

# **Česká morfologie a korpusy**

**Klára Osolobě**

Obsah

**Předznamenání**

**Úvod**

**Korpusové manažery**

**Jazykové korpusy z hlediska lemmatizace a morfologického značkování**

*Tokenizace*

*Automatická morfologická analýza*

*Lemmatizace a pozice 1 morfologické značky*

*Pozice 2 – Detailní určení slovního druhu*

*Pozice 3 – Jmenný rod*

*Pozice 4 – Číslo*

*Pozice 5 – Pád*

*Pozice 6 – Přivlastňovací rod*

*Pozice 7 – Přivlastňovací číslo*

*Pozice 8 – Osoba*

*Pozice 9 – Čas*

*Pozice 12 – Aktivum/pasivum*

*Pozice 10 – Stupeň*

*Pozice 11 – Negace*

*Pozice 13 – Nepoužito*

*Pozice 14 – Nepoužito*

*Pozice 15 – Varianta, stylový příznak apod.*

*Pozice 16 – Vid*

**Česká morfologie a korpusy**

**Substantiva**

*Jak vyhledat v korpusu substantiva podle vzoru?*

*Jak lze v korpusech hledat doklady pro výzkum hláskových alternací v rámci substantivní flexe?*

*Další samohláskové alternace v české substantivní flexi*

*Varianty a dublety – korpusy jako zdroje dat pro formulaci pravidel distribuce variantních spisovných koncovek v české substantivní flexi*

**Adjektiva**

*Která adjektiva se v češtině stupňují?*

*Vlastnosti adjektiv na -cí*

### Číslovky

*Slovnědruhové přesahy*

### Slovesa

*Slovesná třída a slovesný vzor – jak lze v korpusu hledat slovesné tvary bez použití tagu?*

*Syntetické futurum v češtině*

*Jakou osobu signalizují tvary by?*

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

### Adverbia a další neohebné slovní druhy

*Přísllovečné spřežky*

*Značkování neohebných slovních druhů*

*Ještě několik slov ke značkování neohebných slovních druhů s ohledem na slovnědruhové transpozice*

### Korpus a některé pravidelné derivace

*Deverbativa od slovesného kmene a jejich vyhledávání v korpusech*

*Ještě jednou k adjektivům na -cí*

*Hláskové alternace ve slovtvorbě*

### MWE – gramatika a slovník

*Opisné stupňování*

*Slovesné frazémy jako typ MWE*

*Závěr*

**Bibliografie**

**Přílohy**

**Algoritmus určování slovesných tříd a vzorů v češtině**

**Algoritmus tvoření přechodníků v češtině**

## **Předznamenání**

Formulovat pravidla jazyka je mnohem složitější pro mateřský jazyk než pro jazyk, kterému se učíme v době, kdy jsme schopni vnímat jeho gramatiku na pozadí gramatiky jazyka (jazyků), které nějakým způsobem ovládáme. Přesto právě od rodilých mluvčích většinou požadujeme informace o tom, co jak má být a hlavně proč. Fyzik a popularizátor fyziky Jiří Grygar napsal, že zatímco vědci odpovídají na otázky *jak?*, filozofové a teologové by měli dávat odpovědi na otázky *proč?* Tento text nabízí čtenářům, především studentům českého jazyka, řadu návodů, jak lze vyhledávat v jazykových korpusech relevantní data, jejichž prostřednictvím mohou pozorovat svou mateřštinu, a také řadu postupů, jak lze vyvozovat závěry o tom, jak se jazyk užívá.

Kniha má název **Česká morfologie a korpusy** a omezuje se hlavně na otázky spojené s formálním tvaroslovím a některými dalšími otázkami morfologie i tvoření slov. Z pedagogického hlediska necháváme mnohde na čtenářích samotných, aby hledali odpovědi na otázky *proč*. V řadě případů lze najít vysvětlení v systému jazyka. Doufáme, že podnítíme zvědavost, která povede k pátrání po příčinách. Těšilo by nás, kdyby čtenáři byli schopni aktivovat informace z různých jazykovědných disciplín a naučili se je propojovat. Proto dáváme přednost kladení otázek před dáváním odpovědí.

Snažíme se také rozvinout u čtenářů jistý typ myšlení. Jak už jsme naznačili, automatismy, které doprovázejí užívání mateřského jazyka, brání mnohdy jistému odstupu, který je třeba, abychom uměli vymezit některé jevy. Cesta dedukce může být pro jisté typy čtenářů schůdnější než jiné metody učení. Všem budeme vděční za jakékoliv připomínky k tomuto textu.

## Úvod

Předložený text vznikl na základě dlouhodobých zkušeností s výukou seminářů zaměřených na korpusovou lingvistiku na straně jedné a zkušeností s přípravou formálních popisů pro potřeby strojového zpracování přirozeného jazyka – češtiny na straně druhé.

Výklady i cvičení mají pomoci studentům lingvistických oborů, bohemistům, korpusovým lingvistům i dalším zájemcům o češtinu naučit se dívat na svoji mateřštinu jinak, než jsou tomu zvyklí.

Měli by se naučit:

1. Slovně formulovaná tvrzení o jazyce převést do podoby formálních pravidel, která lze zapsat např. jako posloupnost dotazů pro korpusový manažer.
2. Efektivně pozorovat korpusová data tak, aby na základě těchto pozorování byli schopni slovně zformulovat zákonitosti fungování jazyka vyplývající z učiněných pozorování jazyka v textech, z nichž jsou složeny korpusy.
3. Syntetizovat oba výše uvedené postupy tak, aby byli schopni co nejefektivněji používat korpusové nástroje ke shromáždění relevantního materiálu pro vlastní lingvistický výzkum.

Tomuto účelu poslouží jednotlivé kapitoly, v nichž se budeme zabývat jednoduššími i složitějšími otázkami z české morfologie a slovtvorby.

Při práci s korpusy se pro přístup k elektronicky uloženým jazykovým datům používají programy, tzv. korpusové manažery, které umožňují data vyhledávat, zobrazovat, třídít, tvořit frekvenční seznamy a vyhledaná data ukládat. Stručně představíme práci s korpusovými manažery a odkážeme k manuálům, které mohou čtenáři pomoci ke zvládnutí těch stávajících variant korpusových vyhledávačů, s nimiž lze přistupovat ke korpusům Českého národního korpusu.

Následuje oddíl věnovaný podrobnějšímu výkladu problematiky strojového zpracování přirozeného jazyka na rovině lemmatizace a morfologického značkování. Pokusíme se ukázat, jak jsou řešeny některé problémy související s mnohoznačností jednotek přirozeného jazyka a potřebou desambiguace.

Ve 4. kapitole podáváme komentovaný přehled značek použitých pro značkování korpusů řady SYN (synchronní korpusy psané češtinou budované v ÚČNK<sup>1</sup>). K podrobnějšímu

---

<sup>1</sup> Stručný návod je na webových stránkách Ústavu Českého národního korpusu FFUK: <http://ucnk.ff.cuni.cz/bonito/znacky.php>. Jedná se o korpusy popsané tamtéž (<http://ucnk.ff.cuni.cz/struktura.php>), a sice SYN2000, SYN2005, SYN2006PUB, SYN2009PUB a SYN2010.

proniknutí do tematiky, které přesahuje rámec potřebný k výkladu v této knize, odkazujeme k další literatuře.

Jádro knihy tvoří kapitola se stejným názvem jako celek knihy **Česká morfologie a korpusy**, která zahrnuje oddíly týkající se jednotlivých slovních druhů, oddíl s přesahem do slovtvorby a oddíl věnovaný tzv. MWE (Multiword Expressions – víceslovným výrazům). Budeme se v ní zabývat těmito otázkami:

1. Jak získat materiál pro výzkum hláskových alternací, které doprovázejí tvoření tvarů substantiv (skloňování) v češtině.
2. Jak získat materiálovou základnu pro výzkum variantních a dubletních koncovek substantiv.
3. Jak získat podklady pro výzkum stupňování adjektiv.
4. Jak získat přehled o skutečném stavu některých okrajových jevů, např. syntetického futura.
5. Jak najít v korpusech materiál pro výzkum adverbializace.
6. Jak se v lingvistickém výzkumu obejít bez morfologického značkování a lemmatizace.
7. Jak získat z korpusů podklady pro výzkum slovtvorby.
8. Jak se v korpusově založeném výzkumu vypořádat s takzvanými MWE.

Jednotlivá témata budou probírána následujícím způsobem:

- A. Motivační úvod
- B. Nastínění problému
- C. Otázky
- D. Formulace dotazu pro získání dat z korpusů
- E. Třídění a pozorování dat získaných z korpusů
- F. Formulace závěrů
- G. Formulace dalších otázek vypluvších ze zkoumání daného jevu
- H. Zadání cvičení, v nichž lze uplatnit analogické postupy.

V rámci textu se pokusíme doplnit některá fakta z oblasti korpusové lingvistiky, která jsou zřejmá obci korpusových lingvistů, ale nejsou explicitně formulována v běžně dostupných manuálech pro uživatele jazykových korpusů. Zároveň nebudeme podrobně probírat to, co je dobře a přehledně popsáno v dostupné literatuře a dalších zdrojích, na které odkážeme a které budeme citovat. Jako úvod do korpusové lingvistiky doporučujeme z kratších česky psaných příruček *Český národní korpus – úvod a příručka uživatele* (Kocěk – Kopřivová – Kučera

2000), nebo anglicky psanou učebnici *Corpus Linguistics* (McEnery – Wilson 1996). K hlubšímu studiu pak překladový sborník *Studie z korpusové lingvistiky* (Čermák – Klímová – Petkevič 2000) a monografie a sborníky řady *Studie z korpusové lingvistiky* vydávané Nakladatelstvím Lidové noviny a Ústavem českého národního korpusu (viz též [ucnk.ff.cuni.cz/publikace](http://ucnk.ff.cuni.cz/publikace)). Aktuální informace lze čerpat z příručky na dostupné z webových stránek ČNK (<http://wiki.korpus.cz/doku.php>).

Na konci knihy jsou připojeny přílohy. Jedná se o praktické pomůcky, které mají pomoci osvěžit znalosti z české gramatiky.

Text je doplněn bibliografií. V poznámkách odkazujeme na položky bibliografie, jejichž studium je žádoucí pro hlubší proniknutí do dané problematiky.

## Korpusové manažery

Korpusový manažer je soubor programů, které umožňují efektivně využívat jazykové korpusy. Pro práci s korpusy budovanými v rámci Ústavu Českého národního korpusu je možno pracovat buď s verzí webového rozhraní **NoSketch Engine** na adrese <http://korpus.cz/corpora/> nebo s aplikací **KonText** na adrese <http://korpus.cz/kontext>.

Z hlediska uživatele je dobré vědět, že technické ovládání programů lze i bez speciálního školení zvládnout díky uživatelsky přívětivé příručce, elektronicky dostupnému **Manuálu práce s ČNK** (<http://wiki.korpus.cz/doku.php/manual>), kde lze nalézt podrobný popis funkcí aplikace **KonText** i jednotlivých starších verzí korpusových manažerů.

Příklady uvedené níže v textu budou popsány tak, aby je čtenář mohl zopakovat s použitím obou verzí, ilustrační materiál (screenshoty) odpovídají aplikaci **KonText** uveřejněné na začátku roku 2014.

Na jednotlivých místech nebudeme podrobně popisovat práci s korpusovým manažerem. Uživatelské příručky přístupné on-line při práci s manažerem jsou natolik podrobné a dobře pedagogicky zpracované, že je nemíníme „opisovat“. Celkově vzato pracujeme s poměrně úzkým repertoárem zadání dotazů. Kromě zmíněného manuálu lze dále doporučit i soubor cvičení s klíčem Blatná, R. – Čermák, F.: *Jak využívat český národní korpus*. Praha : NLN, 2005.



## **Jazykové korpusy z hlediska lemmatizace a morfologického značkování**

Jazykový korpus je elektronicky zpracovaný a přístupný soubor jazykových textů. Od sbírky textů se liší tím, že je promyšleně a záměrně sestaven ze vzorků jazyka tak, aby byl reprezentativní, tedy aby co možná nejpřesněji ilustroval ty rysy jazyka, k jejichž zkoumání má sloužit. Z tohoto aspektu rozlišujeme typy korpusů psaných versus mluvených, korpusů dle časového období, žánru, autora, atd. Texty, které tvoří jazykový korpus, musí být uživatel korpusu schopen identifikovat. K tomu účelu slouží standardizované vněttextové anotace, které se u různých korpusů liší. Řada korpusů navíc obsahuje také interpretace jednotlivých částí textů, z nichž je korpus složen (vnitrotextové anotace). Pro potřeby tohoto textu upozorňujeme na anotace vět (vyznačení začátku a konce věty) a především na anotace slovních jednotek typu **word** (jednoduchých slovních tvarů). Na lingvistické rovině popisu grafické realizace jazyka odpovídají jednotkám typu **word** nejmenší jednotky textu – slovní tvary definované jako řetězce znaků mezi mezerami, ale i interpunkční znaky, číslice apod. Těmto jednotkám je pak buď automaticky, nebo ručně přiřazena interpretace na úrovni lemmatu a tagu. Běžně se pak hovoří o gramatickém/morfologickém značkování a lemmatizaci.

### ***Tokenizace***

Prvním krokem automatické analýzy je vyčlenění jednotek, z nichž je text z hlediska programu automatické analýzy složen. V případě automatického zpracování korpusů se v prvním kroku jedná o tokenizaci – tj. rozčlenění textu na jednotky (pozice), které budou předmětem další analýzy. Pro potřeby automatické morfologické analýzy se pracuje s lingvisticky zjednodušujícím, nicméně automaticky dobře zpracovatelným pojetím slovního tvaru v textu, který je definován jako řetězec znaků dané abecedy oddělený z obou stran oddělovači (mezery, některé znaky). Takto technicky omezená definice slovního tvaru má při další interpretaci (značkování) slovních tvarů automatickou morfologickou analýzou své důsledky na všech úrovních (srov. níže).

### ***Automatická morfologická analýza<sup>2</sup>***

Ve druhém kroku je každé z takto definovaných jednotek (token) přiřazena interpretace.<sup>3</sup>

---

<sup>2</sup> K historii automatické morfologické analýzy češtiny srov. též Osolsobě 2007<sup>2</sup>, Jelínek 2008.

<sup>3</sup> Morfologické analyzátoři pracují nad databází slovních tvarů a jejich možných interpretací. Tyto databáze byly zpracovány na základě algoritmických popisů flexe (srov. Hajič 1994, 2004, Osolsobě 1996). V databázích

Při aplikaci na jazykový materiál korpusů se ukázalo, že celá řada interpretací, které byly přiřazeny jednotkám na úrovni strojových slovníků, se plně nekryje s bohatstvím přirozeného jazyka, jak je prezentuje korpus. Ukázalo se, že s ohledem na zkušenosti z konkrétní praxe, je třeba některé interpretace zpětně verifikovat.

K automatickému značkování a lemmatizaci se používá programů (automatických morfologických analyzátorů). Ty většinou testují každou jednotku (token) proti „slovníku“ ve formátu **word** + **lemma** + **tag**, kde **word** je jednoduchý slovní tvar, **lemma** je základní tvar odpovídající jednoduchému slovnímu tvaru a **tag** je morfologická značka, a přiřazují jí interpretace nalezené ve slovníku.

Příklady:

Mějme tvary jako *který, je, má, spíš*.

U tvaru *který* jsou ve slovníku ponechána stranou funkční rozlišení (zájmeno vztažné, tázací atd.), nicméně existují tři možné interpretace na rovině spisovného úzu a řada dalších možných interpretací substandardních (viz <<http://ucnk.ff.cuni.cz/bonito/znacky.php>>).

Standardní interpretace:

word:	lemma:	tag:
který	který	P4MS1-----
který	který	P4IS1-----
který	který	P4IS4-----

Substandardní interpretace:

word:	lemma:	tag:
který	který	P4MP1-----6-
který	který	P4MP4-----6-
který	který	P4IP1-----6-
který	který	P4IP4-----6-
který	který	P4NS1-----6-
který	který	P4NS4-----6-
který	který	P4NP1-----6-
který	který	P4NP4-----6-
který	který	P4FS2-----6-
který	který	P4FS3-----6-
který	který	P4FS6-----6-
který	který	P4FP1-----6-
který	který	P4FP4-----6-

---

jsou uloženy potenciaální (kontextově nevázané) interpretace bez ohledu na frekvenční, stylistická i jiná omezení jejich výskytu. Na tomto místě ponecháme stranou rozbor jednotlivých problémů různých přístupů. Pro naše potřeby je důležité si uvědomit, že desambiguátory/desambiguátoři pracují především s těmi interpretacemi, které nabízí automatický morfologický analyzátor.

Podobně tvary *je, má, spíš* mají více interpretací, přičemž formální homonymie se týká jak interpretace na úrovni slovního druhu, tak jednotlivých slovnědruhově závislých gramatických významů.

Standardní interpretace:

word:	lemma:	tag:
je	být	VB-S---3P-AA---I
je	on	PPXP4--3-----
je	on	PPNS4--3-----

word:	lemma:	tag:
má	mít	VB-S---3P-AA---I
má	můj	PSFS1-S1-----1-
má	můj	PSFS5-S1-----1-
má	můj	PSNP1-S1-----1-
má	můj	PSNP4-S1-----1-
má	můj	PSNP5-S1-----1-

word:	lemma:	tag:
spíš	spíš	TT-----
spíš	spíše	Dg-----2A-----
spíš	spát	VB-S---2P-AA---I

Takto prováděná automatická morfologická analýza je obecně nejednoznačná. Většinu jednotek je přiřazena více než jedna interpretace.

Druhým krokem je desambiguace<sup>4</sup> (disambiguace, zjednoznačnění). Desambiguaci je opět možno provádět buď ručně, nebo pomocí automatických nástrojů. Pokud je automatizována, rozlišujeme různé metody, které se pro zjednoznačnění používají. Rozšířené a užívané jsou metody matematické statistiky. Na opačném pólu stojí metody, které se opírají o pravidla fungování přirozeného jazyka.

Výsledky desambiguace jsou sice velmi uspokojivé a mohou dobře sloužit uživatelům korpusů, nejsou ovšem nikdy zcela bezchybné.

Chybnou desambiguaci vidíme na následujících příkladech z korpusů ČNK, a sice SYN2000 a SYN2010. Vidíme, jak je tvar <má> v kontextu <má> *láska* v řadě zobrazených vyhledaných dokladech mylně interpretován jako tvar slovesa *mít*. Od chybné desambiguace na úrovni lemmatu se pak odvíjí též chybná desambiguace na úrovni morfologické značky. Tvar je označen za 3. osobu singuláru indikativu přítomného času aktiva (*VB-S---3P-AA---*, resp. *VB-S---3P-AA---I*).<sup>5</sup>

<sup>4</sup> K problematice desambiguace korpusů ČNK srov. Hajič 2004, Petkevič 2006, Spoustová et al. 2007, Jelínek 2008, Skoumalová 2011.

<sup>5</sup> Jak číst morfologickou značku srov. níže.

**Korpus:**

**Typ dotazu:**

**CQL:**

. Láska přitahuje lásku . Proto , můj Ježíši , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- se vrhá k tobě , chtěla by naplnit propast , slyším každý den . Má síla , má naděje , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- , mé činy to všechno mě drží a dobrým mě Dnes už jsem to věděla . Pochopila jsem , že **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- k Loganovi má kořeny hluboko zapuštěné v realitě a že k Řekům , pokud chceme nějak pojmenovat lásku . Tam **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- tří dimenze . Éros , kde dvojici spojuje pouhý erotismus , co přivede žižkovského kluka k básnění ? Žižkov byl **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- . Do sedmnácti let . Pak jsme se přestěhovali na varování je třeba přistupovat s odpovídající vážností . Paříž , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- je melodramatický útvar , v němž hraje rovnocennou roli hudba otci . Bylo to , jak řekl , stejně jako **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- k matce . Pomalu jsme prošli sály a ven do celý . Věra O vánocích 1939 Žiji , pokud žije **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- , pravím . Zeslábl jsem , ale jsem silnější , její Zděneček . Načež se dveře rozevíely a vstoupila jimi **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- . Věděla jsem to , samozřejmě , že Mirek na opojení životem . Snad se i ptal , jakou cenu **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- tvora , který se považuje a má považovat za nejubožejšího zase sama doma - hity semaforové éry 14.00 Mia , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- ( 2 / 4 ) 15.40 Změna je život ( . Jednu dobu jsme byli hodně pohromadě , ale jinak **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- zůstala neopětována . Až jsem se pak setkala s Petrem pevnou mříží zas vodu pít , až roky zmizí , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- ryzí k Mary se vrátí , budu s ní žít vraťte , vraťte se , vy sívě kroky moje , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- nezklame , když všechno mohlo zmást , má osa prastará tak dobré a tak pohostinné spočinutí . Každým dnem rostla **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- k této zemi a nikde bych nebyl raději znovu vystavěl ) ( 5 . března 19.30 ) , Hirošima , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- - režie Alain Resnais ( 1959 ) ( 6 . Ze zašlých klenotů a sychravého šera tam na mne vanula **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- , atmosféra . STŘÍBRNÉ KŘEPELKY Noc je smuteční prapor který Ráj milenců III , erotická obrazová fantazie 1.55 Max , **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- , drama 3.30 Interaktivní hry NEDĚLE 27 . října ČT přicházely , některé dokonce potají . Myslím , že jim **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- k Marietě působila nekonečné potěšení , a v té jedině ŘÍKEJTE NAHLAS : Toto je můj klid , toto je **má** /mit/VB-S---3P-AA--- **láska** /láska/NNFS1-----A----- . Žehnám tomuto prostoru , aby působil k vyššímu dobru

**Korpus:**

**Typ dotazu:**

**CQL:**

. Leč emoce mají neméně výraznou odvrácenou tvář . Jako <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- svůj rub v nenávisti , v kterou se snadno zvrtné
poklidné krajíně . Žiji zde sám v nebeském vzduchu , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- , má píseň blaží mě . I zde se jedná
eso . Je jím v pořadí čtvrtý koncert Provence - <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- , kvůli němuž přijede z Prahy povědomý pár Milan Dvořák
něco nezajímavého a možná i nepříjemného , za což nás <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- kompenzovat . Jenže ona to za nás neudělá . Udělá
... Nenápadně , do ř idkých vánkú <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- vyvát jak dým ; v hebkých vlnách , polehounku a
jako by byla obsažena v tom svém střešnici . A <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- , má touha , byly v té vážce ; z
svém věku snad již víš , slečinko , že ačkoli <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- nejedno jméno , nakonec vždy muže se ženou vzájemně přitahuje
( zasměje se ) . * Jste čerstvě zamilovaná , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- vliv na váš šatník ? Ano , jsem opravdu hodně
se podíváte , narazíte na psa nebo kočku . Odtud <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- ke zvířatům . Kam se dají nacpat , jsou knihy
ano . David , přemítala horečně Claudia . David , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- . David , který mne naučil milovat lásku , skutečnou
řekl po svém , že věci dechu . Tak jako <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- k bližnímu a Bohu , smíří - li se oba
Prostějovských dnů hudby . Komponovaný pořad Provence - <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- klavíristy Milana Dvořáka a zpěvačky Evy Kriz - Lifkové začíná
pobřeží ( 1958 ) , významná Resnaisova díla Hirošima , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- ( 1958 ) a Loni v Marienbadu ( 1961 )
v PŘEPADENÍ 13 . OKRSKU , zjistíme , jakou cenu <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- v romantické komedii DOKONALÁ PARTIE , pokusíme se zachránit největší
jsem mluvit , tuším o filmu Alaina Resnaisa Hirošima , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- . Později , když už jsme se spřátelili , jedna
v kinu Mír stane dokument z roku 1992 Japonsko , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- . Do tajů východoasijské kultury návštěvníky zasvětil japanolog Martin Vačkář
sebezdokonalování ! Může vás spasit , může vám pomoci - <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- ! Ona je spásná ! Moje láska ! OLGA :
rám ? Řekni mi , svatební noci , zda mě <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- v své moci , nebo jen touha a klam ?
nemohu ani usilovat . Má druhá a drahá polovička , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- , je dána stále stejnou představou , nebo podobou ,
doby se Karel stal pro můj život tak důležitým a <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- k němu tak silným citem , jaký jsem nikdy k
s výjimkou Lisbeth Cookeové - kterou často nazývá Lisbeth , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- . " Takže žádný záznam o jiné ženě ?
dobry život . Pak mi to jednou večer Betty , <b>má</b> /mit/VB-S---3P-AA---I <b>láska</b> /láska/NNFS1-----A----- , nandala hned u prvního panáka : " Hanku ,

V tomto textu se budeme snažit upozornit čtenáře na některé typy chyb a hlavně ukážeme na jednotlivých příkladech, jak je možné kombinacemi vyhledávacích strategií vyloučit zkreslení obrazu jazyka v důsledku chyb v anotacích.

Popis morfologických značek používaných v synchronních anotovaných korpusech ČNK (SYN2000, SYN2005, SYN2010, SYN2006PUB, SYN2009PUB) uvedený na webových stránkách ČNK (viz výše) zachycuje pouze přehled možných vyplnění příslušných pozic se stručnou (řádkovou) charakteristikou vysvětlující, co se pod jednotlivými slovními charakteristikami značek vlastně skrývá.

Teoretik korpusové lingvistiky G. Leech sestavil „sedmero“ anotačních schémat (Leech 1993), ve kterém mimo jiné uvádí, že značkování nesmí být poslední instancí výzkumu, ale má být praktickou pomůckou, která napomáhá uživatelům v rychlejší orientaci v obrovských datech. Na tomto místě bychom rádi uvedli některá fakta, která mohou uživatelům jazykových korpusů pomoci orientovat se ve výsledcích vyhledávací praxe pomocí tagů.

Každá značka je řetězcem 16 pozic (v korpusu SYN2000 je pozic pouze 15). Každá z pozic odpovídá více méně nějaké kategorii známé z gramatiky (slovní druh, jmenný rod, osoba, stupeň). Pozice jsou vyplněny (nebo nevyplněny) ve vzájemných souvislostech. Vyplnění pozice z lingvistického hlediska odpovídá konkrétním gramatickým významům příslušných kategorií. Výsledky anotační praxe jsou ovšem závislé na tom, jak jsou jednotky ve slovníku automatického morfologického analyzátoru označovány. Tato praxe je někdy jedním z možných řešení složitějšího problému.

Naším cílem bude poukázat na to, jak některá ze zvolených řešení mohou být svým způsobem omezená vzhledem k bohatství jazyka, jak jej zachycují korpusy. Budeme postupovat

systematicky a probereme jednotlivé pozice tak, aby bylo patrné, jaké informace obsahují, jaké skutečnosti zachycují a které naopak ponechávají stranou. Budeme si všimnout ryze technických řešení, záměrných zjednodušení i patrných opomenutí.

### ***Lemmatizace a pozice 1 morfologické značky***

Podrobnější komentář vyžaduje 1. pozice. Ta nese název „slovní druh“ a lze podle ní vyhledávat i tehdy, zvolíme-li jako **Typ dotazu** pro vyhledávání v korpusech atribut **pos** (**part of speech**), nebo **tag**, přičemž vyplníme právě pouze 1. pozici.

Na 1. pozici může jako charakteristika slovního druhu figurovat a) značka pro jeden z 10 běžně školsky uváděných slovních druhů, b) X – neznámý slovní druh a c) Z – interpunkce.

Běžný uživatel korpusu by si měl být vědom toho, že slovnědruhová kategorizace je provedena na základě automatické lemmatizace, značkování a desambiguace. Charakteristika slovního druhu je taková a pouze taková, jaká je u přiřazeného lemmatu ve slovníku.

Za příklad poslouží tvary slov *jiný* a *druhý*. V souladu s českými výkladovými slovníky se *jiný* chápe jako adjektivum, přestože např. v Mluvnici češtiny 2 (Dokulil a kol. 1987) je řazeno k zájmenům (alterátorům), *druhý* buď jako adjektivum, nebo jako číslovka řadová (viz níže).

Podobných jevů je celá řada. Problematické jsou zejména případy slovnědruhových přechodů mezi neohebnými slovními druhy (např. adverbii a částicemi, viz výše tvar *spíš*, též níže prepozicionalizace). Desambiguační manuály pro ruční práci jsou složité a pro mnohé badatele sporné. Praktickým důsledkem pro běžného uživatele by měla být ostražitost. V řadě případů jde o jednotlivá slova. Pokud je uživatel chce zkoumat z aspektu slovnědruhové charakteristiky, může postupovat bez použití morfologických anotací, popřípadě se zřetelem k tomu, že anotace mohou obsahovat chyby, popřípadě řešení, s nimiž nesouhlasí.

Chyby v lemmatizaci v naprosté většině případů korespondují s chybami ve značce. V zásadě platí, že je-li něco v nepořádku s lemmatem, je něco v nepořádku i s morfologickou značkou. Z tohoto pravidla se vyděluje jedna velká skupina a dále několik menších skupinek anomálií.

Pro velkou skupinu slovních tvarů neexistuje ve slovníku morfologického analyzátoru žádná interpretace. Těmto tvarům je automaticky jako lemma přiřazen jejich tvar a jako značka X (neznámý, nerozpoznaný slovní druh).

Příklad:

Zadáme-li např. v korpusu SYN2010 dotaz na vyhledání slov, která mají na první pozici ve značce X, dostaneme seznam více než milionu slovních tvarů (cca. 1 % všech tvarů), které nebyly identifikovány ve slovníku automatického morfologického analyzátoru.

Z frekvenčního seznamu je patrné že jde a) o slova cizího jazyka (zejména anglická), b) propria a c) ostatní. Velké procento slov má frekvenci 1. Z hlediska korpusové lingvistiky je třeba mít na zřeteli, že s každým novým korpusem je pravděpodobné, že takový seznam nebude prázdný. Oprávněnost tohoto předpokladu je založena na znalostech o výskytu tzv. hapax legomena (slov s frekvencí 1), který zůstává konstantní s nárůstem rozsahu textů.

Vidíme, že problémem není na rozdíl od případů výše uvedených chyb v desambiguaci mnohoznačnost analyzovaného tvaru z hlediska mnohočetných slovníkových interpretací, ale naopak nedostatečnost slovníku.

Tuto skupinu slov lze dobře použít například pro výzkum okrajových jevů morfologie i slovotvorby (viz níže).

Jednu z malých skupin tvoří slova označovaná tzv. guessery. Guesser neboli hadač je program, který na základě různých postupů přiřazuje interpretace slovům, která nebyla zachycena v prvním kroku automatické morfologické analýzy, protože nejsou ve slovníku automatického analyzátoru. Některé důsledky testování hadačů lze vidět ve značkování a lemmatizaci korpusu SYN2005. Řada slov má přiřazeno lemma a morfologickou značku, přičemž prokazatelně nemůže jít o problém desambiguace (tj. neexistuje kontext, v němž by slovní tvar mohl mít uvedené lemma a značku). Chyby hadačů (zejména těch, které používají statistické metody) lze poměrně těžko odhalit.

Příklad:

Naprostou náhodou při vyhledávání dokladů na slovotvorný typ substantiv na -č jsme si všimli vysokého procenta hledaných slov označovaných v korpusu SYN2005 jako adverbia (D). Uvádíme jejich seznam:

lemma:	tag:	##	
šikmookáč	Db-----		6
překlápěč	Db-----8-		4
šikmookáč	Db-----8-		3
maskáč	Db-----8-		2
svážeč	Db-----		2
cibuláč	Db-----8-		2
spoluspáč	Db-----		2
skupináč	Db-----8-		2
spoluspáč	Db-----8-		1
Překlápěč	Db-----		1
Ceckáč	Db-----		1
procházeč	Db-----8-		1
šikmookáč	Db-----		1
Rychlovyvíječ	Db-----		1
skupináč	Db-----		1
hrobník-kopáč	Db-----		1

sedmispáč	Db-----	1
doprovazeč	Db-----8-	1
autor-vypravěč	Db-----	1
básník-vyprávěč	Db-----	1
bodlináč	Db-----	1
mrkváč	Db-----	1
inženýr-svářeč	Db-----	1
gambáč	Db-----8-	1
řemenáč	Db-----	1
závináč	Db-----	1
kucháč	Db-----8-	1
ceckáč	Db-----	1
on-hráč	Db-----8-	1
superdžič	Db-----8-	1
zaražeč	Db-----	1
tutáč	Db-----	1
bobkáč	Db-----	1
čajpíč	Db-----	1
neženáč	Db-----	1
pruháč	Db-----8-	1
širokokloboukáč	Db-----8-	1
odbíječ	Db-----8-	1
pobízeč	Db-----	1
propouštěč	Db-----8-	1
agent-hráč	Db-----	1
doprovazeč	Db-----	1
pojízďeč	Db-----8-	1
rozjížděč	Db-----8-	1
vegáč	Db-----	1

Povšimněme si také nesrovnalostí v lemmatizaci a značkování slov, kterých se tato evidentně chybná anotace týká.

brzdění na příkrém terénu to však nestačí , a tak	<b>svážeč</b> /svážeč/Db-----	musí brzdit " tlapkami " neboli postranními dřevěnými pákami .
přes něj se položí další bun a s touto soupravou	<b>svážeč</b> /svážeč/Db-----	vyradí rychlou jízdou do údolí . Na mírných , pomalejších
vystřelovala jako bitva a páčila jako oheň . Kolem projeli	<b>svážeči</b> /svážeč/NNMP1-----A-----	sena , s nimiž se večer scházela sestra ( proti
Jenže v tu chvíli se dveře s rachotem otevřely a	<b>šikmookáč</b> /šikmookáč/Db-----8-	řekl : " Dejte ruce nahoru ! " " Jak
. místnosti ? " " Mluvte tišeji ! " upozornil	<b>šikmookáč</b> /šikmookáč/Db-----8-	skoro šeptem . No ano , vždyť za jeho zády
" " Berte a buďte rád ! " okřikl ho	<b>šikmookáč</b> /šikmookáč/Db-----8-	. A tak Volodin prvně v životě začal šít .
okamžitě tvrdě usnul . Netrvalo však ani dvě minuty a	<b>šikmookáč</b> /šikmookáč/NNFS4-----A-----	znova s rámusem vrazil do dveří a zvolal : "
cementovou podlahu velké studené umývárny , dveře byly zamčené a	<b>šikmookáč</b> /šikmookáč/Db-----	ho nechal na pokoji . Snad je to taky člověk
pocestný , že bigošu je bál . Vzali blembák ,	<b>maskáč</b> /maskáč/NNXX-----A---8-	, polní , k tomu kanady a kvér , vyrazili
to Němec , nebo Rus , jestli se mu ošoupal	<b>maskáč</b> /maskáč/Db-----8-	na loktech . . . A teď jde o to
štreku jako nic , já na tvých březích sušival svůj	<b>maskáč</b> /maskáč/NNIS4-----A-----	zmáčený , když houkal vlak od Dolních Kralovic . Nad
mnohý z nás udělal nedobrou zkušenost , protože pod "	<b>maskáč</b> /maskáč/NNIS4-----A-----	" trampa se může schovat leckdo . Nedejme se však
obecná má krásné oči . Ropuchu zelenou chrání dokonaly "	<b>maskáč</b> /maskáč/Db-----8-	" . Program SAPARD Setkal jsem se s pojmem "
necháme ještě chvíli odležet . Podáváme s knedlíky . Králík	<b>cibuláč</b> /cibuláč/Db-----8-	Rozkrájenou cibuli osmažíme na rozpuštěné slanine , na ni vložíme
10 minut vaříme . Knedlíčky jsou výborné k zající	<b>cibuláci</b> /cibuláč/NNMP1-----A-----	" nebo samotné se zelným salátem či špenátem . Špekový
. Podáváme k ní bramborový nebo houskový knedlík . Zajíc	<b>cibuláč</b> /cibuláč/Db-----8-	Zajíce protáhneme špekem , ale nesolíme , pak jej husí
vstojte v unikajícím teple telefonních budek , nevysněná úzkost ,	<b>spoluspáč</b> /spoluspáč/NNXX-----A---8-	strach . Vymýšlím nějaké účastné slovo , lucerničku v pralesě
, vzbuzovaly důvěru . " Jsem John Werner , váš	<b>spoluspáč</b> /spoluspáč/Db-----	, uvidíte mě jen večer . Pěkně vás vítám !
zkoušely nosnost jeho žeber , ale s tím si již	<b>spoluspáč</b> /spoluspáč/Db-----	uměl poradit : štípl vždy ( podle přátelské dohody z
nocí , když ho volali k chroptícímu panu Aronovi ,	<b>spoluspáč</b> /spoluspáč/Db-----8-	se jen obrátil na boku , ale oči neotevřel .



Další malou skupinku tvoří chyby, jejichž vznik je nepochopitelný pro toho, kdo neví nic o historii vývoje nástrojů automatického zpracování přirozeného jazyka. Na následujícím obrázku vidíme doklady poměrně řídké „chyby“, kdy substantivům rodu ženského vzniklým přechylováním od substantiv rodu mužského je připojena značka odpovídající kategorii rodu slovního tvaru a lemma odpovídající fundujícímu maskulinu. Domníváme se, že tento stav je důsledkem aplikace pravidel pro automatické generování pravidelných derivací při výstavbě slovníku automatického morfologického analyzátoru.

---

náměstí, " podotkl Tetter . Nedostatek toalet hodnotí českobudějovická **zastupitelkyně** /zastupitel/NNFS1.....A..... za stranu Důchodci za životní jistoty Helena Kořenská jako jeden

V praxi se jednalo o vybrané typy paradigmatických derivací jako podstatná jména slovesná tvořená od základů shodných s pasivním přičestím, adjektiva tvořená od těchto základů, adjektiva tvořená od přechodníků, tvary II. a III. stupně adjektiv a adverbii, slovesné (a nepravidelně i další) tvary negativní tvořené pravidelně prefixem *ne-*, posesivní adjektiva tvořená od maskulin a feminin (názvů osob) sufixy *-ův* a *-in*.

Ve výše uvedených případech lze ovšem sledovat jednotnou praxi lemmatizace a morfologického značkování. Tak např. u sloves mají tvary s prefixem *ne-* jako lemma sloveso bez prefixu *ne-*, tvary II. a III. stupně adjektiv a adverbii mají (až na výjimky) lemma tvar pozitivu. Lemmatem deverbativních adjektiv a substantiv je příslušné adjektivum (substantivum). Lemmatem posesivních adjektiv je posesivní adjektivum. Z tohoto hlediska je ponechání lemmatu – fundujícího slova – odchylkou od běžné praxe.

Poslední velmi těžce zjištělnou skupinou anomálií jsou případy nesrovnalostí, které se dostaly do anotovaných korpusů ručními zásahy do automaticky zpracovaných dat na různých úrovních. Na úrovni tagu si některé pozice odpovídají. Platí, že jestliže na pozici A je B, pak na pozici X musí být Y nebo Z. Chyby způsobené ručními opravami mohou být ovšem i v souladu s pravidly platnými pro formu značky, pak je lze odhalit velmi těžko.

Tato poslední skupina je pro většinu uživatelů nezajímavá, uvádíme ji pro úplnost přehledu možných příčin chyb v lemmatizaci a anotaci.

### ***Pozice 2 – Detailní určení slovního druhu***

Na 2. pozici je uveden poněkud nepřehledný popis tzv. „Detailního určení slovního druhu“. Oč jde?

Pod touto nálepkou se skrývá a) subklasifikace tvarů slovesných (slovesných subparadigmat), b) subklasifikace adjektiv dle typu skloňování a slovtvorných charakteristik (koresponduje

s pozicí 10 Stupeň a pozicí 6 Přivlastňovací rod), c) subklasifikace druhů zájmen (koresponduje s pozicí 6 a 7), d) subklasifikace druhů číslovek, e) subklasifikace adverbii dle +/- derivace komparativu a superlativu (koresponduje s pozicí 10 Stupeň), f) různé.

Pro lepší přehled uvedeme tabulky pro a) – e<sup>6</sup>.

#### a) Detailní určení slovního druhu – klasifikace slovesných tvarů

POS	Detailní určení slovního druhu (SUBPOS)
V	[Bcefimpqst]
J	,

značka (tag)	slovní druh (1. pozice)	slovesný tvar (2. pozice)
Vf.*	sloveso	infinitiv
VB.*	sloveso	prézent/futurum (indik.)
Vt.*	sloveso	prézent/futurum arch. tvar (indik.)
Vi.*	sloveso	imperativ
Vp.*	sloveso	l-ové přičestí (vč. tvarů s -s)
Vq.*	sloveso	l-ové přičestí (vč. tvarů na -t')
Vs.*	sloveso	pasivní přičestí (vč. tvarů s -s)
Ve.*	sloveso	přechodník přítomný
Vm.*	sloveso	přechodník minulý
Vc.*	sloveso	kondicionál sl. být ( <i>bych, ...</i> )
J.*	spojka	spojky podřadicí vč. <i>aby, ... kdyby, ...</i>

#### b) Detailní určení slovního druhu – klasifikace adjektiv

POS	Detailní určení slovního druhu (SUBPOS)
A	[ACGMOU]

značka (tag)	slovní druh (1. pozice)	
AA.*	adjektivum	adjektivum obyčejné
AC.*	adjektivum	adjektivum jmenný tvar
AG.*	adjektivum	adjektivum odvozené od přech. přít.
AM.*	adjektivum	adjektivum odvozené od přech. min.
AU.*	adjektivum	adjektivum přivlastňovací (na „-ův“ i „-in“)
AO.*	adjektivum	samostatně stojící zájmena

<sup>6</sup> Podrobné popisy vztahů gramatických značek v různých tagsetech používaných pro lematizaci a značkování českých korpusů srov. např. Pořízka – Schäfer 2009, Osolobě 2007<sup>1</sup>). Vylepšená verze open-source webového rozhraní pro vyhledávání v korpusech NoSketch Engine na adrese <http://www.korpus.cz/corpora/> nabízí uživatelům při volbě **Typ dotazu tag** „uživatelsky přítulnější“ přístup k volbě značky na 2. pozici.

		„svůj“, „nesvůj“, „tentam“
--	--	----------------------------

### c) Detailní určení slovního druhu – klasifikace druhů zájmen

POS	Detailní určení slovního druhu (SUBPOS)
P	[01456789DEHJKLPQSWYZ]

POS&SUBPOS	tvary – příklady	popis
P0	<i>naň</i>	spřežka předložka+osobní zájmeno <i>on</i>
P1	<i>jehož</i>	vztažné zájmeno <i>jehož</i>
P4	<i>jaký, který</i>	tázací zájmeno <i>čí, čípak, jaký, jakýpak, jakýž, jakýže, který, kterýpak, kterýž, ...</i>
P5	<i>něj</i>	osobní zájmeno <i>on</i> tvary po předložce ( <i>n-</i> )
P6	<i>sebe</i>	zvratné zájmeno tvary <i>sebe, sobě, sebou</i>
P7	<i>se, si</i>	zvratné zájmeno tvary <i>se, si, ses, sis</i>
P8	<i>svůj</i>	přivlastňovací zvratné zájmeno <i>svůj</i>
P9	<i>něhož</i>	vztažné zájmeno <i>jehož</i> tvary po předložce ( <i>n-</i> )
PD	<i>tento</i>	ukazovací zájmena <i>ten, tento, takový, tehle, onen, týž, tentýž, takovýto, takovýhle, tehleten, toť, tamten, taký, tamhleten, tadyten, tuhleten</i>
PE	<i>což</i>	vztažné zájmeno <i>což</i>
PH	<i>mě</i>	krátké (příklonné) tvary osobních zájmen <i>mi, mě, ti, tě, ji, je, ...</i>
PJ	<i>jenž</i>	vztažné zájmeno <i>jenž</i>
PK	<i>kdo</i>	vztažné/tázací zájmeno <i>kdo, kdopak, kdožpak, kdož, kdos</i>
PL	<i>všechn</i>	zájmena vymežovací (limitativa) <i>všechno, všecek, sám, samý, veškery</i>
PP	<i>ty</i>	osobní zájmena <i>já (my), ty (vy), on, tvar tys</i>
PQ	<i>co</i>	vztažné/tázací zájmeno <i>co, copak, cožpak, cos, což</i>
PS	<i>můj</i>	přivlastňovací osobní zájmena <i>můj, tvůj, jeho, náš, váš</i>
PW	<i>nic</i>	záporná zájmena <i>nic, žádný, nikdo, pranic, nijaký, pražádný, nižádný</i>
PY	<i>oč</i>	spřežka vztažné/tázací zájmeno předložka+č ( <i>oč, nač, zač, več, ...</i> )
PZ	<i>nějaký, něco</i>	neurčitá zájmena <i>některý, něco, nějaký, někdo, jakýsi, jakýkoli, jakýkoliv, cosi, cokoliv, málokdo, kdosi, kdokoli, kterýkoli, leccos, kdokoliv, ničí, kterýkoliv, všelijaký, kdekdo, málokterý, leckdo, leckterý, něčí, ledacos, kdejaký, kterýsi*, jakýs*, kdecos, máloco, čísi, takýs*, bůhvíjaký, ledajaký, bůhvíco, lecjaký, všelicos, kdovíjaký, lecco, kdekterý, kdože, kdovíco, ledasco, ký, ledaco, ledaskdo, nevímjaký, bůhvíkdo, kdovíkdo, všelico, čertvíkdo, čertvíco, číkoliv, nevímkdo, číkoli, nevímčí, ledakdo, kdovíčí, zřídkakdo, ledakterý, čertvíjaký, všelikterý</i>

### d) Detailní určení slovního druhu – klasifikace druhů číslovek

POS	Detailní určení slovního druhu (SUBPOS)
C	3=?adhjklnouvwyz}

POS&SUBPOS	tvary – příklady	popis
C=	<i>1</i>	arabské číslice
C}	<i>XIV</i>	římské číslice
Ca	<i>mnoho</i>	tvary „číslovky“ <i>mnoh-o,-a, ...</i>
Cd	<i>čtvery</i>	druhové číslovky <i>dvoji, obojí, troji</i> a další tvořené sufixem <i>-erý</i>
Ch	<i>jedny</i>	druhová číslovka <i>jedny</i>
Cj	<i>čtvero</i>	úhrnné číslovky <i>dvé, obé, tré</i> a další tvořené sufixem <i>-ero</i>
Ck	<i>čtvery</i>	souborové číslovky <i>dvoje, oboje, troje</i> a další tvořené sufixem <i>-ery</i>
Cl	<i>tři</i>	základní číslovky <i>jeden, dva, oba, tři, čtyři</i>
Cn	<i>pět</i>	základní číslovky <i>pět</i> a výše
Co	<i>tolikrát</i>	číslovka zájmenná ukazovací násobná <i>tolikrát</i>
Cr	<i>druhý</i>	číslovky řadové
Cu	<i>kolikrát</i>	číslovka zájmenná tázací násobná <i>kolikrát</i>
Cv	<i>sedmkrát</i>	číslovky určité násobné <i>.*-krát</i>
Cw	<i>nejeden</i>	nejeden
Cy	<i>desetina</i>	číslovky dílové vyjadřující určitý počet <i>.*-ina</i>
Cz	<i>kolikátý</i>	číslovka zájmenná tázací/vztažná řadová <i>kolikátý</i>
C?	<i>kolik</i>	číslovka <i>kolik</i>

#### e) Detailní určení slovního druhu – klasifikace adverbii

POS	Detailní určení slovního druhu (SUBPOS)
D	bg

POS&SUBPOS	tvary – příklady	popis
Db	<i>nahoru</i>	všechna příslovce, která nelze stupňovat
Dg	<i>rychle</i>	příslovce, která lze stupňovat

#### Pozice 3 – Jmenný rod

- neurčuje se

**F** femininum (ženský rod)

**H** femininum nebo neutrum (tedy nikoli maskulinum)\*

**I** maskulinum inanimatum (rod mužský neživotný)

**M** maskulinum animatum (rod mužský životný)

**N** neutrum (střední rod)

**Q** femininum singuláru nebo neutrum plurálu (pouze u příděstí a jmenných adjektiv)\*

**T** maskulinum inanimatum nebo femininum (jen plurál u příděstí a jmenných adjektiv)\*

**X** libovolný rod (F/M/I/N)

**Y** maskulinum (animatum nebo inanimatum)\*

**Z** ‚nikoli femininum‘ (tj. M/I/N; především u příslovci)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2000, SYN2005, SYN2006PUB, ORWELL.

