola“ jednotlivých slov.

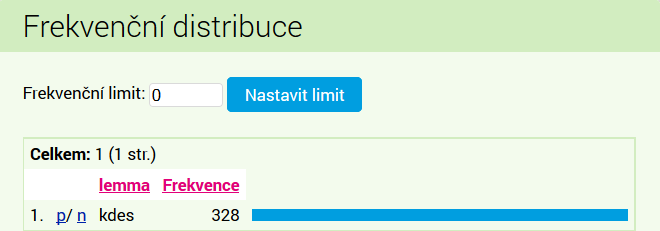
Spolehlivá je anotace slovesných tvarů l-ového příčestí (viz obrázek). Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme *[tag=“V[pq].....2.\*“ & word=“.\*s“]*.



Nespolehlivá je naopak anotace tvarů potenciálních příčestí pasivních (viz výše).

Následující konkordance ukazují, že v případě homonymie tvaru neurčitého adverbia a tvaru vztažného adverbia s volným morfémem *-s* je výsledek desambiguace neuspokojivý.

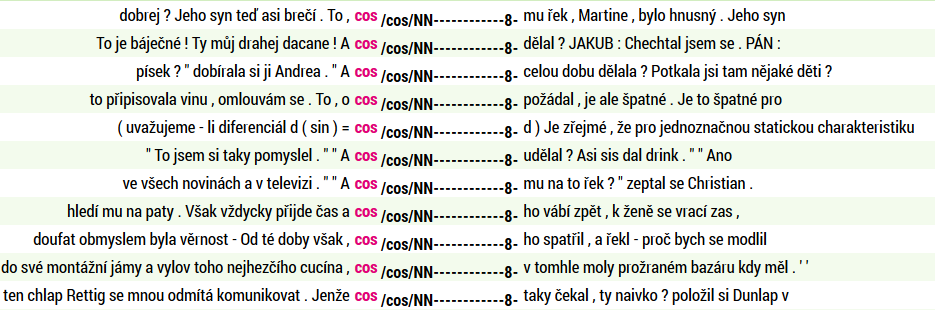
Zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme *kdes*. Podíváme-li se na frekvenční distribuci lemmat, vidíme, že lemma je ve všech případech *kdes* (tedy tvar neurčitého zájmenného příslovce *kdesi*), přičemž podíváme-li se na seznam konkordancí na následujícím obrázku, je z něj patrné, že ve všech případech jde o adverbium (tázací/vztažné) *kde* s volným morfémem *-s* a nikoli o adverbium neurčité (*kdes<kdesi*). Chyba mohla vzniknout buď na úrovni automatického značkování, nebo na úrovni desambiguace.



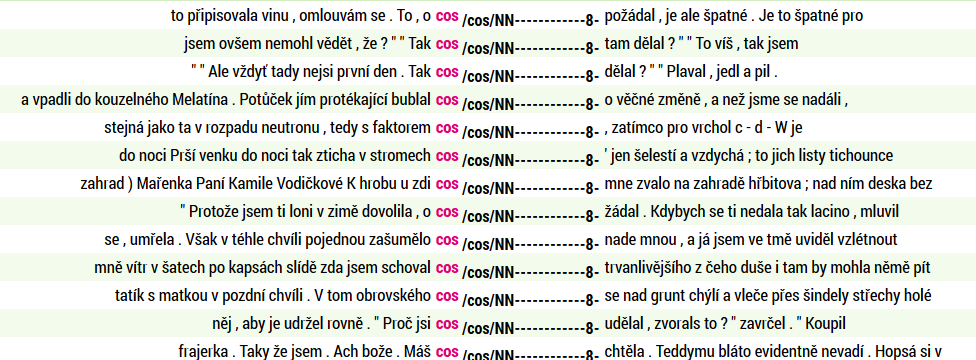


Disambiguace v případě homonymie vztažného zájmena *co* s volným morfémem *-s* a substantivní zkratky goniometrické funkce vykazuje také poměrně špatné výsledky.

Zvolíme **Implicitní atribut Typ dotazu cql** a do dotazovacího řádku napíšeme *[word=“cos“ & tag=“NN.\*“]*.