Kromě značek jednoznačně určujících rod tak, jak jsme na to zvyklí z běžné školské praxe, tedy [MIFN], jsou v některých starších korpusech hodnoty nejednoznačné.<sup>7</sup> V následující tabulce uvedeme přehled vztahu nejednoznačných hodnot kategorie rodu k obsazení první pozice (ke slovnímu druhu) a následně vysvětlíme, pokud možno, proč je tomu tak (viz níže).

POS	Jmenný rod
[PCV]	H
[AV]	T
[AV]	Q
[NAPCV]	X
[APCV]	Y
[PC]	Z

#### ***Pozice 4 – Číslo***

- neurčuje se

**D** duál (pouze 7. pád feminin)

**P** plurál (množné číslo)

**S** singulár (jednotné číslo)

**W** pouze v kombinaci s jmenným rodem ‘Q’ (singulár pro feminina, plurál pro neutra)\*

**X** libovolné číslo (P/S/D)

\* Tato značka je k dispozici pouze v korpusech: SYN2000, SYN2005, SYN2006PUB, ORWELL.

Následující část textu může čtenář, kterého (alespoň prozatím) nezajímají podrobnosti týkající se proměn morfologické anotace, ponechat stranou. Do textu byly zařazeny proto, že odborníkům podílejícím se na automatickém značkování a desambiguaci korpusů ČNK jsou tato fakta rámcově známa, široké veřejnosti ovšem nikoli. Výklad se snaží nezabředat do přílišných detailů, chce jen upozornit na to, co v běžně dostupných manuálech není explicitně uvedeno. Lepší přehled mohou samozřejmě podat odborné články autorů, kteří se na anotační praxi v průběhu let podíleli.

Vzhledem k tomu, že popisy případů, kdy je ve značce ponechána nejednoznačnost, jsou poněkud lakonické, rádi bychom upozornili na důvody, které k zavedení těchto nejednoznačných hodnot vedly. Především si opět musíme připomenout, že automatická

---

<sup>7</sup> Značky s nedesambigovanými významy rodu, čísla, pádu aj. se v korpusech řady SYN od roku 2009 nepoužívají (desambiguují se).

morfológická analýza je vysoce nejednoznačná. Některé typy víceznačností jsou systémové (platí např. pro všechna slovesa/slovesné tvary). Tak např. všechny přechodníky v češtině mají pouze tři tvary a všechna participia I-ová a pasivní pouze pět tvarů pro vyjádření osmi (čtyři rody x dvě čísla) významů reprezentovaných jednoznačnou morfológickou značkou. To, co platí pro přičestí, platí i pro některá adjektiva, tedy pro jejich jmenné tvary, pokud budeme brát v úvahu pouze tvary nominativu a rezignujeme na akuzativy. Velká část číslovek je ve vztahu ke kategorii rodu a čísla defektní a desambiguovat je představuje problém i pro manuální analýzu, jak jsme v praxi ověřili při ruční desambiguaci Korpusu soukromé korespondence (srov. více Hladká a kol. 2005 a Osolsobě 2006). Zájmena a do jisté míry i číslovky představují soubor velmi frekventovaných jednotek (lemmat) s vysoce ambiguitní zájmenou, číslovkovou, či adjektivní flexí.

Za překlep v popisu morfológických značek pokládáme vysvětlující text k hodnotě Z na třetí pozici rod, a sice „především u příslovcí“, značka se totiž ve všech přístupných korpusech řady SYN vyskytuje pouze u zájmen a číslovek zájmené flexe (vzory *ten* a *náš*) a tvrdé složené adjektivní flexe (vzor *mladý*), zřejmě tedy zde má stát „především u zájmen“.

Vztahy mezi prvními čtyřmi pozicemi podrobněji ukazují následující tabulky.

V první tabulce je dle pozic (sloupce) uvedeno obsazení prvních čtyř pozic ambiguitní hodnotou kategorie rodu. Druhá tabulka mapuje frekvenční zastoupení lemmat v jednotlivých korpusech. Následuje interpretace frekvenční distribuce lemmat.

Rod X – tabulka 1

POS		Jmenný rod	Číslo
N	N	X	[SPX]
A	A	X	[PX]
A	U	X	X
P	[DJLPSWZ14589]	X	[PX]
C	[dhl]	X	[SPX]
V	[em]*	X	X

Vp.X.\* – výsledky guesserů chili/chit/ ...

Rod X – tabulka 2

	SYN2000	SYN2005	SYN2006PUB
NNX.*	3	44625	153
AAX.*	179	45338	937
AUX.*	3105	2	4
P.X.*	57	634	60
C.X.*	14	1949	20

V[em]X.*	875	1166	697
----------	-----	------	-----

Počty lemmat substantiv (NN.\*), obyčejných adjektiv (AA.\*), zájmen a číslovek v korpusu SYN2005 vysoce převyšují počty lemmat v ostatních korpusech. Je to způsobeno testováním hadačů, které bylo součástí značkování korpusu SYN2005. Naopak k dramatickému rozdílu ve značkování posesivních adjektiv (AU.\*) přispěla zřejmě systémová úprava morfologického analyzátoru.

S ohledem na rozdíl zastoupení žánrů v korpusu SYN2000 a SYN2005 nepřekvapí, že v obou vyvážených korpusech je vyšší počet lemmat tvarů přechodníků než v korpusu žánrově kompaktním, ani to, že nejvyšší počet lemmat těchto tvarů má právě korpus s nejvyšším zastoupením beletrie (SYN2005).

Rod Y – tabulka 1

POS		Jmenný rod	Číslo
A	C	Y	S
A	O	Y	S
P	[DJLPSWZ48]	Y	[SP]
C	[dhlwz]	Y	[SP]
V	[emqps]	Y	[SP]

Rod Y – tabulka 2

	SYN2000	SYN2005	SYN2006PUB
ACY.*	64	89	86
AOY.*	2	3	3
P.Y.*	57	220	63
C.Y.*	6	16	13
V.Y.*	14006	17936	14309

Rozdíly v počtech lemmat odpovídají rozdílnému lexiku. Není tak patrný podíl gesserů.

Rod T – tabulka 1

POS		Jmenný rod	Číslo
A	C	T	P
V	[pqs]	T	P

Rod T – tabulka 2

	SYN2000	SYN2005	SYN2006PUB
ACTP.*	75	79	84
V.TP.*	9680	11058	10473

Rozdíly v počtech lemmat odpovídají rozdílnému lexiku. Hadače žádnou významnou roli nehrají.

Rod Q – tabulka 1

POS		Jmenný rod	Číslo
A	C	Q	W
V	[pqs]	Q	W

Rod Q – tabulka 2

	SYN2000	SYN2005	SYN2006PUB
ACQW.*	87	704	92
V.QW.*	11856	14475	12236

Femininum singuláru nebo neutrum plurálu (pouze u přičestí a jmenných adjektiv). Hadače jsou opět důvodem pro patrný nepoměr počtu tvarů označovaných jako adjektiva v korpusu SYN2005. Prohlédneme-li si označované tvary, zjistíme, že jde o řadu překlepů, kdy řetězec znaků (token pro automatickou analýzu) končí na *y*. Toto zakončení z lingvistického hlediska nutně implikuje buď slovo s adjektivním skloňováním tvrdé složené flexe, nebo částice (*prý*), či citoslovce (*hý*, *čehý*). Pokud má jít o zakončení tvaru skloňovaného adjektivně, není tento tvar jednoznačný z hlediska rodu. Rozhodně ale nejednoznačnost v rodu neodpovídá nejednoznačnosti značky Q. Důvodem bude pravděpodobně to, že hadače nebyly založeny na žádné, nebo jen velmi málo propracované lingvistické bázi (kategorii slovního druhu dle zakončení jedním znakem lze v češtině jednoznačně určit pouze v minimu případů), a že naopak byly budovány na základě metod matematické statistiky (více Hlaváčová 2001).

Rod H – tabulka 1

POS		Jmenný rod	Číslo
P	[S8]	H	[SP]
C	l	H	P
V	[em]	H	S

Rod H – tabulka 2

	SYN2000	SYN2005	SYN2006PUB
P.H.*	3 (můj, tvůj, svůj)	3 (můj, tvůj, svůj)	2 (můj, tvůj)
C.H.*	2 (dvě, oba)	3 (dvě, oba, voba)	3 (dvě, oba, voba)
V.H.*	1051	1266	566

S ohledem na rozdíl zastoupení žánrů v korpusu SYN2000 a SYN2005 nepřekvapí, že v obou vyvážených korpusech je vyšší počet lemmat tvarů přechodníků než v korpusu žánrově kompaktním.



Rod Z – tabulka 1

POS		Jmenný rod	Číslo
P	[DJLHPSWZ14589]	Z	S
C	[lwz]	Z	S

Rod Z – tabulka 2

	SYN2000	SYN2005	SYN2006PUB
P.ZS.*	57	68	60
C.ZS.*	4	33	7

Vyšší počet číslovkových lemmat v korpusu SYN2005 je opět (dle dat) možno přičíst na vrub testování guesserů.

Korpus SYN (není referenční a je na něm průběžně testována desambiguace řízená pravidly) a korpusy SYN2009PUB i SYN2010 ambiguitní hodnoty pro rod a číslo nemají. Byly odstraněny systematicky (více Jelínek 2008, Jelínek – Petkevič 2011, Skoumalová 2011).

**Pozice 5 – Pád**

- neurčuje se

**1** nominativ (1. pád)

**2** genitiv (2. pád)

**3** dativ (3. pád)

**4** akuzativ (4. pád)

**5** vokativ (5. pád)

**6** lokativ (6. pád)

**7** instrumentál (7. pád)

**X** libovolný pád (1/2/3/4/5/6/7)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2000, SYN2005, SYN2006PUB, ORWELL.

Pád X – tabulka 1

POS		Jmenný rod	Číslo
N	N	[MIFNX]	[SPX]
A	[AU]	[IFX]	[PX]
P	[S14]	[FX]	[SX]
C	[an]	-	[-S]
R	R	-	-

Pád X – tabulka 2

	SYN2000	SYN2005	SYN2006PUB
NNMSX.*	2	3033	7
NNMPX.*	10	8166	24
NNMXX.*	140	6940	363
NNISX.*	14	3802	32
NNIPX.*	36	6563	62
NNIXX.*	999	22333	1512
NNFSX.*	12	4778	71
NNFPX.*	58	8475	85
NNFXX.*	1332	47544	2332
NNNSX.*	6	2042	225
NNNPX.*	9	1418	28
NNNXX.*	1221	21268	2350
NNXXX.*	3	42532	148
AAIXX.*	1	1	1
AAFXX.*	3	3	3
AAXPX.*	1	2	2
AAXXX.*	178	45336	935
AUXXX.*	3105	8	4
PSXXX.*	1 (jeho)	339	4
PSFSX.*	1 (jeho/její)	1	1
P1XXX.*	1(jenž)	1	1
P1FSX.*	1 (jejíž/jenž)	1	1
P4XXX.*	1 (co)	1	1
Ca--X.*	1 (moc)	2 (mnoho, moc)	2 (moc, mnoho)
Cn-SX.*	1 (polovic)	46	1
RR--X.*	8	4177	15

Značku X – libovolný pád mají substantiva a adjektiva nesklonná, zkratky a některá další. U zájmen ji mají nesklonná zájmena vztažná *jenž* a *co* a nesklonná slova *moc* a *polovic* hodnocená jako číslovky. Hodnota X na pozici 5 se vyskytuje u předložek z cizojazyčných textů. Vyšší počet lemmat v korpusu SYN2005 je opět (dle dat) možno přičíst na vrub testování guesserů.

### **Pozice 6 — Přivlastňovací rod**

Rody mužský neživotný a střední se nikdy nevyskytují samostatně. 'M' se může vyskytnout jen u přivlastňovacích adjektiv (ne u příslovčí).<sup>8</sup>

- neurčuje se

**F** femininum (ženský rod)

**M** maskulinum animatum (rod mužský životný)

**X** libovolný rod (F/M/I/N)

**Z** 'nikoli femininum' (tj. M/I/N; u přivlastňovacích adjektiv)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2000, SYN2005, SYN2006PUB, ORWELL.

Přivlastňovací rod – tabulka 1

POS		Jmenný rod	Číslo	Pád	Přív. r.
A	U	[MIFNX]	[SPX]	[1234567X]	[MF]
P	[1S]	[MIFNZX]	[SPXD]	[123467X]	[FXZ]

U adjektiv posesivních je přivlastňovací rod maskulinum u derivátů na *-ův* a u lemmatu *Pučálkovic*, přivlastňovací rod femininum mají deriváty na *-in*. U zájmen se jedná o přivlastňovací zájmena *jeho, její, jejich*.

### **Pozice 7 — Přivlastňovací číslo**

- neurčuje se

**P** plurál (množné číslo)

**S** singulár (jednotné číslo)

Uvádí se pouze u posesivních (na druhé pozici ve značce je uvedeno *S*) a vztažných posesivních (na druhé pozici ve značce je uvedeno *I*) zájmen, P[S1].\* (*můj, tvůj, náš, váš, její, jejich, jenž, jeho* a překvapivě též u některých tvarů *my*<sup>9</sup>, které jsou prokazatelně dle kontextu tvary českého osobního zájmena 1. osoby plurálu).

### **Pozice 8 — Osoba**

- neurčuje se

<sup>8</sup> Jde patrně o překlep.

<sup>9</sup> Dle přehledu slovních tvarů jde o chybu desambiguace. Ve slovníku automatického analyzátoru je totiž přiřazeno lemma *my* též tvarům *my* a *our*, tedy anglickým zájmenům přivlastňovacím.

**1** 1. osoba

**2** 2. osoba

**3** 3. osoba

**X** libovolná osoba (1/2/3)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2000, SYN2005, SYN2006PUB, ORWELL.

Osoba X – tabulka 1

POS		osoba
V	[pqs]	X

Značku libovolná osoba mají všechna participia l-ová a pasivní.

### ***Pozice 9 – Čas***

- neurčuje se

**F** futurum (budoucí čas)

**H** minulost nebo přítomnost (P/R)\*

**P** prézens (přítomný čas)

**R** minulý čas

**X** libovolný čas (F/R/P)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2000, SYN2005, SYN2006PUB, ORWELL.

Čas X – tabulka 1

POS		čas
V	s	X

Značku libovolný čas mají všechna participia pasivní.

### ***Pozice 12 – Aktivum/pasivum***

- neurčuje se

**A** aktivum nebo 'nikoli pasivum'

**P** pasivum

Na tomto místě se pokusíme rekapitulovat stručně pro potřeby běžných uživatelů korpusů to, co jsme už publikovali o vztazích hodnot morfologických značek sloves v širě zaměřené studii srovnávající tagsety používané pro značkování českých a slovenských korpusů (Osolsobě 2007<sup>1</sup>).

Pro větší přehlednost opět uvedeme tabulky, v nichž tentokrát poukážeme na vztah mezi pozicí 2 (Detailní určení slovního druhu, tedy konkrétně těmi hodnotami, které se týkají značkování dílčích slovesných subparadigmat), pozice 9 (Čas) a pozice 12 (Způsob).

POS	Slovesný tvar	Čas	Aktivum/pasivum
V	[Bt]	F	A
V	[Bt]	P	A
V	[pq]	R	A
V	s	H	P
V	s	X	P

V anotaci korpusů řady SYN mají značku VB.....F.\*, tedy budoucí čas (F), vyznačeny tvary budoucího času slovesa *být* používané též pro tvoření analytických tvarů futura (*budu, budeš, bude, ...*) a tvary některých nedokonavých sloves, které mohou tvary pomoci tvarů s prefixem *po-* + indikativu přítomnosti aktiva tvořit tzv. syntetické futurum.

Ostatní tvary zařazené značkou jako „slovesný tvar přítomného nebo budoucího času“ (VB. \*) mají u atributu čas vyplněnou hodnotu *prézens* (přítomný čas).

Problematické se nám zdá, že ve všech korpusech je nejednoznačně vyřešeno značkování celé řady tvarů syntetického futura.

Značku tag=“VB.....F.\*” a lemma bez prefixu *po-* mají totiž pouze tvary cca 20 sloves, jako např. *jít/půjde, jet/pojede, nést/ponese, běžet/poběží, letět/poletí, téct/poteče, vézt/poveze, hrnout/pohne (se), lézt/poleze, plout/popluje, cestovat/pocestuje, stěhovat/postěhuje, řídit/pořídí (se), trvat/potrvá a plazit/poplazi (se)*. Značku tag=“Vi.....-.\*” a lemma bez *po-* mají tvary sloves *jít/pojď, slyšet/poslyš, jet/pojed, lézt/polez, běžet/poběž, letět/polet, nést/pones, stěhovat/postěhuj*.

Ze sond do materiálu všech sledovaných korpusů je ovšem patrné, že by sem měla patřit ještě řada dalších slovesných tvarů. Problémy v lemmatizaci i značkování jsou patrně způsobeny nedostatky ve slovníku automatického morfologického analyzátoru. Jedná se totiž, jak je zřejmé, o otevřenou skupinu sloves, která takto futurum tvoří. Doplnění slovníku by ovšem patrně velmi znesnadnilo desambiguaci, protože v řadě případů se vyskytuje homonymie slovesných tvarů tvořených prefixem *po-* se slovesnými tvary tvořícími futurum prefixem *po-* (např. *tak večer <pomažou/mazat> do díry ... x ... které <pomažeme/pomazat> pikantní marmeládou ...*).

V korpusech SYN2009PUB a SYN2010 mají F na pozici 9 navíc ještě slovesa v pasivním přičestí, a to jak dokonavá, tak nedokonavá (více Petkevič 2010).

Minulý čas (R) mají vyplněny výhradně tvary minulého přičestí (včetně tvarů s volným morfémem -s signalizujícím hodnotu 2. osoby sg.), tedy ty, které mají na úrovni slovesného tvaru vyplněno **Vp.\*** nebo **Vq.\***. Zdá se nám, že jde o redundantní informaci.

V SYN2000/SYN2005/SYN2006PUB mají kategorii času vyznačena n-/t-ová participia. Příslušná značka se realizuje následovně: na 9. pozici (ČAS) je hodnota **X** (libovolný čas) u participií n-/t-ových. Odlišná hodnota **H** (minulost nebo přítomnost) je uvedena u následujících slovních tvarů: *Vitas/vít/VsFS---2H-AP---/5*, *litas/lít/VsFS---2H-AP---/4*, *Vitos/vít/VsNS---2H-AP---/3*, *Plutos/plout/VsNS---2H-AP---/2*, *minutos/minout/VsNS---2H-AP---/1*, *nadrženos/nadržít/VsNS---2H-AP---/1*, *bitas/bit/VsFS---2H-AP---/1*, *rytas/rýt/VsFS---2H-AP---/1*, *velenos/velet/VsNS---2H-AP---/1*, *přenos/přít/VsNS---2H-AP---/1*, *Kutas/kovat/VsFS---2H-AP---/1*, *Ryotos/rýt/VsNS---2H-AP---/1*, *Jatas/jmout/VsFS---2H-AP---/1*.

Jde zřejmě o pokus zachytit na úrovni automatické morfologické analýzy hypotetické tvary typu *bitas-li byla = byla jsi bita*, *minutos rozumem? = minuto jsi rozumem?* *jatas byla = byla jsi jata*, ... Z nahlédnutí do konkordančního seznamu je patrné, že se o takové případy nejedná. Jde o chybné značkování a automatická morfologická analýza „pregenerovává“. (Generuje tvary z hlediska systému „správné“, ale z hlediska úzu periferní, takže se může stát, že tvary jsou homonymní s náhodně se vyskytnuvšími jinými periferními či nesprávnými tvary. Typickým příkladem může být obecně např. značkování tvaru *der*, které je v daném kontextu součástí cizojazyčného – německého – textu jako tvaru imperativu slovesa *drát*: *der/drát/Vi-S---2-A---*, nebo zde uvedené přiřazení neadekvátní interpretace vlastním jménům *Vitas* a *Kutas*, nebo překlepům.)<sup>10</sup>

<sup>10</sup> Výše popsání řešení zůstávají do jisté míry na půli cesty a nejsou průhledná. Mají-li a) všechny tvary l-ového přičestí s volným -s (*přišels*) na úrovni atributu čas hodnotu minulost, b) tvary pasivního participia s volným -s hodnotu přítomnost nebo minulý čas (*minutos rozumem = minuto jsi rozumem*, *minutos bylo rozumem = minuto jsi bylo rozumem*) a c) všechny tvary *ty* s volným -s (*tys*) hodnotu přezens, pak by mělo být explicitně řečeno, proč se u všech ostatních tvarů s volným -s význam (hodnota) čas „nevyplňuje“, ačkoliv bychom mohli předpokládat přinejmenším ambiguitní hodnotu minulost nebo přítomnost ( *...žes vyhoštěn ze sedadel knížecích...*, *?...to by ses před ní ukázal v pěkném světle...*, *?... přesto by ses byl se mnou vyspal ...*, srov. SYN2000). Pokud by platilo pouze a) a b), pak by se nabízela tato odpověď: Protože ostatní tvary nejsou tvary, které mají na první pozici ve značce vyznačenu hodnotu slovní druh *V* (sloveso). Vzhledem k tomu, že platí také c), je patrné, že se hodnoty slovesných atributů neurčují vždy jen u tvarů, které mají hodnotu atributu slovní druh sloveso. Z toho plyne, že v některých případech se u tvarů s volným -s v hodnotách uvedených na úrovni morfologické značky odrazí hodnoty implikované jeho přítomností a v jiných ne. Hodnota atributu čas je u tvarů *tys* vždy přezens, ačkoliv bychom opět očekávali přinejmenším ambiguitní hodnotu minulost nebo přítomnost

V korpusech SYN2009PUB a SYN2010 mají pasivní přičestí (*Vs.\**) vyplněnu na pozici 9 hodnotu P nebo R, přičemž toto značkování je výsledkem pravidlové desambiguace (více Petkevič 2010). Jakožto případy potencionálních výskytů pasivních přičestí s volným morfémem *-s* jsou opět označovány pouze překlepy.

Korpus:

Typ dotazu:

CQL:

LYD 116.088 Lichtenštejnsko Švýcarský frank CHF 118.005 Korunový kurz Litva	<b>Litas</b> /lit/VsFS---2R-AP---I	LTL 17.750 Lotyšsko Lat LVL 137.952 Macao Pataca MOP 12.526
letišť je taxi, za které nezaplatíte víc než 20	<b>litas</b> /lit/VsFS---2P-AP---I	. Taxi lze objednat na telefonním čísle 1446 ( anglicky
) - pořádáno ve spolupráci s galerií FONS 17.15 Agnes	<b>Kutas</b> /kout/VsFS---2R-AP---I	( Hu ) + Jaroslav Kořán : koncert - 2
syna Jiřího . Žije v Praze . Ilustrace : Nenad	<b>Vitas</b> /vit/VsFS---2P-AP---I	Dva alfa samci , dva svůdci Další věc , v
bankomaty jsou ve Vilniusu doslova na každém kroku . Litevský	<b>litas</b> /lit/VsFS---2R-AP---I	se dělí na 100 centas , aktuální kurs vůči české
je nevinný , ale vy ne ! Ta jeho ' <b>zvrátenos</b> /zvrátit/VsNS---3R-AP---P		' se nemůže rovnat svinstvu , které pácháte vy .
některých z nich ( Metro , 20 Minuten , 20	<b>Minutos</b> /minout/VsNS---3R-AP---P	) prozrazují , proč koncept funguje : ekonomicky aktivní čtenáři
CIA , že neležíme v oceánu . Ilustrace : Nenad	<b>Vitas</b> /vit/VsFS---2P-AP---I	Číslo týdne 39 mld. \$ O tolik dolarů méně se
Saaremaa , Hiiumaa a Muhu , Měna : Litva 1	<b>litas</b> /lit/VsFS---2R-AP---I	( LTL ) - 9 Kč , Lotyšsko 1 lat
se můžete stát účastníky tohoto neobvyklého koncertu . 16.15 Agnes	<b>Kutas</b> /kout/VsFS---2R-AP---I	( Maďarsko ) + Jaroslav Kořán : koncert - 1
- 1. část ( Zelená brána ) - Agnes	<b>Kutas</b> /kout/VsFS---2R-AP---I	je svérázná hudebnice pocházející z Budapešti . Zpívá a hraje

## Pozice 10 — Stupeň

- neurčuje se

- 1 1. stupeň
- 2 2. stupeň
- 3 3. stupeň

Pozice 10 – Stupeň koresponduje s pozicí 1 a 2. Hodnotu [123] mohou mít vyplněnu pouze „obyčejná adjektiva“ (*AA.\**) a „příslovce (s určením stupně a negace; „velký“, „zajímavý“, ...)“ (*Dg.\**). Drobnou nekonzistenci lze spatřit v tom, že tvary *rád*, *ráda*, *rádo*, *rády*, *rádi* mají lemma *rád* stejně jako tvary *raděj(i)*, *nejraděj(i)*, ale značku v případě pozitivu *AC.....-.\** a v případě komparativu a superlativu *Dg.....[23].\** Pokusíme-li se aplikovat substituční test (nahradit tvary označené jako adjektiva tvary označenými jako adverbia) zdá se nám, že

---

(...*tys* přišel ..., ...*tys* tam byl přijat..., ... *tys* nejlepší filozof ..., ... *tys* ta jediná ... doklady jsou z korpusu SYN2000). Nemýlíme-li se, nejedná se o chybu statistické desambiguace, ale o chybu způsobenou již na úrovni automatické morfologické analýzy.

značkování poněkud mechanicky aplikuje technickou zásadu, podle níž se jmenné tvary adjektiv v češtině nestupňují.

" souhlasil Jack a kradmo pohlížel na posádku . "	<b>Rádi</b> /rád/ACMP-----A-----	bychom vystoupili , prosím . " Než Aven odpověděla ,
, který si chůva přilepila na patro . " Nemám	<b>ráda</b> /rád/ACFS-----A-----	vodu , Jemmo , " svěřovala se vážně s víčky
1999 Mezi kandidáty je člověk , kterého já mám docela	<b>rád</b> /rád/ACMS-----A-----	. A to je předseda Bursík . To je takový
Jen nejmladší sestra se prohlížení neúčastní . " Já se	<b>radši</b> /rád/Dg-----2A---1-	ani nebudu dívat , " říká . Pak starší sestře
do půl pasu . Japonci milují svou vlast Každý má	<b>rád</b> /rád/ACMS-----A-----	, když je jeho národ chválen . Nejinak Japonec .
Včera večer jsem měla dojem , že máš svou práci	<b>rád</b> /rád/ACMS-----A-----	. " " Dokazovat pomalý zánik Země ? Na tom
houbami Několik posledních let se stavuji v Kladně . Přicházím	<b>rád</b> /rád/ACMS-----A-----	a s ulehčením odcházím . Na Rozdělově , vysokém "
Je jak z kamene . Ptám se , jestli má	<b>rád</b> /rád/ACMS-----A-----	to , co dělá , kolik mu platěj a jestli
Kam se ta stará dobrá poděla ? To bych taky	<b>rád</b> /rád/ACMS-----A-----	věděl . . . Snažíme se nějak hrát , dostat
to , v jaké jste dimenzi . Možná že byste	<b>rádi</b> /rád/ACMP-----A-----	kontaktovali někoho , kdo je už mrtvý . Je to
na Rickyho . Tak strašně moc , že bys mu	<b>nejraději</b> /rád/Dg-----3A-----	ublížila . Takže se vydáš do města , abys tam
, " přeje si . Na výsledek z Číny nebude	<b>rád</b> /rád/ACMS-----A-----	vzpomínat , jinak si Hry pochvaluje . " Organizátoři zvládli
na rozmarýnu , pečeme mazance a beránka . Taky máme	<b>rádi</b> /rád/ACMP-----A-----	balkánské jídlo tzv. kapamu . Kapama se dělá z lístků

Jisté pochybnosti může vzbudit též lemmatizace některých adjektiv na *-oucí* (více níže a též Osolsobě 2009<sup>4</sup>).

### **Pozice 11 — Negace**

- neurčuje se
- A** afirmativ (bez negativní předpony "ne-")
- N** negace (tvar s negativní předponou "ne-")

Pozice 11 – Negace působí rozpaky u značkování některých negativ tantum v SYN2000 a SYN2005. Technické řešení spočívající v praxi, kdy lemmatem tvaru s *ne-* je tvar bez *ne-*, nepůsobí větší problémy v případě lemmatizace slovesných tvarů (tato praxe odpovídá praxi tištěných slovníků, jde o tvary víceméně paradigmatické). Má ale svá úskalí na úrovni lemmatizace některých negativ tantum (*nezbytnost/zbytnost, nezbytný/zbytný, nezbytně/zbytně, ...*). Tyto chyby byly v anotacích korpusů posléze odstraněny (srov. anotace v korpusu SYN2000, SYN2005 a v korpusech SYN2006PUB a dalších).

Atribut negace s hodnotou afirmace je v SYN2000/SYN2005 uveden ve značce tvaru *tys*. Tvary *ses, sis, zes, kohos, ...* mají u negace vyznačenu hodnotu „neurčuje se“. Jde opět o nejednotné řešení analogických případů.

### **Pozice 13 — Nepoužito**

- neurčuje se



## Pozice 14 — Nepoužito

- neurčuje se

## Pozice 15 — Varianta, stylový příznak apod.

- neurčuje se („základní“ tvar pro kategorie v pozicích 1–14)
- 1 varianta, víceméně rovnocenná („méně častá“)
- 2 řídká, archaická nebo knižní varianta
- 3 velmi archaický tvar, též hovorový
- 4 velmi archaický nebo knižní tvar, pouze spisovný (ve své době)
- 5 hovorový tvar, ale v zásadě tolerovaný ve veřejných projevech
- 6 hovorový tvar (koncovka standardní obecné češtiny)
- 7 hovorový tvar (koncovka standardní obecné češtiny), varianta k '6'
- 8 zkratky
- 9 speciální použití (tvary zájmen po předložkách apod.)

Pozice 15 je vyplněna na základě slovníku automatického morfologického analyzátoru. Nelze se absolutně spolehnout na to, že při vyhledávání pomocí této pozice tagu zachytíme všechny případy, které se skutečně budeme snažit vyhledat, protože ve slovníku nejsou zahrnuty všechny substandardní jevy, ale jenom některé. Z následující konkordance je patrné, že nerozpoznané tvary (ty, které nejsou ve slovníku a mají na první pozici značku X) mají na pozici 15 uvedeno „neurčuje se („základní“ tvar pro kategorie v pozicích 1–14)“.

**Korpus:**

**Typ dotazu:**

**Slovní tvar:**

---

, řekl jsem mu : " Neměl byste tý ženský	<b>podlejtat</b> /podlejtat/X@-----	. Má oplzlý myšlenky . Polovina americkéjch matek s drahocennýma
Ať ty zelený sračky daj tolik , až mi budou	<b>vylejtat</b> /vylejtat/Vf-----A----I	oči z důlk ů. Obyčejně se z takový potrawy poseru
je bumerang . " Škorpil pokračuje : " Musí latku	<b>podlejtat</b> /podlejtat/X@-----	místo přeskakování . " Víte už teď , kolik budete
kotlí a fandil . Slavii respektuju jako soupeře , ale	<b>podlejtat</b> /podlejtat/X@-----	jí nebudu . Slavii ve finále vyzve Sparta Znojmo historicky
bys našla nějaký , ze kterých ti tvý půlky budou	<b>vylejtat</b> /vylejtat/Vf-----A----I	ještě víc . " " Jo , to mi povídej
že se promění v hada , prokřistapána , a bude	<b>oblejtat</b> /oblejtat/X@-----	a chřestit . Řekni , jestli to ale není pecka
" To mi za to nestojí , kvůli pěti dolarům	<b>vylejtat</b> /vylejtat/Vf-----A----I	do toho hnusnýho povětří . " " Vidělas to v
Na Mary zařival , ať jí ani nenapadne z auta	<b>vylejtat</b> /vylejtat/Vf-----A----I	, že je tam bahno , a to byla poslední

### *Pozice 16 – Vid*

Tato pozice byla k původní sadě doplněna Miroslavem Spoustou na základě slovníku morfologické analýzy. Tato pozice není k dispozici v korpusech SYN2000 a ORWELL.

**P** perfektivum (dokonavé sloveso)

**I** imperfektivum (nedokonavé sloveso)

**B** obouvidé sloveso

Vid je na úrovni morfologického značkování chápán (v souladu s běžnou praxí českých výkladových slovníků) jako klasifikační kategorie.

## Česká morfologie a korpusy

V úvodu jsme uvedli otázky, kterými se v této centrální kapitole našeho textu chceme zabývat. Zopakujme si je:

1. Jak získat materiál pro výzkum hláskových alternací, které doprovázejí tvoření tvarů substantiv (skloňování) v češtině.
2. Jak získat materiálovou základnu pro výzkum variantních a dubletních koncovek substantiv.
3. Jak získat podklady pro výzkum stupňování adjektiv.
4. Jak získat podklady pro výzkum hláskoslovných alternací při tvoření slovesných tvarů.
5. Jak získat přehled o skutečném stavu některých okrajových jevů, např. syntetického futura.
6. Jak najít v korpusech materiál pro výzkum adverbializace.
7. Jak se v lingvistickém výzkumu obejít bez morfologického značkování a lemmatizace.
8. Jak získat z korpusů podklady pro výzkum slovotvorby.
9. Jak se v korpusově založeném výzkumu vypořádat s takzvanými MWE<sup>11</sup>.
10. Jak využívat korpus jako on-line slovník.

Následující text rozčleníme nejdříve podle jednotlivých slovních druhů, dále bude následovat kapitola věnovaná problematice slovnědruhových přesahů a přechodů a kapitola zaměřená na problematiku tvoření slov a víceslovných výrazů.

Níže zobrazené výsledky konkordančních seznamů a frekvenční distribuce jsou, není-li explicitně uvedeno jinak, z korpusu SYN2010.

---

<sup>11</sup> MWE – Multiword Expression je zastřešující termín pro heterogenní skupinu víceslovných jednotek zahrnující složené mluvnické formy (slovesné analytické tvary, analytické stupňování, víceslovné číslovky, propria, ale také frazémy, idiomy atd.). Morfologické značkování korpusů nebere vždy v úvahu interpretaci, která odpovídá tomu, že daný tvar je součástí MWE.

## **Substantiva**

V následujících podkapitolách se zaměříme na otázky spojené s formální morfologií substantiv. Budeme se zabývat substantivními vzory z hlediska možnosti jejich vyhledávání v korpusech. Všimneme si formálních vlastností slov řazených pod jednotlivé vzory. Druhým okruhem budou hláskoslovné alternace v rámci systému české substantivní flexe. Ukážeme si, jak lze na základě pozorování dat získaných z jazykových korpusů hledat odpovědi na otázky týkající se pravidel distribuce sledovaných alternací. Třetí oblastí, na niž se chceme zaměřit, bude výskyt a pravidla distribuce spisovných variantních a dubletních koncovek v české deklinaci podstatných jmen. Na konkrétním příkladu ukážeme, jak lze z korpusů získat data a jak lze na základě jejich pozorování formulovat empiricky založená pravidla fungování jazykového systému.

### ***Jak vyhledat v korpusu substantiva podle vzoru?***

#### **Motivační úvod**

Vzor (flektivní typ) není gramatická kategorie v terminologickém smyslu. Jde o kategorii formální. Položíme-li si otázku, proč se české děti ve čtvrté třídě učí, ke kterému vzoru patří to které podstatné (i přídavné) jméno, konstatujeme, že důvodem je vytvoření opory pro morfologický pravopis. Mohli bychom říci, že proto, aby věděly, jaké [iy] mají napsat po obojetných souhláskách [bpfvmszl].

K čemu ještě mohou být tyto informace užitečné? Deklinační systém současné češtiny je velmi složitý. Běžný mluvčí, pro nějž je čeština mateřštinou, si neklade otázky, na které by odborník (bohemista) měl umět odpovídat. Měl by mít přehled o tom, jaká platí pravidla, kde se vyskytují výjimky a můžeme-li říci, kolik jich je. Systém jazyka (např. tvarosloví) je explicitně popsán v mluvnicí a slovníku (langue). Implicitně je přítomen v promluvách a textech (parole). Jazykové korpusy umožňují přímo zkoumat reprezentativní vzorek jazyka (parole) a na základě pozorování dat vytvářet a ověřovat hypotézy o systému pravidel a výjimek (popis systému langue).

Pokud tedy chceme odpovídat na otázky jako: co je pravidelné a co je nepravidelné, nebo kolik je té a té nepravidelnosti, lze zvolit dvojí přístup. Buď hledáme odpovědi v příslušné literatuře, nebo pozorujeme korpusová data (empirie) a vyvozujeme (dedukce).

#### **Nastínění problému**

Dle klasických mluvnic pro rodilé mluvčí má čeština 14 vzorů pro jmennou substantivní flexi. Počítáme-li se čtyřmi rody, pak maskulina životná se mohou skloňovat podle 4 vzorů (*pán*,

muž, předseda, soudce), neživotná podle 2 (hrad, stroj), feminina podle 4 (žena, růže, píseň, kost) a neutra taktéž podle 4 (město, moře, kuře, stavení). Kromě rodu je hlavním kritériem pro zařazení substantiva k příslušnému vzoru zakončení, a to a) bez koncovky (s koncovkou - 0) x s koncovkou a b) kvalitativní charakteristika finální souhlásky kmene, ať už za ní následuje koncovka, nebo ne. Tabulková paradigmata uvádějí navíc a) podvzory (např. pán – hoch, hrad – les, ...), b) paragrafy hovořící o výjimkách (jednotlivá slova, např. host, kůň, ... nebo skupiny slov např. den, týden, kámen, kořen, kmen, hřeben, ...) a c) případy kolísání mezi vzory (kolísání mezi tvrdými a měkkými vzory např. paňáca, skica, ..., kolísání mezi vzorem píseň a kost). Kromě toho mají zvláštní flexi některá propria a existují též slova neskloňná. Substantiva se také mohou skloňovat podle adjektivních vzorů a naopak jako substantiva se skloňují slova, která se řadí do tzv. nastavbových slovních druhů (číslovky např. tisíc, milion, miliarda).

## Otázky

Pokusme se nyní nejdříve slovně formálně popsat operaci, kterou je třeba provést, chceme-li přiřadit slovo – substantivum ke vzoru. Nejdříve si položíme otázku, zda je zakončení slova (formální vlastnost) relevantní pro zařazení ke slovnímu druhu.

Mějme řetězce písmen, např. kos, sál, pila, žeň, chudě, ... Všimněme si, že ani o jednom z uvedených tvarů nemůžeme s určitostí tvrdit, že jde o podstatné jméno!

Jak to? A jaké jsou důsledky pro disambiguaci?

**Korpus:**

**Typ dotazu:**

**Slovní tvar:**

. Než jsem přijela do Anglie, požár zničil hodovní	sál /sál/NNIS4-----A-----	ve Whitehallu a Inigo Jones navrhl nový . Byl sice
BUDĚJOVICE Zastupitelé Českých Budějovic v minulém týdnu rozhodli, že	sál /sál/NNIS1-----A-----	za dvě miliardy korun může vyrůst na plánovaném území ,
zvolit epidurální analgezii , a to bez omezení . PORODNÍ	SÁL /Sála/NNFP2-----A-----	A POROD Na druhou dobu porodní se rodička " stěhuje
ale nakonec ušetřilo . Pokud by totiž existoval kvalitní symfonický	sál /sál/NNIS1-----A-----	v Janáčkově divadle , nemuselo by se stavět Janáčkovo kulturní
, veze samý vožralý ... " plně ovládla	sál /sál/NNIS4-----A-----	. Při této písni i šedovlasí kmeti vyrazili na parket
písníkáře Arnošta Frauenbergera . Do obce míří heligonkáři UDÁLOSTI HVOŽDANY	Sál /Sála/NNFP2-----A-----	Kulturního domu Hvožďany na Rožmitálsku hostí zítřa od 16 hodin
, ale nabídne v centru Prahy ojedinelý , moderně vybavený	sál /sál/NNIS1-----A-----	pro pořádání konferencí , plesů nebo koncertů . Pojme asi
při tréninku . Téměř okamžitě putoval ze sjezdovky na operační	sál /sál/NNIS4-----A-----	. " To se může stát . Je to drobnost
. Areál pod jednou střechou nabídne soudní místnosti , dražební	sál /sál/NNIS1-----A-----	pro dvě stovky osob , studovnu určenou k nahlížení do
poslechla . Sklonil se a chvíli jí jenom líbal a	sál /sál/NNIS1-----A-----	bradavky . Pak jí jednou rukou vklouzl pod boky a
poznámenal Pokorný s tím , že na střední Moravě obdobný	sál /sál/NNIS1-----A-----	zatím k dispozici není . S častějším pořádáním i jiných

Ptáme-li se tedy, jak poznáme substantivum, musíme říci, že je na základě pouhé formy mnohdy poznat nemůžeme.

Většinou se ovšem tvrdívá, že vzor poznáme podle zakončení substantiva. Proč jsou pak ale *sráz* i *mráz* maskulina neživotná, a *hráz* je femininum? Proč je *kroj* maskulinum neživotné, a *kůň* životné, a proč *zbroj* i *tůň* jsou feminina? Jak to, že *choť* je buď maskulinum životné, nebo femininum a *jeřáb* je buď maskulinum životné, nebo neživotné?

Přidejme opěrný tvar genitivu (ten bývá uveden ve slovnících bezprostředně za heslovým slovem). Vidíme, že zůstávají případy (*kroj* x *zbroj*, *kůň* x *tůň*), kdy musíme znát rod, respektive zařazení ke vzoru.

To, že jde o substantivum, musíme tedy buď vědět, nebo předpokládat. Totéž platí i o rodu. Ti, kteří se učí češtině jako druhému jazyku, jsou tedy na tom podobně jako my, když se třeba učíme německy (u substantiv se musíme naučit i rod). Je to ale nutné ve všech případech? Co nám může pomoci?

Podíváme-li se na všechny výše uvedené příklady, pak můžeme konstatovat, že mají jednu společnou vlastnost týkající se formy. Touto vlastností je shoda koncové části řetězce. Toto pozorování platí především v případě slovnědruhově víceznačných tvarů. Uděláme-li ze *zbroje výzbroj* a z *kroje stejnokroj*, problém zůstane stejný. Při „prodloužení slova“ vpředu (prefixace, kompozice) se rod nemění. A co budeme-li srovnávat *mechorost* a *soběstačnost*? Suffixy sice mohou být formálně shodné s koncovými řetězci slov neutvořených, nicméně platí, že slova tvořená stejným sufixem se (až na výjimky – maskulina životná a neživotná např. <*politický*> *činitel* / <*slohotvorný*> *činitel*) skloňují podle téhož vzoru.

Vraťme se nyní k původnímu zadání a zrekapitulujme, o co se vlastně opíráme, když určujeme vzor podstatného jména.

Můžeme např. říci, že podle vzoru *pán* nebo *muž* by se měla skloňovat maskulina životná, která končí v nominativu singuláru na souhlásku, kdežto podle vzoru *předseda* ta, která končí na *-a* a podle *soudce* ta, která končí na *-[eě]*. To, zda se substantivum skloňuje podle vzoru *pán* nebo podle vzoru *muž*, poznáme podle tvaru genitivu singuláru zakončeného na *-a* nebo *-e/-ě*, což v naprosté většině případů koresponduje s kmenovou finálou (tvrdou nebo měkkou souhláskou).

Podobný postup lze popsat i pro další rody a příslušné vzory.

## Formulace dotazu pro získání dat z korpusů

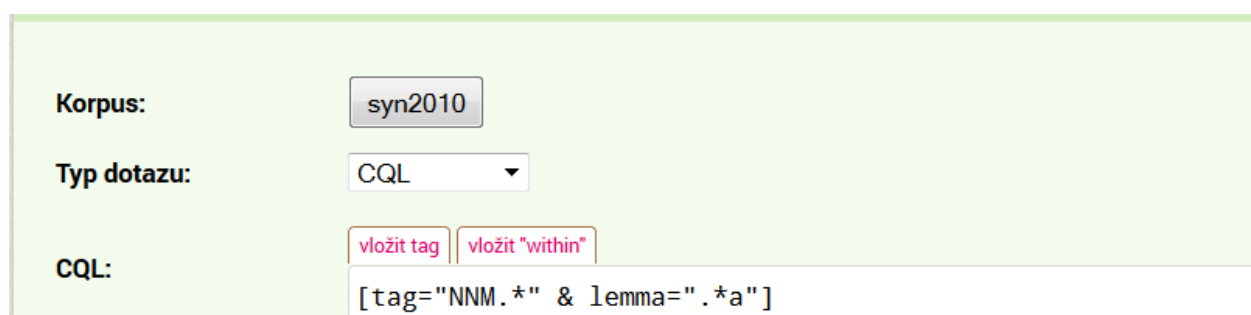
Shrňme tedy ještě jednou to, k čemu jsme došli. Budeme-li chtít v korpusu vyhledat slova skloňovaná podle určitého vzoru, budeme muset použít informace o lemmatu a morfologické značce a kombinovat je s formálními vlastnostmi – zakončením lemmatu.

Podle morfologické značky snadno vyhledáme substantiva stejného rodu.

V menu **Korpus** zvolíme **Typ dotazu tag** a do dotazovacího řádku napíšeme dotaz ve formě odpovídající morfologické značky.

Tedy *NNM.\** hledáme-li maskulina životná, *NNI.\**, hledáme-li maskulina neživotná, *NNF.\**, hledáme-li feminina, *NNN.\** hledáme-li neutra.

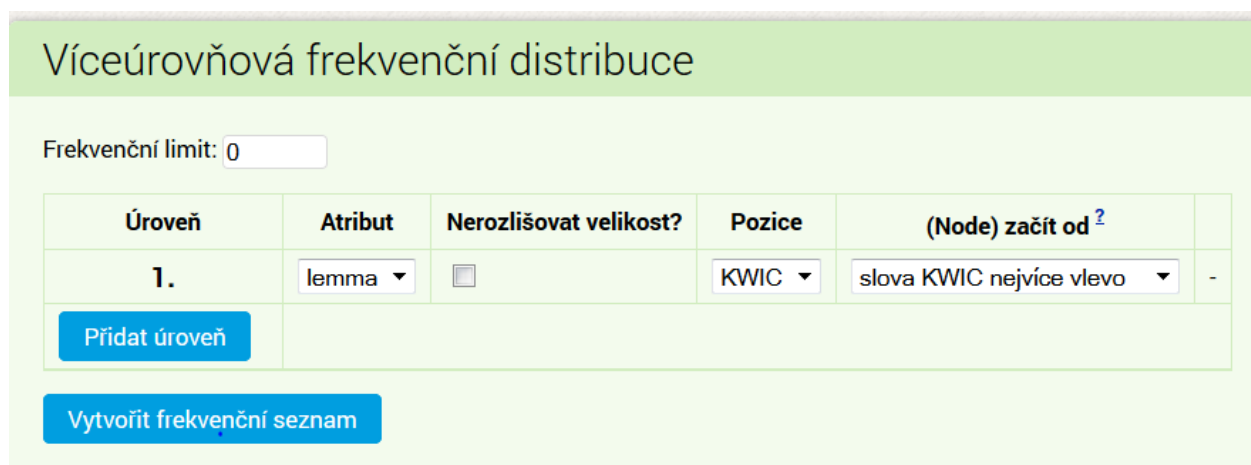
Hledáme-li substantiva skloňovaná podle vzoru *předseda*, hledáme všechna maskulina životná, jejichž lemmata končí na *-a*. Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě `[tag="NNM.*" & lemma="*.a"]`.



The screenshot shows a search configuration interface with the following elements:

- Korpus:** A text box containing the value "syn2010".
- Typ dotazu:** A dropdown menu set to "CQL".
- CQL:** A text box containing the query `[tag="NNM.*" & lemma="*.a"]`. Above this box are two buttons: "vložit tag" and "vložit 'within'".

Seznam substantiv skloňovaných podle vzoru *předseda* získáme pomocí menu **Frekv. distribuce** zvolíme-li **Atribut lemma**.



The screenshot shows the "Víceúrovňová frekvenční distribuce" interface with the following elements:

- Header: "Víceúrovňová frekvenční distribuce"
- Input: "Frekvenční limit: 0" with a text box.
- Table with columns: "Úroveň", "Atribut", "Nerozlišovat velikost?", "Pozice", "(Node) začít od?", and a final column with a dash.
- Row 1: "1.", "lemma", a checkbox, "KWIC", "slova KWIC nejvíce vlevo", and "-".
- Buttons: "Přidat úroveň" and "Vytvořit frekvenční seznam".



Obdobně jednoduše vyhledáme substantiva skloňovaná podle vzorů *soudce*, *žena*, *růže*, *město* a *stavení*. Při vyhledávání substantiv skloňovaných podle vzorů *soudce* a *růže* nesmíme zapomenout na dvě možné grafické realizace koncovky *-e* (*[eě]*).

Složitější bude postup pro rozlišení maskulin neživotných skloňovaných podle vzoru *hrad* nebo *stroj*, maskulin životných skloňovaných podle vzoru *pán* nebo *muž*, feminin skloňovaných podle vzoru *píseň* nebo *kost* a neuter skloňovaných podle vzoru *moře* nebo *kuře*.

U maskulin, kde jde o odlišení tzv. tvrdých a měkkých vzorů, může pomoci kmenová finála – souhláska, na niž končí lemma. Problematické mohou být pouze tzv. souhlásky obojetné, ale nikoli vždy a všechny.

U feminin najdeme rovněž některá zakončení (finály) typické pouze pro jeden ze vzorů.

U neuter lze kombinovat zakončení kmene a koncovku.

Slovotvorná utvářenost slova skýtá možnost vyhledávat (filtrovat) podle formy derivačního sufixu.



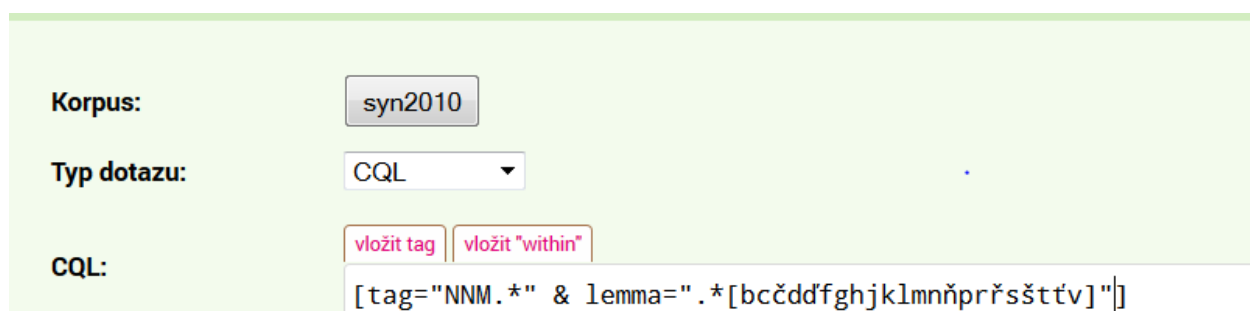
## Třídění a pozorování dat získaných z korpusů

Pokusme se nyní na základě pozorování dat formulovat pravidla pro rozlišení substantiv jednoho rodu skloňovaných podle různých vzorů, které se v nominativu singuláru neliší koncovkou.

Jako příklad poslouží substantiva rodu mužského životná, která nemají v nominativu singuláru koncovku a která se mohou skloňovat buď podle vzoru *pán*, nebo podle vzoru *muž*.

Vyhledejme nejdříve maskulina životná, která v nominativu singuláru končí na souhlásku.

Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě `[tag="NNM.*" & lemma=".*[bcčdd'fghjklmnňprřsšt'tv]" ]`.



The screenshot shows a search interface with a light green background. It contains the following elements:

- Korpus:** A text input field containing "syn2010".
- Typ dotazu:** A dropdown menu currently set to "CQL".
- CQL:** A text input field containing the query `[tag="NNM.*" & lemma=".*[bcčdd'fghjklmnňprřsšt'tv]" ]`. Above this field are two buttons: "vložit tag" and "vložit 'within'".

V dalším kroku lze pomocí **Filtru** vybrat substantiva zakončená na `.*[dghknrt]`, která by se měla skloňovat podle vzoru *pán*<sup>12</sup>, a substantiva zakončená na `.*[cčd'jňřžž]` která by se měla skloňovat podle vzoru *muž*<sup>13</sup>. Užitečnější bude zaměřit se na substantiva zakončená na tzv. obojetnou souhlásku a podívat se, můžeme-li na základě pozorování dat formulovat nějaké pravidlo, podle kterého se tato substantiva řadí k jednomu ze dvou vzorů (*pán/muž*). Zvolíme **Filtr** (pozitivní), rozsah hledání `<0,0>`, **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz ve formě `.*[bflmpsvz]`. Poté se podíváme na frekvenční distribuci lemmat.

<sup>12</sup> Všimněte si skloňování vlastního jména *Smith* i jemu podobných.

<sup>13</sup> Všimněte si skloňování vlastního jména *Balzac* i jemu podobných.

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC


**Typ dotazu:**

**Lemma:**

Filtrovat konkordanci





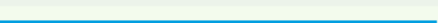
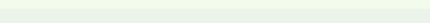

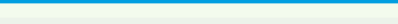










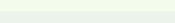
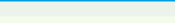
Vymazat formulář

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 7506 (151 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> ředitel	24 415	
2.	<a href="#">p/ n</a> přítel	23 968	
3.	<a href="#">p/ n</a> obyvatel	19 459	
4.	<a href="#">p/ n</a> šéf	19 434	
5.	<a href="#">p/ n</a> pes	17 701	
6.	<a href="#">p/ n</a> manžel	17 278	
7.	<a href="#">p/ n</a> Pavel	16 942	
8.	<a href="#">p/ n</a> majitel	15 626	
9.	<a href="#">p/ n</a> Karel	15 597	
10.	<a href="#">p/ n</a> Josef	14 868	
11.	<a href="#">p/ n</a> král	14 744	
12.	<a href="#">p/ n</a> Václav	14 198	
13.	<a href="#">p/ n</a> učitel	10 943	
14.	<a href="#">p/ n</a> podnikatel	9 283	
15.	<a href="#">p/ n</a> Jaroslav	8 868	
16.	<a href="#">p/ n</a> chlap	8 637	
17.	<a href="#">p/ n</a> uživatel	8 373	
18.	<a href="#">p/ n</a> spisovatel	7 544	
19.	<a href="#">p/ n</a> nepřítel	7 377	
20.	<a href="#">p/ n</a> Miroslav	7 167	

Mezi frekventovanými substantivy převažují ta, která končí na *. \*l*, přičemž jsou mezi nimi jak substantiva skloňovaná podle vzoru *muž* (*ředitel, přítel, obyvatel*), tak podle vzoru *pán* (*Pavel, Karel, Michal*), ale i substantivum *manžel*, které mezi oběma vzory kolísá. Jiné zakončení mají substantiva *šéf, pes, chlap* skloňovaná podle vzoru *pán*, a *Klaus* skloňované podle vzoru *muž*. Takto můžeme pokračovat v prohledávání seznamu substantiv s nižšími až nejnižšími frekvencemi.

V dalším kroku můžeme formulovat předpoklad, že některé z tzv. obojetných souhlásek se v případě vzoru *pán* chovají jako „tvrdé“ [pf] a jiné jako „obojetné“ [ls]. Jak takový předpoklad ověříme?

Postupně pomocí **Filtr pozitivní** získáme relevantní data (např. všechna substantiva rodu mužského životného, která končí na *. \*b* atd.), budeme je sledovat a zjišťovat, zda můžeme (na základě empirie – pozorování korpusových dat) vyslovit nějaké závěry.

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**  ▼


**Lemma:**

Frekvenční seznam lemmat získáme opět pomocí menu **Frekv. Distribuce**, zvolíme-li **Atribut lemma** (viz výše).

Poté, co si seznam lemmat na *. \*b* prohlédneme, vrátíme se zpět (3 kroky) a celou operaci postupně zopakujeme pro lemmata na *. \*p*, *. \*f*, *. \*v*, *. \*m*.

Při procházení seznamů lemmat si budeme všimnout, ke kterému vzoru životných maskulin (*pán/muž*) vyhledaná substantiva patří, abychom na základě pozorování dat mohli formulovat pravidla o tom, jak se substantiva s tzv. obojetnou finální souhláskou řadí k těmto vzorům.

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 159 (4 str.)
















	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> Jakub	4 822	
2.	<a href="#">p/ n</a> Arab	1 295	
3.	<a href="#">p/ n</a> Bob	1 210	
4.	<a href="#">p/ n</a> holub	1 139	
5.	<a href="#">p/ n</a> Jákob	942	
6.	<a href="#">p/ n</a> Holub	579	
7.	<a href="#">p/ n</a> Rob	577	
8.	<a href="#">p/ n</a> jestřáb	510	
9.	<a href="#">p/ n</a> Kocáb	453	
10.	<a href="#">p/ n</a> krab	445	
11.	<a href="#">p/ n</a> sob	410	
12.	<a href="#">p/ n</a> Webb	409	
13.	<a href="#">p/ n</a> sáhib	360	
14.	<a href="#">p/ n</a> Dub	341	
15.	<a href="#">p/ n</a> mikrob	324	
16.	<a href="#">p/ n</a> šváb	297	
17.	<a href="#">p/ n</a> Srb	288	
18.	<a href="#">p/ n</a> jeřáb	285	
19.	<a href="#">p/ n</a> Jakob	266	
20.	<a href="#">p/ n</a> Jacob	258	

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 267 (6 str.)

	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> šéf	19 434	
2.	<a href="#">p/ n</a> Josef	14 868	
3.	<a href="#">p/ n</a> Rudolf	3 237	
4.	<a href="#">p/ n</a> fotograf	2 997	
5.	<a href="#">p/ n</a> filozof	2 267	
6.	<a href="#">p/ n</a> Adolf	1 356	
7.	<a href="#">p/ n</a> šerif	1 217	
8.	<a href="#">p/ n</a> Jeff	1 012	
9.	<a href="#">p/ n</a> Lief	958	
10.	<a href="#">p/ n</a> Ralf	881	
11.	<a href="#">p/ n</a> Kryštof	880	
12.	<a href="#">p/ n</a> Wolf	871	
13.	<a href="#">p/ n</a> elf	577	
14.	<a href="#">p/ n</a> Jozef	544	
15.	<a href="#">p/ n</a> Rolf	352	
16.	<a href="#">p/ n</a> Gandalf	348	
17.	<a href="#">p/ n</a> Volf	284	
18.	<a href="#">p/ n</a> Graf	252	
19.	<a href="#">p/ n</a> Grof	231	
20.	<a href="#">p/ n</a> Šlouf	229	




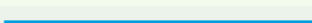
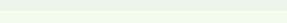






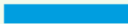



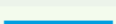




# Frekvenční distribuce

strana 1

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 401 (9 str.)

	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> Adam	5 421	
2.	<a href="#">p/ n</a> Rom	3 474	
3.	<a href="#">p/ n</a> Tom	3 273	
4.	<a href="#">p/ n</a> Sam	2 824	
5.	<a href="#">p/ n</a> Jim	2 571	
6.	<a href="#">p/ n</a> Kim	2 286	
7.	<a href="#">p/ n</a> ekonom	2 277	
8.	<a href="#">p/ n</a> muslim	2 043	
9.	<a href="#">p/ n</a> William	2 025	
10.	<a href="#">p/ n</a> Bém	1 273	
11.	<a href="#">p/ n</a> Vilém	1 247	
12.	<a href="#">p/ n</a> Tím	1 107	
13.	<a href="#">p/ n</a> Radim	1 090	
14.	<a href="#">p/ n</a> Jarolím	999	
15.	<a href="#">p/ n</a> Mandelštam	993	
16.	<a href="#">p/ n</a> Jašim	963	
17.	<a href="#">p/ n</a> Maxim	840	
18.	<a href="#">p/ n</a> Abrahám	761	
19.	<a href="#">p/ n</a> astronom	707	
20.	<a href="#">p/ n</a> Saddám	655	







ontext.korpus.cz/fun.cgi/view?q=aword,[tag="NNM.\*"+&+lemma=".\*[bcčdd'fghjklmnñpršst'tv]"]&q=f&q=p

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 188 (4 str.)

	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> chlap	8 637	
2.	<a href="#">p/ n</a> biskup	3 797	
3.	<a href="#">p/ n</a> Filip	3 204	
4.	<a href="#">p/ n</a> arcibiskup	1 000	
5.	<a href="#">p/ n</a> Prokop	965	
6.	<a href="#">p/ n</a> Philip	697	
7.	<a href="#">p/ n</a> filip	687	
8.	<a href="#">p/ n</a> Soukup	672	
9.	<a href="#">p/ n</a> Šíp	638	
10.	<a href="#">p/ n</a> lidoop	459	
11.	<a href="#">p/ n</a> Dunlap	454	
12.	<a href="#">p/ n</a> lump	435	
13.	<a href="#">p/ n</a> čáp	411	
14.	<a href="#">p/ n</a> sup	381	
15.	<a href="#">p/ n</a> pop	297	
16.	<a href="#">p/ n</a> Popp	267	
17.	<a href="#">p/ n</a> op	229	
18.	<a href="#">p/ n</a> Depp	195	
19.	<a href="#">p/ n</a> Philipp	192	
20.	<a href="#">p/ n</a> zlatokop	188	










# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 986 (20 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> Václav	14 198	
2.	<a href="#">p/ n</a> Jaroslav	8 868	
3.	<a href="#">p/ n</a> Miroslav	7 167	
4.	<a href="#">p/ n</a> Ladislav	4 554	
5.	<a href="#">p/ n</a> Stanislav	3 690	
6.	<a href="#">p/ n</a> detektiv	2 996	
7.	<a href="#">p/ n</a> lev	2 367	
8.	<a href="#">p/ n</a> Miloslav	1 342	
9.	<a href="#">p/ n</a> Bohuslav	1 285	
10.	<a href="#">p/ n</a> červ	927	
11.	<a href="#">p/ n</a> Vítězslav	829	
12.	<a href="#">p/ n</a> Vladislav	826	
13.	<a href="#">p/ n</a> Gustav	814	
14.	<a href="#">p/ n</a> Udalov	747	
15.	<a href="#">p/ n</a> Rostislav	689	
16.	<a href="#">p/ n</a> Lev	666	
17.	<a href="#">p/ n</a> Vratislav	497	
18.	<a href="#">p/ n</a> Břetislav	387	
19.	<a href="#">p/ n</a> Medveděv	357	
20.	<a href="#">p/ n</a> Gorbačov	327	

## Formulace závěrů

Hledáme-li v korpusu nějaký jev, pak pouze na základě toho, že jej nenalezneme, nemůžeme tvrdit, že v jazyce není možný. Nulový výskyt spolehlivě signalizuje malou frekvenci nějakého jevu, popřípadě to, že hledaný jev nebyl zachycen na úrovni automatické morfologické analýzy (to platí především tehdy, používáme-li při vyhledávání morfologické anotace). Může ale též upozornit na existenci pravidla, které tento jev vylučuje.

Pozorováním korpusově doložených dat zjistíme, že podle vzoru *pán* nebo *muž* se mohou skloňovat substantiva zakončená na *-[lsz]*. Substantiva maskulina životná zakončená na *-[bfmpv]* se mohou skloňovat pouze podle vzoru *pán*.

### **Formulace dalších otázek vyplývajících ze zkoumání daného jevu**

Platí obecně, že souhlásky *[bfmpv]* se jakožto finály vyskytují pouze u substantiv tzv. tvrdých vzorů?

Zadání cvičení, v nichž lze uplatnit analogické postupy

1. Navrhněte postup pro rozlišení substantiv skloňovaných podle vzoru *hrad* a *stroj*.
2. Navrhněte postup pro rozlišení substantiv skloňovaných podle vzoru *píseň* a *kost*.
3. Navrhněte postup pro rozlišení substantiv skloňovaných podle vzoru *moře* a *kuře*.
4. Porovnejte distribuci kmenových finál *[bfmpv]* substantiv skloňovaných podle vzoru *hrad* a *stroj*.
5. Porovnejte distribuci kmenových finál *[bfmpv]* substantiv skloňovaných podle vzoru *žena* a *růže*.
6. Porovnejte distribuci kmenových finál *[bfmpv]* substantiv skloňovaných podle vzoru *město* a *moře*.
7. Existuje nějaký postup, jak oddělit substantiva na *-[lsz]* skloňovaná podle vzoru *pán* od substantiv skloňovaných podle vzoru *muž*?
8. Jakými pravidly se řídí distribuce grafické varianty *e/ě* po tzv. obojetných souhláskách  
a) u tvrdých vzorů, b) u měkkých vzorů a c) u vzoru *píseň*.

### ***Jak lze v korpusech hledat doklady pro výzkum hláskových alternací v rámci substantivní flexe?***

#### **Motivační úvod**

Hláskové alternace jsou definovány jako pravidelné střídání hlásek (dvojic hlásek), na němž je založena alomorfie. Alternace provázejí jak tvoření tvarů slov (skloňování a časování), tak odvozování slov. Každý rodilý mluvčí ví, že chce-li po něm malé dítě „*leva*“, nežádá asi o bulharskou měnu, ale o plyšového mazlíčka, a pozná, že je chyba, raduje-li se dítě, že napadlo hodně „*sníhu*“. Přesto mu bude asi činit jisté potíže, položí-li mu dítě, které opraví, otázku proč? Na tomto místě se spokojíme s tím, že se na základě analýzy dostupných korpusových dat pokusíme odpovědět na otázky kde a jak. V této kapitole se podíváme na to, kdy se v genitivu plurálu feminin skloňovaných podle vzoru *žena* vkládá mezi dvě poslední

souhlásky *e*. Tedy trochu s humorem řečeno, proč je hodně *barev* a málo *želv*? Půjde tedy o alternace *o/e* u feminin skloňovaných podle vzoru *žena*.

### Nastínění problému

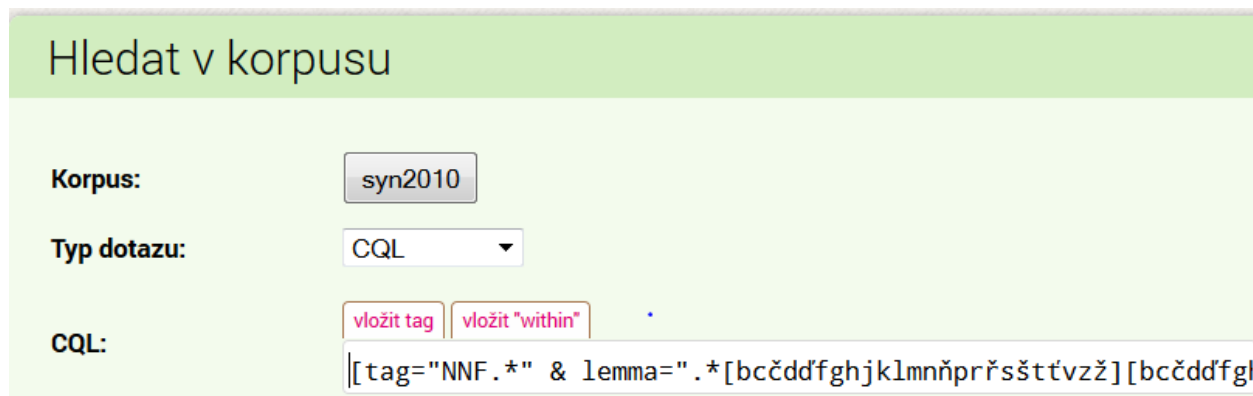
V *Mluvnici současné češtiny* (Cvrček a kol. 2010, s. 172) čteme: „U řady jmen s kmenem zakončeným skupinou dvou nebo více souhlásek, např. *hra, kresba, látka, astra, ondatra*, se v genitivu plurálu do skupiny vkládá *e*: *her, kreseb, ...*“ Tabulka, která následuje a která si „neklade nároky na úplnost“, obsahuje bohužel chyby (více Kosek – Křístek – Osolsobě – Ziková – Vojtová 2011). S ohledem na proklamativně korpusový charakter *MSC* se pokusíme o vlastní analýzu korpusových dat.

### Otázky

Jak získáme relevantní data pro výzkum alternací *o/e* u feminin skloňovaných podle vzoru *žena*? Je třeba vyhledat všechna lemmata substantiv skloňovaných podle vzoru *žena* taková, že u lemmatu před koncovkou *-a* předcházejí alespoň dvě souhlásky. U těchto lemmat pak dále sledujeme tvary genitivu plurálu a snažíme se vypátrat, kdy se mezi dvě souhlásky vkládá *e* a kdy nikoli.

### Formulace dotazu pro získání dat z korpusů

Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě `[tag="NNF.*" & lemma=".*[bcčdd'fghjklmnňprřsšt'tvzž][bcčdd'fghjklmnňprřsšt'tvzž]a"]`.



Hledat v korpusu

Korpus:

Typ dotazu:

CQL:

Zvolíme **Filtr pozitivní**, rozsah hledání `<0,0>`, **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme dotaz ve formě `NNFP2.*` (hledáme všechny tvary genitivu plurálu).

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**

**CQL:**

Následně můžeme a) ve **Filtr konkordancí** zvolit **Typ dotazu slovní tvar** a do dotazovacího řádku napsat dotaz ve formě `.*[bcčdd'fghjklmnňprřsšt'tvzž]e[bcčdd'fghjklmnňprřsšt'tvzž]` a vrátíme-li se o krok zpět, pak b) do dotazovacího řádku napsat dotaz ve formě `.*[bcčdd'fghjklmnňprřsšt'tvzž][bcčdd'fghjklmnňprřsšt'tvzž]`. Získáme tak postupně feminina skloňovaná podle vzoru *žena*, která a) mají v genitivu plurálu alternaci *0/e* a b) nemají v genitivu plurálu alternaci *0/e*.

a)

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**

**Slovní tvar:**

V menu **Frekv. distribuce** zvolíme na první úrovni **Atribut lemma** a na druhé úrovni **Atribut word (slovní tvar)**.

## Víceúrovňová frekvenční distribuce

Frekvenční limit:

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>?</sup>	
1.	lemma ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	-
2.	word ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	<input type="button" value="x"/>

# Frekvenční distribuce

strana 1

Frekvenční limit: 0

Nastavit limit

Celkem: 4253 (86 str.)

	lemma	word	Frekvence	
1.	p/ n služba	služeb	9 571	
2.	p/ n firma	firem	7 646	
3.	p/ n podmínka	podmínek	6 551	
4.	p/ n desítka	desítek	4 290	
5.	p/ n volba	voleb	3 828	
6.	p/ n látka	látek	3 609	
7.	p/ n hra	her	3 251	
8.	p/ n otázka	otázek	3 163	
9.	p/ n barva	barev	3 039	
10.	p/ n stavba	staveb	2 890	
11.	p/ n jednotka	jednotek	2 815	
12.	p/ n stránka	stránek	2 376	
13.	p/ n stovka	stovek	2 301	
14.	p/ n myšlenka	myšlenek	2 047	
15.	p/ n forma	forem	2 013	
16.	p/ n vrstva	vrstev	1 871	
17.	p/ n památka	památek	1 834	
18.	p/ n karta	karet	1 777	
19.	p/ n dívka	dívek	1 730	
20.	p/ n značka	značek	1 623	

b)

## Filtr konkordancí

Filtr:  pozitivní  negativní

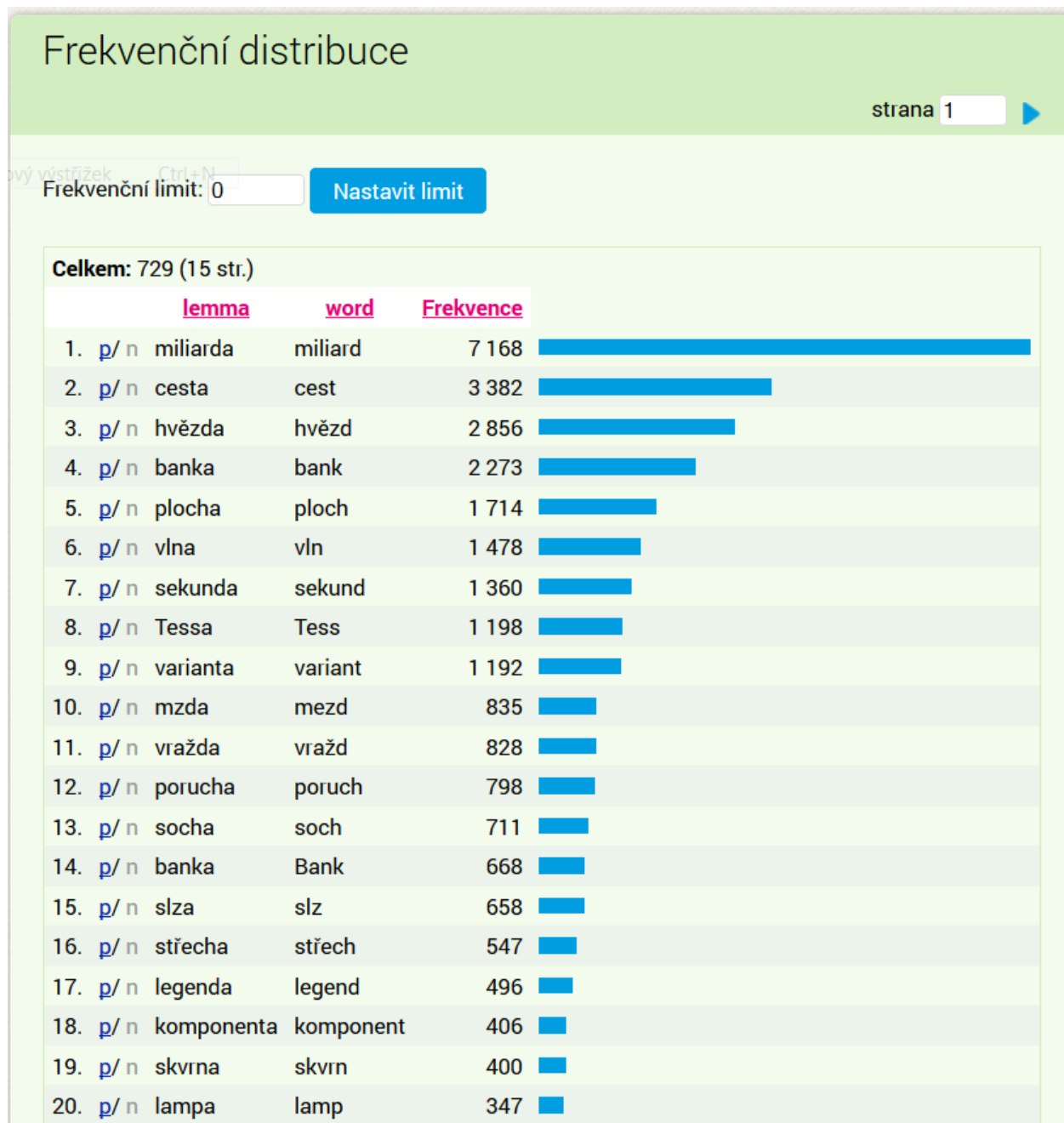
Vybraný token:  první  poslední

Rozsah hledání: od 0 do 0  včetně KWIC

Typ dotazu: Slovní tvar

Slovní tvar:


Frekvenční seznam lemmat a tvarů získáme stejně jako u předešlé úlohy.



### Třídění a pozorování dat získaných z korpusů







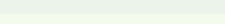













Vrátíme se zpět a vytvoříme frekvenční seznamy lemmat a tvarů a pozorujeme je. (V menu **Frekv. distribuce** zvolíme na první úrovni **Atribut lemma** a na druhé úrovni **Atribut word** (slovní tvar).)

# Frekvenční distribuce

strana 1 

Frekvenční limit:

Celkem: 5034 (101 str.)

	lemma	word	Frekvence	
1.	p/ n služba	služeb	9 571	
2.	p/ n firma	firem	7 646	
3.	p/ n miliarda	miliard	7 168	
4.	p/ n podmínka	podmínek	6 551	
5.	p/ n desítka	desítek	4 290	
6.	p/ n volba	voleb	3 828	
7.	p/ n látka	látek	3 609	
8.	p/ n cesta	cest	3 382	
9.	p/ n hra	her	3 251	
10.	p/ n otázka	otázek	3 163	
11.	p/ n barva	barev	3 039	
12.	p/ n stavba	staveb	2 890	
13.	p/ n hvězda	hvězd	2 856	
14.	p/ n jednotka	jednotek	2 815	
15.	p/ n stránka	stránek	2 376	
16.	p/ n stovka	stovek	2 301	
17.	p/ n banka	bank	2 273	
18.	p/ n buňka	buněk	2 051	
19.	p/ n myšlenka	myšlenek	2 047	
20.	p/ n forma	forem	2 013	

V dalších krocích by bylo s ohledem na početná data dobré postupovat po menších celcích, např. podle finální souhlásky. (Můžeme se vrátit zpět a vybrat pomocí pozitivního filtru pouze lemmata na *. \*ba*, *. \*da*, *. \*ka*, *. \*ma* atd.).<sup>14</sup>

<sup>14</sup> Vyhledaná data bychom mohli dále filtrovat. Je např. zřejmé, že u mnoha cizích slov končících na *-ll* nejde o doklady, které by pro náš výzkum mohly být relevantní. Podobně jsou na tom případy, kdy dva grafické znaky označují jediný konsonant (*ch*). Vyloučit bychom měli také případy, kdy druhá souhlásky od konce je slabikotvorná [*rl*] (např. *vrba*, *brva*, *slza*, ...). Při zobrazení frekvenční distribuce by bylo také možné místo atributu **word** volit atribut **lc**.



## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**


**Lemma:**

Filtrovat konkordanci

Vymazat formulář

















Seznamy lemmat a slovních tvarů (genitivu plurálu) získáme výše popsaným způsobem. Pozorujeme data a všímáme si lemmat i tvarů slov, u nichž dochází/nedochází k alternacím *o/e* v genitivu plurálu a hledáme pravidlo (pravidla), kterými se řídí distribuce této alternace.

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 143 (3 str.)

	<u>lemma</u>	<u>word</u>	<u>Frekvence</u>	
1.	<u>p/ n</u> služba	služeb	9 571	
2.	<u>p/ n</u> volba	voleb	3 828	
3.	<u>p/ n</u> stavba	staveb	2 890	
4.	<u>p/ n</u> skladba	skladeb	1 008	
5.	<u>p/ n</u> vazba	vazeb	897	
6.	<u>p/ n</u> sazba	sazeb	814	
7.	<u>p/ n</u> tržba	tržeb	690	
8.	<u>p/ n</u> platba	plateb	601	
9.	<u>p/ n</u> hradba	hradeb	501	
10.	<u>p/ n</u> chodba	chodeb	491	
11.	<u>p/ n</u> kresba	kreseb	452	
12.	<u>p/ n</u> modlitba	modliteb	280	
13.	<u>p/ n</u> bomba	bomb	277	
14.	<u>p/ n</u> malba	maleb	243	
15.	<u>p/ n</u> tužba	tužeb	227	
16.	<u>p/ n</u> hrozba	hrozeb	214	
17.	<u>p/ n</u> bohoslužba	bohoslužeb	150	
18.	<u>p/ n</u> novostavba	novostaveb	120	
19.	<u>p/ n</u> svatba	svateb	109	
20.	<u>p/ n</u> klenba	kleneb	99	





















# Frekvenční distribuce

strana 1 


Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 156 (4 str.)





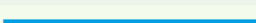
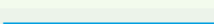
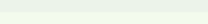
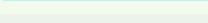
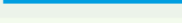











	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> miliarda	miliard	7 168	
2.	<a href="#">p/ n</a> hvězda	hvězd	2 856	
3.	<a href="#">p/ n</a> sekunda	sekund	1 360	
4.	<a href="#">p/ n</a> mzda	mezd	835	
5.	<a href="#">p/ n</a> vražda	vražď	828	
6.	<a href="#">p/ n</a> legenda	legend	496	
7.	<a href="#">p/ n</a> jízda	jízd	282	
8.	<a href="#">p/ n</a> Golda	Gold	218	
9.	<a href="#">p/ n</a> brzda	brzd	212	
10.	<a href="#">p/ n</a> sonda	sonď	168	
11.	<a href="#">p/ n</a> pravda	pravď	163	
12.	<a href="#">p/ n</a> sebevražďa	sebevražď	153	
13.	<a href="#">p/ n</a> Rosalinda	Rosalind	153	
14.	<a href="#">p/ n</a> dividenda	dividend	151	
15.	<a href="#">p/ n</a> garda	gard	120	
16.	<a href="#">p/ n</a> křívďa	křívď	113	
17.	<a href="#">p/ n</a> agenda	agend	99	
18.	<a href="#">p/ n</a> banda	band	87	
19.	<a href="#">p/ n</a> bunda	bund	72	
20.	<a href="#">p/ n</a> Minda	Mind	66	

# Frekvenční distribuce


strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 3546 (71 str.)

	<u>lemma</u>	<u>word</u>	<u>Frekvence</u>	
1.	p/ n podmínka	podmínek	6 551	
2.	p/ n desítka	desítek	4 290	
3.	p/ n látka	látek	3 609	
4.	p/ n otázka	otázek	3 163	
5.	p/ n jednotka	jednotek	2 815	
6.	p/ n stránka	stránek	2 376	
7.	p/ n stovka	stovek	2 301	
8.	p/ n banka	bank	2 273	
9.	p/ n buňka	buněk	2 051	
10.	p/ n myšlenka	myšlenek	2 047	
11.	p/ n památka	památek	1 834	
12.	p/ n dívka	dívek	1 730	
13.	p/ n značka	značek	1 623	
14.	p/ n deska	desek	1 590	
15.	p/ n složka	složek	1 586	
16.	p/ n branka	branek	1 467	
17.	p/ n zakázka	zakázek	1 381	
18.	p/ n zkouška	zkoušek	1 291	
19.	p/ n vzpomínka	vzpomínek	1 192	
20.	p/ n dávka	dávek	982	

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 61 (2 str.)

	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> cihla	cihel	1 024	
2.	<a href="#">p/ n</a> perla	perel	257	
3.	<a href="#">p/ n</a> Bulla	Bull	225	
4.	<a href="#">p/ n</a> Della	Dell	216	
5.	<a href="#">p/ n</a> jehla	jehel	108	
6.	<a href="#">p/ n</a> Hella	Hell	46	
7.	<a href="#">p/ n</a> bedla	bedel	42	
8.	<a href="#">p/ n</a> modla	model	33	
9.	<a href="#">p/ n</a> kukla	kukel	27	
10.	<a href="#">p/ n</a> truhla	truhel	22	
11.	<a href="#">p/ n</a> palla	Pall	22	
12.	<a href="#">p/ n</a> Bella	Bell	22	
13.	<a href="#">p/ n</a> cella	Cell	19	
14.	<a href="#">p/ n</a> palla	pall	18	
15.	<a href="#">p/ n</a> modla	Model	16	
16.	<a href="#">p/ n</a> jáhla	jáhel	15	
17.	<a href="#">p/ n</a> Stella	Stell	11	
18.	<a href="#">p/ n</a> Anabella	Anabell	11	
19.	<a href="#">p/ n</a> metla	metel	10	
20.	<a href="#">p/ n</a> Polla	Poll	10	





















ontext.korpus.cz/fun.cgi/view?q=aword,[tag="NNP.\*"&+lemma=".\*[bcčdd'fghjklmnňpršst'tvzž][bcčdd'fghjklmnňpr

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 66 (2 str.)

	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> firma	firem	7 646	
2.	<a href="#">p/ n</a> forma	forem	2 013	
3.	<a href="#">p/ n</a> norma	norem	1 587	
4.	<a href="#">p/ n</a> reforma	reforem	628	
5.	<a href="#">p/ n</a> šelma	šelem	307	
6.	<a href="#">p/ n</a> palma	palem	287	
7.	<a href="#">p/ n</a> uniforma	uniforem	223	
8.	<a href="#">p/ n</a> farma	farem	184	
9.	<a href="#">p/ n</a> platforma	platforem	70	
10.	<a href="#">p/ n</a> firma	FIREM	59	
11.	<a href="#">p/ n</a> drachma	drachem	35	
12.	<a href="#">p/ n</a> tma	tem	34	
13.	<a href="#">p/ n</a> terma	Therm	16	
14.	<a href="#">p/ n</a> krčma	krčem	16	
15.	<a href="#">p/ n</a> helma	helem	13	
16.	<a href="#">p/ n</a> terma	THERM	12	
17.	<a href="#">p/ n</a> firma	Firem	12	
18.	<a href="#">p/ n</a> Gamma	Gamm	10	
19.	<a href="#">p/ n</a> ekofarma	ekofarem	9	
20.	<a href="#">p/ n</a> vědma	vědem	8	








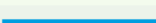
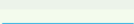

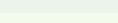









# Frekvenční distribuce

strana 1

Frekvenční limit: 0

Nastavit limit

Celkem: 438 (9 str.)

	lemma	word	Frekvence	
1.	p/ n vlna	vln	1 478	
2.	p/ n elektrárna	elektráren	1 090	
3.	p/ n pojišťovna	pojišťoven	1 014	
4.	p/ n prodejna	prodejen	858	
5.	p/ n knihovna	knihoven	540	
6.	p/ n továrna	továren	498	
7.	p/ n dílna	dílen	429	
8.	p/ n skvrna	skvrn	400	
9.	p/ n pokladna	pokladen	334	
10.	p/ n kavárna	kaváren	329	
11.	p/ n tiskárna	tiskáren	299	
12.	p/ n bedna	beden	290	
13.	p/ n šatna	šaten	212	
14.	p/ n lékárna	lékáren	211	
15.	p/ n tepna	tepen	208	
16.	p/ n základna	základen	205	
17.	p/ n herna	heren	204	
18.	p/ n kachna	kachen	175	
19.	p/ n provozovna	provozoven	171	
20.	p/ n koupelna	koupelen	153	

# Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 39 (1 str.)

	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	p/ n lampa	lamp	347	
2.	p/ n pumpa	pump	167	
3.	p/ n Alpa	Alp	138	
4.	p/ n rampa	ramp	89	
5.	p/ n pulpa	Pulp	61	
6.	p/ n culpa	Culp	39	
7.	p/ n pampa	pamp	20	
8.	p/ n žumpa	žump	19	
9.	p/ n pompa	Pomp	17	
10.	p/ n alpa	Alp	15	
11.	p/ n chrpa	chrp	14	
12.	p/ n malpa	malp	12	
13.	p/ n pumpa	Pump	6	
14.	p/ n výspa	výsep	5	
15.	p/ n Alpa	ALP	5	
16.	p/ n zácpa	zácp	4	
17.	p/ n alpa	ALP	4	
18.	p/ n šerpa	šerp	3	
19.	p/ n rampa	RAMP	3	
20.	p/ n lampa	Lamp	3	























# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 126 (3 str.)


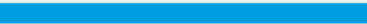


















	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	p/ n hra	her	3 251	
2.	p/ n libra	liber	1 598	
3.	p/ n sestra	sester	999	
4.	p/ n výhra	výher	236	
5.	p/ n Tigra	Tiger	175	
6.	p/ n jiskra	jisker	158	
7.	p/ n kostra	koster	115	
8.	p/ n prohra	proher	94	
9.	p/ n hra	Her	92	
10.	p/ n šifra	šifer	80	
11.	p/ n kra	ker	72	
12.	p/ n nozdra	nozder	68	
13.	p/ n katedra	kateder	59	
14.	p/ n čakra	čaker	54	
15.	p/ n zebra	zeber	47	
16.	p/ n videohra	videoher	45	
17.	p/ n jikra	jiker	43	
18.	p/ n sestra	Sester	42	
19.	p/ n Barbra	Barber	41	
20.	p/ n ambra	Amber	40	

# Frekvenční distribuce


Frekvenční limit:

[Nastavit limit](#)

Celkem: 36 (1 str.)





















	lemma	word	Frekvence	
1.	p/ n Tessa	Tess	1 198	
2.	p/ n kapsa	kapes	754	
3.	p/ n římsa	říms	39	
4.	p/ n Hansa	Hans	24	
5.	p/ n Opsa	Ops	20	
6.	p/ n elipsa	elips	19	
7.	p/ n Impresa	Impress	17	
8.	p/ n Elsa	Els	14	
9.	p/ n Hansa	HANS	13	
10.	p/ n Fatsa	Fats	9	
11.	p/ n Elsa	ELS	9	
12.	p/ n Opsa	OPS	6	
13.	p/ n missa	miss	4	
14.	p/ n fossa	Foss	4	
15.	p/ n Mensa	Mens	4	
16.	p/ n Kossa	Koss	4	
17.	p/ n missa	mess	3	
18.	p/ n apokalypsa	apokalyps	3	
19.	p/ n Pinsa	Pins	3	
20.	p/ n Issa	Iss	3	

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 230 (5 str.)