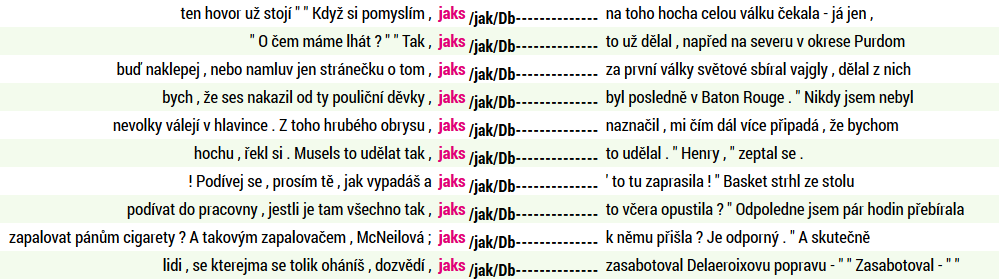


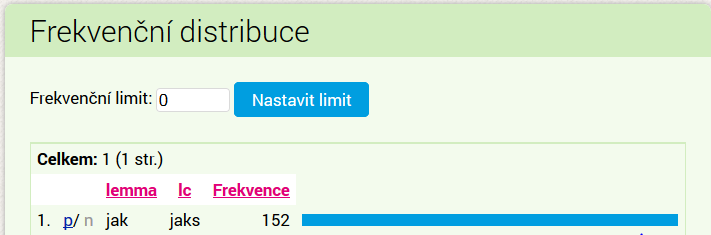
Zvolíme filtr (negativní), rozsah hledání <–1,–1>, **Typ dotazu tag** a do dotazovacího řádku napíšeme *[ZJ].\**. Tímto způsobem odstraníme jasné případy chybné desambiguace vztažného *cos* a můžeme sledovat 75 zbylých řádků.



V řadě případů (26) jde o chybnou desambiguaci vztažného/tázacího zájmena *co* s volným morfémem *-s.* Bez správné interpretace se ocitají ovšem i případy neurčitého zájmena *cos* (ruční analýzou lze zjistit, že v 30 případech z 75 jde o neurčité zájmeno *cos<cosi*). Pouze v 19 případech se skutečně jedná o název goniometrické funkce.

Podobně můžeme postupovat dále. Zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme *jaks*. Podíváme-li se na frekvenční distribuci lemmat, lc a tagů, vidíme, že lemma je ve všech případech *jak* a tag *Db------.\**.





Všimněme si taky, že na úrovni morfologické značky není patrné, že na tvaru se formálně demonstruje tvar 2. osoby přísudkového slovesa (na 8. pozici není *„2“*, ale *„-“*). Tento fakt souvisí s chybnou lemmatizací. Tvar není chápán jako tvar s volným morfémem *–s*, ale jako neurčité adverbium *jaks<jaksi*. To se v korpusových dokladech patrně nevyskytuje, pokud bychom to ovšem chtěli dokázat, museli bychom projít ručně všech 152 řádků.

**4.4.H Zadání cvičení, v nichž lze uplatnit analogické postupy**

Za použití analogických postupů vyhledejte tvary analytického futura (*budu spát*), kondicionálu přítomného (*přišel bych*), kondicionálu minulého (*byl by zaspal*), pasiva (*jsem chválen, budu chválen, byl jsem chválen*), přechodníků pasiva (*jsa stíhán, byv poslán*).

Podívejte se, jak jsou v korpusech lemmatizována, označkována a popř. desambiguována slova jako *tos, komus, kams, pročs, začs, prýs, …*. Všimněte si, kdy má tvar s volným morfémem *-s* ve značce uvedeno na 8. pozici (osoba), že jde o 2. osobu, a kdy nikoli. Má toto řešení nějakou logiku?

Projděte konkordance tvaru *jaks* (ve všech korpusech řady SYN) a podívejte se, v kolika případech jde o adverbium s volným morfémem *-s* a v kolika případech o básnické *„jaksi“*.

Vyhledejte tvary s volným morfémem *-s* v Korpusu soukromé korespondence (ksk-dopisy) a porovnejte jejich frekvenci (relativní) s frekvencí ve velkých synchronních korpusech řady SYN.

1. V tabulce nejsou zahrnuty varianty reflexiv tantum. [↑](#footnote-ref-1)
2. Rovněž neprobíráme případy, kdy je v 1. osobě vynechaný tvar pomocného slovesa nahrazen příslušným osobním zájmenem (*... <já> mu klíč <vrátil> ... ... a <my> tu <čekali> v naprosté tmě ...*). [↑](#footnote-ref-2)
3. Nebereme v úvahu ani archaismy typu (*... on <jest> mi <uletěl> ... ... Pirát Ockham <byl jest přepadl> naši flotilu ...*). [↑](#footnote-ref-3)
4. V případě, že bychom používali starší korpusy (SYN2000, SYN2005, SYN2006PUB, ORWELL), bude mít poslední dotaz podobu *V[pq].\**, nebo *V[pq].....X.\** (srov. výše oddíl ***Pozice 8***). [↑](#footnote-ref-4)
5. Požadavek desambiguace (zjednoznačnění) naráží na grafický úzus, který zaznamenává dva významy tečky jediným znakem (jedinou tečkou). Obdobně vidíme, jak v případě požadavku jednoznačné lemmatizace adjektiv/adverbií 2. a 3. stupně narážíme na problém, kdy několik adjektiv/adverbií tvoří tvary 2. a 3. stupně stejně(homonymie), takže při desambiguaci nelze zjistit, ke kterému lemmatu (první stupeň) se vztahují (např. *více – mnoho/moc/hodně/tuze, horší – zlý/špatný*). [↑](#footnote-ref-5)
6. Např. plurálové formy anglických slov. [↑](#footnote-ref-6)