	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> cesta	cest	3 382	
2.	<a href="#">p/ n</a> karta	karet	1 777	
3.	<a href="#">p/ n</a> varianta	variant	1 192	
4.	<a href="#">p/ n</a> komponenta	komponent	406	
5.	<a href="#">p/ n</a> Investa	Invest	321	
6.	<a href="#">p/ n</a> fakulta	fakult	309	
7.	<a href="#">p/ n</a> Projekta	Projekt	222	
8.	<a href="#">p/ n</a> plachta	plachet	201	
9.	<a href="#">p/ n</a> sekta	sekt	163	
10.	<a href="#">p/ n</a> fronta	front	143	
11.	<a href="#">p/ n</a> pošta	pošt	127	
12.	<a href="#">p/ n</a> šachta	šachet	126	
13.	<a href="#">p/ n</a> konstanta	konstant	88	
14.	<a href="#">p/ n</a> pocta	poct	87	
15.	<a href="#">p/ n</a> dominanta	dominant	86	
16.	<a href="#">p/ n</a> lišta	lišť	82	
17.	<a href="#">p/ n</a> jachta	jachet	70	
18.	<a href="#">p/ n</a> Santa	Sant	69	
19.	<a href="#">p/ n</a> Testa	TEST	63	
20.	<a href="#">p/ n</a> nevěsta	nevěšť	61	

# Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 38 (1 str.)

	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	p/ n barva	barev	3 039	
2.	p/ n vrstva	vrstev	1 871	
3.	p/ n bitva	bitev	300	
4.	p/ n výzva	výzev	187	
5.	p/ n rezerva	rezerv	186	
6.	p/ n želva	želv	161	
7.	p/ n jizva	jizev	132	
8.	p/ n konzerva	konzerv	131	
9.	p/ n larva	larev	74	
10.	p/ n kotva	kotev	74	
11.	p/ n barva	BAREV	34	
12.	p/ n salva	salv	30	
13.	p/ n pitva	pitev	29	
14.	p/ n kurva	kurev	28	
15.	p/ n břitva	břítev	11	
16.	p/ n pochva	pochev	10	
17.	p/ n odezva	odezev	10	
18.	p/ n Selva	SELV	7	
19.	p/ n bulva	bulv	6	
20.	p/ n brva	brv	6	

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

Celkem: 19 (1 str.)

	lemma	word	Frekvence
1.	p/ n slza	slz	658
2.	p/ n burza	burz	105
3.	p/ n slza	slzí	61
4.	p/ n traverza	traverz	14
5.	p/ n Schwarza	Schwarz	12
6.	p/ n pizza	pizz	10
7.	p/ n menza	menz	10
8.	p/ n menza	mens	10
9.	p/ n sanza	Sanz	9
10.	p/ n slza	Slz	3
11.	p/ n llza	llz	3
12.	p/ n slza	SLZ	2
13.	p/ n sanza	sanz	2
14.	p/ n hanza	Hanz	2
15.	p/ n menza	Menz	1
16.	p/ n menza	MENS	1
17.	p/ n lanza	Lanz	1
18.	p/ n bonanza	bonanz	1
19.	p/ n Lanza	Lanz	1

### Formulace závěrů

V následujícím tabulkovém přehledu shrneme pozorování korpusových dat.

finála-koncovka	s alternací	bez alternace
<i>Cb-a</i>	<i>služba/služeb</i>	<i>bomba/bomb</i>
<i>Cd-a</i>	----- <sup>15</sup>	<i>miliarda/miliard</i>
<i>Ck-a</i>	<i>podmínka/podmínek</i>	<i>banka/bank</i>

<sup>15</sup> Jediným dokladem alternace 0/e slova zakončeného na konsonant + -da je substantivum *mzda*. Jde ovšem o alternaci ve skupině CCC, nikoli CC. Obdobou by mohla být alternace u substantiva *msta*. Tvar *mest* je v korpusech doložen pouze v překlapech nebo slovensky psaných textech (gen. pl. slova *mesto*). Ani slovníky tvar genitivu plurálu neuvádějí. V korpusech je řídce doložen genitiv plurálu od prefigované varianty *pomsta/pomst* bez alternace. První případ (substantivum *mzda*) nebere tabulka v Mluvnici současné češtiny (Cvrček a kol. 2010 : 172–173) v úvahu, nicméně se na ně upozorňuje v textu pod tabulkou.

<i>Cl-a</i>	<i>cihla/cihel</i>	-----
<i>Cm-a</i>	<i>firma/firem</i>	<i>termy/term</i> <sup>16</sup>
<i>Cn-a</i>	<i>elektrárna/elektráren</i>	<i>skvrna/skvrn</i>
<i>Cp-a</i>	<i>výspa/výsep</i>	<i>lampa/lamp</i>
<i>Cr-a</i>	<i>hra/her</i>	-----
<i>Cs-a</i>	<i>kapsa/kapes</i>	<i>římsa/říms</i>
<i>Ct-a</i>	<i>karta/karet</i>	<i>cesta/cest</i>
<i>Cv-a</i>	<i>barva/barev</i>	<i>želva/želv</i>
<i>Cz-a</i>	-----	<i>burza/burz</i>
<i>Cf-a</i>	----- <sup>17</sup>	<i>nymfa/nymf</i>
<i>Cg-a</i>	-----	<i>sfinga/sfing</i>
<i>Ch-a</i>	-----	<i>volha/volh</i> <sup>18</sup>
<i>Cch-a</i>	-----	<i>střemcha/střemch</i>
<i>Cš-a</i>	-----	<i>gejša/gejš</i>

Z pozorování dat plyne, že alternace *o/e* v genitivu plurálu feminin skloňovaných podle vzoru *žena* jsou obligatorní pouze v případě substantiv na *C[rl]a* (srov. Caha – Scheer 2007). Obligatorní je také před sufiky *-b(a) služ-b(a)*, *-k(a) bran-k(a)*, *-n(a) kněž-n(a)*, *-v(a) žat-v(a)*, nikoli *-d(a)* srov. *sran-d(a)*.

Naopak mezi substantivy na *-za*, *-fa*, *-ga*, *-ha*, *-cha*, *-ša* případ alternace není doložen. Slova na *C[cč]a* nejsou dle kontextu feminina (*Franca*, *Korča*, ...), jde o chybnou lemmatizaci a tagování.

Výše uvedená tabulka uvádí na základě sledování korpusových dat na pravou míru závěry předložené v *Mluvnici současné češtiny* (viz výše).

### Formulace dalších otázek vypluvších ze zkoumání daného jevu

Je možné tvrdit, že distribuce *o/e* je vázána na potřebu usnadnit výslovnost souhláskové skupiny? Je možné tvrdit, že alternace *o/e* je vázána na morfémovou skladbu slova (tautomorfni/heteromorfni)? Je možné tvrdit, že alternace v tautomorfni pozici je vázána na jednotlivé lexémy?

### Zadání cvičení, v nichž lze uplatnit analogické postupy

<sup>16</sup> Srv. SSČ i SSJČ. Tabulka v *Mluvnici současné češtiny* (Cvrček a kol. 2010 : 172–173) nebere tuto alternaci v úvahu.

<sup>17</sup> Na jednom z běhů LŠSS v Brně položili zvědaví zahraniční studenti dotaz lektorovi, jakpak že zní genitiv plurálu substantiva *Fa* (značka deodorantu). Lektor po chvíli váhání uvedl, že *Ef*.

<sup>18</sup> Doklady genitivu plurálu u slov jako *myrha*, *Volha*, *mha*, ... chybějí v korpusu SYN2010, v korpusu SYN najdeme doklady genitivu pl. *volh* (značka automobilu). K alternaci *o/e* nedochází v případě slabikotvorných sonant (*mlha*, *vlha*, *sprcha*, ...).

Pozoruj doklady jako *rezerv, sekt, želv, žvejlk*. Najdi v korpusu doklady takové, že ve formálně stejném řetězci dochází k alternaci *e/0*.

S podobnými alternacemi se setkáváme také u neuter skloňovaných podle vzoru *město (městeček, pater, divadel, ...)*. Popište, jak budete postupovat, budete-li chtít z korpusu vybrat data pro další interpretaci alternace *0/e*.

Tato alternace se vyskytuje též u některých maskulin. Vyhledejte v korpusu příklady a popište je.

Podívejte se v korpusu na doklady genitivu plurálu substantiva *tma* a analyzujte je.

Alternace *e/0* mají též substantiva skloňovaná podle vzorů *píseň* a *kost*. Vyhledejte v korpusu příklady a popište je.

### ***Další samohláskové alternace v české substantivní flexi***

#### **Motivační úvod**

K samohláskovým alternacím dochází u některých substantiv skloňovaných podle vzoru *žena*, a to a) obligatorně v genitivu plurálu a b) fakultativně v některých dalších pádech (*tráva – trav, síla – sil, míra – měř, houba – hub, ...*). Podíváme-li se do *SSJČ* a *SSČ*, pak zjistíme, že *SSČ* pokládá na rozdíl od *SSJČ* varianty (s/bez alternace) ve všech pádech (kromě obligatorní v genitivu) za rovnocenné. Hodnocení v mluvnicích (např. *PMČ*) i *SSJČ* je diferencovanější.

**ssjc** Slovník spisovného jazyka českého

**síla**

-y ž. (j. 7. sílou, mn. 2. síl, 3. sílám, 6. sílách, 7. sílami, zř. a zast. též sílou, sílám...)

**SSC** Slovník spisovné češtiny

**síla**

-y ž. (2. mn. síl, 7. j., 3., 6., 7. mn. sí- i sí-)

#### **Nastínění problému**

Relevantní data pro analýzu tohoto jevu lze z korpusů získat poměrně jednoduše. Je třeba najít všechna feminina, která se a) skloňují podle vzoru *žena*, b) mají v genitivu plurálu krátkou samohlásku ve kmeni a c) mají lemma s dlouhou samohláskou (diftongem) ve kmeni.

#### **Otázky**

Korpusy mohou poskytnout data pro diferencovanější analýzu tohoto jevu české morfologie. Můžeme najít v korpusech všechna substantiva, která mají tento typ alternace v genitivu plurálu i dalších pádech? Jedná se opravdu ve všech případech o rovnocenné varianty, jak naznačuje SSČ? Jak lze postupovat, abychom zjistili, jaká pravidla platí pro jejich užití?

### Formulace dotazu pro získání dat z korpusů

V jedné z předchozích kapitol jsme ukázali postup, jehož pomocí můžeme hledat v korpusech substantiva skloňovaná podle určitého vzoru. Stručně tedy zopakujme, že naším cílem bude vyhledat substantiva skloňovaná podle vzoru *žena* taková, jejichž kmenová samohláska se v genitivu plurálu obligatorně a v některých dalších pádech fakultativně krátí (alternuje).

Hledáme tedy feminina, která se skloňují podle vzoru *žena* a mají dlouhou kmenovou samohlásku (popř. diftong *ou* ve kmeni), takže je možné, aby tato samohláska alternovala.

Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě `[tag="NNF.*" & lemma="((.*[áéíýóů]|(. *ou.))a"]`. Bereme v úvahu, že takto zachytíme pouze ty případy, kdy mezi koncovkou *-a* a dlouhou kmenovou samohláskou stojí jeden grafém (*síla*), nikoli grafémy dva (*moucha*). Dotaz je ovšem možno přeformulovat tak, aby slova jako *moucha* zachycena byla, tedy `[tag="NNF.*" & lemma="((.*[áéíýóů]|(.ch))|(. *ou(. |ch)))a"]`.

## Hledat v korpusu

**Korpus:**

**Typ dotazu:**

**CQL:**

[vložit tag](#) [vložit "within"](#)

**Z Frekvenční distribuce lemmat** (viz výše) získáme seznam kandidátů lemmat hledaných substantiv.



# Frekvenční distribuce

strana 1

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 1135 (23 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> vláda	36 367	
2.	<a href="#">p/ n</a> síla	33 489	
3.	<a href="#">p/ n</a> zpráva	27 104	
4.	<a href="#">p/ n</a> smlouva	25 637	
5.	<a href="#">p/ n</a> míra	16 136	
6.	<a href="#">p/ n</a> třída	15 758	
7.	<a href="#">p/ n</a> scéna	12 689	
8.	<a href="#">p/ n</a> ztráta	12 677	
9.	<a href="#">p/ n</a> brána	12 639	
10.	<a href="#">p/ n</a> správa	12 072	
11.	<a href="#">p/ n</a> půda	11 931	
12.	<a href="#">p/ n</a> armáda	11 683	
13.	<a href="#">p/ n</a> rána	10 354	
14.	<a href="#">p/ n</a> dráha	9 759	
15.	<a href="#">p/ n</a> atmosféra	8 731	
16.	<a href="#">p/ n</a> analýza	8 520	
17.	<a href="#">p/ n</a> víra	8 397	
18.	<a href="#">p/ n</a> kariéra	8 213	
19.	<a href="#">p/ n</a> dáma	7 868	
20.	<a href="#">p/ n</a> krása	7 592	

Zvolíme filtr (pozitivní) a **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě `[tag="NNFP2.*" & word!=".*([áéíóů](.|ch))|(.ou(.|ch))"]`.

## Filtr konkordancí

Filtr:  pozitivní  negativní

Vybraný token:  první  poslední

Rozsah hledání: od  do   včetně KWIC

Typ dotazu:

Dotaz:

V menu **Frekvenční distribuce** (lemma a word) získáme seznam lemmat hledaných substantiv.

## Víceúrovňová frekvenční distribuce

Frekvenční limit:

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>?</sup>	
1.	lemma	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	-
2.	word	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	X

Přidat úroveň

Vytvořit frekvenční seznam

Pozorujme lemmata a odpovídající slovní tvary.

# Frekvenční distribuce

strana 1

Frekvenční limit:

[Nastavit limit](#)

Celkem: 169 (4 str.)

	lemma	word	Frekvence	
1.	p/ n síla	sil	6 205	
2.	p/ n smlouva	smluv	1 783	
3.	p/ n houba	hub	1 694	
4.	p/ n dráha	drah	1 191	
5.	p/ n rána	ran	989	
6.	p/ n skála	skal	781	
7.	p/ n dáma	dam	656	
8.	p/ n brána	bran	636	
9.	p/ n čára	čar	534	
10.	p/ n díra	děr	519	
11.	p/ n kráva	krav	503	
12.	p/ n žíla	žil	399	
13.	p/ n pára	par	338	
14.	p/ n váha	vah	300	
15.	p/ n moucha	much	272	
16.	p/ n žába	žab	247	
17.	p/ n tráva	trav	231	
18.	p/ n lípa	lip	209	
19.	p/ n louka	luk	201	
20.	p/ n bouda	bud	137	

Třídění a pozorování dat získaných z korpusů

Při srovnávání obou seznamů si povšimneme, že ne všechna lemmata substantiv skloňovaných podle vzoru *žena*, která mají ve kmeni dlouhou samohlásku nebo diftong, mají v češtině alternaci (*půda*, *armáda*, *atmosféra*, *kariéra*, *krása*, *sezóna*, *hrůza*, ...). Můžeme tedy tvrdit, že alternace není obligatorní.

Dále můžeme sestavit seznam lemmat, která v genitivu plurálu alternaci mají. Tento seznam můžeme porovnávat např. se slovníky, mluvnicemi, údaji z *Internetové jazykové příručky*

(dále IJP).

Lze se zamýšlet nad tím, zda se v případě dvojic jako *sezóna/sezona*, *disciplína/disciplin*, *turbína/turbin*, *premiéra/premier*, ... , ale třeba i *bába/bab* jedná o hledaný jev. Vidíme také některé nedostatky v zadáních dotazů (bylo by lépe namísto implicitního atributu **word/slovní tvar** volit **lc**). Projdeme-li frekvenční seznam a vyloučíme-li omyly, získáme seznam substantiv s alternací v genitivu plurálu. Nyní je třeba, abychom ověřili, zda a za jakých podmínek může/musí docházet k alternacím i v jiných pádech.

Nejdříve si položíme otázku, o které pády jde. Můžeme vycházet a) z jazykové kompetence rodilých mluvčích, b) z jazykových příruček (mluvnice, slovníky) a c) z korpusových dat. Pokud budeme vycházet z prvních dvou, pak budeme cíleně v korpusech vyhledávat u lemmat substantiv ze seznamu (těch, která mají v genitivu plurálu příslušnou alternaci) tvary podle tagu, tedy pádové formy, v nichž může také docházet k alternacím. Můžeme ale také postupovat čistě deduktivně a pády, v nichž k alternacím dochází, získat na základě pozorování korpusových dat.

Posledně zmíněný postup může vypadat následovně. Vyhledáme postupně lemmata *síla* (*smlouva*, *houba*, *dráha*, ...). Poté se podíváme na **Frekvenční distribuci** (lemma, lc, tag) a získáme seznam lemmat, tvarů a značek, z něhož je patrné, ve kterých pádech a s jakými frekvencemi dochází/nedochází k příslušné alternaci.

### Víceúrovňová frekvenční distribuce

Frekvenční limit:


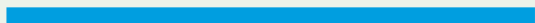


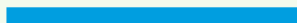


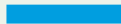
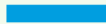








Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>2</sup>	
1.	lemma ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	-
2.	lc ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	<input checked="" type="checkbox"/>
3.	tag ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	<input checked="" type="checkbox"/>

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

Celkem: 17 (1 str.)

	lemma	lc	tag	Frekvence	
1.	p/ n síla	sil	NNFP2---A---	6 304	
2.	p/ n síla	silu	NNFS4---A---	6 086	
3.	p/ n síla	síla	NNFS1---A---	4 457	
4.	p/ n síla	silou	NNFS7---A---	3 684	
5.	p/ n síla	síly	NNFP4---A---	3 464	
6.	p/ n síla	síly	NNFS2---A---	3 251	
7.	p/ n síla	síly	NNFP1---A---	1 926	
8.	p/ n síla	síle	NNFS6---A---	1 382	
9.	p/ n síla	silami	NNFP7---A---	1 181	
10.	p/ n síla	silách	NNFP6---A---	896	
11.	p/ n síla	síle	NNFS3---A---	474	
12.	p/ n síla	silám	NNFP3---A---	293	
13.	p/ n síla	silou	NNFS7---A--1-	76	
14.	p/ n síla	sílo	NNFS5---A---	10	
15.	p/ n síla	silách	NNFP6---A--1-	2	
16.	p/ n síla	silami	NNFP7---A--1-	2	
17.	p/ n síla	silám	NNFP3---A--1-	1	

Výsledky pak můžeme zapsat do tabulky například takto:

<i>síla</i>	gen. pl. (-0)	instr. sg. (-ou)	dat. pl. (-ám)	lok. pl. (-ách)	instr. pl. (-ami)
<i>sil-</i>	6304	3684	293	896	1181
<i>síl-</i>	0	76	1	2	2
<i>smluv-</i>	1796	0	0	0	0
<i>smlouv-</i>	0	919	85	330	180
<i>drah-</i>	11	45	106	196	112
<i>dráh-</i>	0	243	3	22	48

### Formulace závěrů

Na základě analýzy korpusových dat lze např. tvrdit, že alternace v jiných pádech než v genitivu plurálu nemají feminina s kmenovým diftongem *-ou-* (*smlouva, houba, bouda, ...*). V korpusu jsou doloženy tvary *hrudách, hrudu, hrudy*, které nebyly rozpoznány automatickou morfologickou analýzou. Tyto nálezy lze ovšem spíše interpretovat jako doklady teritoriálně podmíněného neprovedení přehlásky *u-ou* (viz příklady nominativu plurálu *hrudy*). Totéž platí i pro varianty *mucha, muchy*.

srpnovém podvečeru . Siluety lidí paběrkují po poli . Na <b>hrudách</b> /hrudách/X@----- sbírají slepice červíky , u plotu dokvétá kozlíku okolík .
půjdu . Oheň ve sněhu Žebřík ke slunci Hvězdokopy V <b>hrudách</b> /hrudách/X@----- země je obráceno všechno kdysi živé , a proto její
i lodyhy nať . . . A k návratu pod <b>hrudu</b> /hrudu/X@----- za jedno měř , kvetla - li růže či vrtiče
hlavě plamen -- sípavý brook rudý v skrytu rozpálené <b>hrudy</b> /hrudy/X@----- -- tetelí se nedočkavě . Neustane moucha vzteká .
často krásnou bez velkého trudu . Na měkkou přišel obyčejně <b>hrudu</b> /hrudu/X@----- a jejím přáním snadno rozuměl . S Helenou první potíž
třicátých letech mu vyšly básnické sbírky Spiewajuco piasc , Hlos <b>hrudy</b> /hrudy/X@----- a v roce 1936 stihl ještě vydat Wybrane wérsze .
já ) a tak jsem mulesa být v kraťasech a <b>muchy</b> /muchy/X@----- mě furt ohlodávaly . Část vedení byla včera v kině
dokazuje svou sílu a teploměr šplhá nad třicítku Žerou mě <b>muchy</b> /muchy/X@----- , potok šustí a země volá po zálivce . Ale

Na základě frekvenční analýzy můžeme tvrdit, že v korpusech nelze najít oporu pro to, že by varianty s alternacemi *á-a, í-i, í-ě* v instrumentálu singuláru a dativu, lokálu a instrumentálu plurálu byly ve všech případech rovnocenné s variantami bez alternací, jak naznačuje interpretace *SSČ*. Liší se na první pohled frekvencí, přičemž rozdíly jsou vázány na jednotlivé lexémy a na kolokace.

Ve korpusech lze najít dokonce i okrajovou alternaci *ů-o*, a to u substantiva *fůra/for*.

**ssjc** Slovník spisovného jazyka českého

### fůra

-y ž. (2. mn. -fůr, -for) (z něm.)

1. vůz s těžkým nákladem, vůz na těžký náklad; množství nákladu, které se na takový vůz vejde: f. uhlí tažená párem koní; nevešlo se to ani na pět for; - koupil několik fůr dříví, [x] ob. expr. mít rozumu na fůry velmi mnoho

2. ob. expr. velké množství: mít fůru peněz, práce, času; -> zřob. fůrka, forka, expr. fůrečka, -y ž.

**psjc** Příruční slovník jazyka českého

**fůra**, -y f. (fůrou i forou, fůr i for, fůrám, fůrách, fůrami) vůz s těžším nákladem. Přijela fůra s nábytkem. Šmil. Naloží fůru jako horu vysokou. Baar. D povoz na těžší náklad. Chasa přijela s fůrami, aby seno odvezla. Klost. Měli jsme vás znát, na fůru vás dát, za hranice vyjet, tam vás vysypat. Havl. D těžší náklad na voze. Fůru pšenice [na půdu] složím. Rais. Kdo si může najednou koupit fůru, aspoň mu ve sklepě vyschně [uhlí]. Herrm. D *Expr. mnoho, velmi mnoho.* Jenom ne se otírat o hejmana, z toho by byla fůra nepřijemných opletaček. Vrba. Probreptáte tady fůru času a za to vám má mistr platit? Krás.

Nenajdeme ji ovšem výše popsaným postupem, protože doklad je chybně interpretován automatickou morfologickou analýzou, což musíme u okrajových jevů vždy předpokládat a doklady se objeví až v celém korpusu SYN (pro nalezení okrajových jevů je vhodný maximální rozsah korpusu). (Možný výskyt genitivu plurálu substantiva *fůra* lze hledat tak, že nejdříve vyhledáme slovní tvar *for*, a poté prostřednictvím filtru vybereme případy, kdy se bezprostředně před předpokládaným substantivem vyskytuje např. číslovka.) Ručně pak vytřídíme hledané doklady.

### **Formulace dalších otázek vypluvších ze zkoumání daného jevu**

Samohláskové alternace se vyskytují také u dalších vzorů (*pán : vůl, muž : kůň, hrad : mráz, stroj : déšť, ...*). U kterých se nevyskytují? Popište pravidla, jimiž se řídí distribuce alternací v jednotlivých flektivních tvarech u substantiv vzorů *pán, hrad, muž, stroj, píseň, kost, ....*

### **Zadání cvičení, v nichž lze uplatnit analogické postupy**

K podobným alternacím dochází také v genitivu plurálu několika substantiv skloňovaných podle vzoru *růže*. Která substantiva to jsou? Pokuste se je najít v korpusu. Mají obdobné alternace také v ostatních pádech, tak jako substantiva skloňovaná podle vzoru *žena*?

### ***Varianty a dublety – korpusy jako zdroje dat pro formulaci pravidel distribuce variantních spisovných koncovek v české substantivní flexi***

#### **Motivační úvod**

Korpus slouží jako zdroj dat pro výzkum variantních a dubletních deklinačních koncovek (srov. více Tušková 2006, 2012). Na tomto místě se pokusíme naznačit, jak budeme postupovat, abychom shromáždili relevantní data, a jak na základě jejich pozorování můžeme dedukcí budovat pravidla distribuce variant a dublet. Vycházíme ze zkušenosti dlouholeté výuky povinných kurzů věnovaných morfologii českého jazyka a z pozorování, z nichž plyne, že pro rodilé mluvčí je velmi obtížné zobecnit jazykovou kompetenci do podoby nějakých pravidel. Mluvnice, které slouží studentům jako výchozí studijní materiály, podávají již hotové výsledky. Chceme proto studentům nabídnout jinou možnost studia: místo aby se učili poučky, uplatní vlastní pozorování a dedukci. Pokusíme se uplatnit deduktivní metodu používanou v rámci tzv. DDL (Data Driven Learning) většinou při výuce angličtiny jako druhého jazyka, jejíž použití pro výuku češtiny jako cizího jazyka se úspěšně rozvíjí (více Osolsobě – Vališová 2012, Vališová 2011).

#### **Nastínění problému**

U všech substantivních vzorů kromě vzorů neuter *kuře* a *stavení* existují v českých gramatikách tzv. podvzory. Co se skrývá pod termínem podvzor? Jde o snahu zachytit v rámci tabulkových paradigmat fakt, že část (většinou má jít o nějakou větší skupinu) slov

skloňovaných podle základního vzoru má u některých tvarů a) koncovku/koncovky navíc, b) jinou koncovku/koncovky. Jako podvzor lze pak chápat také např. skupinu slov, která má větší počet variant. Jde většinou o skupiny slov, u nichž buď lze vymezit jejich počet (uvést jejich úplný výčet), nebo je lze definovat na základě jejich slovtvorné utvářenosti. Krajní možností je chápat takto i jednotlivé slovo (podvzor pro tvoření tvarů jednoho jediného slova). V takových případech mluvnice spíše uvádějí výjimky. Algoritmické popisy české flexe pracují většinou s více podvzory (více Osolsobě 1996, Osolsobě – Pala – Rychlý 1998).

### Otázky

Odpověď na otázku, zda se to a to řekne tak a tak, tedy spontánní užití tvarů s variantními a dubletními koncovkami, je pro nás běžné. Neumíme však mnohdy odpovědět na otázky kde a jak<sup>19</sup>. Za těchto okolností se velmi často setkáváme s tím, že studenti používají v odpovědích na otázky týkající se vztahů variantních a dubletních koncovek definice kruhem.

Např. na otázku, jaká jsou pravidla pro distribuci variantní koncovky *-u/-e* u vzoru *pán*, odpovídají, že koncovku *-e* mají substantiva skloňovaná podle vzoru *pán* a koncovku *-u* substantiva skloňovaná podle vzoru *hoch*. Odpověď na otázku, podle jakého kritéria se řadí substantiva k jednotlivému podvzoru, bývá ovšem někdy kamenem úrazu.

### Formulace dotazu pro získání dat z korpusů

V jedné z předchozích kapitol jsme ukázali postup, jehož pomocí můžeme hledat v korpusech substantiva skloňovaná podle určitého vzoru. Stručně tedy zopakujme, že naším cílem bude vyhledat substantiva skloňovaná podle vzoru *pán* ve vokativu singuláru a formulovat pravidla, jimiž se řídí distribuce variantních/dubletních koncovek.

Víme, že maskulina skloňovaná podle vzoru *pán* lze vyhledat pomocí morfologické značky a kmenové finály. Dále víme, že kmenové finály *[lsz]* mají jak substantiva (maskulina životná) skloňovaná podle vzoru *pán*, tak substantiva skloňovaná podle vzoru *muž*.

Hledáme-li tedy substantiva skloňovaná podle vzoru *pán* ve vokativu singuláru, hledáme všechna maskulina životná, jejichž lemmata končí na *[dtnhgkrbpfvm]* (a mohou končit na *[lsz]*) a která mají značku *NNMS5.\**. Zvolíme tedy **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag="NNMS5.\*" & lemma=".\*[dtnhgkrbpfvm]"]*.

---

<sup>19</sup> Ještě složitější jsou pak odpovědi na otázku proč. Systém jazyka v jeho synchronní podobě je výsledkem historického vývoje a řadu jevů z oblasti morfologie, které se ze synchronního pohledu jeví jako výjimky, lze objasnit diachronně. Existují ovšem i teoretické koncepty ryze synchronní.



## Hledat v korpusu

Korpus:

syn2010

Typ dotazu:

CQL

CQL:

vložit tag

vložit "within"

[tag="NNMS5.\*" & lemma=".\*[dtnhgkrbpfvm]"]

V dalším kroku se podíváme na frekvenční distribuci lemma a tvarů (viz výše). Pozorujeme, která lemmata mají koncovku *-e* a která koncovku *-u* a hledáme pravidla distribuce.

## Víceúrovňová frekvenční distribuce


Frekvenční limit: 0

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>?</sup>	
1.	lemma	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	-
2.	word	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	x

Přidat úroveň





















Vytvořit frekvenční seznam

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 2777 (56 str.)

	lemma	word	Frekvence	
1.	p/ n pan	pane	7 056	
2.	p/ n bůh	Bože	2 762	
3.	p/ n pan	Pane	2 048	
4.	p/ n doktor	doktore	1 160	
5.	p/ n bůh	bože	1 127	
6.	p/ n kamarád	kamaráde	845	
7.	p/ n kapitán	kapitáne	813	
8.	p/ n David	Davide	618	
9.	p/ n člověk	člověče	564	
10.	p/ n šéf	šéfe	538	
11.	p/ n Tom	Tome	465	
12.	p/ n John	Johne	435	
13.	p/ n bratr	bratře	403	
14.	p/ n Robert	Roberte	338	
15.	p/ n major	majore	335	
16.	p/ n člověk	Člověče	313	
17.	p/ n miláček	miláčku	301	
18.	p/ n profesor	profesore	281	
19.	p/ n Adam	Adame	267	
20.	p/ n poručík	poručíku	262	

## Třídění a pozorování dat získaných z korpusů

Můžeme postupně pomocí filtru (pozitivní) vytvořit seznam maskulin s koncovkou a) *-u* a b) *-e*.

a)

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC


**Typ dotazu:**  ▼

**Slovní tvar:**

Filtrovat konkordanci

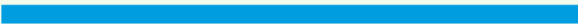






Vymazat formulář

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 677 (14 str.)

	<b>lemma</b>	<b>word</b>	<b>Frekvence</b>	
1.	p/ n miláček	miláčku	301	
2.	p/ n poručík	poručíku	262	
3.	p/ n hoch	hochu	141	
4.	p/ n syn	synu	134	
5.	p/ n drahoušek	drahoušku	134	
6.	p/ n plukovník	plukovníku	105	
7.	p/ n brach	brachu	75	
8.	p/ n strýček	strýčku	56	
9.	p/ n soudruh	soudruhu	53	
10.	p/ n Jack	Jacku	53	
11.	p/ n synek	synku	52	
12.	p/ n tatínek	tatínku	38	
13.	p/ n chlapeček	chlapečku	37	
14.	p/ n mladík	mladíku	36	
15.	p/ n Frank	Franku	36	
16.	p/ n náměstek	náměstku	34	
17.	p/ n Hank	Hanku	34	
18.	p/ n strýček	Strýčku	31	
19.	p/ n kluk	kluku	30	
20.	p/ n Jindřich	Jindřichu	29	

b)

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**  ▼

**Slovní tvar:**

Filtrovat konkordanci

Vymazat formulář


# Frekvenční distribuce

strana 1

Frekvenční limit:

[Nastavit limit](#)

Celkem: 2081 (42 str.)

	lemma	word	Frekvence	
1.	p/ n pan	pane	7 056	
2.	p/ n bůh	Bože	2 762	
3.	p/ n pan	Pane	2 048	
4.	p/ n doktor	doktore	1 160	
5.	p/ n bůh	bože	1 127	
6.	p/ n kamarád	kamaráde	845	
7.	p/ n kapitán	kapitáne	813	
8.	p/ n David	Davide	618	
9.	p/ n člověk	člověče	564	
10.	p/ n šéf	šéfe	538	
11.	p/ n Tom	Tome	465	
12.	p/ n John	Johne	435	
13.	p/ n bratr	bratře	403	
14.	p/ n Robert	Roberte	338	
15.	p/ n major	majore	335	
16.	p/ n člověk	Člověče	313	
17.	p/ n profesor	profesore	281	
18.	p/ n Adam	Adame	267	
19.	p/ n prezident	prezidente	250	
20.	p/ n mylord	mylorde	248	

V dalším kroku se budeme zabývat substantivy s lemmaty na *[lsz]*, která mají značku *NNMS5.\**. Zvolíme tedy **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag="NNMS5.\*" & lemma=".\*[lsz]" ]*.


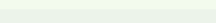



Podíváme se na frekvenční distribuci lemmat a slovních tvarů a budeme přitom postupovat stejně jako při řešení předchozích případů.

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 740 (15 str.)

	lemma	word	Frekvence	
1.	p/ n přítel	příteli	406	
2.	p/ n Kristus	Kriste	217	
3.	p/ n hajzl	hajzle	208	
4.	p/ n Fontinus	Fontine	207	
5.	p/ n Bill	Bille	175	
6.	p/ n generál	generále	148	
7.	p/ n Will	Wille	133	
8.	p/ n velitel	veliteli	124	
9.	p/ n Saul	Saule	114	
10.	p/ n ředitel	řediteli	96	
11.	p/ n Holmes	Holmesi	96	
12.	p/ n učitel	učiteli	62	
13.	p/ n Michael	Michaeli	59	
14.	p/ n Karel	Karle	54	
15.	p/ n Cecil	Cecile	48	
16.	p/ n Hastings	Hastingsi	46	
17.	p/ n Cadfael	Cadfaele	46	
18.	p/ n James	Jamesi	45	
19.	p/ n Davis	Davisi	44	
20.	p/ n král	králi	43	

Pozorujeme, že substantiva skloňovaná podle vzoru *pán* mají koncovku *-e* (*hajzle, Bille, generále, Wille, Saule, Karle, ...*), zatímco substantiva skloňovaná podle vzoru *muž* mají koncovku *-i* (*příteli, veliteli, řediteli, Holmesi, Michaeli, ...*). Pokusíme se tedy získat přehled o distribuci variant u substantiv na *[lsz]* tak, že pomocí filtru (negativní) odstraníme ze

seznamu slovní tvary zakončené na *. \*i* (odstraníme tvary vokativu singuláru substantiv skloňovaných podle vzoru *muž* na  $-[lsz]^{20}$ ).

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**  ▼

**Slovní tvar:**

<sup>20</sup> Toho, koho bude trápit otázka, zda neexistují náhodou substantiva skloňovaná podle vzoru *muž* na  $-[lsz]$ , která mají ve vokativu singuláru koncovku *-e*, vyzýváme, aby na základě pozorování korpusových dat vyvodil odpověď na otázku, která jej trápí. Jakými pravidly se řídí distribuce variantních koncovek *-i/-e* ve vokativu singuláru maskulin životných skloňovaných podle vzoru *muž*? Jak lze z hlediska pravidel distribuce hodnotit tvary vokativu substantiva *kněz* doložené v korpusu SYN2010?









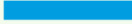













# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 535 (11 str.)

	lemma	word	Frekvence	
1.	p/ n Kristus	Kriste	217	
2.	p/ n hajzl	hajzle	208	
3.	p/ n Fontinus	Fontine	207	
4.	p/ n Bill	Bille	175	
5.	p/ n generál	generále	148	
6.	p/ n Will	Wille	133	
7.	p/ n Saul	Saule	114	
8.	p/ n Karel	Karle	54	
9.	p/ n Cecil	Cecile	48	
10.	p/ n Cadfael	Cadfaele	46	
11.	p/ n Paul	Paule	40	
12.	p/ n los	Lose	39	
13.	p/ n Julianus	Juliane	35	
14.	p/ n Honorius	Honorie	33	
15.	p/ n Vasil	Vasile	32	
16.	p/ n generál	Generále	29	
17.	p/ n Sméagol	Sméagole	29	
18.	p/ n Cyril	Cyrile	29	
19.	p/ n Pavel	Pavle	28	
20.	p/ n pes	pse	27	

## Formulace závěrů

Na základě pozorování dat můžeme formulovat pravidla. Koncovka *-u* se s výjimkou substantiva *syn* vyskytuje pouze u substantiv s kmenovou finálou, kterou je velára *.\*([gkh]|(ch)).*

Tento závěr ověříme tak, že u konkordance zahrnující maskulina životná, která mají ve vokativu singuláru koncovku *-u*, odstraníme pomocí negativního filtru slovní tvary s kmenovou finálou, jíž je velára (tedy ty, které končí na *.\*[ghk]u*).

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 1 (1 str.)

	<b>lemma</b>	<b>lc</b>	<b>Frekvence</b>	
--	--------------	-----------	------------------	--

1.	p/ n	syn	synu	136
----	------	-----	------	-----

Na základě pozorování dat můžeme dále tvrdit, že koncovku *-e* mají substantiva, která nekončí na veláru. Výjimkou jsou substantiva jako *Bože, člověče, ...*, jejichž lemmata sice na veláru končí, ta ale před koncovkou *-e* alternuje. Druhou skupinu výjimek představují některá vlastní jména cizího původu, která končí grafémem *h* většinou ve skupině *[tp]h* (např. *Josephe*).

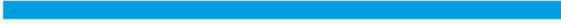

Tento závěr ověříme tak, že z konkordance zahrnující maskulina životná, která mají ve vokativu singuláru koncovku *-e*, vybereme pomocí pozitivního filtru lemmata s kmenovou finálou, jíž je velára (tedy lemmata končící na *.\*[ghk]*). Dále zobrazíme frekvenční distribuci lemmat a tvarů.

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

Celkem: 36 (1 str.)

	lemma	lc	Frekvence	
1.	p/ n bůh	bože	3 902	
2.	p/ n člověk	člověče	888	
3.	p/ n Ralph	ralphe	25	
4.	p/ n člověk	čoveče	19	
5.	p/ n člověk	čéče	16	
6.	p/ n Seth	sethe	11	
7.	p/ n Astaroth	astaroth	11	
8.	p/ n poutník	poutníče	10	
9.	p/ n Wordsworth	wordsworthe	10	
10.	p/ n Goldsmith	goldsmithe	10	
11.	p/ n pánbůh	panebože	7	
12.	p/ n Worth	worthe	6	
13.	p/ n Matzerath	matzerathe	6	
14.	p/ n Woolworth	woolworthe	5	
15.	p/ n Griffith	griffithe	5	
16.	p/ n Forsyth	forsythe	5	
17.	p/ n Joseph	josephe	3	
18.	p/ n Barth	barthe	3	
19.	p/ n Adolph	adolphe	3	
20.	p/ n Whitworth	whitworthe	2	

Všimneme si, že v seznamu se objevují slova cizího původu (většinou propria z angličtiny) zakončená na *-th*, které odpovídá frikativní dentále<sup>21</sup>, nebo *-ph*, které vyslovujeme jako [f], takže grafické *h* nereprezentuje veláru. Kromě toho si všimneme kompozit s druhým členem – *člověk*, substandardního tvaru *čéče* a tvarů *nešťastníče*, které zní v uších milovníků opery *Prodaná nevěsta*, a *poutníče* připomínající bitvu u Thermopyl.

### Formulace dalších otázek vypluvších ze zkoumání daného jevu

<sup>21</sup> Mezi nimi se evidentně vyskytují chyby v desambiguaci. Srv. např. tvary *Goldsmithe*, které jsou ve všech případech mylně desambiguovány jako tvary vokativu. Vlastní jména jako *Smith*, *Goldsmith* kolísají mezi vzory *pán* a *muž*. Tento fakt u nich zvyšuje míru tvarové homonymie, což negativně ovlivňuje desambiguaci.

Podmiňuje hláskové okolí distribuci variant vždy? Pokud ano, pak jak? Pokud nikoli, čím se distribuce řídí v jednotlivých případech?

V následujícím přehledu uvedeme tabulky substantivní flexe a vyznačíme barevně případy variant/dublet, čímž naznačíme, kterých dalších případů si lze při zkoumání daného jevu všimnout.

<i>pán-0</i>	<i>muž-0</i>	<i>předsed-a</i>	<i>soudc-e</i>
<i>pán-a</i>	<i>muž-e</i>	<i>předsed-y</i>	<i>soudc-e</i>
<i>pán-u/-ovi</i>	<i>muž-i/-ovi</i>	<i>předsed-ovi</i>	<i>soudc-i/-ovi</i>
<i>pán-a</i>	<i>muž-e</i>	<i>předsed-u</i>	<i>soudc-e</i>
<i>pane-e vrah-u</i>	<i>muž-i otč-e</i>	<i>předsed-o</i>	<i>soudc-e</i>
<i>pán-u/-ovi</i>	<i>muž-i/-ovi</i>	<i>předsed-ovi</i>	<i>soudc-i/-ovi</i>
<i>pán-em</i>	<i>muž-em</i>	<i>předsed-ou</i>	<i>soudc-em</i>
<i>pán-i/-ové</i>	<i>muž-i/-ové</i>	<i>předsed-i/-ové</i>	<i>soudc-i/-ovi</i>
<i>občan-é</i>	<i>obyvatel-é</i>	<i>husit-é</i>	
<i>pán-ů lid-i</i>	<i>muž-ů kon-i</i> <i>obyvatel-0</i>	<i>předsed-ů</i>	<i>soudc-ů</i>
<i>pán-ům lid-em</i>	<i>muž-ům kon-ím</i>	<i>předsed-ům</i>	<i>soudc-ům</i>
<i>pán-y lid-i</i>	<i>muž-e</i>	<i>předsed-y</i>	<i>soudc-e</i>
<i>pán-i/-ové</i>	<i>muž-i/-ové</i>	<i>předsed-i/-ové</i>	<i>soudc-i/-ovi</i>
<i>občan-é</i>	<i>obyvatel-é</i>	<i>husit-é</i>	
<i>pán-ech vraz-ích</i>	<i>muž-ích</i>	<i>předsed-ech</i>	<i>soudc-ích</i>
<i>pán-y lid-mi</i>	<i>muž-i koň-mi</i>	<i>předsed-y</i>	<i>soudc-i</i>

Kolik je v češtině substantiv, která se skloňují podle typu *obyvatel*? Kolik českých substantiv má variantní koncovky jako substantivum *kůň* a kolik jako substantivum *lidé*?

<i>hrad-0</i>	<i>stroj-0</i>
<i>hrad-u les-a dn-e</i>	<i>stroj-e</i>
<i>hrad-u dn-i</i>	<i>stroj-i</i>
<i>hrad-0</i>	<i>stroj-0</i>
<i>hrad-e stolec-k-u dn-i</i>	<i>stroj-i kopč-e</i>
<i>hrad-u les-e dn-i</i>	<i>stroj-i</i>
<i>hrad-y</i>	<i>stroj-i</i>

<i>hrad-y dn-i/-y/-ové</i>	<i>stroj-e</i>
<i>hrad-ů dn-í Loun-0</i>	<i>stroj-ů peněz-0</i>
<i>hrad-ům</i>	<i>stroj-ům</i>
<i>hrad-y</i>	<i>stroj-e</i>
<i>hrad-y dn-i/-ové</i>	<i>stroj-e</i>
<i>hrad-ech les-ích domečk-ách</i>	<i>stroj-ích</i>
<i>hrad-y</i>	<i>stroj-i</i>

Která česká neživotná maskulina mají v genitivu plurálu tvar bez koncovky? Kolik českých substantiv má stejné variantní koncovky jako substantivum *den*?

<i>žen-a</i>	<i>růž-e</i>	<i>píseň-0</i>	<i>kost-0</i>
<i>žen-y Mán-i</i>	<i>růž-e</i>	<i>písn-ě</i>	<i>kost-i</i>
<i>žen-ě</i>	<i>růž-i</i>	<i>písn-i</i>	<i>kost-i</i>
<i>žen-u</i>	<i>růž-i</i>	<i>píseň-0</i>	<i>kost-0</i>
<i>žen-o</i>	<i>růž-e</i>	<i>písn-i</i>	<i>kost-i</i>
<i>žen-ě</i>	<i>růž-i</i>	<i>písn-i</i>	<i>kost-i</i>
<i>žen-ou</i>	<i>růž-í</i>	<i>písn-í</i>	<i>kost-í</i>
<i>žen-y Mán-i</i>	<i>růž-e</i>	<i>písn-ě</i>	<i>kost-i</i>
<i>žen-0</i>	<i>růž-i</i>	<i>písn-í mil-0</i>	<i>kost-í</i>
<i>žen-ám</i>	<i>růž-ím</i>	<i>písn-ím</i>	<i>kost-em</i>
<i>žen-y</i>	<i>růž-e</i>	<i>písn-ě</i>	<i>kost-i</i>
<i>žen-y</i>	<i>růž-e</i>	<i>písn-ě</i>	<i>kost-i</i>
<i>žen-ách</i>	<i>růž-ích</i>	<i>písn-ích</i>	<i>kost-ech</i>
<i>žen-ami</i>	<i>růž-emi</i>	<i>písn-ěmi</i>	<i>kost-mi</i>

Kolik substantiv skloňovaných podle vzoru *růže* má v genitivu plurálu variantní koncovku *-0*? Některá substantiva mohou mít tvary jak podle vzoru *růže*, tak podle vzoru *píseň* a jak podle vzoru *píseň*, tak podle vzoru *kost*. Navrhňte postup, jak z korpusu získáte relevantní data, porovnejte své vyhledávky s tím, co se o dané problematice píše např. v *Příruční mluvnici češtiny*, nebo v *Internetové jazykové příručce*.

<i>měst-o</i>	<i>moř-e</i>	<i>kuř-e</i>	<i>staven-í</i>
<i>měst-a</i>	<i>moř-e</i>	<i>kuř-et-e</i>	<i>staven-í</i>

<i>měst-u břemen-i</i>	<i>moř-i</i>	<i>kuř-et-i</i>	<i>staven-í</i>
<i>měst-o</i>	<i>moř-e</i>	<i>kuř-e</i>	<i>staven-í</i>
<i>měst-o</i>	<i>moř-e</i>	<i>kuř-e</i>	<i>staven-í</i>
<i>měst-ě (po) rán-u břemen-i</i>	<i>moř-i</i>	<i>kuř-et-i</i>	<i>staven-í</i>
<i>měst-em</i>	<i>moř-em</i>	<i>kuř-et-em</i>	<i>staven-ím</i>
<i>měst-a</i>	<i>moř-e</i>	<i>kuř-at-a</i>	<i>staven-í</i>
<i>měst-0</i>	<i>moř-í vajec-0</i>	<i>kuř-at-0</i>	<i>staven-í</i>
<i>měst-ům</i>	<i>moř-ím</i>	<i>kuř-at-ům</i>	<i>staven-ím</i>
<i>měst-a</i>	<i>moř-e</i>	<i>kuř-at-a</i>	<i>staven-í</i>
<i>měst-a</i>	<i>moř-e</i>	<i>kuř-at-a</i>	<i>staven-í</i>
<i>měst-ech jablc-ích městečk-ách</i>	<i>moř-ích</i>	<i>kuř-at-ech</i>	<i>staven-ích</i>
<i>měst-y</i>	<i>moř-i</i>	<i>kuř-at-y</i>	<i>staven-ími</i>

Jak se skloňují substantiva na *-um*, *-eum*, *-ium*, *-eon*? Kolik substantiv skloňovaných podle vzoru *moře* má v genitivu plurálu koncovku jako *vejce*? Kolik substantiv skloňovaných podle vzoru *město* má variantní koncovky jako substantivum *břemeno* (*břímě*)?

### Zadání cvičení, v nichž lze uplatnit analogické postupy

Která maskulina s kmenovou finálou *r* mají ve vokativu singuláru alternaci *r>ř* a která nikoli?

Lze formulovat obecně platné pravidlo, nebo je třeba uvést výčet?

Pokuste se formulovat pravidla, podle nichž se v české substantivní flexi distribuuje grafická varianta *e/ě* v koncovkách. Jde o jednoduchá pravidla? Je distribuce závislá na předcházející finále, nebo ne?

V jakém pádě je substantivum *anděl* ve spojení *<Anděle Boží, strážce můj>* a v jakém ve spojení *<Ó Lucie, můj krásný anděli>* ...? V jakém pádě je substantivum *pán* ve spojení *<anděl Páně>*, *<den Páně>*, *<večeře Páně>*?

V jakém pádě je substantivum *den* ve spojení *<ve dne>* a v jakém *<v tomto dni>*?

Najděte v korpusu SYN tvar *Vojtěše* a podívejte se, jak je interpretován na úrovni lemmatu a morfologické značky.

## **Adjektiva**

V následujících podkapitolách se zaměříme na otázky spojené se stupňováním adjektiv. Popisy tvoření syntetických tvarů komparativu a superlativu adjektiv bývají v lingvistických příručkách řazeny jednak k pojednáním o formální morfologii adjektiv (např. *Česká mluvnice* Bohuslava Havránka a Aloise Jedličky nebo *Mluvnice současné češtiny* Václava Cvrčka, toto pojetí sledují i výkladové slovníky a taktéž systémy značkování korpusů), jednak k oddílům zaměřeným na tvoření slov (např. *Mluvnice češtiny 1*, *PMČ*, *ČŘJ*). V jednotlivých oddílech budeme sledovat výskyt tvarů komparativu a superlativu v českých korpusech. Zaměříme se i na to, jak mohou data získaná z korpusů upřesnit tvrzení o stupňovatelnosti některých typů adjektiv tradovaná v české jazykovědné literatuře.

### ***Která adjektiva se v češtině stupňují?***

#### **Motivační úvod**

Meze tvoření syntetických tvarů komparativu sufixy *-í*, *-ší*, *-e/ější* a superlativu prefixem *nej-* jsou v mluvnících naznačeny obecnými formulacemi. Východiskem je sémantická klasifikace adjektiv a jejich dělení na kvalifikační (+hodnotící) a relační (Dokulil – Komárek 1986), popř. na adjektiva s příznakem +/- graduálnost (Karlík – Nekula – Pleskalová 2002 : s. 447).

#### **Nastínění problému**

Ve značkových korpusech ČNK mají na 10. pozici stupeň uvedenu hodnotu [123] pouze adjektiva, která mají na druhé pozici – detailní určení slovního druhu – hodnotu A (obyčejné adjektivum). Všechna ostatní adjektiva mají na 10. pozici vyplněnu hodnotu nestupňuje se (-). Mezi nestupňovatelná adjektiva se tedy implicitně řadí jmenné tvary adjektiv (AC.\*), adjektivizované přechodníky přítomné (AG.\*), minulé (AM.\*) a subjektivně posesivní adjektiva tvořená od životných maskulin a feminin sufixy *-ův*, *-in* (AU.\*).

Korpusy byly a jsou budovány s cílem zkoumat jazyk empiricky tak, jak skutečně funguje. Proto se autoři nejrůznějších popisů a modelů jazyka snaží ověřit platnost popisu/modelu na korpusových datech. Podívejme se tedy, jak mohou korpusy potvrdit/vyvrátit výše formulovaná pravidla a intuice včetně těch, z nichž vyšli autoři automatických morfologických analyzátorů.

Na prvním místě je třeba konstatovat, že i stamilionové korpusy dokládají výskyt tvarů II. a III. stupně jen sporadicky. Zobecnění závěrů opřených o data z nich získaná je obtížné.

## Otázky

Jaký je procentuální poměr tvarů, které mají na 10. pozici (stupeň) uvedenu hodnotu 1, 2 a 3?

Existují v korpusech tvary na *.\*ší, .\*[eě]jší*, které nemají na 10. pozici hodnotu 2 nebo 3?

Co je na nich zajímavé?

## Formulace dotazu pro získání dat z korpů

Nejdříve se podívejme na procentuální zastoupení tvarů označovaných jako stupňovatelné dle hodnoty stupeň (pozice 10). Jak budeme postupovat?

Chceme-li získat přehled o tvarech/lemmatech adjektiv a o tom, kolik adjektivních tvarů/lemmat pozitivu, komparativu, superlativu se v korpusech vyskytuje, budeme postupovat takto. Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag="A.....[123].\*"]*.

Poté pomocí pozitivního filtru postupně vybereme tvary označované jako tvary pozitivu, komparativu a superlativu a podíváme se na frekvenční distribuci lemmat a zapíšeme si počet výskytů slovních tvarů (word) a lemmat.

### Hledat v korpusech

**Korpus:**

**Typ dotazu:**

**CQL:**

[vložit tag](#) [vložit "within"](#)



## Filtr konkordancí

**Filtr:**  pozitivní  negativní  
**Vybraný token:**  první  poslední  
**Rozsah hledání:** od 0 do 0  včetně KWIC  
**Typ dotazu:** CQL

**CQL:**

[vložit tag](#) [vložit "within"](#)  
[tag="A.....1.\*"]

Filtrovat konkordanci

Vymazat formulář

Výskytů: 9 933 690 | i.p.m.?: 81 646,27 (vztaženo k celému syn2010)

## Frekvenční distribuce

strana 1

Frekvenční limit: 0 [Nastavit limit](#)

**Celkem:** 49883 (998 str.)

Dále se podíváme na lemmata měkkých adjektiv, která končí na *\*ší* (formálně se podobají tvarům komparativu a superlativu). Zvolíme **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz ve formě *\*ší*. Poté pomocí filtru (pozitivní) vybereme adjektiva a podíváme se na frekvenční distribuci lemmat.


## Hledat v korpusu

**Korpus:** syn2010

**Typ dotazu:** Lemma

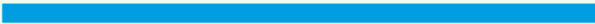








**Lemma:** *\*ší*

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 464 (10 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> další	147 982	
2.	<a href="#">p/ n</a> vnější	7 662	
3.	<a href="#">p/ n</a> zdejší	7 558	
4.	<a href="#">p/ n</a> tehdejší	6 305	
5.	<a href="#">p/ n</a> někdejší	4 732	
6.	<a href="#">p/ n</a> včerejší	4 610	
7.	<a href="#">p/ n</a> vedlejší	4 298	
8.	<a href="#">p/ n</a> nynější	2 783	
9.	<a href="#">p/ n</a> sebemenší	2 017	
10.	<a href="#">p/ n</a> protější	1 729	
11.	<a href="#">p/ n</a> pěší	1 539	
12.	<a href="#">p/ n</a> dřívější	1 520	
13.	<a href="#">p/ n</a> zítřejší	873	
14.	<a href="#">p/ n</a> tamější	719	
15.	<a href="#">p/ n</a> postarší	567	
16.	<a href="#">p/ n</a> myší	521	
17.	<a href="#">p/ n</a> zazší	354	
18.	<a href="#">p/ n</a> přeživší	254	
19.	<a href="#">p/ n</a> sebelepší	181	
20.	<a href="#">p/ n</a> sebevětší	135	

## Třídění a pozorování dat získaných z korpusů

Rozložení výskytu tvarů pozitivu, komparativu a superlativu v korpusu SYN2010 zachycuje následující tabulka.

	<i>A.....[123].*</i>	<i>A.....1.*</i>	<i>A.....2.*</i>	<i>A.....3.*</i>
word	10428947	9933690	280941	214316
	(100 %)	(95,25 %)	(2,69 %)	(2,06 %)

lemma	49914	49883	3149	2338
-------	-------	-------	------	------

Lemmata, která končí na *-ší* (viz výše), jsou např. deadverbiální adjektiva (*zdejší, tehdejší, někdejší, ...*), o nichž se někdy hovoří jako o „formálních komparativech“, dále adjektiva druhově posesivní (*myší, bleší, ...*) a adjektiva z přechodníků minulých (*přeživší, odstoupivší, ...*), ale také komparativ *zazší* a komparativy tantum jako *sebechytřejší, sebelepší, ...*. Zajímavé je sledovat značkování jednotlivých tvarů (tvary *dřívější* mají ve značce uvedeno, že se jedná o tvary 2. stupně).

V dalším kroku můžeme zjistit, která další adjektiva jsou označována jako komparativy/superlativy. Pomocí filtru (pozitivní) vybereme tvary označované *A.....[23].\** a podíváme se na frekvenční přehled lemmat/ slovních tvarů.

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**  ▼

**CQL:** vložit tag vložit "within"

[tag=".....[23].\*"]

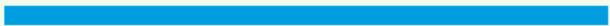











Filtrovat konkordanci
Vymazat formulář

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

Celkem: 12 (1 str.)

	lemma	lc	Frekvence	
1.	p/ n dřívější	dřívějších	814	
2.	p/ n dřívější	dřívějšího	303	
3.	p/ n dřívější	dřívějším	255	
4.	p/ n zazší	nejzazší	169	
5.	p/ n zazší	nejzazším	98	
6.	p/ n dřívější	dřívějšímu	84	
7.	p/ n dřívější	dřívějšími	61	
8.	p/ n zazší	nejzazších	40	
9.	p/ n zazší	nejzazšího	36	
10.	p/ n zazší	nejzazšímu	11	
11.	p/ n dřívější	nejdřívější	2	
12.	p/ n dřívější	nejdřívějším	1	

Je patrné, že jde o „technická řešení“, která lze přičíst na vrub automatické morfologické analýze.

### Formulace závěrů

V českých korpusech je výskyt užití tvarů komparativu a superlativu velmi vzácný (necelých 5 % všech tvarů adjektiv, která jsou označovaná jako „stupňovatelná“).

Navíc je patrné, že při použití automatické morfologické analýzy se nebere zřetel na některé výjimky a anomálie. Jednou (okrajovou) je značkování a lemmatizace tvarů *(nej)zazší.\** a *(nej)dřívější.\** (viz výše) a tvarů komparativů tantum derivovaným prefixoidem *sebe-*.<sup>22</sup>

Vrátíme se zpět k seznamu lemmat adjektiv na *.\*ší*. a) Za pomoci filtru (pozitivní) vybereme tvary s lemmatem *sebe.\** a podíváme se na jejich seznam.

b) Poté se vrátíme a tvary s lemmatem *sebe.\** pomocí filtru (negativní) odstraníme a podíváme se na jejich seznam.

<sup>22</sup> Vyjadřování komparativu i superlativu genitivem (*lepší sebe>sebelepší*) bylo běžné už ve staroslověněštině. V řadě jazyků se setkáváme s etymologickou figurou tzv. hebrejského superlativu (*král králů, kniha knih, píseň písní*), v níž se genitivem vyjadřuje superlativ.

a)

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**  ▼

**Slovní tvar:**





















# Frekvenční distribuce

strana 1

Frekvenční limit: 0


Nastavit limit

Celkem: 106 (3 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> sebemenší	2 017	
2.	<a href="#">p/ n</a> sebelepší	181	
3.	<a href="#">p/ n</a> sebevětší	135	
4.	<a href="#">p/ n</a> sebekrásnější	21	
5.	<a href="#">p/ n</a> sebesilnější	18	
6.	<a href="#">p/ n</a> sebehorší	17	
7.	<a href="#">p/ n</a> sebedokonalejší	17	
8.	<a href="#">p/ n</a> sebeslabší	16	
9.	<a href="#">p/ n</a> sebekratší	12	
10.	<a href="#">p/ n</a> sebetěžší	8	
11.	<a href="#">p/ n</a> sebedelší	8	
12.	<a href="#">p/ n</a> seberychlejší	7	
13.	<a href="#">p/ n</a> sebedražší	7	
14.	<a href="#">p/ n</a> sebevzdálenější	6	
15.	<a href="#">p/ n</a> sebevyšší	6	
16.	<a href="#">p/ n</a> sebebodivnější	6	
17.	<a href="#">p/ n</a> sebedůmyslnější	6	
18.	<a href="#">p/ n</a> sebedůležitější	6	
19.	<a href="#">p/ n</a> sebechytřejší	6	
20.	<a href="#">p/ n</a> sebestrašnější	5	

b)

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 358 (8 str.)

	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> další	147 982	
2.	<a href="#">p/ n</a> vnější	7 662	
3.	<a href="#">p/ n</a> zdejší	7 558	
4.	<a href="#">p/ n</a> tehdejší	6 305	
5.	<a href="#">p/ n</a> někdejší	4 732	
6.	<a href="#">p/ n</a> včerejší	4 610	
7.	<a href="#">p/ n</a> vedlejší	4 298	
8.	<a href="#">p/ n</a> nynější	2 783	
9.	<a href="#">p/ n</a> protější	1 729	
10.	<a href="#">p/ n</a> pěší	1 539	
11.	<a href="#">p/ n</a> dřívější	1 520	
12.	<a href="#">p/ n</a> zítřejší	873	
13.	<a href="#">p/ n</a> tamější	719	
14.	<a href="#">p/ n</a> postarší	567	
15.	<a href="#">p/ n</a> myší	521	
16.	<a href="#">p/ n</a> zazší	354	
17.	<a href="#">p/ n</a> přeživší	254	
18.	<a href="#">p/ n</a> vezdejší	122	
19.	<a href="#">p/ n</a> bleší	117	
20.	<a href="#">p/ n</a> hroší	103	

Obdobou adjektiv s prefixoidem *sebe-* jsou adjektiva derivovaná od komparativu prefixem *po-* (*pomenší, postarší, povětší, ...*), do této skupiny bychom mohli zařadit i tvary *znejmilejší*. Obdobou tvarů (*nej*)*zazší* jsou tvary (*nej*)*hořejší* (technické řešení není jednotné). Značkování a lemmatizace tvarů *nastavší*. \* je v rozporu se značkováním adjektivizovaných přechodníků minulých. Ty mají pravidelně značku *AM.....-*. \*. Uvedený doklad je výjimečný a zdá se, že se jedná o inkonzistenci slovníku automatického morfologického analyzátoru.

nás tvářili co nejvlastenečtěji, smířovali se poměrně nejspíše s	<b>nastavším</b> /nastavši/AAIS7----1A-----	hrůzným převratem . Skutečnost , že národ tak vynikající svou
aspoň desetkrát větší než francouzského . Po krátké pauze ,	<b>nastavši</b> /nastavši/AMFS6-----A-----	dotazem generálovým , požádal jej tehdejší ministerský předseda Tusar ,
, kterými museli zaplatit za ulovené živobytí . A v	<b>nastavším</b> /nastavši/AANS6----1A-----	mlčení zřizavý lišák naplno rozvíjel své názory o tom ,
a že pravý opak byl pravdou atd . Obecnou kočičinou	<b>nastavši</b> /nastavši/AAFS7----1A-----	po roce 1918 však nijak nezmoudřeli , nýbrž celé toto
, ač byla již částečně vysázena . Její publikaci zmařily	<b>nastavši</b> /nastavši/AAFP4----1A-----	politické události . Vysázené torso vyšlo jako " prozatímní ukázka
svou bigotností , vandalismem , nekulturností , nesvobodomyšlností podírájí nohy	<b>nastavšimu</b> /nastavši/AAIS3----1A-----	kulturnímu vývoji , a dokazuje , že není podstatného rozdílu
Bílákem * * * Psát se nezačíná v euforii náhle	<b>nastavši</b> /nastavši/AAIS1----1A-----	To činí jen začátečník zelenáč ucho neznalec snad jen do
, jako na příklad u advokátů ) a možnosti skutečně	<b>nastavši</b> /nastavši/AMFS6-----A-----	po únoru roku 1948 . Snad bude mně prominuto ,
těchto vepřů , tlačících se ke korytům , aby v	<b>nastavši</b> /nastavši/AAFS6----1A-----	nové velké tlačeničky buď uchránili aspoň to , co na
( žlutá ) a časově následující nutností ( zelená	<b>nastavšihó</b> /nastavši/AAIS2----1A-----	účinku . Kromě toho uvádí vzájemné působení ( anebo vzájemnost
truchlivou naši minulost , těšící se , že v červancích	<b>nastavších</b> /nastavši/AAIP6-----1A-----	po světovém zápění válečném zvítězila dějinná spravedlnost , která dala
představa , že mohli se nalézt jednotlivci , kteří v	<b>nastavším</b> /nastavši/AAIS6-----1A-----	úpadku mého životního díla počali pro svou osobu věřití různě
nezaštěká pes . Odkud zaštěká , odtud přijde ženich .	<b>Nastavši</b> /nastavši/AAFS1----1A-----	situace : " Plote , plote třepím tě . .

## Formulace dalších otázek vyplynulých ze zkoumání daného jevu

Na korpusech je možné dále zkoumat například poměr syntetických tvarů komparativu a superlativu a forem opisných ((*nej*)více + *pozitiv*), a to zejména v případě stupňování směrem „dolů“ ((*nej*)méně + *pozitiv*). Co se týče „strategie vyhledávání“ potřebných dokladů, je třeba odfiltrovat případy, kdy vedle sebe stojí tvary (*nej*)více/(*nej*)méně + *pozitiv*, ale dohromady netvoří opisný tvar komparativu/superlativu (viz níže).

Je zcela v pořádku předpoklad, že „obyčejná adjektiva“ (AA. \*) se mohou stupňovat, zatímco jmenné tvary adjektiv (AC. \*), adjektivizované přechodníky (AG. \*, AM. \*) a adjektiva subjektivně posesivní (AU. \*) se stupňovat za žádných okolností nemohou?

## Zadání cvičení, v nichž lze uplatnit analogické postupy

Porovnejte frekvenční distribuci tvarů pozitivu, komparativu a superlativu adverbíí. Vytvořte frekvenční seznam adverbíí, která mají značku *Db. \** (nestupňovatelná), a podívejte se, která adverbia jsou do této skupiny zařazena a zda mezi nimi nelze nalézt taková, která podle vašeho názoru stupňovat lze. Podívejte se, zda vámi předpokládané tvary naleznete v korpusu. Porovnejte kontexty, v nichž se vyskytují tvary *řád, ráda, rádo, ...* označované jako *AC. \** a tvary (*nej*)raději, (*nej*)radši, ... označované jak *Dg. \** a zamyslete se nad možnými důvody pro zvolenou interpretaci, popřípadě se pokuste formulovat argumenty proti zvolenému postupu.

## Vlastnosti adjektiv na -cí

### Motivační úvod

Václav Jan Rosa (1672) uvádí formální omezení stupňování adjektiv a tvrdí, že adjektiva se sufixy -cý, -zý, -sý a -ší nelze (synteticky) stupňovat.



Šmilauer připouští možnost stupňování adjektiv na *-cí* přidáním *-n-* : *vroucí – vroucnější* (srov. Šmilauer 1971 : 127).

*Encyklopedický slovník češtiny* (Karlík – Nekula – Pleskalová 2002 : 447) uvádí, že adjektiva na *-cí* a subjektivě posesivní adjektiva na *-ův/-in* stupňovat nelze, míní se, že nelze tvořit příslušné tvary příslušnými prostředky (sufixy *-í, -ší, -(e/ě)jší* a prefixem *nej-*). Otázkám stupňování dezaktualizovaných adjektiv a tvarům typu *nejvzrušující* se věnuje Alexandr Stich (Stich 1969 : 64). K lemmatizaci a značkování adjektiv na *-cí* srovnej Osolsobě 2009<sup>4</sup>.

Podívejme se, jak lze hledat podklady pro objasnění zmíněného problému v dostupných korpusech češtiny.

### Nastínění problému

Jak bylo naznačeno v předcházející kapitole, rozdělení adjektiv na stupňovatelná a nestupňovatelná nemá zcela ostré hranice. O jejich stanovení se pokusili autoři slovníku a tagsetu analyzátoru použitého pro značkování korpusů ČNK (více Hajič 1994, 2004). Vedla je k tomu především nutnost explicitního popisu. Tagset analyzátoru *ajka* používaný pro značkování korpusů vzniklých na FF MU (část Korpusu soukromé korespondence, více Hladká a kol. 2005) a FI MU (více Osolsobě – Pala – Sedláček 2006) explicitní údaj o možnosti/nemožnosti tvoření tvarů komparativu a superlativu neobsahuje. Přesto je ve slovníku automatického analyzátoru u každého adjektiva informace (vzor) o tvoření/netvoření příslušných tvarů (Osolsobě 2008<sup>1</sup>).

V českých gramatikách se věnuje pozornost pouze stupňování syntetickému se zvláštním zřetelem k některým nepravidelnostem, jako je např. supletivní tvoření tvarů komparativu a superlativu (*dobrý – (nej)lepší*), anomáliím s dopadem na častější pravopisné chyby (*sladký – (nej)sladší*), dubletám (*trpký – (nej)trpč(ejš)i*). V těchto případech bývá informace o tvoření komparativu a superlativu součástí gramatické informace u motivujícího adjektiva v českých výkladových slovnících.

**ssjc** Slovník spisovného jazyka českého

**hladký**

příd. (2. st. hladší)

K tomu, abychom mohli pomocí rozsáhlého jazykového materiálu empiricky zkoumat vlastnosti adjektiv zakončených na *-cí*, která, jak se zdá na základě uvedených citátů, představují skupinu adjektiv s problematickým vztahem k +/- stupňovatelnosti (přesněji tvoření tvarů komparativu a superlativu synteticky), potřebujeme nejdříve shromáždit potřebná data. K tomuto účelu si položíme několik otázek a budeme se snažit hledat na ně odpovědi.

## Otázky

Která adjektiva mohou v češtině končit na *-cí*? Jsou mezi nimi nějaké výraznější skupiny, jejichž identifikace by nám mohla pomoci při řešení problému stupňovatelnosti? Pokud ano, je zachyceno dělení na takové skupiny na úrovni morfologického značkování?

## Formulace dotazu pro získání dat z korpusů

Pro získání odpovědi na první z otázek použijeme jednoduchý postup. Chceme vyhledat všechna adjektiva, která končí na *ci*. Zvolíme tedy **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě `[tag="A.*" & lemma="*.ci"]`. Podíváme se na frekvenční distribuci lemmat.


### Hledat v korpusu

Korpus:

Typ dotazu:






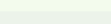
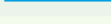













CQL: 
vložit tag vložit "within"

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 6656 (134 str.)

	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> domácí	24 256	
2.	<a href="#">p/ n</a> následující	13 290	
3.	<a href="#">p/ n</a> vedoucí	8 348	
4.	<a href="#">p/ n</a> budoucí	8 225	
5.	<a href="#">p/ n</a> vynikající	5 437	
6.	<a href="#">p/ n</a> rozhodující	4 993	
7.	<a href="#">p/ n</a> rostoucí	4 731	
8.	<a href="#">p/ n</a> odpovídající	4 244	
9.	<a href="#">p/ n</a> stávající	4 180	
10.	<a href="#">p/ n</a> žijící	3 714	
11.	<a href="#">p/ n</a> týkající	3 713	
12.	<a href="#">p/ n</a> cestující	3 233	
13.	<a href="#">p/ n</a> žádoucí	3 095	
14.	<a href="#">p/ n</a> zbývajících	3 031	
15.	<a href="#">p/ n</a> vzdělávací	2 755	
16.	<a href="#">p/ n</a> stojící	2 712	
17.	<a href="#">p/ n</a> související	2 593	
18.	<a href="#">p/ n</a> ležící	2 577	
19.	<a href="#">p/ n</a> psací	2 471	
20.	<a href="#">p/ n</a> existující	2 360	

## Třídění a pozorování dat získaných z korpusů

Z pozorování seznamu adjektiv můžeme vyvodit odpověď na druhou z otázek.





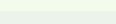
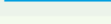













Na třetí otázku odpovíme tak, že v menu **Frekv. distribuce** přidáme druhou úroveň (**lemma**) a třetí úroveň (**tag**). Budeme sledovat, zda a jak odpovídá značkování (především údaj na 2. pozici tagu) vysledovaným skupinám.

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 6656 (134 str.)
















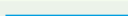




	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n</a> domácí	24 256	
2.	<a href="#">p/ n</a> následující	13 290	
3.	<a href="#">p/ n</a> vedoucí	8 348	
4.	<a href="#">p/ n</a> budoucí	8 225	
5.	<a href="#">p/ n</a> vynikající	5 437	
6.	<a href="#">p/ n</a> rozhodující	4 993	
7.	<a href="#">p/ n</a> rostoucí	4 731	
8.	<a href="#">p/ n</a> odpovídající	4 244	
9.	<a href="#">p/ n</a> stávající	4 180	
10.	<a href="#">p/ n</a> žijící	3 714	
11.	<a href="#">p/ n</a> týkající	3 713	
12.	<a href="#">p/ n</a> cestující	3 233	
13.	<a href="#">p/ n</a> žádoucí	3 095	
14.	<a href="#">p/ n</a> zbývajících	3 031	
15.	<a href="#">p/ n</a> vzdělávací	2 755	
16.	<a href="#">p/ n</a> stojící	2 712	
17.	<a href="#">p/ n</a> související	2 593	
18.	<a href="#">p/ n</a> ležící	2 577	
19.	<a href="#">p/ n</a> psací	2 471	
20.	<a href="#">p/ n</a> existující	2 360	

# Frekvenční distribuce

strana 1

Frekvenční limit:  [Nastavit limit](#)

Celkem: 66289 (1326 str.)

	lemma	tag	Frekvence	
1.	p/ n domácí	AAMP1---1A---	3 968	
2.	p/ n cestující	AGMP2---A---	1 335	
3.	p/ n domácí	AAMS1---1A---	1 277	
4.	p/ n následující	AGIS4---A---	1 227	
5.	p/ n domácí	AAFS2---1A---	1 218	
6.	p/ n domácí	AAMP2---1A---	1 102	
7.	p/ n domácí	AAIS2---1A---	1 096	
8.	p/ n následující	AGIS2---A---	1 082	
9.	p/ n domácí	AAFS6---1A---	1 069	
10.	p/ n cestující	AGMP1---A---	1 004	
11.	p/ n domácí	AAIS1---1A---	994	
12.	p/ n domácí	AAFS1---1A---	930	
13.	p/ n domácí	AANS2---1A---	917	
14.	p/ n domácí	AAFS4---1A---	904	
15.	p/ n domácí	AAIS4---1A---	902	
16.	p/ n obyvatel	AAIS2---1A---	892	
17.	p/ n domácí	AANS6---1A---	888	
18.	p/ n následující	AGIP6---A---	826	
19.	p/ n vedoucí	AGMS1---A---	823	
20.	p/ n domácí	AAIS6---1A---	779	

## Formulace závěrů

Na základě pozorování dat můžeme říci, že adjektiva derivovaná od kmene shodného s kmenem pro tvoření slovesného tvaru přechodníku přítomného (tzv. procesuální adjektiva) jsou označena jako *AG.\** (*cestující, následující, ...*). Všechna ostatní mají značku *AA.\** (*domácí, ...*).

Na základě pozorování dat zjistíme, že kromě adjektiv od „přechodníku přítomného“ jsou zastoupena adjektiva odvozená od sloves vyjadřující účel (účelová), např. *psací, obyvatel, ...* Vně těchto skupin stojí nejfrekventovanější adjektivum *domácí*.

Nyní si položíme otázku, zda na *-cí* mohou v češtině končit pouze dvě velké a na první pohled patrné skupiny deverbativ a adjektivum *domáci*, nebo zda lze najít ještě nějaká další adjektiva (skupiny adjektiv). Jak budeme postupovat? Pokusíme se eliminovat a) procesuální adjektiva („od přechodníků přítomných“) a b) adjektiva účelová.

Procesuální adjektiva lze extrahovat buď pomocí filtru (pozitivní) a morfologické značky, nebo na základě lingvistických znalostí derivačního typu. Porovnejme výsledky obou postupů.

Nastavíme filtr (negativní), **Typ dotazu tag** a do dotazovacího řádku zadáme příkaz *AG.\** a podíváme se na frekvenční distribuci lemmat.

## Filtr konkordancí

**Filtr:**  pozitivní  negativní


**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**





















**CQL:**

# Frekvenční distribuce

strana 1 

Frekvenční limit:

Celkem: 1988 (40 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> domácí	24 256	
2.	<a href="#">p/ n</a> budoucí	8 225	
3.	<a href="#">p/ n</a> stávající	3 834	
4.	<a href="#">p/ n</a> žádoucí	3 095	
5.	<a href="#">p/ n</a> vzdělávací	2 755	
6.	<a href="#">p/ n</a> psací	2 471	
7.	<a href="#">p/ n</a> obývací	2 246	
8.	<a href="#">p/ n</a> parkovací	1 556	
9.	<a href="#">p/ n</a> přijímací	1 487	
10.	<a href="#">p/ n</a> řídicí	1 429	
11.	<a href="#">p/ n</a> hrací	1 382	
12.	<a href="#">p/ n</a> dýchací	1 311	
13.	<a href="#">p/ n</a> zvířecí	1 308	
14.	<a href="#">p/ n</a> čerpací	1 120	
15.	<a href="#">p/ n</a> vyšetřovací	1 090	
16.	<a href="#">p/ n</a> kuřecí	1 090	
17.	<a href="#">p/ n</a> ovládací	957	
18.	<a href="#">p/ n</a> rozhodovací	951	
19.	<a href="#">p/ n</a> poznávací	945	
20.	<a href="#">p/ n</a> smrtící	939	

Druhá možnost eliminace procesuálních adjektiv vychází ze znalosti tvoření přechodníků a derivovaných adjektiv v češtině (viz Příloha). Procesuální adjektiva se v češtině tvoří pravidelně od tvaru přechodníku přítomného pro feminina a neutra singuláru. Z řečeného plyne, že před *ci* může předcházet pouze *i* nebo *ou*.

Nastavíme filtr (negativní), **Typ dotazu lemma** a do dotazovacího řádku zadáme příkaz *.\*[í|ou]cí* a podíváme se na frekvenční distribuci lemmat.

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**


**Dotaz:**

Filtrovat konkordanci

Vymazat formulář













# Frekvenční distribuce

strana 1 

Frekvenční limit:

Celkem: 1803 (37 str.)

	lemma	Frekvence	
1.	<u>p/ n</u> domácí	24 256	
2.	<u>p/ n</u> vzdělávací	2 755	
3.	<u>p/ n</u> psací	2 471	
4.	<u>p/ n</u> obývací	2 246	
5.	<u>p/ n</u> parkovací	1 556	
6.	<u>p/ n</u> přijímací	1 487	
7.	<u>p/ n</u> řídicí	1 429	
8.	<u>p/ n</u> hrací	1 382	
9.	<u>p/ n</u> dýchací	1 311	
10.	<u>p/ n</u> zvířecí	1 308	
11.	<u>p/ n</u> čerpací	1 120	
12.	<u>p/ n</u> vyšetřovací	1 090	
13.	<u>p/ n</u> kuřecí	1 090	
14.	<u>p/ n</u> ovládací	957	
15.	<u>p/ n</u> rozhodovací	951	
16.	<u>p/ n</u> poznávací	945	
17.	<u>p/ n</u> měřicí	903	
18.	<u>p/ n</u> odvolací	856	
19.	<u>p/ n</u> spojovací	801	
20.	<u>p/ n</u> sdělovací	800	

Porovnáme-li oba seznamy, pak nás zaujme a) skupina adjektiv jako *budoucí, stávající, žádoucí, smrtící, okouzlující, ...* a b) skupina adjektiv jako *zvířecí, kuřecí*.

První skupinu tvoří adjektiva tvořená od sloves, která jsou po formální stránce shodná s adjektivy od přechodníku přítomného, takže je třeba pátrat po tom, proč nemají příslušnou značku, tedy proč mají značku *AA.\**, nikoli *AG.\**.

Druhou skupinu tvoří adjektiva odvozená od substantiv. Označují buď přivlastnění druhu, nebo široký vztah. Jsou to adjektiva odvozená od substantiv skloňovaných podle vzoru *kuře (zvířecí, kuřecí, telecí, morčecí)* (více Osolsobě 2009<sup>3</sup>).

Při vyhledávání kandidátů adjektiv, která patří do této skupiny, budeme postupovat následovně. Nastavíme filtr (pozitivní), zvolíme **Typ dotazu lemma**, zadáme příkaz `.*[eě]cí` a podíváme se na frekvenční distribuci lemmat.

## Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**

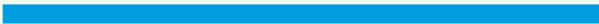
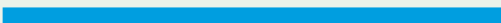



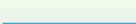
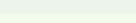













**Lemma:**

# Frekvenční distribuce

strana 1 

Frekvenční limit:

**Celkem:** 128 (3 str.)

	<b>lemma</b>	<b>Frekvence</b>	
1.	<u>p/ n</u> zvířecí	1 308	
2.	<u>p/ n</u> kuřecí	1 090	
3.	<u>p/ n</u> napájecí	604	
4.	<u>p/ n</u> obráběcí	433	
5.	<u>p/ n</u> prováděcí	432	
6.	<u>p/ n</u> knížecí	316	
7.	<u>p/ n</u> telecí	314	
8.	<u>p/ n</u> spouštěcí	237	
9.	<u>p/ n</u> čtecí	191	
10.	<u>p/ n</u> třecí	191	
11.	<u>p/ n</u> promlčecí	181	
12.	<u>p/ n</u> markraběcí	178	
13.	<u>p/ n</u> hraběcí	160	
14.	<u>p/ n</u> vyprávěcí	136	
15.	<u>p/ n</u> natáčecí	117	
16.	<u>p/ n</u> předváděcí	108	
17.	<u>p/ n</u> sklápěcí	100	
18.	<u>p/ n</u> otáčecí	100	
19.	<u>p/ n</u> tvářecí	77	
20.	<u>p/ n</u> zaváděcí	71	

Ručně můžeme oddělit desubstantiva (*zvířecí, kuřecí, knížecí, telecí, markraběcí, hraběcí, batolecí, ...*) a deverbativa účelová (*napájecí, obráběcí, prováděcí, spouštěcí, třecí, ...*).

Podívejme se nyní na skupinu adjektiv, která jsou po formální stránce shodná s deverbativními procesuálními adjektivy.





















Vrátíme se dva kroky zpět. Nastavíme filtr (pozitivní) **Typ dotazu tag** a do dotazovacího řádku zadáme příkaz *AA.\** a podíváme se na frekvenční distribuci lemmat (postup viz výše).

# Frekvenční distribuce

strana 1 

Frekvenční limit:

Celkem: 185 (4 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> budoucí	8 225	
2.	<a href="#">p/ n</a> stávající	3 834	
3.	<a href="#">p/ n</a> žádoucí	3 095	
4.	<a href="#">p/ n</a> smrtící	939	
5.	<a href="#">p/ n</a> okouzlující	874	
6.	<a href="#">p/ n</a> plovoucí	836	
7.	<a href="#">p/ n</a> živoucí	567	
8.	<a href="#">p/ n</a> vroucí	462	
9.	<a href="#">p/ n</a> tisící	420	
10.	<a href="#">p/ n</a> kolemjdoucí	356	
11.	<a href="#">p/ n</a> všemohoucí	354	
12.	<a href="#">p/ n</a> dlouhotrvající	335	
13.	<a href="#">p/ n</a> neutuchající	307	
14.	<a href="#">p/ n</a> plačící	292	
15.	<a href="#">p/ n</a> visící	291	
16.	<a href="#">p/ n</a> vševědoucí	277	
17.	<a href="#">p/ n</a> zničující	270	
18.	<a href="#">p/ n</a> protijedoucí	262	
19.	<a href="#">p/ n</a> mizející	236	
20.	<a href="#">p/ n</a> horoucí	187	

Při procházení seznamu (185 lemmat je kratší seznam, který můžeme projít celý) si všimněme, že kromě deverbativ se zde vyskytují např. řadová číslovka *tisící*, dále kompozita *růžovo-/červeno-lící* a adjektiva jako *letoucí*, *divoucí*, která nejsou deverbativy. Celkově vzato ale deverbativa převažují. Vidíme, že důvodem, proč nejsou označována jako deriváty přechodníků, může být např. to, že řada z nich jsou kompozita, k nimž žádné složené sloveso, a tedy ani přechodník neexistuje (*kolemjdoucí*, *všemohoucí*, *dlouhotrvající*, *všeobjímající*, ...) <sup>23</sup>. Dále vidíme adjektiva jako *žádoucí* nebo *živoucí*, která sice jsou příbuzná se slovesy

<sup>23</sup> V praxi se strojové slovníky pro automatickou morfologickou analýzu buďtovaly tak, že pro některé paradigmatické (pravidelné) derivace existovala pravidla (morfologické vzory), jejichž pomocí se tyto

*žádat, žít*, nicméně nejsou synchronně pravidelně tvořená od přechodníků přítomných těchto sloves, neboť pak by musela znít *žadající, žijící*<sup>24</sup>.

Rýsují se tedy tři další skupiny: a) procesuální adjektiva, která nemají značku *AG.\** včetně kompozit, b) adjektiva deverbální, která nejsou synchronně pravidelně tvořená od přechodníků přítomných (viz Dokulil 1986 : 322, 330), c) desubstantivní adjektiva od podstatných jmen, jejichž kmen náhodně končí na skupinu hlásek *íc* (*tisíc, líce*), d) adjektiva od substantiv tvořená příponou *-oucí* (*divoucí, letoucí, tmoucí ...*).

Nyní se vrátíme ke druhé skupině, a to k adjektivům, která nejsou zakončena formálně shodně s adjektivizovanými přechodníky přítomnými. Jde o adjektiva na *-cí*, která dle našich pozorování představují skupinu potenciálních adjektiv účelových, tedy ke konkordanci, jejímž výsledkem je 1803 lemmat adjektiv (viz výše).

Podívejme se na to, zda lze využít gramatických formálních vlastností adjektiv účelových k tomu, abychom tuto skupinu nějak blíže poznali, aniž by bylo nutné ručně projít seznam o 1803 řádcích. Adjektiva účelová se v češtině tvoří od kmene minulého, konkrétně od tvaru *l*-ového přičestí. Před *ci* mohou předcházet pouze některé vokály. Konkrétně a) vokály, které mohou být českými kmenovými vokály (KmV) pro tvoření tvarů *l*-ového přičestí, a b) vokály, které mohou být kořenovými vokály tvarů *l*-ových přičestí sloves patřících do III. třídy slovesné ke vzoru *krýt* (viz Příloha). Bohužel jak adjektiva derivovaná od vzoru *kuře*, tak adjektiva, jejichž základovým slovem je sloveso I. třídy vzoru *umřít* (např. *třecí*) a IV. třídy vzorů *trpět a sázet* (např. *napájecí, obráběcí, ...*) a adjektivum *čtecí*, mají před *ci* [eě] (viz výše). Konsonanty před *ci* předcházet nemohou, protože tvoří-li se účelové adjektivum od slovesa, které má v *l*-ovém přičestí uzavřený kmen (I. třída slovesná vzory *nést, péci* a II. třída slovesná vzor *tisknout*), pak se mezi uzavřený kořen a derivační sufix vkládá vokalický konekt, jako je tomu např. u adjektiv jako *čtecí, hnětací, pečicí, tiskací* (více Osolsobě 2010, 2011<sup>1</sup>).

Platí tedy pravidlo, že má-li jít o účelové adjektivum, pak podmínkou nutnou, nikoli postačující, je, že před *ci* předchází [aeěiy]. V češtině sice existují *l*-ová přičestí, u nichž před *l* předchází *á*, toto *á* se ovšem při derivaci adjektiv účelových buď krátí (*hrál – hrací*), nebo alternuje (*stát – stojací*). V češtině ovšem existují *l*-ová přičestí, u nichž před *l* předchází *u*. Od sloves II. třídy se účelová adjektiva na *-nucí*, jak se zdá, netvoří (srov. *tiskací*). Neexistuje

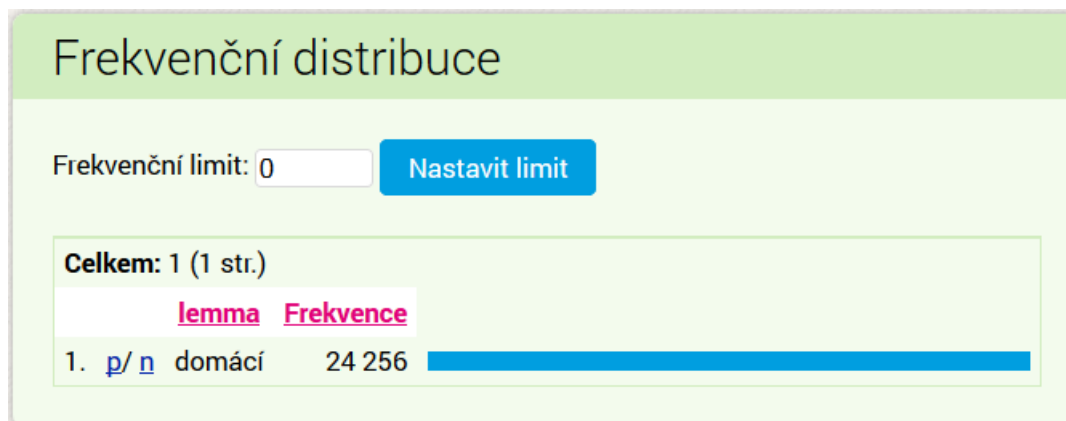
---

pravidelné deriváty tvořily a značkovaly automaticky přímo od množiny základů (např. od sloves nedokonavých se automaticky tvořily nejen tvary přechodníku přítomného, ale i tvary příslušného procesuálního adjektiva). Naopak kompozita s druhým členem procesuálním adjektivem byla do slovníku doplněna samostatně. Je také možné uvažovat o tom, jaká je slovnědruhovává platnost těchto formálních adjektiv, tedy kdy dle kontextu fungují skutečně jako adjektiva a kdy jde o substantivizaci.

<sup>24</sup> K těmto adjektivům srov. podrobněji Osolsobě 2009<sup>4</sup>.

žádný doklad účelového adjektiva od slovesa III. třídy vzoru *kryt* s KoV *u* v l-ovém přičestí (\**kucí*<*kout*).

Podívejme se nyní, zda nám tyto úvahy nějak pomohou při analýze dat. Nastavíme filtr (negativní), **Typ dotazu lemma** a do dotazovacího řádku zadáme příkaz *.\*[aeěiy]cí* a podíváme se na frekvenční distribuci lemat. Jediným výsledkem je adjektivum *domáci*.



Nezbývá tedy nic jiného než projít ručně seznam adjektiv na *.\*[aeěiy]cí* a určit, zda kromě a) deverbativ účelových a b) desubstantiv od vzoru *kuře* najdeme ještě další případy, které nelze zahrnout pod a) nebo b).

Závěrem tedy shrňme, že většina adjektiv na *-cí* jsou deverbativa (procesuální a účelová adjektiva), tedy adjektiva relační, která se dle obecné charakteristiky pravidel opřených o sémantiku těchto adjektiv nestupňují. Druhou výraznou skupinu tvoří desubstantiva, tedy rovněž relační adjektiva. Mimo stojí frekventované deadverbiální (relační) adjektivum *domáci*.

### Formulace dalších otázek vplynuvších ze zkoumání daného jevu

Ze zkušeností s pravidly pro +/- stupňovatelnost adjektiv opřenými o sémantiku víme, že mnohde bývají tato pravidla porušována. Děje se tak v případech, kdy dochází k sémantickému tvoření (přenos významu většinou metaforický), adjektivum nabývá významu spojeného s hodnocením a u tohoto významu vznikne potřeba vyjádřit vyšší míru vlastnosti (srv. např. doklady z internetu <*hrošejší*> *kůži*, <*psejší*> *počasí*, <*dívčejší*> *oči*, ale též doklad z korpusu SYN2006PUB *Basilio je <víc Beaumarchaisův> než Rossiniho ...*).

Můžeme si položit otázku, zda k naznačenému jevu nemůže docházet i v případě námi sledované skupiny adjektiv na *-cí*. Šmilauer a Stich upozorňují na adjektiva na *-cnějši*.

### Zadání cvičení, v nichž lze uplatnit analogické postupy

Vyhledejte v korpusu SYN adjektiva s tvary *\*cnějš.\**. Sledujte je, podívejte se, jak jsou lemmatizována a značkována, a pokuste se vyvodit závěry o platnosti toho, co uvádí Šmilauer a co se píše na citovaném místě v *ESC*.

## Číslovky

Číslovky jsou jako slovní druh vymezeny primárně na základě sémantiky (slova označující určitý/neurčitý počet nebo určité/neurčité množství). Význam množství (vztahu k množství) se vyjadřuje gramaticky prostřednictvím gramatické kategorie čísla. Číslovky (základní) jej vyjadřují lexikálně. Množství ale mohou vyjadřovat i slova, která se na základě jiných vlastností řadí k odlišným slovním druhům. Naopak některé číslovky se používají v jiných významech, než je označení počtu. Tento jev spadá pod pojem tzv. slovnědruhových přesahů a přechodů (případů, kdy jedno slovo může patřit k více slovním druhům). Zaměříme se na to, jak jsou některé takové případy zachyceny na úrovni morfologického značkování. Zatímco v případě slovnědruhových přesahů je značkování otázkou zvoleného popisu na úrovni slovníku automatického morfologického analyzátoru, v případě slovnědruhového přechodu zaznamenaného na úrovni slovníku vícerou interpretací (lemmatu a pos) by ideálně měla existovat desambiguační pravidla, podle nichž lze interpretovat jedno slovo/tvar na úrovni slovního druhu (a tedy i lemmatu). Bude nás především zajímat, zda je možné (na základě pozorování korpusových dat) alespoň v některých případech taková pravidla formulovat.

### *Slovnědruhové přesahy*

#### **Motivační úvod<sup>25</sup>**

Číslovky jsou ohebný slovní druh (skloňují se), patří k autosémantikům (vyjadřují počet/množství lexikálně). Zastávají funkce adjektiv i substantiv (srov. Veselý 2011).

Druhy číslovek v českých mluvnicích se do jisté míry kryjí s klasifikací v tagsetech používaných pro automatickou morfologickou analýzu (viz výše).

Subklasifikace slovního druhu numeralií v tradičních mluvnicích používá kritérium určitosti vyjádření množství. Dalším kritériem je sémantika číslovek. Formálně jde u číslovek vyjadřujících množství, které lze převést na číselnou hodnotu, o slovotvorbu (Osolsobě 1995). Číslovky označované jako dílové skupinové, velikostní a násobné jsou formálními i syntaktickými substantivy nebo adjektivy. Formálními adjektivy jsou i číslovky řadové a druhové.

#### **Nastínění problému**

Číslovky *jeden* a *druhý* plní často funkci zájmenou. Tento aspekt není zahrnut v praxi dosavadního značkování. Nicméně na některá rozlišení funkcí slovních tvarů obou lemmat

---

<sup>25</sup> K probíranému problému srov. též Panevová, Ševčíková 2011.



automatická morfoloická analýza a na ní založené značkování i desambiguace berou zřetel. Podívejme se na to, jak lze v korpusových datech vyhledávat některé relevantní významy užití sledovaných tvarů.

## Otázky

*Slovník spisovné češtiny (SSČ)* heslové slovo *druhý* interpretuje jako řadovou číslovku, v některých kontextech v platnosti přídavného jména, a heslové slovo *jeden* jako číslovku (viz níže).

Ponecháme stranou diskusi o tom, zda by nebylo vhodnější hovořit v případě adjektivní interpretace též o zájmenné funkci adjektiva *druhý* s významem *jiný* (více *MČ 1* a vymezení identifikátorů a alterátorů v rámci zájmen).

SSČ Slovník spisovné češtiny

### jeden

čís. zákl.

1. *označ. počet 1* má jedno dítě (ne víc), ztratil jednu rukavici (z páru), jedna jediná věc mi vadí, dva muži a jedna žena, nepřišel ani jeden divák, *nikdo*; jeden ze sta, jeden z nejlepších, dům o jednom poschodí, dokončil práci za (pouhý) jeden den, dvacet jeden litr, je jedna hodina, o jedné hodině, v *někt. spoj. označ. přibližný malý počet* vrátí se za jeden dva dny.
2. (v platn. před.) *vyznačující se celistvostí tvořící celek* socha z jednoho kusu dřeva, jeden (společný) národ.
3. *společný (pro více jedinců)* jíst z jedné mísy, oba bydlí v jednom domě, v *stejném*, v *témž*; narodil se v jednom roce, jedna a táž osoba, věc.
4. *označ. mnohost, úplnost n. nepřetržitost (toho, co je označeno podst. jménem) samý (symo), 4 jedný (symo), 2* strom je jeden květ, dům byl v jednom plameni, *celý hofel*; jedním tahem (nakreslit).
5. ve dvojici jeden - druhý, jeden - jeden, jeden - jiný, s *význ. ukaz. vyj. vzájemnost n. střídání*: vzpomínat jeden na druhého, jedna krásnější než druhá, jedni souhlasili, jiní ne, přicházeli jeden za druhým, *postupně; rychle za sebou*; volali jeden přes druhého, (dáme-li, uvážíme-li, vezmeme-li) jedno k druhému.
6. s *význ. neurč. nějaký (symo), 1 některý (symo)* přišel do jednoho města, jednoho dne, byl jednou jeden král.
7. ve zvoláních s *citovým zabarvením* ty jeden!, darebáku, kluku jeden (jedna)!, *první, jedenkrát, jednou % (viz)* být (stále) jednou nohou v hrobě, v kriminále ap., v *nebezpečí smrti, zatčení ap.*; být s někým jedna ruka, v *důvěrné shodě*; házet (všechno, všechny) do jednoho pytle, *nerozlišovat mezi rúz. jevy, osobami*; (oběd, dort) jedna báseň, *expr. výborný*; (udělat něco, být hotov) jedna dvě, *brzo, rychle*; (prospívat) jedna radost, *expr. k plné spokojenosti*; jedním dechem (přečíst, povědět), *bez přerušení*; jedním douškem (vypít), *naráz*; jedním slovem (řečeno), *stručně*; jsou jedné myslí, *mají tytéž názory*; jsou (jako) jedno tělo, jedna krev, *blíže příbuzní; naprosto jednotní*; poslouchat (jen) jedním uchem, *nepozorně, roztržitě*; přijímat pod jedinou (způsobou), i *pod jednou; (viz)* táhnout za jeden provaz, *shodně, svorně postupovat*; to je jedno, mně (vám, někomu) je to jedno, *lhostejně, na tom nezáleží*; v jednom kuse (spát, zlobit), *hovor. ustavičně*; vyprat něco z jedné vody načisto, *vyřádit rychle a povrchně*; vystoupit jako jeden muž, (*o kolektivu*) *jednotně*; zabít dvě mouchy jednou ranou, *vyřádit dvě věci najednou*; žít jako jedna rodina, *svorně, ve shodě*

SSČ Slovník spisovné češtiny

### druhý

čís. řad. (ve význ. 2-6 v platn. před.)

1. *k dvě* druhý v řadě, v cíli, druhý díl slovníku, druhého (2.) června, dnes je druhého, kolem druhé (hodiny), je půl druhé, Václav II. (Druhý), za první, za druhé, (*při vyčtu*); po první, po druhé (i poprvé, podruhé), půl druhého (i půldruhého), půl druha (i půldruha) roku, druhý pád, *genitiv*; druhý stupeň, *komparativ*; druhá (mluvnická) osoba, druhý hlas, druhé housle, druhá mocnina, s *mocnitelem* 32; druhý rychlostní stupeň, druhá kosmická rychlost.
2. *jsoucí nižšího stupně* druhý místopředseda, druhá jakost.
3. *následující (symo) příští (symo), 1 další (symo), 1* přišel až druhý den.
4. *opačný (symo), 1 protiější (symo)* přejít na druhý chodník, druhá strana, druhý břeh, druhé potlaví, začít z druhého konce.
5. *ještě jeden jiný (symo), 1* druhého (takového) nenajdeš, ať to dokáže někdo druhý.
6. *mn. ostatní (symo), 1* dostavil se sám, druzí nepřišli.
7. *existující vedle základního* druhý domov, druhé brejle, *rezervní*
8. ve dvojici jeden -- druhý *jeden 5 % (viz)* (zprávy) z druhé ruky, *nepřímé, a tím zprav. zkrácené*

Podívejme se, zda a jak těmto slovnědruhovým interpretacím odpovídá praxe značkování v korpusu SYN2010.

## Formulace dotazu pro získání dat z korpusů

Zvolíme tedy **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz *jeden*. Následně se podíváme na frekvenční distribuci slovních druhů (**pos**).

- No , děti . Jedna věc , Henriku ,	<b>jedna</b> /jeden/CIFS1-----	věc je jistá : na děti jsi vždycky byl hodný
přístroj před poškozením při přenášení . * Nespoléhejte se na	<b>jeden</b> /jeden/CIFS4-----	nabitý akumulátor . Pořídte si druhý a noste jej vždy
. Už jsme se neskrývali v cizí zemi , nejedli	<b>jedno</b> /jeden/CINS4-----	žádná jídla , neviseli na sobě kvůli pocitu bezpečí ,
. Toho bylo dosaženo jen díky mnoha důmyslným vynálezům .	<b>Jedním</b> /jeden/CIFS7-----	z důvodů , proč byla v roce 1984 udělena Nobelova
se dá jasně a dokonale osvětlit z mechanických důvodů vznik	<b>jedné</b> /jeden/CIFS2-----	jediné rostliny nebo housenky . " Přeloženo do našeho jazyka
v tisku objevují inzeráty na " Komorníka , který je	<b>jedním</b> /jeden/CIFS7-----	členem rodiny , ale vydá za osm sluhů " .
jsi mrtvá , co ? " Překřídil jí nohy .	<b>Jedna</b> /jeden/CIFS1-----	ruka se jí sesula z klína na přehoz . Uviděl
příhrál peníze obci , ve které bydlí Bouzov má jako	<b>jedna</b> /jeden/CIFS1-----	z mála obcí jistotu , že si letos ukousne z
vyloučeno . " Kromě toho nabízí zapůjčení malého balonu pro	<b>jednu</b> /jeden/CIFS4-----	osobu za 50 franků za přistání a většího o objemu
dnech mají hrozný hlad . Potom je oba přivede do	<b>jedné</b> /jeden/CIFS2-----	místnosti , ve které je jediný krajíček chleba . "
Je to zvláštní česká melanž tohoto trudnomyslného podzimu - na	<b>jedné</b> /jeden/CIFS6-----	straně blahobyt vrcholící konjunktury , jakou už moje generace mezi
ale o dobré hudbě a příjemné pohodě , " tvrdí	<b>jeden</b> /jeden/CIFS1-----	z pořadatelů festivalu Michal Thomes . Areál řevnického divadla bude
. Hronova rodina bývá často sama . Vladimír je v	<b>jednom</b> /jeden/CINS6-----	kole . Tříkrát týdně trénuje v posilovně se svým osobním
rány zakrýval obvazy napuštěnými tímtež roztokem . Úmrtnost poklesla na	<b>jednu</b> /jeden/CIFS4-----	třetinu původního stavu a s tím , jak propracovával svou
vrstva půdy . Když farmáře postihne sucho , je mu	<b>jedno</b> /jeden/CINS1-----	, že to množství vody , které se ztrácí ve

## Víceúrovňová frekvenční distribuce

Frekvenční limit:

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>?</sup>	
1.	lemma	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	-
2.	pos	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	X

## Frekvenční distribuce

Frekvenční limit:

Celkem: 1 (1 str.)

**lemma** **pos** **Frekvence**

1. **p/n** **jeden** **C** 210 543

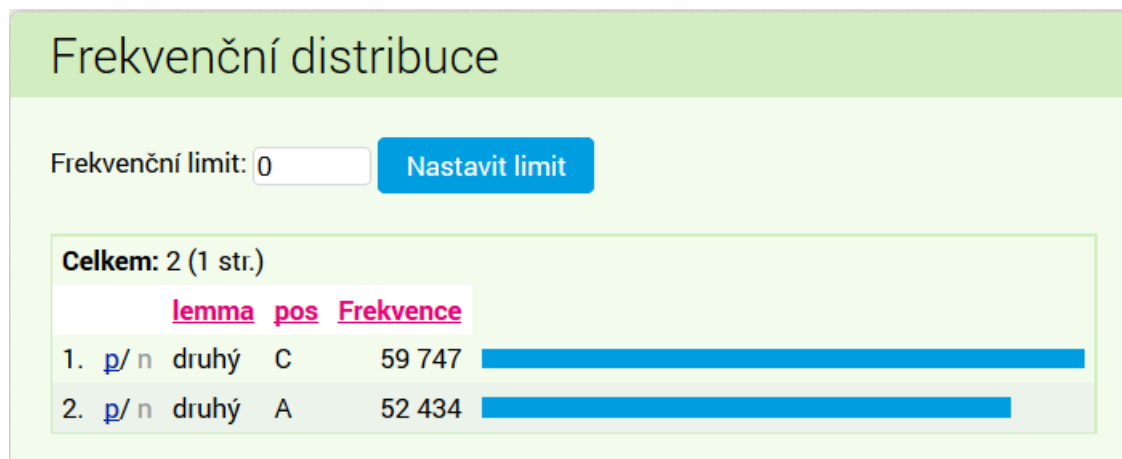
Procházíme-li slovnědruhovými interpretacemi lemmat na jednotlivých konkordančních řádcích, vidíme, že druhá pozice ve značce je vyplněna většinou jako *l* (základní číslovky *jeden, dva, oba, tři, čtyři*), méně jako *h* (druhá číslovka *jedny*).

Nastavíme filtr (pozitivní), zvolíme **Typ dotazu cql** a do dotazovacího řádku zadáme příkaz `[tag="Cl.P.*"]`. Výsledkem je prázdný seznam. Vrátime-li se o krok zpět a do dotazovacího řádku zadáme příkaz ve formě `[tag="Ch.S.*"]`, obdržíme opět prázdný konkordanční seznam. Z tohoto vyplývá, že tvary plurálu jsou na úrovni použitého morfologického

značkování chápány vždy jako číslovky druhové, zatímco tvary singuláru vždy jako tvary číslovky základní.

- No, děti . Jedna věc , Henriku ,	<b>jedna</b> /jeden/CIFS1-----	věc je jistá : na děti jsi vždycky byl hodný
přístroj před poškozením při přenášení . * Nespolehejte se na	<b>jeden</b> /jeden/CIFS4-----	nabitý akumulátor . Pořídte si druhý a noste jej vždy
. Už jsme se neskrývali v cizí zemi , nejedli	<b>jedno</b> /jeden/CINS4-----	žádná jídla , neviseli na sobě kvůli pocitu bezpečí ,
. Toho bylo dosaženo jen díky mnoha důmyslným vynálezům .	<b>Jedním</b> /jeden/CIFS7-----	z důvodů , proč byla v roce 1984 udělena Nobelova
se dá jasně a dokonale osvětlit z mechanických důvodů vznik	<b>jedné</b> /jeden/CIFS2-----	jediné rostliny nebo housenky . " Přeloženo do našeho jazyka
v tisku objevují inzeráty na " Komorníka , který je	<b>jedním</b> /jeden/CIFS7-----	členem rodiny , ale vydá za osm sluhů " .
jsi mrtvá , co ? " Překřídil jí nohy .	<b>Jedna</b> /jeden/CIFS1-----	ruka se jí sesula z klína na přehoz . Uviděl
přihrál peníze obci , ve které bydlí Bouzov má jako	<b>jedna</b> /jeden/CIFS1-----	z mála obcí jistotu , že si letos ukousne z
vyloučeno . " Kromě toho nabízí zapůjčení malého balonu pro	<b>jednu</b> /jeden/CIFS4-----	osobu za 50 franků za přistání a většího o objemu
dnech mají hrozný hlad . Potom je oba přivede do	<b>jedné</b> /jeden/CIFS2-----	místnosti , ve které je jediný krajíček chleba . "
Je to zvláštní česká melanž tohoto trudnomyslného podzimu - na	<b>jedné</b> /jeden/CIFS6-----	straně blahobyt vrcholící konjunktury , jakou už moje generace mezi
ale o dobré hudbě a příjemné pohodě , " tvrdí	<b>jeden</b> /jeden/CIMS1-----	z pořadatelů festivalu Michal Thomes . Areal řevnického divadla bude
kdysi jeho dceři prodala troje šponovky , ačkoliv potřebovala jenom	<b>jedny</b> /jeden/ChFP4-----	. Na jednu stranu mu tvoje finanční motivace imponovala ,
nezdřítla neudrživá stařecká moudrost ! " Čekal na odpověď .	<b>Jedni</b> /jeden/ChMP1-----	řekli : " My bychom rádi šli . Ale jsme
posílil o 16 procent . Indie se tak zařadila mezi	<b>jedny</b> /jeden/ChIP4-----	z nejrychleji rostoucích trhů na světě . Není tedy divu
A čím byl ? " " Kovářem . "	<b>Jedni</b> /jeden/ChMP1-----	jsou bohatí a druzí chudí . Je snad to ,
, který svými nálezy rozhodnutí zmrazené společenství řídil . Podle	<b>jedněch</b> /jeden/ChFP2-----	překročil hrubě své pravomoci , podle druhých se choval přesně
světě a s holi teplého října ve slepeckých listech ,	<b>jedni</b> /jeden/ChMP1-----	se na lavičce libají a druzí sedí opuštěni . Jedni
disciplině reining . Jezdci Ranče Kostelany jsou již od počátku	<b>jedni</b> /jeden/ChMP1-----	z nejuspěšnějších v republice a o tom svědčí to ,
, už vím . Pan Battle mi ukázal článek v	<b>jedněch</b> /jeden/ChFP6-----	novinách , už nevím v kterých , a tam Hicksona
hladké čedičové hrázi , na kterou se matně pamatuje z	<b>jedněch</b> /jeden/ChFP2-----	prázdnin v dětství . Tu vzpomínku v něm musel vyvolat
k akci obdobné bublinatkám . Nápadný je dimorfismus pastí .	<b>Jedny</b> /jeden/ChIP1-----	měchyřky jsou dlouze stopkaté , ponořené až 3 cm hluboko
, zda by eutanazie měla být v ČR povolena .	<b>Jedni</b> /jeden/ChMP1-----	jsou pro , druzí proti . Zdeněk Dienstbier , předseda
nejvýznamnějším českým spisovatelem S českou literaturou to není jednoduché .	<b>Jedni</b> /jeden/ChMP1-----	volají , že je v krizi , druzí píší ze
čištění grilu vůbec . To vše dělá z lávových grilů	<b>jedny</b> /jeden/ChIP4-----	z nejžádanějších pro zahradní grilování . Další výhodou plynových

Nyní se podíváme na lemma *druhý*. Zvolíme **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz *druhý*. Následně se podíváme na frekvenční distribuci slovních druhů (**pos**).



### Třídění a pozorování dat získaných z korpusů

Pokud bychom tedy chtěli např. zkoumat funkce, kdy tvary lemmatu *jeden* fungují jako zájmena, pak bychom museli pracovat bez použití morfologických značek. Nicméně bychom mohli využít znalostí gramatiky k třídění bohatého jazykového materiálu, který nabízejí korpusy.

Jednou z funkcí druhových číslovek v češtině je zastupovat číslovky základní, je-li počítaným předmětem plurale tantum, ale i tehdy, je-li jím např. substantivum označující párové orgány. Otázka tedy zní: jak vyhledat v korpusu konkordance, kdy počítaný předmět vztahující se ke tvarům lemmatu *jeden* je plurale tantum? Můžeme postupovat následovně. V pravém kontextu na pozici za tvary lemmatu *jeden* označovanými jako *Ch.\** vyhledáme substantiva (počítaný předmět) a zjistíme, zda se jedná o pluralia tantum.

Zvolíme **Typ dotazu cql** a do dotazovacího řádku zadáme příkaz *[lemma="jeden" & tag="Ch.\*"]*. Zvolíme filtr (pozitivní), rozsah hledání <1,1> a do dotazovacího řádku zadáme dotaz *NN.\**. Podíváme se na frekvenční distribuci **lemmat** na první pozici vpravo od **KWIC**.


automaticky rozžihalo , za mnou zhasínalo . Hned tím ušetřili	<b>jedny</b> /jeden/ChFP4-----	<b>pacičky</b> olivrejované . Abych byl spravedlivý : když jsme v
chlapcem , slečno Fellowesová . Do Stasis jedna vedou jen	<b>jedny</b> /jeden/ChFP4-----	<b>dveře</b> a ty jsou dobře uzamčené a hlídané . Chtěl
velký , u dvora prý pobýval jen zřídka . Podle	<b>jedněch</b> /jeden/ChFP2-----	<b>milenc</b> carevny a šedá eminence vládnoucí celé zemi , podle
jsem mu dobře , kráček s kufříkem napřed a u	<b>jedněch</b> /jeden/ChFP2-----	<b>dveří</b> zastavil . Objevila se malá krásná žena s rozdělenými
zdi , v níž byla jen tři malá okénka a	<b>jedny</b> /jeden/ChFP4-----	<b>dveře</b> . Nalevo byly natažené prádelní šňůry . Nevyužité .
v krosně pro všechny případy . Navlékám na sebe ještě	<b>jedny</b> /jeden/ChIP4-----	<b>šustáky</b> , pod bundu přidávám tričko s dlouhými rukávy a
prej vytrh trn z paty , dostanu jedny prváky a	<b>jedny</b> /jeden/ChMP4-----	<b>čtvrtáky</b> a začít můžu hned odpoledne . Když mi zazvonilo
Keogh . Kroky ji následovaly , doháněly ji . Ani	<b>jedny</b> /jeden/ChFP1-----	<b>dveře</b> v celé vetché budově se neotevřely . Železné postele
vozu . Flétnové kvartety jsou napsány tak , že místo	<b>jedněch</b> /jeden/ChFP2-----	<b>houslí</b> zpívá flétna . Ale prim vždycky mít nemusí .
marocké televize . Na filmovou Casablanku nepanují jednotné názory .	<b>Jedni</b> /jeden/ChMP1-----	<b>film</b> zbožňují , druzí mu rádi nasazují nálepku kyče .
podoby a stal se i teoretickou* základnou k násilnému ovládní	<b>jedněch</b> /jeden/ChIP2-----	<b>národů</b> druhými - jednou z nejnebezpečnějších buržoazních ideologií 19 .

### Víceúrovňová frekvenční distribuce

Frekvenční limit:

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>2</sup>	
1.	lemma ▾	<input type="checkbox"/>	1P ▾	slova KWIC nejvíce vlevo ▾	-
<input type="button" value="Přidat úroveň"/>					
<input type="button" value="Vytvořit frekvenční seznam"/>					

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 259 (6 str.)

	<b>lemma</b>	<b>Frekvence</b>	
1.	<a href="#">p/ n dveře</a>	167	
2.	<a href="#">p/ n noviny</a>	40	
3.	<a href="#">p/ n šaty</a>	28	
4.	<a href="#">p/ n bota</a>	17	
5.	<a href="#">p/ n člověk</a>	15	
6.	<a href="#">p/ n ústa</a>	12	
7.	<a href="#">p/ n ruka</a>	12	
8.	<a href="#">p/ n Vánoce</a>	12	
9.	<a href="#">p/ n hra</a>	9	
10.	<a href="#">p/ n volba</a>	8	
11.	<a href="#">p/ n manžel</a>	8	
12.	<a href="#">p/ n peníze</a>	7	
13.	<a href="#">p/ n kamna</a>	7	
14.	<a href="#">p/ n kalhoty</a>	7	
15.	<a href="#">p/ n rodič</a>	6	
16.	<a href="#">p/ n národ</a>	6	
17.	<a href="#">p/ n hodinky</a>	6	
18.	<a href="#">p/ n znamení</a>	5	
19.	<a href="#">p/ n vrata</a>	5	
20.	<a href="#">p/ n soused</a>	4	

Sledujeme-li seznam, vidíme, že řada substantiv jsou pluralia tantum (*dveře, noviny, šaty, boty, kamna, Vánoce, kalhoty, hodinky, ...*), vidíme i názvy párových orgánů (*ruce, oči, ...*). Zajímavé může být z tohoto hlediska substantivum *rodiče*. Ve frekvenčním seznamu klikneme na **p** u lemmatu *rodič* a získáme příslušné konkordanční řádky.

naše společnost je trochu nemravná . Říká se , že <b>jedni</b> /jeden/ChMP1-----	<b>rodiče</b> užíví deset dětí , ale deset dětí neužíví jedny
.. čím vlastně ? Dalšími rodiči navíc ? Ještě <b>jedni</b> /jeden/ChMP1-----	<b>rodiče</b> . Práví rodiče . Biologičtí rodiče ? Ano ,
nebylo nikomu sděleno , policie to musí respektovat . " <b>Jedni</b> /jeden/ChMP1-----	<b>rodiče</b> tak oznámili pohřešování syna s tím , že by
už o tom , prosím tě , nemluv ! Děti <b>jedněch</b> /jeden/ChMP2-----	<b>rodičů</b> můžou být přece úplně rozdílné . Vůbec nemusí znamenat
souhlasili . " To bude v pořádku , " říkali <b>jedni</b> /jeden/ChMP1-----	<b>rodiče</b> a " Tady se už nedá nic dělat "
jedni rodiče užíví deset dětí , ale deset dětí neužíví <b>jedny</b> /jeden/ChMP4-----	<b>rodiče</b> . Většina starých lidí - a o ty se

Na uvedených dokladech můžeme dobře sledovat užití ve významu číslovky a zájmena (... Říká se, že <jedni> rodiče užíví deset dětí, ale deset dětí neužíví <jedny> rodiče ..., ... Děti <jedněch> rodičů můžou být přece úplně rozdílné ...). Jak je to ale s dalšími významy? Jedná se ve všech případech o význam druhu, tak jako např. v dokladu ... a proti <jedněm> paprskům vysíláme jiné ...?

Podívejme se nyní na kolokace s lemmatem *druhý*. Vrátime se zpět, ponecháme filtr (pozitivní) i rozsah hledání <1,3> a do dotazovacího řádku zadáme příkaz *druhý*.

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**

**Lemma:**

podoby a stal se i teoretickou* základnou k násilnému ovládní <b>jedněch</b> /jeden/ChIP2-----	<b>národů druhými</b> - jednou z nejnebezpečnějších buržoazních ideologií 19 .
zrovna této rodiny a musíš s ní táhnout od štace <b>jedněch</b> /jeden/ChFP2-----	<b>Vánoc k druhým</b> tak učila ses podvazovat obloze trpkých starostí
Stála v ložnici a probírala se svou garderobou . Zkoušela <b>jedny</b> /jeden/ChIP4-----	<b>šaty po druhých</b> , všechny , které měla . Ptát
smilování ! Vy , kteří věříte ! Nechť neposmívají se <b>jedni</b> /jeden/ChMP1-----	<b>lidé druhým</b> , možná že tito jsou lepší než oni
částku na opravu . Hrdý , osamělý , chodil od <b>jedněch</b> /jeden/ChFP2-----	<b>dveří k druhým</b> s nataženou rukou . To ho ranilo
velkých feudálů nebo dobročinnost klášterů . Většina rolníků žila od <b>jedněch</b> /jeden/ChFP2-----	<b>žni k druhým</b> z výnosů , jež byly i v
čtenáři může dočíst o tom , jak sociální demokracie prohrávala <b>jedny</b> /jeden/ChFP4-----	<b>volby za druhými</b> a neuměla komunikovat s lidmi . Vše
na tu hru , se začne proplítat a přelítávat z <b>jedny</b> /jeden/ChFP4-----	<b>strany na druhou</b> , prozkoumávat , chápeš , jako to
Mouřeninu benátském , aby ji zahráli pro potěchu a poctu <b>jedněch</b> /jeden/ChMP2-----	<b>umělců druhým</b> , pro radost z přátelství mezi národy ,
naším je Bůh ! " A kdyby byl Bůh nezahnal <b>jedny</b> /jeden/ChMP4-----	<b>lidí druhými</b> , věru by byly bývaly zničeny poustevny ,
. Ten druhý konec byl zjevně využíván aktivněji . Od <b>jedněch</b> /jeden/ChFP2-----	<b>dveří ke druhým</b> se táhl dlouhý pracovní pult téměř přes

Procházíme-li jednotlivé konkordance, zajisté nás napadá i možná interpretace zájmenná/adjektivní s významem *ten/nějaký*. Rozlišování jednotlivých významů na úrovni

praxe automatické morfologické analýzy by ovšem přineslo pouze další problémovou, a tudíž i chybami zatíženou desambiguaci, jak je patrné i z případu interpretací lemmatu *druhý*.

jednoho typu indexů , jejichž řadu známe , získáme indexy	<b>druhého</b> /druhý/CrIS2-----	typu a posoudíme vývoj jevu , aniž bychom museli znát
ke zdi a s rukama zkříženýma za hlavou jednoho po	<b>druhém</b> /druhý/AAMS6----1A-----	postřílel ranou do týla ukrajinský kat . Když tak vyřídil
otevřenému oknu a vyhlédl do vlhkého nočního klidu . Na	<b>druhém</b> /druhý/AAIS6----1A-----	konci trávníku se proti zlehka blednoucím nebi rýsoval
stromy a křoviny jsou neupravené . O něco lépe vypadá	<b>druhá</b> /druhý/AAFS1----1A-----	polovina parku blíže k vlakovému nádraží . " Je dobře
, dárkových košů , selete a dalších zajímavých cen .	<b>Druhá</b> /druhý/CrFS1-----	cena - minibike je předána a nastává ta osudová vteřina
kus betonu . Batty se jednou rukou držel armatury a	<b>druhou</b> /druhý/CrFS7-----	šmátral nad sebou , až se mu podařilo zachytit se
rozvíjí a rozšiřuje . Prý je z jednoho konce na	<b>druhý</b> /druhý/CrNS4-----6-	sto kilometrů ! Ale nedaleko města na sever , tam
šestibodové vedení po úvodní čtvrtině , ale Severočeši v závěru	<b>druhě</b> /druhý/CrFS2-----	desetiminutovky manko rychle smazali . Byla to zásluha
konci . Vyjekl , ale než se probral úplně ,	<b>druhě</b> /druhý/CrNS1-----	škubnutí ho povalilo na břeh . Strašná síla vlekla muže
povzbuzovat své svěřence a dávat jim pokyny . Jedni i	<b>druzi</b> /druhý/AAMP1----1A-----	se směli obléci až po vyhlášení vítěze . Na nejstarších
za ní chtěl rozběhnout , aby ji chytil , ale	<b>druhý</b> /druhý/AAMS1----1A-----	ho zastavil . " Radši ji nech , je nějaká
Kaščák . Hanačky si v nadstavbě připsaly další porážky O	<b>druhou</b> /druhý/CrFS4-----	výhru se pokusí doma s Karlovými Vary Sport Další dva
jsem tě ! Naopak dával najevo , že má toho	<b>druhého</b> /druhý/AAMS4----1A-----	rád a těší ho pobývat v jeho společnosti . Sáhl
až když země někdo , kdo se jmenuje tím jeho	<b>druhým</b> /druhý/CrNS7-----	jménem . Mnohem větší binec , a to i z
obětovat . Když jeden muž zabije svou ženu , tak	<b>druhý</b> /druhý/AAMS1----1A-----	mu prokazuje přátelství tím , že ji s ním sní
jím uživit ? Je . Kamarád Adam Ondra je oficiálně	<b>druhým</b> /druhý/CrMS7-----	nejlepším lezcem světa , teprve v patnácti letech . Má
Divoce máchali oběma nohama a jednou rukou - - tu	<b>druhou</b> /druhý/AAFS4----1A-----	jím Bůh nadělil , aby si s ní mohli zacpat
úsměvy , přátelská gesta ) , které má získat sympatie	<b>druhých</b> /druhý/AAMP2----1A-----	, - v předstírání emocí : zájmu , soucitu ,

## Formulace závěrů

V případě lemmatizace a tagování tvarů lemmatu *jeden* jde o jev, který na lingvistické rovině spadá do oblasti obtížně řešitelných slovnědruhových přesahů. Při anotacích korpusů se některé problémy slovnědruhových přesahů neřeší. Pokud existují přechody a pokud jsou zaznamenány ve slovníku automatického morfologického analyzátoru, jako je tomu v případě lemmatu *druhý*, jsou výsledky desambiguace sporné. Uživatel může ovšem pozorovat relevantní data, aniž by k jejich třídění používal výsledky desambiguace. Tak může na základě pozorování hledat odpovědi na nejrůznější otázky, mj. i takové, které by vedly k odhalení pravidel pro úspěšnou desambiguaci.

Spolehlivost slovnědruhové desambiguace lemmatu *druhý* by samozřejmě bylo třeba ověřit na větším vzorku. Zůstává ale zřejmé, že se na desambiguaci nelze spoléhat absolutně.

V obou případech je patrné, že ruční třídění je východiskem pro ty uživatele korpusu, kteří se nechtějí spokojit s tím, jak je slovnědruhová interpretace (převzatá v zásadě z tištěných slovníků) vtělena do značek automatické morfologické analýzy a jak se odráží v desambiguaci.

## Formulace dalších otázek vypluvších ze zkoumání daného jevu

Obdobným způsobem lze sledovat použité značkování dalších lemmat, která mohou mít různou slovnědruhovou platnost. Nabízí se např. tvary *mnoho/mnoha*, které jsou

interpretovány jako tvary číslovek, ve srovnání se synonymními výrazy jako např. *hodně*, *moc*, které jsou na úrovni slovního druhu interpretovány jako adverbia<sup>26</sup>.

### **Zadání cvičení, v nichž lze uplatnit analogické postupy**

Vyhledejte v korpusu tvary označované jako druhové číslovky (*Cd. \**). Popište postup práce s korpusovým manažerem, kterým byste získali potřebná data pro sledování případů, kdy číslovka druhová plní funkci číslovky základní (označuje prostý počet).

---

<sup>26</sup> Technicky je jasné, že tvary *mnoho/mnoha* se chovají jako základní číslovky od 5 výše (soustava dvou tvarů, syntaktické vlastnosti na úrovni gramatické shody podmětu a přísudku). Jak se z tohoto hlediska chovají výrazy *hodně*, *moc* aj., pokud jimi ve stejných kontextech nahradíme tvary *mnoho/mnoha*?



## Slovesa

V následujících podkapitolách se zaměříme na otázky spojené s formální morfologií sloves, konkrétně s těmi vlastnostmi sloves, na nichž je založeno třídění do slovesných tříd a vzorů. Ukážeme, jak lze tyto vlastnosti využít pro vyhledávání některých slovesných tvarů bez použití morfologického značkování, přičemž se nejdříve na základě empirie (pozorování korpusových dat) pokusíme formulovat pravidla hláskové podoby některých slovesných tvarů a pak výsledky pozorování srovnáme s pravidly, která můžeme vyvodit ze znalostí fungování gramatického systému. K tomuto tématu se vrátíme i v některých dalších kapitolách věnovaných tvoření deverbativ. Druhou oblastí, na niž chceme zaměřit pozornost, je výzkum mezi a možností tvoření syntetického futura v češtině. Ukážeme některé nedostatky morfologického značkování tvarů budoucího času tvořeného prefixem *po-* od nedokonavých sloves a naznačíme, jak lze v korpusech hledat kandidáty na tvary syntetického futura. Dále budeme sledovat jeden (dle našeho názoru poněkud opomíjený) případ vyjadřování 2. osoby u některých slovesných tvarů. Ukážeme, jak vyhledat relevantní data pro pozorování tohoto jevu v korpusech. Nakonec se zaměříme na analytické slovesné tvary. Ukážeme, jak lze na základě pozorování korpusových dat odhalovat pravidla slovosledných variant složených slovesných tvarů.

### ***Slovesná třída a slovesný vzor – jak lze v korpusu hledat slovesné tvary bez použití tagu?***

#### **Motivační úvod**

V každé mluvnici češtiny najdeme pasáže věnované formální morfologii českého slovesa, které operují s termíny jako kmen, přičemž rozlišují kmen přítomný minulý/infinitivní, a dále kmenotvorná přípona (téma, tematický vokál). Na pojeticích podstaty těchto termínů je pak založeno třídění českých sloves do slovesných tříd a jejich řazení ke vzorům. V oddílech věnovaných tvoření slov se pak většinou nějakým způsobem odkazuje ke třídění zavedenému na úrovni tvaroslovné, a to především na úrovni rozlišení derivací od kmene a od kořene, popř. od slovesného tvaru.

Určování slovesných tříd a vzorů pokládá většina studentů za duchamorný dril. Přesto může dobrý přehled v této oblasti účinně napomoci pochopení, jak vyhledat v korpusech doklady řady jevů v oblasti morfologie a tvoření slov. Segmentaci slovesných tvarů, při níž bereme v úvahu přítomnost kmenotvorné přípony, lze totiž dobře využít, chceme-li formálně definovat různé tvary sloves i některých deverbativ (srov. výše).

Pro přehlednost v příloze uvádíme algoritmus určování slovesných tříd podle kmene přítomného a vzorů podle kmene minulého v češtině. Vycházíme zde z Komárkova modelu (Komárek a kol. 1987 : 427–495).

### **Nastínění problému**

Pokud chceme vyhledávat v korpusech nejen jednotlivé slovní tvary, uvítáme, máme-li k dispozici morfologicky označovaný korpus. Ten ale vždy k dispozici nemáme. Máme-li jej, pak v každém korpusu existují tvary, které jsou označovány „nesprávně“, což souvisí s chybami na úrovni automatické morfologické analýzy (tvaru není přidělena žádná interpretace), nebo nedostatky v desambiguaci (z přidělených interpretací nebyla vybrána správná interpretace). Mohou tedy nastat situace, kdy se při práci s jazykovými korpusy chceme, či jsme nuceni bez použití lemmatizace a tagování obejít.

### **Otázky**

Naším cílem bude vyhledat kandidáty na l-ové přičestí. Záměrně ponecháme stranou substandardní tvary sloves s uzavřeným kmenem minulým (vzory *nést, péct, tisknout*) realizované bez koncového *-l* (*nes, ved, upek, spad, ...*).

Při řešení takového i podobných úkolů můžeme stavět na předpokladech, které vycházejí právě ze znalostí pravidel utváření slovesného tvaru. Můžeme ale zvolit postup, kdy pravidelnosti odhalíme na základě pozorování korpusových dat a až nakonec propojíme empiricky zjištěná pravidla s pravidly vycházejícími z gramatických popisů.

### **Formulace dotazu pro získání dat z korpusů**

Nejdříve se pokusme formálně definovat tvary l-ových přičestí v češtině. Co o nich lze po formální stránce s jistotou tvrdit?

Nejjednodušší odpověď zní, že tvar musí končit na řetězec (*l|l[aoiy]*).

Zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme dotaz ve formě *.\*(l|l[aoiy])*.

## Hledat v korpusu

Korpus:

syn2010

Typ dotazu:

Slovní tvar ▾

Slovní tvar:

.\*(1|l[aoiy])

sekcí , kterou tehdy s Karlem Srpem spoluzaložil . "	<b>Bavilo</b> /bavit/VpNS---3R-AA---I	mě to , protože jsem měl možnost vytvořit její výtvarnou
. Jmenuje se Sebastian Vettel . O vikendu v Monze	<b>přepsal</b> /přepsat/VpIS---3R-AA---P	kus historie formule 1 . V sobotu se stal nejmladším
, tolik se změnil . Zídka a plot kolem parku	<b>zmizely</b> /zmizet/VpIP---3R-AA---P	, zůstal jen zelený ovál uprostřed domu , jež se
němu otočila a naklonila hlavu . " A copak ti	<b>řekla</b> /řici/VpFS---3R-AA---P	mamka o nakukování do aury cizích lidí bez jejich souhlasu
nich laciné natahovací hračky , z kterých při druhém natažení	<b>padala</b> /padat/VpNP---3R-AA---I	kolečka a vyskakovala péra . Byly tam ohybné sexuální pomůcky
o překážky v práci . Uvedené nedostatky v zásadě odstranila	<b>novela</b> /novela/NNFS1-----A-----	zákoníku práce , pro níž se vžilo označení technická -
speciální okénko vydavatele a spisovatele Martina Reinera , který se	<b>díval</b> /dívat/VpMS---3R-AA---I	na svět jako unuděný boháč . Svrázné glosy a fejetony
skutků ze ztratila a všichni tvrdí , že tvá ruka	<b>zestárla</b> /zestárnout/VpFS---3R-AA--1P	. Odpusť , že ti to říkám - dodal Satan
do nich vešel gravitatorový pohon a kompenzátor setrvačnosti , a	<b>byly</b> /být/VpFP---3R-AA---I	ozbrojené . Ve srovnání s pravidelnými válečnými loděmi sice jejich


Podíváme se na frekvenční distribuci **slovních tvarů** z hlediska slovních druhů (a na druhé úrovni zadáme atribut **pos**).

## Víceúrovňová frekvenční distribuce

Frekvenční limit: 0

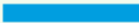
Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>?</sup>	
1.	lc ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	-
2.	pos ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	X
<input type="button" value="Přidat úroveň"/>					

## Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 106711 (2135 str.)

	lc	pos	Frekvence	
1.	p/ n byl	V	240 868	
2.	p/ n bylo	V	178 118	
3.	p/ n byla	V	172 366	
4.	p/ n řekl	V	104 307	
5.	p/ n měl	V	104 290	
6.	p/ n byly	V	69 578	
7.	p/ n měla	V	61 037	
8.	p/ n byli	V	50 117	
9.	p/ n mohl	V	46 771	
10.	p/ n jestli	J	46 356	
11.	p/ n chvíli	N	44 933	
12.	p/ n měli	V	43 415	
13.	p/ n li	J	43 079	
14.	p/ n kvůli	R	39 116	
15.	p/ n řekla	V	37 775	
16.	p/ n dál	D	35 689	
17.	p/ n zcela	D	30 793	
18.	p/ n nebyl	V	30 142	
19.	p/ n nebylo	V	29 385	
20.	p/ n chtěl	V	29 144	

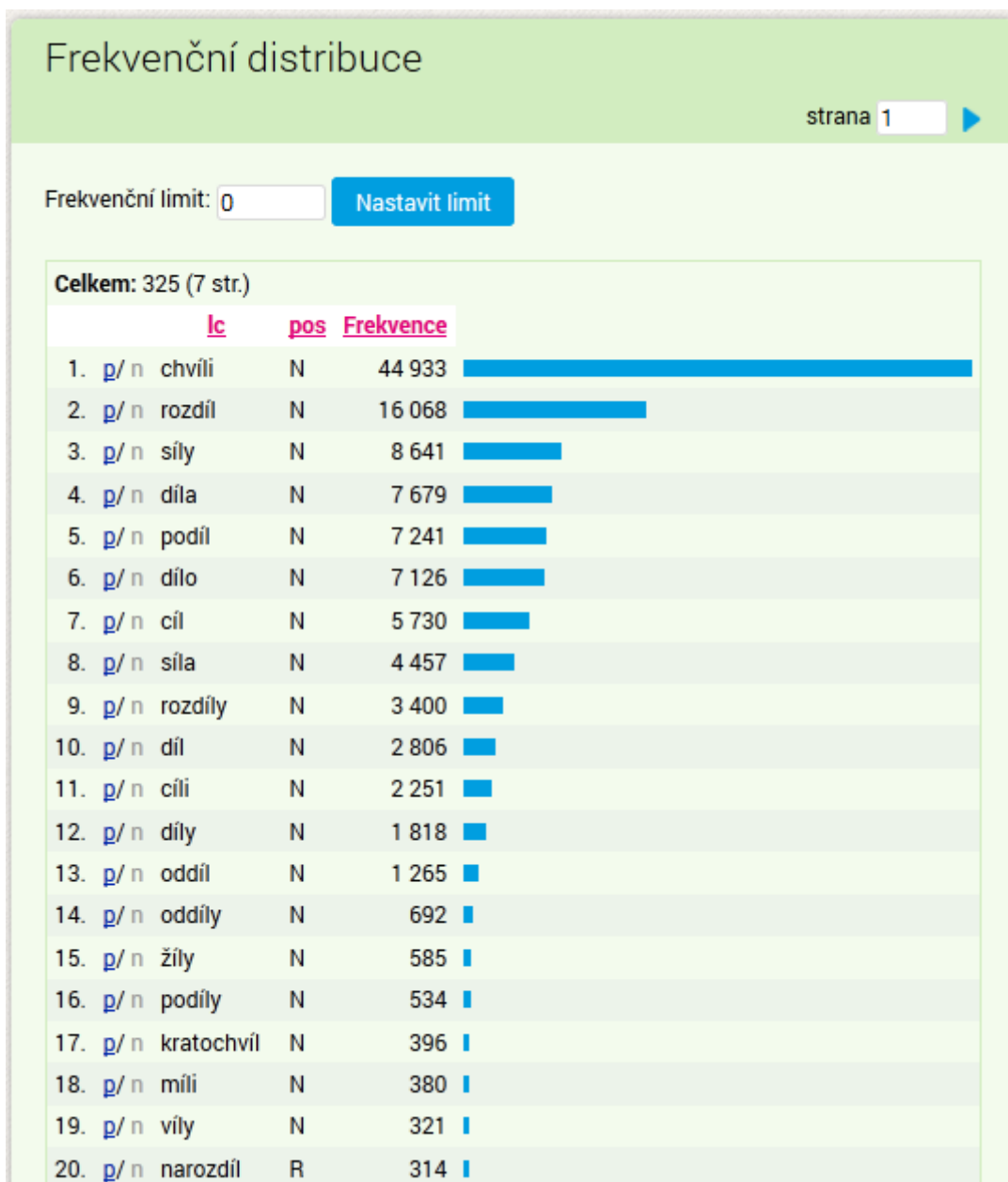
### Třídění a pozorování dat získaných z korpusů

Již mezi velmi frekventovanými formálně definovanými tvary nacházíme nežádoucí doklady: *li, chvíli, jestli, dál, kvůli, zcela, ...* Procházet ručně 106 711 tvarů není řešení, které by nás lákalo. Vede k cíli nějaká efektivnější cesta?

V úvodu jsme hovořili o tom, že budeme brát v úvahu segmentaci slovesného tvaru v češtině, a to především vydělení kmenotvorné přípony. Kmenotvornou příponu lze pojímat jako segment stojící před osobní nebo infinitivní koncovkou, popřípadě před komplexem tvarotvorná přípona participia + rodová koncovka. Při řešení našeho problému pracujeme s tvary l-ových přičestí, tedy tvary zakončenými komplexem tvarotvorná přípona *-l-* a rodová koncovka *-o, -a, -o, -i, -y*. Pokud zapomeneme úplně všechno, co jsme kdy slyšeli o

kmenotvorné příponě, můžeme minimálně předpokládat, že půjde o „něco“, co bude předcházet před tímto komplexem.

Otázka, kterou si můžeme položit, pak bude znít takto: Může před *-l*, má-li být tvarovou koncovkou *l*-ového přičestí, předcházet cokoliv (třeba i mezera, srov. *li*), nebo existují nějaká omezující (vymezující) pravidla? Dále můžeme experimentovat, hledat postupně pomocí filtru a dívat se na tvary stejně zakončené jako ty, které prokazatelně slovesnými tvary nejsou (viz výše). Nastavíme filtr (pozitivní), zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme postupně dotazy ve formě *.\*í(l|l[aoiy])*, *.\*t(l|l[aoiy])*, *.\*á(l|l[aoiy])*, *.\*ů(l|l[aoiy])*, *.\*e(l|l[aoiy])*. Sledujeme frekvenční distribuce slovních tvarů a slovních druhů.



## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 5 (1 str.)

	<u>pos</u>	<u>Frekvence</u>	
1. <a href="#">p/n</a> N		118 377	<div style="width: 100%;"></div>
2. <a href="#">p/n</a> X		411	<div style="width: 0.34%;"></div>
3. <a href="#">p/n</a> R		314	<div style="width: 0.26%;"></div>
4. <a href="#">p/n</a> D		45	<div style="width: 0.038%;"></div>
5. <a href="#">p/n</a> V		1	<div style="width: 0.0008%;"></div>

## Frekvenční distribuce

strana  [▶](#)

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 1628 (33 str.)







	<u>lc</u>	<u>pos</u>	<u>Frekvence</u>	
1. <a href="#">p/n</a> jestli	J		46 356	<div style="width: 28.5%;"></div>
2. <a href="#">p/n</a> světlo	N		9 595	<div style="width: 5.9%;"></div>
3. <a href="#">p/n</a> světla	N		8 353	<div style="width: 5.1%;"></div>
4. <a href="#">p/n</a> odmítl	V		4 747	<div style="width: 2.9%;"></div>
5. <a href="#">p/n</a> četl	V		3 886	<div style="width: 2.4%;"></div>
6. <a href="#">p/n</a> odmítla	V		2 147	<div style="width: 1.3%;"></div>
7. <a href="#">p/n</a> přečetl	V		2 049	<div style="width: 1.25%;"></div>
8. <a href="#">p/n</a> ocitl	V		1 975	<div style="width: 1.2%;"></div>
9. <a href="#">p/n</a> četla	V		1 768	<div style="width: 1.08%;"></div>
10. <a href="#">p/n</a> poskytl	V		1 568	<div style="width: 0.96%;"></div>
11. <a href="#">p/n</a> odmítli	V		1 341	<div style="width: 0.82%;"></div>
12. <a href="#">p/n</a> namítl	V		1 311	<div style="width: 0.8%;"></div>
13. <a href="#">p/n</a> vzrostl	V		1 235	<div style="width: 0.76%;"></div>
14. <a href="#">p/n</a> ocitla	V		1 168	<div style="width: 0.72%;"></div>
15. <a href="#">p/n</a> ocitli	V		1 092	<div style="width: 0.67%;"></div>
16. <a href="#">p/n</a> poskytla	V		1 008	<div style="width: 0.62%;"></div>
17. <a href="#">p/n</a> vyrostl	V		906	<div style="width: 0.55%;"></div>
18. <a href="#">p/n</a> rostly	V		875	<div style="width: 0.53%;"></div>
19. <a href="#">p/n</a> přečetla	V		824	<div style="width: 0.50%;"></div>
20. <a href="#">p/n</a> vyrostla	V		822	<div style="width: 0.50%;"></div>

## Frekvenční distribuce


Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 6 (1 str.)

	<b>pos</b>	<b>Frekvence</b>	
1.	<a href="#">p/n</a> V	63 148	
2.	<a href="#">p/n</a> J	46 356	
3.	<a href="#">p/n</a> N	20 613	
4.	<a href="#">p/n</a> X	717	
5.	<a href="#">p/n</a> A	20	
6.	<a href="#">p/n</a> D	3	

## Frekvenční distribuce

strana  

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 1816 (37 str.)

	<b>lc</b>	<b>pos</b>	<b>Frekvence</b>	
1.	<a href="#">p/n</a> dál	D	35 689	
2.	<a href="#">p/n</a> málo	D	13 830	
3.	<a href="#">p/n</a> stál	V	12 890	
4.	<a href="#">p/n</a> zdálo	V	11 657	
5.	<a href="#">p/n</a> stála	V	7 790	
6.	<a href="#">p/n</a> král	N	7 014	
7.	<a href="#">p/n</a> hrál	V	6 385	
8.	<a href="#">p/n</a> usmál	V	5 948	
9.	<a href="#">p/n</a> vyhrál	V	4 963	
10.	<a href="#">p/n</a> materiál	N	4 827	
11.	<a href="#">p/n</a> stálo	V	4 569	
12.	<a href="#">p/n</a> usmála	V	4 453	
13.	<a href="#">p/n</a> hráli	V	3 927	
14.	<a href="#">p/n</a> stáli	V	3 412	
15.	<a href="#">p/n</a> přál	V	3 401	
16.	<a href="#">p/n</a> materiály	N	3 227	
17.	<a href="#">p/n</a> málo	C	3 186	
18.	<a href="#">p/n</a> hrála	V	2 952	
19.	<a href="#">p/n</a> areál	N	2 893	
20.	<a href="#">p/n</a> signál	N	2 843	

## Frekvenční distribuce

Frekvenční limit:


[Nastavit limit](#)

**Celkem:** 7 (1 str.)

	<b>pos</b>	<b>Frekvence</b>	
1.	<a href="#">p/</a> <a href="#">n/</a> V	142 918	
2.	<a href="#">p/</a> <a href="#">n/</a> N	73 527	
3.	<a href="#">p/</a> <a href="#">n/</a> D	54 230	
4.	<a href="#">p/</a> <a href="#">n/</a> C	4 414	
5.	<a href="#">p/</a> <a href="#">n/</a> X	1 113	
6.	<a href="#">p/</a> <a href="#">n/</a> I	326	
7.	<a href="#">p/</a> <a href="#">n/</a> A	6	



## Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

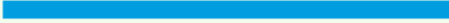




Celkem: 70 (2 str.)

	lc	pos	Frekvence		
1.	p/ n	kvůli	R	39 116	
2.	p/ n	půl	N	20 431	
3.	p/ n	stůl	N	7 227	
4.	p/ n	vůli	N	3 483	
5.	p/ n	sůl	N	3 113	
6.	p/ n	napůl	D	3 057	
7.	p/ n	půli	N	2 425	
8.	p/ n	půl	C	1 630	
9.	p/ n	hůl	N	1 175	
10.	p/ n	smůla	N	652	
11.	p/ n	důl	N	471	
12.	p/ n	vůl	N	466	
13.	p/ n	kůl	N	193	
14.	p/ n	kůly	N	164	
15.	p/ n	smůly	N	135	
16.	p/ n	vejpůl	D	76	
17.	p/ n	zvůli	N	72	
18.	p/ n	libovůli	N	64	
19.	p/ n	nevůli	N	34	
20.	p/ n	vpůli	R	28	


## Frekvenční distribuce

Frekvenční limit:  [Nastavit limit](#)

Celkem: 5 (1 str.)

	pos	Frekvence		
1.	p/ n	N	40 163	
2.	p/ n	R	39 144	
3.	p/ n	D	3 161	
4.	p/ n	C	1 630	
5.	p/ n	X	65	

## Frekvenční distribuce

strana 1 

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 10083 (202 str.)


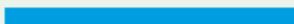




	lc	pos	Frekvence		
1.	p/ n	zcela	D	30 793	
2.	p/ n	musel	V	23 183	
3.	p/ n	přišel	V	19 881	
4.	p/ n	docela	D	16 616	
5.	p/ n	ředitel	N	14 876	
6.	p/ n	šel	V	13 120	
7.	p/ n	bohužel	T	12 082	
8.	p/ n	pavel	N	12 073	
9.	p/ n	museli	V	11 853	
10.	p/ n	musela	V	11 762	
11.	p/ n	karel	N	9 640	
12.	p/ n	myslel	V	9 628	
13.	p/ n	obyvatel	N	9 348	
14.	p/ n	našel	V	9 166	
15.	p/ n	slyšel	V	8 078	
16.	p/ n	zemřel	V	7 498	
17.	p/ n	otevřel	V	7 399	
18.	p/ n	odešel	V	6 820	
19.	p/ n	model	N	6 342	
20.	p/ n	vyšel	V	5 925	

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 6 (1 str.)

	pos	Frekvence		
1.	p/ n	V	522 735	
2.	p/ n	N	282 579	
3.	p/ n	D	47 868	
4.	p/ n	T	12 417	
5.	p/ n	X	12 325	
6.	p/ n	A	56	

### Formulace závěrů

Na základě empirie lze konstatovat, že existují některé kombinace, jako např.  $.[i\acute{u}](l|l[aoiy])$ , pro něž v korpusu nebyl nalezen jediný doklad l-ového přičestí.

V dalším postupu lze postupně testovat další možnosti. Můžeme pokračovat po dílčích krocích. Testovat třeba vokály, přičemž můžeme začít od dlouhých (kromě *á*, protože ve dvou případech  $[i\acute{u}]$  jsme zjistili pozitivní výsledky) a pokračovat krátkými. Může nás napadnout, že nebude asi možné kombinovat s *l* libovolné konsonanty (přinejmenším *ř*, *l*, ale třeba také *d'*, *t'*, *ň*, či další).

Na tomto místě nebudeme jednotlivé kroky s ohledem na rozsah našeho textu dokumentovat, a ukážeme výsledky výše naznačených postupů. Empiricky lze dokázat, že mezi tvary končícími řetězcem  $.[\acute{e}i\acute{o}\acute{o}\acute{u}\acute{y}](l|l[aoiy])$  se nevyskytl ani jeden spisovný tvar českého l-ového přičestí. Jak?

Vrátíme se zpět ke konkordanci tvarů na  $.(l|l[aoiy])$ , zvolíme filtr (pozitivní), **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme dotaz ve formě  $.[\acute{e}i\acute{o}\acute{o}\acute{u}\acute{y}](l|l[aoiy])$  a podíváme se na frekvenční distribuci slovních tvarů a slovních druhů (postup viz výše).

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 2993 (60 str.)

	lc	pos	Frekvence		
1.	p/ n	chvíli	N	44 933	
2.	p/ n	kvůli	R	39 116	
3.	p/ n	půl	N	20 431	
4.	p/ n	školy	N	19 656	
5.	p/ n	rozdíl	N	16 068	
6.	p/ n	roli	N	14 223	
7.	p/ n	síly	N	8 641	
8.	p/ n	nikoli	T	8 631	
9.	p/ n	díla	N	7 679	
10.	p/ n	podíl	N	7 241	
11.	p/ n	stůl	N	7 227	
12.	p/ n	dílo	N	7 126	
13.	p/ n	kola	N	7 095	
14.	p/ n	ačkoli	J	6 712	
15.	p/ n	kolo	N	6 475	
16.	p/ n	okolo	R	6 440	
17.	p/ n	úkol	N	6 435	
18.	p/ n	škol	N	6 283	
19.	p/ n	škola	N	6 031	
20.	p/ n	cíl	N	5 730	

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

Celkem: 11 (1 str.)

	pos	Frekvence	
1.	<a href="#">p/ŋ</a> N	304 534	<div style="width: 100%;"></div>
2.	<a href="#">p/ŋ</a> R	51 544	<div style="width: 17%;"></div>
3.	<a href="#">p/ŋ</a> P	18 959	<div style="width: 6%;"></div>
4.	<a href="#">p/ŋ</a> D	16 094	<div style="width: 5%;"></div>
5.	<a href="#">p/ŋ</a> J	10 859	<div style="width: 3.5%;"></div>
6.	<a href="#">p/ŋ</a> T	8 631	<div style="width: 2.8%;"></div>
7.	<a href="#">p/ŋ</a> X	5 433	<div style="width: 1.8%;"></div>
8.	<a href="#">p/ŋ</a> C	1 630	<div style="width: 0.5%;"></div>
9.	<a href="#">p/ŋ</a> V	566	<div style="width: 0.18%;"></div>
10.	<a href="#">p/ŋ</a> I	51	<div style="width: 0.017%;"></div>
11.	<a href="#">p/ŋ</a> A	36	<div style="width: 0.012%;"></div>


Slovesné tvary jsou pouze imperativy sloves (např. *dovol, osol, zvol, povol, přisol, ohol, ...*).

Vyskytly se též tvar l-ového přičestí slovesa *být* ve slovenštině (*nebol*) mylně interpretovaný jako imperativ slovesa *nebolet*, nebo tvar l-ového přičestí ruského slovesa *ubit'* přepsaný jako *ubili* interpretovaný automatickou morfologickou analýzou jako imperativ slovesa *ubilit*.

Dále lze empiricky dokázat, že mezi tvary končícími řetězcem *.\*[bdfghkmnprtv]e(l|l[aoiy])* se nevyskytl ani jeden spisovný tvar českého l-ového přičestí.






Vrátíme se o krok zpět, zvolíme filtr (pozitivní), zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme dotaz ve formě *.\*[bdfghkmnprtv]e(l|l[aoiy])* a podíváme se na frekvenční distribuci slovních tvarů a slovních druhů (postup viz výše).

## Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)



**Celkem:** 4034 (81 str.)

	lc	pos	Frekvence	
1.	p/ n ředitel	N	14 876	
2.	p/ n pavel	N	12 073	
3.	p/ n karel	N	9 640	
4.	p/ n obyvatel	N	9 348	
5.	p/ n model	N	6 342	
6.	p/ n majitel	N	5 592	
7.	p/ n posteli	N	5 112	
8.	p/ n kostela	N	4 734	
9.	p/ n přítel	N	4 652	
10.	p/ n učitel	N	3 463	
11.	p/ n pravidel	N	3 366	
12.	p/ n přátel	N	3 355	
13.	p/ n kostel	N	3 198	
14.	p/ n velitel	N	3 153	
15.	p/ n postel	N	3 136	
16.	p/ n modely	N	3 073	
17.	p/ n hotel	N	3 049	
18.	p/ n vozidel	N	3 026	
19.	p/ n kapely	N	2 809	
20.	p/ n kapela	N	2 759	

## Frekvenční distribuce

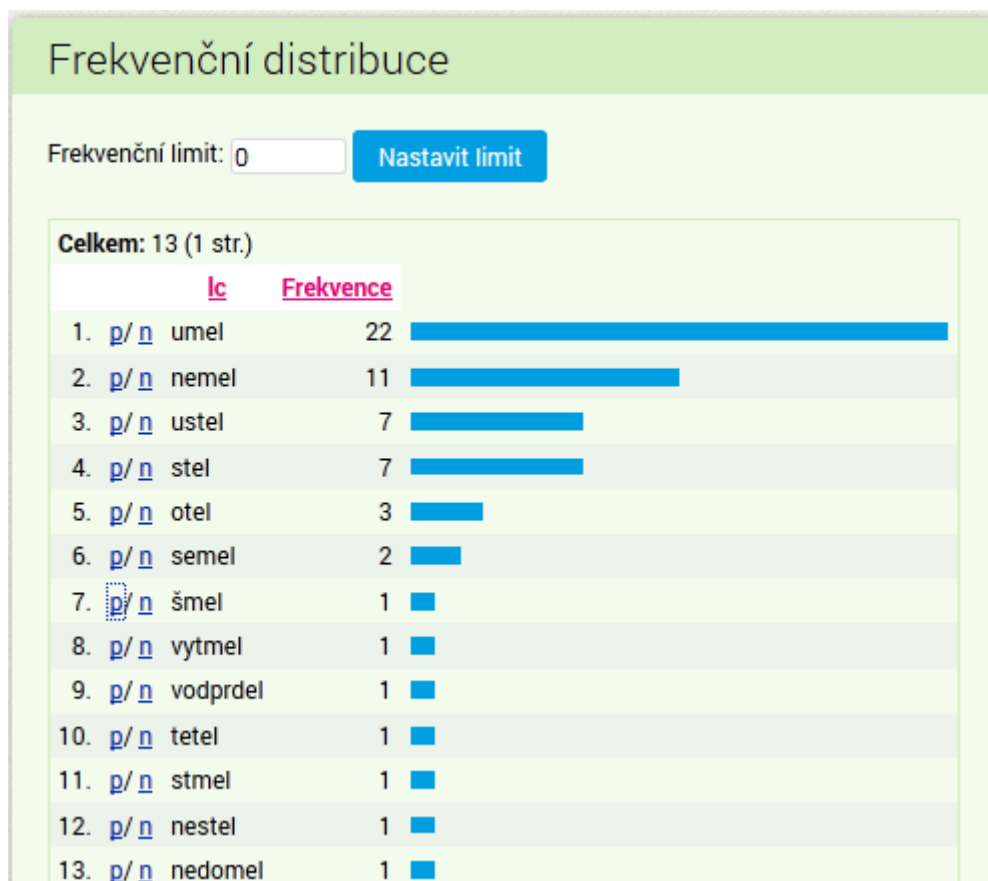
Frekvenční limit:  [Nastavit limit](#)

**Celkem:** 4 (1 str.)

	pos	Frekvence	
1.	p/ n N	218 871	
2.	p/ n X	7 175	
3.	p/ n V	59	
4.	p/ n A	5	

Pokud se tvary označované jako slovesa vyskytly, pak šlo o tvary označované jako tvary imperativu (*umel, nemel, ustel, vytmel, ...*), přičemž v některých případech se vyskytly chyby v morfologickém značkování (Tvary *tetel, šmel* nejsou v daných kontextech imperativy sloves *tetelit se, šmelit*).


Kliknutím na **p** na příslušném řádku získáme konkordanční seznam s výskyty tvarů označovaných jako slovesa a opět se podíváme se na frekvenční distribuci slovních tvarů a slovních druhů (postup viz výše).



Empiricky lze konečně taktéž dokázat, že mezi tvary končícími řetězcem *.\*[ááééěiioóúúůůýý][ááéěiu](l|l[aoiy])* se nevyskytl ani jeden spisovný tvar českého l-ového příčestí.

Vrátíme se zpět ke konkordanci tvarů na *.\*(l|l[aoiy])*, zvolíme filtr (pozitivní), **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme dotaz ve formě *.\*[ááééěiioóúúůůýý][ááéěiu](l|l[aoiy])* a podíváme se na frekvenční distribuci slovních tvarů a slovních druhů (postup viz výše).

## Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)




Celkem: 1990 (40 str.)

	<u>lc</u>	<u>pos</u>	<u>Frekvence</u>	
1.	p/ n	materiál	N	4 827 
2.	p/ n	materiály	N	3 227 
3.	p/ n	michael	N	2 921 
4.	p/ n	areál	N	2 893 
5.	p/ n	paul	N	2 320 
6.	p/ n	detaily	N	2 198 
7.	p/ n	daniel	N	2 146 
8.	p/ n	michaela	N	1 930 
9.	p/ n	e-mail	N	1 907 
10.	p/ n	seriál	N	1 901 
11.	p/ n	potenciál	N	1 824 
12.	p/ n	duel	N	1 719 
13.	p/ n	medaili	N	1 480 
14.	p/ n	izrael	N	1 441 
15.	p/ n	daniela	N	1 278 
16.	p/ n	cadfael	N	1 075 
17.	p/ n	detail	N	1 007 
18.	p/ n	paula	N	904 
19.	p/ n	izraeli	N	892 
20.	p/ n	real	N	844 

## Frekvenční distribuce

Frekvenční limit:  [Nastavit limit](#)

Celkem: 3 (1 str.)

	<u>pos</u>	<u>Frekvence</u>	
1.	p/ n	N	61 328 
2.	p/ n	X	7 638 
3.	p/ n	V	3 



Pokud se tvary označované jako slovesa vyskytly, pak šlo o nesprávně desambiguované tvary imperativu, jak je patrné, klikneme-li na **p** u **pos V**. Vidíme jasně, že jde o doklady chyb v morfologickém značkování.

několik postranních uliček na boulevard Saint-Michel - " Na **Boul** /boulit/Vi-S---2--A----I ' Mich ' - " A po Boul '


a jedou naším směrem nahoru . Stoupáme pod Vf . **Boul** /boulit/Vi-S---2--A----I , jde to celkem rychle , bereme také ze silného

Na Boul ' Mich ' - " A po **Boul** /boulit/Vi-S---2--A----I ' Mich ' kolem Cluny - " Po naší

Empiricky lze dokázat, že mezi tvary končícími řetězcem *.\*[d'g]lnňřřžqw](l|l[aoiy])* se nevyskytl ani jeden spisovný tvar českého l-ového přičestí.

Vrátíme se zpět ke konkordanci tvarů na *.\*(l|l[aoiy])*, zvolíme filtr (pozitivní), **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme dotaz ve formě *.\*[d'g]lnňřřžqw](l|l[aoiy])* a podíváme se na frekvenční distribuci slovních tvarů a slovních druhů (postup viz výše).





















# Frekvenční distribuce

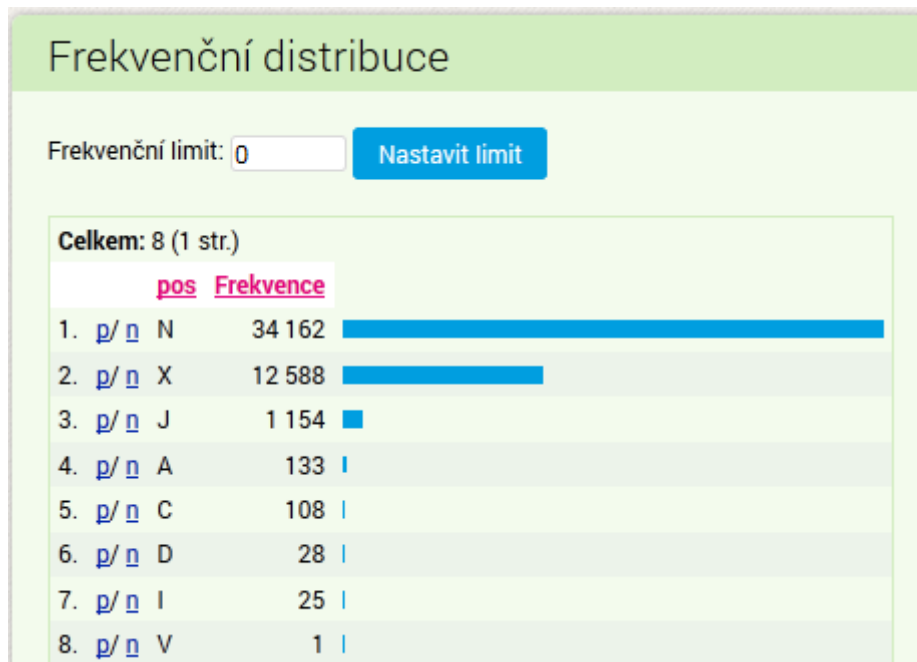
strana 1 

Frekvenční limit:

[Nastavit limit](#)

**Celkem:** 3815 (77 str.)

	lc	pos	Frekvence		
1.	p/ n	sally	N	1 925	
2.	p/ n	bill	N	1 795	
3.	p/ n	billy	N	1 240	
4.	p/ n	nežli	J	1 154	
5.	p/ n	kelly	N	1 105	
6.	p/ n	polly	N	975	
7.	p/ n	hall	N	724	
8.	p/ n	billa	N	696	
9.	p/ n	jill	X	631	
10.	p/ n	churchill	N	604	
11.	p/ n	will	N	603	
12.	p/ n	hill	N	519	
13.	p/ n	novella	N	505	
14.	p/ n	wall	X	489	
15.	p/ n	all	X	467	
16.	p/ n	carella	N	466	
17.	p/ n	koktejl	N	415	
18.	p/ n	džungli	N	415	
19.	p/ n	moll	N	413	
20.	p/ n	ella	N	399	



Pokud se tvar označovaný jako sloveso vyskytl, pak šlo o chybu v lemmatizaci i značkování, která patrně spadají na vrub automatického značkování a lemmatizace (přegenerování).<sup>27</sup>

CITY Snímek Padre Nuestro ( Náš otec ) režiséra Christophera Zally /zalnout/VpiP---3R-AA--11 překvapivě zvítězil v hlavní soutěži nejslavnějšího světového festivalu nezávislých filmů

V korpusu byly ovšem při podrobném procházení konkordančních řádků nalezeny též doklady nerozpoznané automatickou morfologickou analýzou. Jedná se o některé pro češtinu méně obvyklé kombinace.

se a pomalu umírají . Za mé roční působení tu **exli** /exli/X@----- dva , " říká Vlastimil Chromčák , někdejší primář oddělení s nima pomohla na postel . Saniták : Prej málem **exli** /exli/X@----- na sále ? Sestra : Dejte pokoj . Víte ,

Výše uvedená pravidla založená na empirii (korpusová data) lze podložit důkazem opřeným o pravidla fungování systému české slovesné flexe (konjugace a derivace tvarů přičestí).

<sup>27</sup> Sloveso *zalnout* uvádí jak PSJČ, tak SSJČ jako zastaralé ve významu *přilnout*, sloveso *přemnout* uvádějí oba slovníky ve významu „jemně promnout“. Možnost interpretovat tvar *zally* jako tvar slovesa *zalnout*, spadá na vrub přegenerování automatické morfologické analýzy, respektive pravidlu uplatněnému při sestavování slovníku pro automatickou morfologickou analýzu. Toto pravidlo připouští, že všechna slovesa vzoru *tisknout* mohou tvořit tvary l-ového přičestí (n-ového přičestí) jak s kmenotvornou příponou *nu-*, tak bez ní (tedy *tiskl/tištěn* i *tisknul/tisknut*), což je sporné zejména pro sloveso *lnout* (ale např. také pro *mnout*) i jejich prefigované deriváty. Pro ostatní slovesa (např. *hnout*, *schnout*, *usnout*. ...) neexistují v mluvnických popisech pravidla, která by bylo možné jednoduše zobecnit a především formálně popsat. To byl zřejmě důvod pro zařazení zjednodušujícího pravidla do popisu pro automatickou morfologickou analýzu. V korpusu SYN jsou tudíž tvary *Zally* (mužské jm., nom. je *Zalla*, nebo ženské jm. nesklonné) a *Zalli* (ženské jm. patrně nesklonné, nejde-li ovšem o překlep) interpretovány jako l-ová přičestí slovesa *zalnout*. Vlastní jména se ve slovnících automatických morfologických analyzátorů uvádějí více méně ad hoc. Předpokládáme, že tvar neprošel desambiguací. (Srv. též lemmatizaci a značkování překlepu *přemlo*<*přemlouval*.)

Mějme tři tvrzení:

1. Tvar končí na řetězec  $l|l[aoiy]$ , před nímž mohou předcházet pouze některé vokály a konsonanty.
2. Před řetězcem  $el$  nemohou předcházet některé konsonanty.
3. Před vokály stojícími před řetězcem  $l$  nemohou předcházet žádné vokály.

Diskuse: První tvrzení vychází ze znalostí pravidel struktury slovesného tvaru. Mezi vokály, které mohou předcházet před  $l$ , patří pouze ty, které a) mohou figurovat jako kmenotvorná přípona (dále KmV) a b) mohou být vokálem v kořeni (KoV) sloves podle vzoru *krýt* a některých nepravidelných sloves (*být, mít, chtít, ...*). Pod a) patří vokály  $-a-$  (*děl-a-l*),  $-e/-ě-$  (*drž-e-l, trp-ě-l*),  $-i-$  (*kouř-i-l*),  $-u-$  (*tisk-nu-l*), pod b) patří  $-á-$  (*hrá-0-l*),  $-e/-ě-$  (*kle-0-l, pě-0-l*),  $-i-$  (*pi-0-l*),  $-u-$  (*zu-0-l*),  $-y-$  (*kry-0-l*). Mezi konsonanty mohou patřit všechny kromě  $[dʒlnřřtʃqw]$ , naopak mezi ně mohou patřit konsonanty  $[gž]$ , o čemž svědčí doklady z dalších korpusů (typicky moravské *rožl*) nebo z internetu (*nagrgl*)<sup>28</sup>.

tom obleku vypadal jak mafioso . </s><s> Nahmatal jsem vypínač a rožl /rožl/X@----- . </s><s> Petr ležel na loktech a žmural na nás očima  
mazaní a neviděli na krok . </s><s> " Póó-zor , " rožlo /rožnout/VpNS---3R-AA---P sa světlo a velitelovi při tem pohledu cukalo v kútkoch

Ahojky brusky, ..tak Silvia uz stacila obecne vylicit holou skutecnost,..on je vazne nepolepsitelny a dneska se zas tak nagrgl, ze mi mezi dvermi zahlasil, ze jestli ho opet vytocim, tak do mesice a do dne koncime a at se potom nedivim, kdyz me posle zpatky k nasim...a zmizel ve dverich, pak se vratil s tim, ze se prihlasil do kurzu boxovani (tak to jsem mu hned rekla, ze to byl dobry napad..a at si tu svou zlost vybijí na pytlí a ne na me, jsem si pomyslela..hi..hi). Jinak se snazi jak muze, uz me alespon nemlati, sem tam se vytoci, ale tak snad mu to boxovani bude ku prospechu...  
Na ten registracni urad jsem se nedostala...  
koncim, bezim, uz je tu  
pa Markina

 [Nahlásit](#)

[Napsat příspěvek](#)

Druhé tvrzení vychází z toho, že česká  $l$ -ová přičestí zakončená na  $-el$  musí být tvořena od sloves některých tříd a vzorů, a to a) I. třída vzor *umřít*, b) III. třída vzor *krýt*, c) IV. třída vzory *trpět, sázet*, d) nepravidelná. Pod a) patří pouze slovesa, u nichž před  $el$  předchází  $[řl]$ , pod b) patří pouze slovesa, u nichž před  $el$  předchází  $[lsz]$ , pod c) patří pouze slovesa, u nichž před  $el$  předchází  $[cčjlršsžž]$ , pod d) patří slovesa *jel, šel*, u nichž před  $el$  předchází  $[jš]$ . Mezi

<sup>28</sup> Emimino.cz [online]. 2011. Cit. 20.3. 2011, dostupné z: <http://www.emimino.cz/denicky/denicek-maminek-v-zahranici-ix-svokra-verzus-matka-3474/strankovani/2/>. Ke tvarům *rožl*, ... srv. též Osolsobě 2011<sup>1</sup>: 43n.

konsonanty, po kterých nemůže následovat *el*, patří [dtnbpfvm] (protože ve spisovné češtině neexistují slovesa, pro něž by platilo, že kořenová finála je [dtnbpfvm], a zároveň KmV je *e*, neboť po kořenových finálách realizovaných graficky jako [dtnbpfvm] následuje v psané podobě nutně grafické *ě*, takže ve zvukové podobě jazyka se tyto finály vlastně nerealizují), dále [kgh|(ch)] (protože ve spisovné češtině neexistují slovesa, pro něž by platilo, že kořenová finála je velára a zároveň KmV je *e*) a *r* (protože ve spisovné češtině neexistují slovesa, pro něž by platilo, že kořenová finála je *r* a zároveň KmV je *e*). Tímto způsobem lze vyloučit deverbativní substantiva činitelská na *-tel*, dále např. vlastní jména (*Karel, Pavel, Michel, ...*), nebo substantiva/tvary jako *vyvrhel, chrchel, pekel, plevel, boubel, koupel, cumel, ...*

Třetí tvrzení vychází z toho, že předchází-li před *l* vokál, pak jde buď o kmenový vokál KmV, nebo o kořenový vokál KoV (viz výše). KmV nemůže být v *l*-ovém přičestí diftong (srov. pravidelnou alternací *ou>u* u *l*-ových přičestí sloves patřících do II. třídy, která mají v *l*-ovém přičestí příslušný KmV). V češtině existují slovesa, jejichž KoV je diftong a zároveň se tvar *l*-ového přičestí tvoří od otevřeného kořene. Jde o slovesa III. třídy vzoru *krýt* (*obout, dout, kout, plout, zout, ...*), u nichž KoV diftong pravidelně alternuje v *l*-ovém přičestí s *u*. Tímto způsobem lze vyloučit např. propria typu *Soul* a cizí slova zakončená např. na *-aul, -ael* aj.

Obdobné postupy lze navrhnout pro usnadnění analýzy neoznačkových korpusů, popř. tehdy, chceme-li postupovat bez použití morfologických značek. Zároveň je ale patrné, že analogické lingvistické úvahy nám mohou posloužit při vyhledávání kandidátů na určitý derivační typ, přičemž nasnadě se jeví aplikovat je a prověřit vhodnost navržené metody pro deverbativa (viz níže).

### **Formulace dalších otázek vypluvších ze zkoumání daného jevu**

Lze předpokládat, že existují slovesné tvary, u nichž bude možné uplatnit analogické postupy. U kterých slovesných tvarů by mohlo být uplatnění analogických postupů snazší a u kterých obtížnější?

### **Zadání cvičení, v nichž lze uplatnit analogické postupy**

Vyhledejte podobným způsobem kandidáty na tvary infinitivu. Uplatněte podobný postup při vyhledávání názvů osob (vč. proprií) tvořených konverzí z minulých přičestí (typ *kutil, čumil, Málodělal, Bral, Musil, Drbal, ...*) (srv. též Osolsobě 2011<sup>1</sup> : 61n ). Všimněte si chyb ve značkování.

## *Syntetické futurum v češtině*

### **Motivační úvod**

Termínem syntetické futurum se v českých gramatikách označují tvary několika desítek českých nedokonavých sloves tvořených od tvarů indikativu přítomnosti aktiva prefixem *po-* (*pů-*) většinou paralelně ke tvarům opisného futura (kromě *jet*, *jít*). Tato slovesa jednotlivé mluvnické a monografické práce 1) významově vymezují a 2) uvádějí jejich výčty. Ve slovnících jsou zachyceny nejednotně. *SSJČ* uvádí celkem 48 sloves (*SSČ* pouze 21), která mohou tvořit tvary syntetického futura. V gramatikách najdeme stručné výčty a v odborné literatuře (srv. Čermák 1990 : 160, Kopečný 1962) výčty rozsáhlejší.

### **Nastínění problému**

Pokud připustíme, že se jedná o otevřenou skupinu sloves, pak je nanejvýš pravděpodobné, že korpusy jakožto velká data mohou přispět k doplnění obrazu o tomto způsobu tvoření budoucího času: a) může v nich být doloženo užití těchto tvarů od sloves, u nichž slovníky tvoření neuvádějí, b) lze zkoumat na základě rozsáhlejší materiálové základny, o jaká slovesa jde (běžně se uvádí, že jde o slovesa pohybu, nebo taková, která pohyb implikují, a navíc, že jde o pohyb určený směrem), c) s ohledem na to, že jde (kromě sloves *jít*, *jet*) o vyjádření budoucího času formou syntetickou paralelně s formou analytickou, můžeme v korpusech zkoumat distribuci (frekvence, významové odstíny) obou paralelních možností.

### **Otázky**

Nejdříve je třeba zjistit, zda jsou v korpusu označkovány tvary syntetického futura tak, aby je bylo možné vyhledávat pomocí morfologické značky. Z popisu morfologických značek zjistíme, že ke gramatickým významům kategorie času se vztahuje 2. a 9. pozice morfologické značky, konkrétně k futuru na 2. pozici značka *B* (sloveso, tvar přítomného, nebo budoucího času) a na 9. pozici značka *F* (futurum (budoucí čas)). Dále je třeba zjistit, kolik a jakých lemat/slovních tvarů syntetického futura lze vyhledat pomocí morfologické značky. Další otázky mohou vyplynout z tohoto šetření.

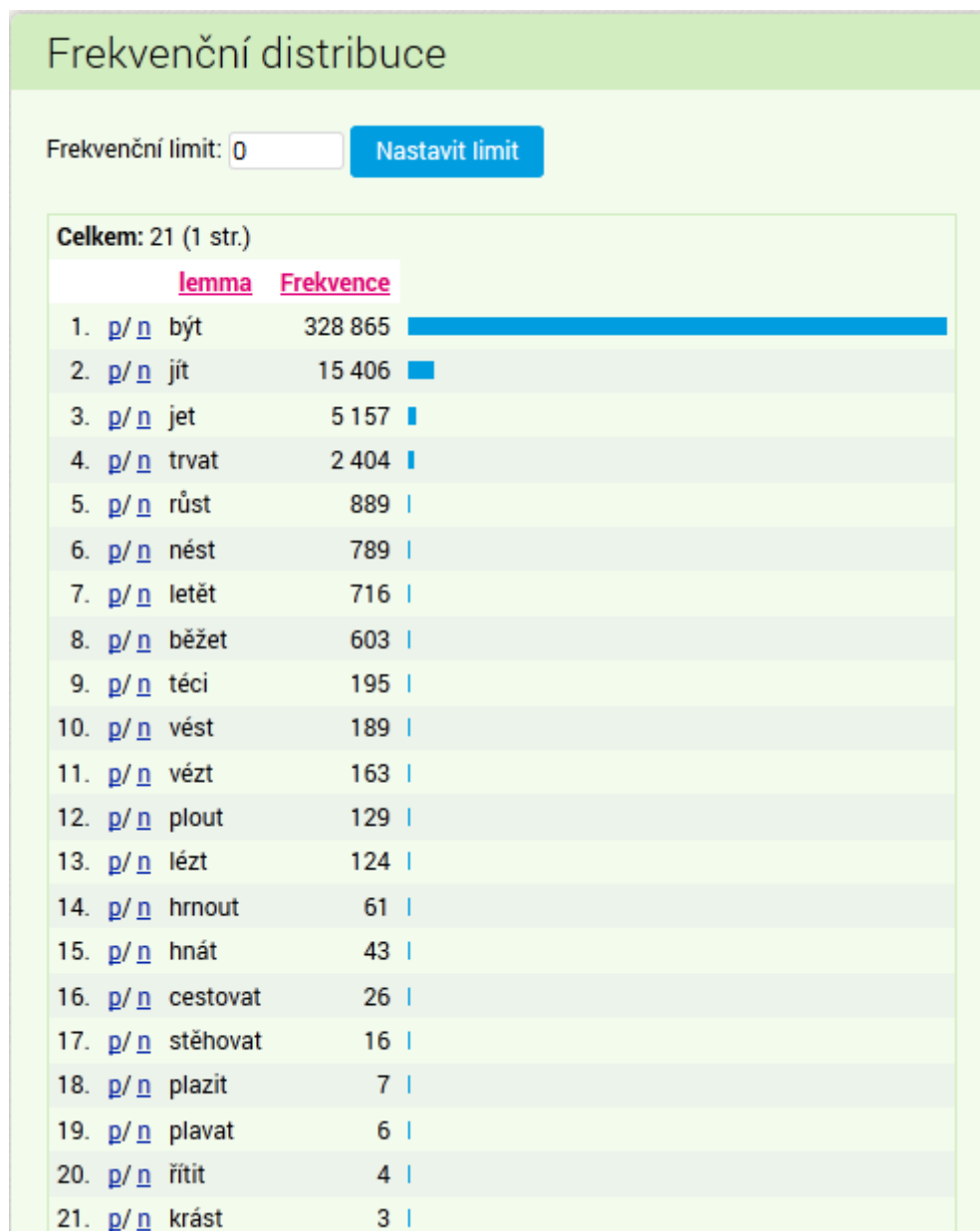
### **Formulace dotazu**

Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[tag="VB.....F.\*"]*.

Kromě tvarů *budu*, *budeš*, ... si všimneme tvarů slovesa *jít*.

u svaté Anny v Brně . Stamilióny korun sem ale <b>půjdu</b> /jit/VB-P---3F-AA---I	v příštích letech , kdy se začne stavět Mezinárodní centrum
přece ! šeptá Bridel a klesá na kolena , už <b>nepůjde</b> /jit/VB-S---3F-NA---I	za Šimonem , modlí se , po modlitbě , pokorný
kombinace užití : Odpoledne půjdu do hospody/na pivo/k Vejvodům . <b>Půjdu</b> /jit/VB-S---1F-AA---I	do kina Blaník/do Blaníku . 3 . Dále je třeba
úsměvu . " Ale steward se zřejmě někam ztratil . <b>Půjdu</b> /jit/VB-S---1F-AA---I	se po něm podívat , " slíbil jí a zmizel
tatínek . " To je tedy tvůj problém . Já <b>půjdu</b> /jit/VB-S---1F-AA---I	příkladem . Mně nikdo nemusí říkat , kde jsou lihoviny
že byla nezávislá ona . " Zešílím , jestli tam <b>půjdu</b> /jit/VB-S---1F-AA---I	. V životě jsem neslyšel takovou blbost . Stejně mám

Podíváme se na frekvenční distribuci lemmat.



Projdeme-li tento seznam, jistě nás napadne, že není úplný. Není pravděpodobné, že by se v korpusu některá lemmata nevyskytovala. Zřejmě se jedná o otevřenou třídu sloves, takže ve slovníku automatického morfologického analyzátoru nejsou zaznamenána všechna slovesa, která mohou v češtině takto tvořit futurum. Naopak ve slovníku automatického morfologického analyzátoru jsou zaznamenána slovesa, která jsou tvořena prefixem *po-* a

mají s námi sledovaným typem sloves homonymní některé tvary od kmene přítomného. Vyjdeme z předpokladu, že u řady tvarů s prefixem *po-* se automaticky uvádí lemma s tímto prefixem, přičemž se nerozlišují případy, kdy jde o tvarotvorný prefix syntetického futura a kdy jde o slovotvorný prefix. Z toho důvodu chybí ovšem také rozlišení na úrovni morfologické značky (na 9 pozici není vyznačeno, že jde o tvar futura, je zde uvedeno P). Podívejme se, jak můžeme testovat, kolik a jakých tvarů syntetického futura se v korpusu vyskytuje bez odpovídajícího tagu. Nejdříve se pokusíme vzpomenout si na nějaké sloveso, které používáme ve tvaru syntetického futura a které zároveň chybí v uvedeném seznamu. Vezměme např. sloveso „*valit (se) někam*“ ve významu „*pospíchat, rychle spět někam*“, které může mít tvary syntetického futura (asi připustíme, že je možné říci obojí, tedy že „*se při záplavách vody povalí/ budou valit na Prahu*“), ty jsou ovšem homonymní s tvary od přítomného kmene slovesa „*povalit koho/co*“. Jak budeme postupovat? Budeme hledat tvary slovesa *povalím, povalíš, ...*. Můžeme vyjít z předpokladu, že některé tyto tvary budou a) chybně lemmatizovány (lemma *povalit*), a b) budou tudíž i mylně označovány (*VB.....P.\** nikoli *VB.....F.\**). Jak budeme postupovat?

Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[lemma="povalit" & tag="VB.....P.\*"]*. Výsledkem je konkordance, jejíž část vidíme na následujícím obrázku.

z množství lidí, kteří se přes Polsko a Slovensko	<b>povalí</b> /povalit/VB-S---3P-AA---P	nelegálně na Západ po rozšíření EU, se nenaplnily.
náručě větví, povalené kmeny, nejsou - li,	<b>povalí</b> /povalit/VB-S---3P-AA---P	se stojící . Prostor mezi novou, moderní technikou a
tváře se jí nahrmla krev . Co teď udělá ?	<b>Povalí</b> /povalit/VB-P---3P-AA---P	ji a rozřeže , jak to navrhoval ten chlap v
, tající snih bude kapat ze stromů a strouhami se	<b>povalí</b> /povalit/VB-S---3P-AA---P	voda . Kdyby nebylo dlouhých nocí , člověk by myslel
brány ani ve starém tunelu ! Možná že se tu	<b>povalí</b> /povalit/VB-S---3P-AA---P	voda - a chvíli to bude pořádně špinavá voda ,
, které zvažovaly nepodpořit rozpočet . Obě potvrdily, že	<b>nepovalí</b> /povalit/VB-S---3P-NA---P	vládu při hlasování o nedůvěře , ale o rozpočtu se
od něj dětinství se domnívat , že pomocí několika kamarádů	<b>povalí</b> /povalit/VB-S---3P-AA---P	" to Rakousko zpuchřelé " . Zatím přišla do vinárny
se Arly , tak sebou radši hoď , nebo se	<b>povalíš</b> /povalit/VB-S---2P-AA---P	tou řekou jako oni . Když se Bert nepohnul ,
. jak s tím může žít ? ! ) a	<b>povalí</b> /povalit/VB-S---3P-AA---P	mě na záda . Seberu v sobě veškerou sílu ,
!! Marilyn Monroe !!! Chytne Olgu a	<b>povalí</b> /povalit/VB-S---3P-AA---P	ji na postel . TMA . obraz . Olga a
. Při dnešním hlasování o nedůvěře sice unionisté Grossův kabinet	<b>nepovalí</b> /povalit/VB-S---3P-NA---P	, ale poté je čeká ostrá debata o účasti v
vyškřábat oči , ale protože je zakovaná v řetězech ,	<b>povalíš</b> /povalit/VB-S---2P-AA---P	ji prostě na zem a . . - Ale já
připlíží až těsně ke mně , skočím po něm ,	<b>povalím</b> /povalit/VB-S---1P-AA---P	ho na zem a sednu si na něj . Začne
předehře vůbec neuvažovala , protože předpokládala , že ji prostě	<b>povalí</b> /povalit/VB-S---3P-AA---P	, a až bude hotový , zase se z ní
" Někam spadnu . Do něčeho vrazím . Někdo mě	<b>povalí</b> /povalit/VB-S---3P-AA---P	. " Ne , " řekl Ron . "
. Vrhá se k Pelcovi a ten ho jedním rozmachem	<b>povalí</b> /povalit/VB-S---3P-AA---P	do sněhu . Puška mu vypadne z ruky a zaboří

Chyby vidíme na prvním, čtvrtém, pátém i osmém řádku, neboť se jedná o tvary lemmatu *valit* v syntetickém futuru.

Nyní stručně ukážeme postup hledání kandidátů syntetického futura. Namátkovými sondami jsme zjistili, že v lemmatizovaných korpusech jsou tvary, které jsou/mohou být tvary syntetického futura lemmatizovány tak, že lemma je skutečné nebo utvořitelné sloveso se



slovotvorným prefixem *po-*. Těto lemmatizaci odpovídá značkování, tedy hodnota P na pozici 9. Podívejme se např. na tvary lemmatu *pořčet*.


bych mohla jít pracovat do školství . Tak zítra zase <b>pořčím</b> /pořčet/VB-S---1P-AA---P do Prostějova . Začínám si uvědomovat stereotyp prac. zařazení .
takříkajíc lepší start . No, Quinne, tak já <b>pořčím</b> /pořčet/VB-S---1P-AA---P . " Počkej , " řekl Quinn naléhavě .
Row a dokonce vozičkářům - Potom si koupí mercedes a <b>pořčí</b> /pořčet/VB-S---3P-AA---P si to do Mexico City po dálnici El-Paso , "
12 . září hostí Vítkovice ( 17.30 ) . Veteráni <b>pořčí</b> /pořčet/VB-P---3P-AA---P S kopce a do kopce VETERÁNI Přehlídka nablýskaných automobilových a
Tak mám těžkej dojem , že ke konci příštího týdne <b>pořčím</b> /pořčet/VB-S---1P-AA---P do Černý Hory , aby mi nezlenivěly zadní svaly .
Mži , Pattonův most projde generálkou , do Třemošné se <b>pořčí</b> /pořčet/VB-S---3P-AA---P po čtyřproudové výpadovce , na Karlovarské přibude další pruh ,
vystoupení mohou užít od 17.30 hodin . ZÁBAVA Spanilá jízda <b>pořčí</b> /pořčet/VB-S---3P-AA---P v Kladruzech Městský úřad Kladruby , kulturně školská komise ,
ně dohlížet . " Já vám podám flintu a <b>pořčím</b> /pořčet/VB-S---1P-AA---P na ty brambory . Kluci přece rádi bramborovej salát ,

Vidíme, že ve všech případech je chybná lemmatizace, a tedy i značkování.

Jak budeme hledat kandidáty obdobných sloves s chybnou lemmatizací?


Nejdříve je třeba vytvořit seznam lemmat začínajících na *po-* slovesných tvarů označovaných jako přítomný tvar (*VB.....P.\**), popř. *X.\** (neznámý tvar). V menu **Korpus** zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[lemma="po.\*" & tag="(VB.\*)|(X.\*)"]* (chceme získat přehled o přítomných tvarech sloves, jejichž lemma mají na začátku řetězec *po*). Podíváme se na frekvenční distribuci lemmat.

# Frekvenční distribuce

strana 1 

Frekvenční limit:  [Nastavit limit](#)

Celkem: 7368 (148 str.)

	lemma	Frekvence	
1.	<a href="#">p/ n</a> potřebovat	24 900	
2.	<a href="#">p/ n</a> používat	12 300	
3.	<a href="#">p/ n</a> považovat	11 725	
4.	<a href="#">p/ n</a> počítat	8 286	
5.	<a href="#">p/ n</a> pomoci	8 085	
6.	<a href="#">p/ n</a> pomáhat	6 878	
7.	<a href="#">p/ n</a> pocházet	6 819	
8.	<a href="#">p/ n</a> pohybovat	6 794	
9.	<a href="#">p/ n</a> pokračovat	6 363	
10.	<a href="#">p/ n</a> poskytovat	5 929	
11.	<a href="#">p/ n</a> podařit	5 739	
12.	<a href="#">p/ n</a> poznat	4 804	
13.	<a href="#">p/ n</a> popisovat	4 709	
14.	<a href="#">p/ n</a> podporovat	4 681	
15.	<a href="#">p/ n</a> povídat	4 395	
16.	<a href="#">p/ n</a> podávat	4 113	
17.	<a href="#">p/ n</a> <b>pověst</b>	<b>4 060</b>	
18.	<a href="#">p/ n</a> podívat	4 030	
19.	<a href="#">p/ n</a> potvrzovat	3 880	
20.	<a href="#">p/ n</a> pořádat	3 828	

Všimněme si lemmatu *pověst* (kliknutím na **p** získáme seznam konkordančních řádků a mnoha chyb v lemmatizaci slovesa *vést*).

ale zkušený "piráti" už tušili, že plavba	<b>povede</b> /pověst/VB-S---3P-AA---P	kolem východofríských ostrovů . Samotnou akci únosu proto načasovali na
ještě nikdy nezažila tak slunečné a zářivé jaro . Kam	<b>povedou</b> /pověst/VB-P---3P-AA---P	mé další kroky ? Snad bude lepší to ani nevědět
tak dlouhá , aby to nevyhrál mansaft , kterému se	<b>povede</b> /pověst/VB-S---3P-AA---P	začátek . Nejdřív se musíme do play-off dostat . A
juniori . Vzal také lovosického Jelinka . " Tým teď	<b>povede</b> /pověst/VB-S---3P-AA---P	Venca Toman . Pauza je 14 dní , pak nás
Kerstin bude hodná a tak se jim v životě dobře	<b>povede</b> /pověst/VB-S---3P-AA---P	. Proč jsme na to , Kerstin , přistoupili ?
rozhodnutí soudu o regulérnosti voleb . Pokud žádá ze stížností	<b>nepovede</b> /pověst/VB-S---3P-AA---P	k vyhlášení nových voleb v Mostě , musí Bártl svolat
Dobry večer , žáci : k vidině vrcholů vás napříště	<b>povede</b> /pověst/VB-S---3P-AA---P	někdo jiný než já . Dobry večer , budoucí učitelé
asi ani sám moc nevěřil tomu , že se mu	<b>povede</b> /pověst/VB-S---3P-AA---P	přimět ČSSD , aby jeho trojkoaliční vládu tolerovala . V
účelové jednání operátora s významnou tržní silou , které dočasně	<b>povede</b> /pověst/VB-S---3P-AA---P	ke zdání rovnováhy na trhu , následnému zrušení jeho povinností
, že nás mladí navštěvují , a snad se to	<b>povede</b> /pověst/VB-S---3P-AA---P	i s dětským domovem , " uvedl ředitel domova pro
akce čeká několik zajímavostí . " Poměrně velká část trasy	<b>povede</b> /pověst/VB-S---3P-AA---P	po naučné stezce Paměti Vitorazska , " řekl Michal Jarolímek

Ručně procházíme poměrně dlouhý seznam slovesných lemmat. Musíme zapojit jazykovou představivost a ptát se, u kterých z nich by mohlo jít o hledaný případ nesprávné lemmatizace.

Nalezené příklady si zapisujeme. Všimneme si mj. (viz níže) lemmat *pokvést* a *pomazat* (kliknutím na **p** získáme seznam konkordančních řádků a můžeme sledovat chyby v lemmatizaci a značkování).

234. p/ n pokvést	112
" Ale at je to v únoru , až <b>pokvetou</b> /pokvést/VB-P---3P-AA---P mimózy , " řekla lehkomyšlně , protože věděla , že	
změnila v Hrabalově Majitelce hutí Jarmilčinu říznou odpověď " Až <b>pokvetou</b> /pokvést/VB-P---3P-AA---P hovna " na učesané " Až pokvete kravinec " ;	
, loňské jednoleté výhony silně zkrátíme . Druhy , které <b>pokvetou</b> /pokvést/VB-P---3P-AA---P časně zjara na jednoletých výhonech ( např. zlatice ) ,	
na nevyzpytatelnost letošního počasí je těžké odhadovat , co kdy <b>pokvete</b> /pokvést/VB-S---3P-AA---P . V každém případě na návštěvníky této části zahrady čeká	
pěstovat s minimálním řezem . Po hlubokém řezu začátkem léta <b>nepokvetou</b> /pokvést/VB-P---3P-NA---P , ale koncem léta se obalí květem . C. viticella	
hlíně hnit Tak jako nyní srdce hníjí Z té strany <b>pokvete</b> /pokvést/VB-S---3P-AA---P falešná růže světa Anebo věčný nářek z oné strany Kdo	
základě mohla postavit náramně úspěšnou reklamní kampaň . Obchody jenom <b>pokvetou</b> /pokvést/VB-P---3P-AA---P . Lidské vztahy jsou pro společnost tou nejzákladnější a nejbrutálnější	
ale rostlina rychle regeneruje a vytvoří bujně výhony , které <b>pokvete</b> /pokvést/VB-P---3P-AA---P v dalším roce . B. madagascariensis Vitální stálezelený keř ,	
odpověď " Až pokvetou hovna " na učesané " Až <b>pokvete</b> /pokvést/VB-S---3P-AA---P kravinec " ; usilovně hledal neliterátský styl pro své mimořádné	
Co teď ? V Němci prolezlém Buenos Aires mi pšenice <b>nepokvete</b> /pokvést/VB-S---3P-NA---P . Proto přijímám lodnické místo na dvojtěžníku Olaf , plující	

310. p/ n pomazat	58
Vy a já , " prohlásil šerif , " teď <b>pomažeme</b> /pomazat/VB-P---1P-AA---P do hotelu . Máte tu vůz ? " Odvětil jsem	
Houska se nakrájejí na plátky a namočí v mléce . <b>Pomaží</b> /pomazat/VB-P---3P-AA--1P se zavařeninou a vloží do vymazané formy . První vrstva	
svůj končíř . . . . Tou mastí kord svůj <b>pomaží</b> /pomazat/VB-S---1P-AA--1P , aby sebelehčí zásah znamenal jistotu smrt . " Měření	
. Každý muž dostane svůj díl a všichni si pak <b>pomažou</b> /pomazat/VB-P---3P-AA---P těla tukem . Následuje kremace nebožtíka . Po kremaci se	
. Podle chuti osolíme a opepříme . Jednotlivé plátky sýra <b>pomažeme</b> /pomazat/VB-P---1P-AA---P připravenou pomazánkovou směsí a stočíme je . Podáváme s čerstvým	
a na pomoučném plechu do zlatohněda upečou . Vychladí se <b>pomažou</b> /pomazat/VB-P---3P-AA---P čokoládou , která se povarí s 2 - 3 lžícema	
! S velkejma zvířatama ! Jestli s tím nepřestanete , <b>pomažete</b> /pomazat/VB-P---2P-AA---P až do Jižní Karolíny než se nějaká dobročinná kuchyně nad	
, stal by se impotentním . V den obřizky dívce <b>pomažou</b> /pomazat/VB-P---3P-AA---P hlavu a korále červenou hlínou smíchanou s máslem . Žena	

Zatímco u *pokvést* můžeme předpokládat, že jde pouze o mechanicky utvořený infinitiv, sloveso *pomazat* tvořené prefixem *po-* má prezntní tvary homonymní se syntetickými tvary slovesa *mazat* s významem rychle se někam pohybovat (první a předposlední řádek výše).

### Třídění a pozorování dat získaných z korpusů

Podívejme se na konkordance (viz výše). Vidíme, že v případě slovesných tvarů s lemmatem *pokvést* je lemmatizace (tudíž i značkování) chybná ve všech případech. V případě lemmatu *pomazat* jsou chyby vyznačeny (viz výše). Projdeme-li všechny konkordanční řádky, zjistíme, že z celkového počtu 58 výskytů se v 11 případech jedná o tvary syntetického futura slovesa *mazat*.

### Další otázky

Pokuste se v různých korpusech najít co nejvíce dalších dokladů vyjádření futura gramatickým prefixem *po-*.

Najděte ke slovesu *mazat* případy vyjádření futura opisnou (analytickou) formou (*bude + infinitiv*) a porovnejte frekvenci (11 : ??).

## Formulace dotazu pro získání dat z korpusů

Seznamy sloves, jejichž lemma začíná na *po* a která jsou označována značkou *VB.\**, je třeba projít ručně.

Nyní naznačme, jak lze postupovat, hledáme-li tvary analytického futura slovesa *mazat*, abychom mezi nimi mohli hledat případy užití ve významech, které mohou vyjádřit futurum také synteticky. Postup není úplně jednoduchý. Nejdříve vyhledáme všechny tvary slovesa *mazat* v infinitivu.

Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[lemma="mazat" & tag="Vf.\*"]*.

V dalším kroku zvolíme filtr (pozitivní), nastavíme rozsah hledání <-5,5> (lze nastavit i větší), **Typ dotazu cql** a do dotazovacího řádku napíšeme dotaz ve formě *[lemma="být" & tag="VB.....F.\*"] within <s id="."/.\*>* (vyhledáváme v rámci jedné věty).

v Novigradě . " " Ludé ti **budu** dveře hnojem **mazat** /mazat/Vf-----A----J . " řekl Yarpen . " Okna ti kamením rozbíjat lidu předvést , že on se s nějakými sjetými tanečníky **mazat** /mazat/Vf-----A----J **nebude** . Pronajatý pozemek byl pro CzechTek 2005 na louce dobrovolníků má chuť se zapojit , kdo **bude** v parku **mazat** /mazat/Vf-----A----J chleby " férovou " čokoládovou pomazánkou , kdo **bude** rozdávat . " Jestli si myslíš , že mně tady **budeš** **mazat** /mazat/Vf-----A----J med kolem huby a že se nechám obměkčit a odkážu šesti měsíců nepodaří uspět , **budu** se muset spakovat a **mazat** /mazat/Vf-----A----J zpátky na sever . " D ' Alessandro se odmíchl , je to přece povinnost , ví , že **bude** **mazat** /mazat/Vf-----A----J na plonkovou desítku , což ovšem Viktor netuší a do postupně pročistí a ubude příspěvků , které **budou** administrátoři muset **mazat** /mazat/Vf-----A----J , " říká šéfredaktor iDnes . cz Michal Hanák na Nechci to poslouchat ! A **nebudu** to poslouchat ! Koukejte **mazat** /mazat/Vf-----A----J a pusťte se do Barryho Collinse . Teď hned . a nabrat sil . Těším se , až **budu** večer **mazat** /mazat/Vf-----A----J karty nebo až se u mě na vrátnici sejde parta obměňovat způsoby , kterými s gumou pracujete . Jestliže **budete** **mazat** /mazat/Vf-----A----J gumou vždy ve stejném směru , mohou vaše kresby vyhlížet dojde i na skutečné hrdinství . Ano , **nebudeme** si **mazat** /mazat/Vf-----A----J včelí produkt kolem úst , i tady se objevuje klasický

## Formulace závěrů

U slovesa *mazat* ve významu pohybu se vyskytují výhradně tvary syntetického futura. U dvou výskytů analytického futura slovesa *mazat* se domníváme, že by také bylo možné použít futura syntetického. Jde o význam hrát karty (*až budu večer <mazat> karty : až večer <pomažu> karty*) a zabývat se něčím nepříjemným (*on se s nějakými stejnými tanečníky <mazat nebude> : on se s nějakými stejnými tanečníky <nepomaže>*).

Upozorňujeme dále na to, že pokud je v přísudku modální sloveso, pak se tvar pomocného slovesa *být* váže k tomuto slovesu, nikoli ke slovesu významovému. Není tudíž možné počítat případy jako ... *budu se muset spakovat a <mazat>*....

Závěry pozorování můžeme zanáest do tabulky.

	syntetické futurum	analytické futurum
mazat	11	2

### Formulace dalších otázek vyplynuvších ze zkoumání daného jevu

Podívejte se na další gramatické vlastnosti sloves, která mají tvary syntetického futura homonymní s tvary sloves tvořených prefixem *po-*, a pokuste se zamyslet nad tím, zda by se jich nedalo nějak použít při třídění dokladů.

V případech lemmat jako *pokvést*, *pořčet*, *potrvat*, ... (předpokládáme o nich, že nejsou v úzu) zkontrolujte, zda se v korpusech opravdu nevyskytují tvary infinitivu a l-ových příčestí. Pokud se nějaké takové tvary vyskytnou, popište jejich užití a významy.

### Zadání cvičení, v nichž lze uplatnit analogické postupy<sup>29</sup>

Vyberte ze seznamu kandidátů další slovesa a naznačený postup pro ověření poměru tvarů syntetické : analytické futurum zopakujte.

Doklady syntetického futura můžeme hledat i v neoznačovaných korpusech (mluvené korpuse, korpus soukromé korespondence), hledání bude ovšem trochu složitější. Jak budeme postupovat? Chceme najít slovní tvary, které začínají řetězcem *po* a končí řetězcem *[uielášm]*. V nabídce korpusů zvolme např. korpus ORAL2008. Zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme *po.\*[uielášm]* Podívejme se na frekvenční seznam tvarů. Doklady nechybějí.

to tam ...: dyť už je nemožnej , kampa tam **poleze** . no , to je dobře , eště dyž to  
, ... --- . ... máme upravenou mouchu , která **poleze** po vašem obličejí . ne , vona byla spíš ---  
ňák tam prostě to . a pak právě , že **poleze** třeba NJ . a pak jako dolu , že já  
že prostě to tak jakože našel a že , že **poleze** první , protože NN byl eště na tamtý skále jiný

Možná jste už někdy slyšeli vulgární formulaci v tom smyslu, že ... *se někdo s něčím po...* , tedy že *někdo bude něco dělat příliš pomalu*. Další významovou nuancí vulgarismu v syntetickém futurum najdeme v mluveném korpuse ORAL2008.

nás to je jiný , to zajíždí do Biliny a **posere** se ke mně a já budu na myslivně nebo vobráčeně

Pátrejte po dalších kandidátech v dalších mluvených korpusech a v Korpusu soukromé korespondence.

Je pravda, že slovesa schopná tvořit synteticky futurum mají vždy pouze význam pohybu určeného směrem, nebo lze doložit i další významové nuance?

Jak to, že v korpusech můžeme najít *<bude jet>* a *<bude jít>*?

<sup>29</sup> Více Osolsobě 2007<sup>3</sup>, 2008<sup>1</sup>.

## ***Jakou osobu signalizují tvary by?***

### **Motivační úvod**

Motto: *Aby ses nezbláznila!*

Tabulkové popisy pomocného tvaru pro tvoření kondicionálu v češtině v naprosté většině českých mluvnic včetně popisu v „korpusové“ *Mluvnicí současné češtiny* vypadají zhruba takto:

1. osoba	<i>bych</i>	<i>bychom</i>
2. osoba	<i>bys</i>	<i>byste</i>
3. osoba	<i>by</i>	<i>by</i>

Porovnáme-li morfologické značkování tvarů *by* v korpusech řady SYN, narazíme na to, že v korpusu SYN2000 mají všechny tvary *by* na 8. pozici osoba hodnotu 3 (3. osoba). V korpusech SYN2005 a SYN2006PUB mají tvary *by* na 8. pozici osoba hodnotu – (neurčuje se). V korpusech SYN2009PUB a SYN2010 mají tvary *by* na 8. pozici osoba hodnotu buď 3 (3. osoba), anebo – (neurčuje se).

Mohou tvary *by* implikovat, že jde o jinou než třetí osobu?

### **Nastínění problému**

Z výše uvedeného je patrné, že tabulkové popisy tvarů *by* v mluvnicích a značkování korpusu SYN2000 nebere v úvahu případy, kdy se tvar *by* vyskytuje nejen jako pomocný tvar při tvoření složených slovesných tvarů třetí osoby, ale kdy je za určitých podmínek jedním z tvarů, jimiž se tvoří složený tvar osoby druhé.

### **Otázky**

Které tvary mají v korpusech na 8. pozici osoba hodnotu 2 (2. osoba)?

Zvolíme **Typ dotazu cql**, do dotazovacího řádku napíšeme dotaz ve formě *[tag=".....2.\*"]*

Podíváme se na frekvenční distribuci slovních druhů. V dalším kroku zvolíme filtr (negativní) a do dotazovacího řádku napíšeme dotaz ve formě *V.\**. Podíváme se na frekvenční distribuci lemmat a slovních druhů.

## Hledat v korpusu

Korpus:

syn2010

Typ dotazu:

CQL

CQL:

[vložit tag](#) [vložit "within"](#)  
[tag=".....2.\*."]

## Víceúrovňová frekvenční distribuce

Frekvenční limit: 0

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>?</sup>	
1.	pos	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	-
<a href="#">Přidat úroveň</a>					

[Vytvořit frekvenční seznam](#)

## Filtr konkordancí

Filtr:

pozitivní  negativní

Vybraný token:

první  poslední

Rozsah hledání:

od 0 do 0  včetně KWIC

Typ dotazu:

CQL

CQL:

[vložit tag](#) [vložit "within"](#)  
[tag="V.\*"]

## Víceúrovňová frekvenční distribuce

Frekvenční limit:

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>?</sup>	
1.	lemma	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	-
2.	pos	<input type="checkbox"/>	KWIC	slova KWIC nejvíce vlevo	X

Přidat úroveň




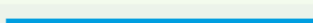






Vytvořit frekvenční seznam

## Frekvenční distribuce

Frekvenční limit:

Nastavit limit

Celkem: 10 (1 str.)

	lemma	pos	Frekvence	
1.	p/ n váš	P	3 857	
2.	p/ n vy	P	3 074	
3.	p/ n ty	P	2 502	
4.	p/ n tvůj	P	1 968	
5.	p/ n se	P	128	
6.	p/ n aby	J	77	
7.	p/ n kdyby	J	47	
8.	p/ n co	P	12	
9.	p/ n že	J	9	
10.	p/ n kdos	P	4	

Pozorujeme, že kromě slovesných tvarů se ve značkování korpusu objevuje na 8. pozici osoba hodnota 2 (2. osoba) u tvarů osobních a přivlastňovacích zájmen 2. osoby a u tvarů lemmat *aby, se, kdyby, že, co, kdos*.

Pozorujte v korpusu SYN2010 tvary lemmat, která mají na 8. pozici osoba hodnotu – (neurčuje se). Pokuste se mezi nimi najít příklady, kdy se nejedná o 3. osobu. Kde všude se může vyskytovat volný morfém *-s* signalizující 2. osobu v případě, že jde o tvar slovesa v kondicionálu, a jak lze doklady najít v korpusech?



## Formulace dotazu pro získání dat z korpusů

Zvolíme **Typ dotazu cql**, do dotazovacího řádku napíšeme dotaz ve formě `[lemma="být" & tag="Vc.....-.*"]`.

### Hledat v korpusu

**Korpus:**

**Typ dotazu:**

**CQL:**

[vložit tag](#) [vložit "within"](#)

Zvolíme filtr (pozitivní), rozsah hledání  $\langle 1,1 \rangle$ , **Typ dotazu slovní tvar** a do dotazovacího řádku napíšeme dotaz ve formě `s[ei]s`.

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**

**Dotaz:**

[Filtrovat konkordanci](#) [Vymazat formulář](#)

ale i tehle postoj vypovídá o něčem , nad čím **by** /být/Vc-----| **ses** asi měl zamyslet . " Alexis se obrátila k je to správné , člověk potřebuje především pořádek , jak **by** /být/Vc-----| **ses** na to díval , kdybch ti já začala stávkovat kam si chodí užívat . Mám nahrávku , ale asi **by** /být/Vc-----| **sis** ji nechtěl poslechnout . Vybral sis opravdu špatně , Tvá nová firma to všechno uhradí . A tři měsíce **by** /být/Vc-----| **ses** mohl jenom věnovat studiu jazyka . " " Já . " " Doufám . " " Noelle , ty **by** /být/Vc-----| **sis** měla najít útěchu . Můj otec by ti ji směrem do kraje , půjdu spát , Jindřichu . Radši **by** /být/Vc-----| **sis** měl také trochu odpočinout . Kdo ví , co , která nepozorovaně vklouzla do mého pokoje . " Neměla **by** /být/Vc-----| **ses** raději učít ? Víš , že to co máš ? Protože tuto lásku nemůžeš získat bez toho , že **by** /být/Vc-----| **sis** předtím neosvojl předešlé čtyři síly . Nemůžeš milovat , zda hra bouchne či ne . Když bouchne , vydělal **by** /být/Vc-----| **sis** pěkněj tolar . Úspěch by vedl k dalším podnikům otec útržkem novin následující citát : Žij tak , jako **by** /být/Vc-----| **ses** měl ihned rozloučiti se životem , jako by čas zazněla jako prosba . " No tak broučku , přece **by** /být/Vc-----| **ses** nezlobil , vždyť víš , že musím ke Kristýně bys byl pro nás na hovno kumpán a do Wyzimy **by** /být/Vc-----| **sis** mohl šlapat sám . Treba hned teď . " kde zní teď vlků vyťí do vzbouřeného žilobití v němž **by** /být/Vc-----| **ses** ztratila Tichý hrad též zaživa pohíbená vychladlý ohně žár , kdybys náhodou mluvil se sirem Thomasem Ullastonem , nemohl **by** /být/Vc-----| **ses** ho mezi řeči zeptat , jestli o té moji

## Třídění a pozorování dat získaných z korpusů

Nejdříve ověříme, zda tvar *s[ei]s* může stát na jiné pozici než <1,1> vzhledem k tvaru *by*. Zjistíme, že toto je jediná slovosledná varianta doložená ve zkoumaném korpusu.

Na otázku, zda jsou v korpusu doloženy případy, kdy zvrtné sloveso v kondicionálu má tvar složen z tvarů *bys* + *se/si* + l-ové přičestí významového slovesa, lze najít odpověď takto. V menu **Konkordance** zvolíme **Typ dotazu fráze**, do dotazovacího řádku napíšeme dotaz ve formě *bys s[ei]*. Podívejme se na jednotlivé řádky a označme případy, kdy nejde o předložku *se* (chyby v desambiguaci). Pomocí filtrů se můžeme pokusit odstranit případy, kdy jde s velkou mírou jistoty o *se* správně označované jako předložka. Zvolíme filtr (negativní), rozsah hledání <2,2>, **Typ dotazu cql** a do dotazovacího řádku napíšeme *[tag="....7.\*"]*. Pozorujeme jednotlivé konkordanční řádky.

raději odpustit prolog . Pokud se Hemingwaye týká , měl <b>bys</b> /být/Vc-S---2-----I <b>si</b> /se/P7-X3----- dát " Death in the Afternoon " a co "
poslal , tím spíš , kdybych ti psal , že <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- ukořenil ( opět vole ) . Frances mi posílá detektivky
se mně stalo , když jsem volala tatkově . Mohla <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- stavit o prázdninách , ale jak Tě znám , budu
byl a mluvil tváří v tvář , možná , že <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- domluvil pomocí zkušeností , psychologie , vzájemného vztahu hovořících ,
otočil . Kdyby byl proti ní méně zaujatý , jistě <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- mu líbila ; jako mnozí velcí lidé , měl i
to nikdy nebyl , tak jsem snad zaujatej . Ty <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- tam asi ukořenil ( od slova koření nikoli kořeny )
to nemělo mít ? Budeš to hrát dlouho ? Nechtěl <b>bys</b> /být/Vc-S---2-----I <b>si</b> /se/P7-X3----- zahrát v tom mém Falstaffovi třeba Lorda Chief of Justice
. Hrál jsi za Kutnou Horu a Čáslav , jak <b>bys</b> /být/Vc-S---2-----I <b>si</b> /se/P7-X3----- zhodnotil dosavadní průběh kariéry ? V Čáslavi je to o
bys mít velké , větší , největší uvítání , jak <b>bys</b> /být/Vc-S---2-----I <b>si</b> /se/P7-X3----- vybral . Mohl <b>bys</b> mít i tiskovou konferenci . Musil
, že dodnes to není do češtiny přeloženo . Měl <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- na to s Janou kouknout . V té adaptaci <b>by</b>
se rovnala hořkosti medicíny interní , v případě , že <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- uvolil . No a ve všech případech , ať onom
piští . A smutek vane těmi samotami , jak dával <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- ve vsi k spáleníšti , z kad ve vzduch trčí
ti závidim Velhartice - a že kdybys chtěl , moh <b>bys</b> /být/Vc-S---2-----I <b>si</b> /se/P7-X3----- klidně zajet do Sázavy . A taky ti to přeju
bych ti to přeložil rovnou z francýny . A pak <b>bys</b> /být/Vc-S---2-----I <b>si</b> /se/P7-X3----- to moh ještě ulíznout , aby ti to leželo osobně
vlastně soused v kapuci . Myslím si stále , že <b>bys</b> /být/Vc-S---2-----I <b>si</b> /se/P7-X3----- měl uplést nějaké filmování v Itálii , já bych tam
Dupontová <b>by</b> tě samozřejmě doprovázela . Jsem přesvědčená , že <b>bys</b> /být/Vc-S---2-----I <b>se</b> /se/P7-X4----- dozvěděla spoustu nového . " Marie - Christina rozpíhla paže

V menu **Konkordance** zvolíme **Typ dotazu fráze**, do dotazovacího řádku napíšeme dotaz ve formě *bys s[ei]s*.

ktej <b>by</b> se zajímal o šaty a módu : Mohla <b>bys</b> /být/Vc-S---2-----I <b>sis</b> /se/P7-X3--2----- obléknout ty hezké nové letní šaty . Ovšem , jestli
podělíme o postel . " Johne , myslím , že <b>bys</b> /být/Vc-S---2-----I <b>sis</b> /se/P7-X3--2----- měl zapnout televizi . . . hned ! " Chtěl
, že se pořádným vydrhnutím pupínků zbavíš , ale tím <b>bys</b> /být/Vc-S---2-----I <b>sis</b> /se/P7-X3--2----- pokožku jenom podráždila a vysušila . To <b>by</b> samozřejmě podnitilo
mohla hodit na letiště , že jo ? A rovnou <b>bys</b> /být/Vc-S---2-----I <b>sis</b> /se/P7-X3--2----- mohla nechat lincolna , protože až budeš pročešávat okolí ,
chromozom . Ale skutečné přátelství , to ano . Mohl <b>bys</b> /být/Vc-S---2-----I <b>ses</b> /se/P7-X4--2----- otočit na druhou stranu ? " Takže <b>by</b> koukal do
počkal trochu déle , řekněme tak půl hodiny , mohl <b>bys</b> /být/Vc-S---2-----I <b>sis</b> /se/P7-X3--2----- ječet , co hrdlo ráčí . " Jak to
záliby zamířily někam jinam . " " Možná , že <b>bys</b> /být/Vc-S---2-----I <b>ses</b> /se/P7-X4--2----- měl poohlédnout po někom , kdo je tě lady víc
na světě se pravděpodobně řídí svými vlastními zákony . Nejspíš <b>bys</b> /být/Vc-S---2-----I <b>ses</b> /se/P7-X4--2----- ukousal nudou , kdyby sis vzal mě . " Faith
. Je v tom jen jeden malinký háček . Měla <b>bys</b> /být/Vc-S---2-----I <b>sis</b> /se/P7-X3--2----- přestat přát , když se máš místo toho prát .
, je v tvém životě nějaká věc , kvůli které <b>bys</b> /být/Vc-S---2-----I <b>ses</b> /se/P7-X4--2----- vzdal úplně všeho ? Jedna jediná věc , kvůli níž
odsekla mu . " Je to strašně legrační . Kéž <b>bys</b> /být/Vc-S---2-----I <b>ses</b> /se/P7-X4--2----- viděl , jak jsi se tvářil , když s tebou
že mě probere až šulec mě milé ! ( Že <b>bys</b> /být/Vc-S---2-----I <b>ses</b> /se/P7-X4--2----- zvednul a pustil slečnu sednout ty ? ) Uff ,

Pokládali byste takováto užití za chybná? Překvapuje vás, že jich není více/méně?

Jak byste hodnotili uvedená užití kondicionálu zvrtných sloves doložená v korpusu SYN2010?

## Formulace závěrů

Z pozorování dat korpusu SYN2010 je patrné, že použitá desambiguace nebere v úvahu při desambiguaci významu osoby obvyklejší případy tvarů kondicionálu 2. osoby singuláru zvrtných sloves. Druhá osoba u zvrtných sloves je většinou (nikoli vždy) signalizována volným morfémem *-s* na tvaru zvrtného *se/si*. Ten stojí (téměř) vždy v pozici <1,1> od tvaru *by*, což odpovídá pravidlům postavení klitik v češtině. Kromě toho lze v korpusech pozorovat také nečetné doklady kondicionálu zvrtných sloves tvořených tvarem *bys se* zvrtným *se/si*, ale i *ses/sis*.

## Formulace dalších otázek vplynuvších ze zkoumání daného jevu<sup>30</sup>

Jak se chová volný morfém *-s* při tvoření tvarů 2. osoby minulého času indikativu aktiva? Existují v korpusech doklady, kdy se volný morfém *-s* vyskytuje v pasivu? Existují v korpusu doklady, kdy volný morfém *-s* zastupuje tvar slovesa *jsi* i v případě, že tvar *jsi* je sponové sloveso?

## Zadání cvičení, v nichž lze uplatnit analogické postupy

Zjistěte, v jakých typech textů se vyskytují řídké případy typu *bys se / bys si / bys ses / bys sis*. Použijte **Menu Možn. zobrazení** zvolte **Reference** (např. opus genre, postupujte dle manuálu ke korpusovému manažeru). Podívejte se, zda se v korpusech vyskytují doklady, kdy mezi tvarem *by* a tvary *s[ei]s* stojí nějaký další slovní tvar (... *ty by/bys* <.\*> *s[ei]s mi svěřila/mě vybrala ...*).

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

### 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 přítomnosti (futura) aktiva, imperativu a infinitivu aktiva 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

---

<sup>30</sup> Srv. Osolobě 2007<sup>1</sup> a též níže.

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.

### Nastínění problému<sup>31</sup>

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.).<sup>32</sup>

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*.<sup>33</sup>

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*.

<sup>31</sup> V tabulce nejsou zahrnuty varianty reflexiv tantum.

<sup>32</sup> 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 ...).

<sup>33</sup> Nebereme v úvahu ani archaismy typu (... on <jest> mi <uletěl> ..., ... Pirát Ockham <byl jest přepadl> naši flotilu ...).

; Ducha posvátné básně, a chraň si, <b>cos</b>	<b>přijal</b> /přijmout/VpMS---3R-AA---P	. " V krátkém komentáři O tajemstvích se Ambrož vrací
. ráno na kolejích, budu hlavně zvědavý, <b>cos</b>	<b>viděl</b> /vidět/VpMS---3R-AA---I	zájímavého přes prázdniny . @ měl bych ještě jednu prosbu
profilu se ho zeptala: " No, a <b>cos</b>	<b>dělal</b> /dělat/VpIS---3R-AA---I	ty ? " S šelmovskými rozzářenými očima si prohlížela jeho
krv a další stopy po ní. " <b>Cos</b>	<b>udělal</b> /udělat/VpMS---3R-AA---P	s jejím vozem ? " ptal se Shanahan . "
Kdo ukradl ve středu v noci ty vejce a <b>tos</b>	<b>udělal</b> /udělat/VpMS---3R-AA---P	zasejíc ty nebo tvůj táta mě za to určitě nepověsej
se tím, že vojáci neplatí účty ? <b>Tos</b> tím	<b>chtěl</b> /chtít/VpMS---3R-AA---I	říct ? Myslels tím, že zas přijde doba,
tohle jsem ukradl z tátovy skříňe. " <b>Tos</b>	<b>neměl</b> /mít/VpMS---3R-NA---I	, " řekla Alíka, " tvůj tatínek přece není
a lapající po vzduchu . - Tak vidíš, <b>cos</b>	<b>udělal</b> /udělat/VpMS---3R-AA---P	, syčel kovboj, - teď jsi spokojený ? -
Dobré odpoledne, pane Davide. " <b>Cos</b> to	<b>provedl</b> /provést/VpMS---3R-AA---P	? " Trochu jsem uklidil. " " Co
Samotný Moers čekal na Maigreta v salonu. " <b>Cos</b>	<b>našel</b> /najít/VpMS---3R-AA---P	? " Samozřejmě otisky, staré i nedávne .
tríčku s ušpiněnou baseballovou čepicí na hlavě. " <b>Tos</b>	<b>byl</b> /být/VpMS---3R-AA---I	ty, Ippestone ? " zeptal se dobráčky Warwick .
někdo vydával za detektiva z New York City. <b>Tos</b>	<b>byl</b> /být/VpIS---3R-AA---I	ty ! Včera v noci tam došlo k několika vraždám
" Tys na to přišla ? " <b>Tos</b> mě	<b>měl</b> /mít/VpMS---3R-AA---I	za tak blbou ? " " Takže vidíš náš cit
. Rozhněvaně jsem se podívala na Dennise. " <b>Tos</b> <b>fotografoval</b>	/fotografovat/VpMS---3R-AA---I	ty ? Měl jsi být u dětí, jenže když
žes byl snad v kómatu . Ptáš se, <b>cos</b>	<b>mohl</b> /moci/VpMS---3R-AA---I	udělat ? Nejjednodušší a často neúčinnější taktika je nechat ho
Koukám, že jsi pěkně nabroušená ! " <b>Tos</b>	<b>uhod</b> /uhodnout/VpMS---3R-AA---6P	, kamaráde . Nemáš tu nic, co by federálové
a zaplatil . Starostlivě přitom vraštil čelo. " <b>Tos</b>	<b>nemusel</b> /muset/VpMS---3R-NA---I	, Jime, " řekl jsem . " Vypij to
, žes ho před patnácti letama strkal někam, <b>kams</b>	<b>neměl</b> /mít/VpMS---3R-NA---I	. " Rozesmála se, ostře a překvapivě silně -
železné pelesti. " Kams dal můj šaty, <b>kams</b>	<b>dal</b> /dát/VpMS---3R-AA---P	mou kabelku, kams to všechno střílel, o cos
. To je můj ! mobil . Hýl : <b>Kams</b>	<b>volal</b> /volat/VpMS---3R-AA---I	? ! Prcek : ( směje se ) Budeš se
už ale sníš napadl, celá zem, <b>kams</b> až <b>dohlédl</b>	/dohlédnout/VpMS---3R-AA--1P	, jím byla pokrytá . Někdy sněžilo celou noc ,
Nádhera, nic lepšího už tě nemohlo napadnout . <b>Kams</b>	<b>dal</b> /dát/VpIS---3R-AA---P	proboha rozum ? Ale na lamentování už je pozdě ,
" Jo, ale - " " Tak <b>kams</b> ji	<b>dal</b> /dát/VpMS---3R-AA---P	? " " Nemáš ji ty ? " " Já
poslední jsi hodil pěkně blízko . " " <b>Kams</b> to	<b>dostal</b> /dostat/VpMS---3R-AA---P	, Tome ? " " Do několika míst . "
do hluboké propasti . A je ticho . Dav <b>kams</b>	<b>zmizel</b> /zmizet/VpIS---3R-AA---P	; na ulicích všední tváře, policajti, hříšné děvky
otloukala jeho hlavu o příčky železných pelesti . " <b>Kams</b>	<b>dal</b> /dát/VpIS---3R-AA---P	můj šaty, kams dal mou kabelku, kams to
: " Cože ? Cos napsal ? A <b>kams</b> to	<b>poslal</b> /poslat/VpMS---3R-AA---P	? " Odpověděl, že napsal na velitele stížnost straně

## Otázky

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

## 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 přičestí a ověřit, že s tvarem pomocného slovesa tvoří hledaný analytický tvar. Zvolíme filtr (pozitivní), rozsah

hledání, v němž chceme vyhledávat tvary l-ových přičestí, **Typ dotazu cql** a do dotazovacího řádku zapíšeme dotaz ve formě  $[tag="V[pq].....[12].*"]$ .<sup>34</sup>

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.

## Třídění a pozorování dat získaných z korpusů

Nejdříve nastavíme rozsah hledání  $\langle -1, -1 \rangle$ .

psaní Smyslu života , na čas jsme přestali , <b>rozešli jsme</b> /být/VB-P---1P-AA---I se a dělali na svých vlastních projektech , znovu se
Tak jakým právem mě tu zadržujete ? " <b>Přišla jsi</b> /být/VB-S---2P-AA---I sama , " potvrdil vážně . " Ne však z
? Věděl jsem , že se něco děje . <b>Poznal jsem</b> /být/VB-S---1P-AA---I ti to na hlase . " " Vypadá to dobře
za neomezené připojení , " říká Neff . " <b>Představovali jsme</b> /být/VB-P---1P-AA---I si ho asi za 700 korun měsíčně . Dneska by
k ničemu , nedaří se mi to skloubit . <b>Byla jsem</b> /být/VB-S---1P-AA---I pitomá , když jsem si myslela , že to zmáknou
akcie ? Ne . A mrzí mě to . <b>Říkal jsem</b> /být/VB-S---1P-AA---I si , že jsem hloupý . Moji známí loni nakupovali
Appletonová , která už tu byla taky . " <b>Šla jsem</b> /být/VB-S---1P-AA---I za Michaelem Kinsolvingem . " " Za tím režisérem ?
vždy na večerního pošťáka ( ano , tak , <b>slyšeli jste</b> /být/VB-P---2P-AA---I dobře , třikrát denně chodila kdysi pošta ! ) ,
pane poručíku ? " " Promiňte nám , " <b>přejel jsem</b> /být/VB-S---1P-AA---I si obličej oběma rukama , " ale dneska už jsme
převlékáme . " Vzdorně se na mě podívala . <b>Vyvalila jsem</b> /být/VB-S---1P-AA---I oči . " Vážně ? " " Samozřejmě v soukromí
jsem se zaměřili hlavně na obrannou fázi , a <b>zkusili jsme</b> /být/VB-P---1P-AA---I si standardní situace . Domácí nás v nich prověřili ,
zabradlí a neslyšně jsme pluli kolem pohádkového lesa , <b>podívala jsem</b> /být/VB-S---1P-AA---I se na jeho ruku , na jeho dlouhé prsty volně

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  $\langle 1,1 \rangle$ , 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 přičestí slovesa *řici*). Na dalších dvou označených řádcích jsou chyby v interpunkci důvodem přegenerování. Chyby v interpunkci vidíme i na dalších řádcích, na nich ovšem k přegenerování nedochází.

<sup>34</sup> 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**).

co si povídají . Ani mě to nezajímalo . <b>Začal jsem</b> /být/VB-S---1P-AA---I <b>byl</b> zoufalý z následujícího týdne . Copak neexistuje lék ,
na malou chvilku budete vy sami Žila jsem a <b>milovala jsem</b> /být/VB-S---1P-AA---I <b>byla</b> milována Netruchlete příliš Vidím žiju dýchám svého Pána Plačte
Sbírka : Stromy u cesty ( 1906 ) . <b>Zpíval jsem</b> /být/VB-S---1P-AA---I <b>Zpíval</b> jsem jen na neschůdných cestách , zpíval jsem v
nedižích ; v té hrůze , kterou ze mne <b>měla jsem</b> /být/VB-S---1P-AA---I <b>byl</b> i ten chirurg s nožem v ruce , který
. Během vánočních svátků , co jsem u ní <b>byla jsem</b> /být/VB-S---1P-AA---I <b>psala</b> každé den dopis Robinovi / po částech / -
článek pravděpodobně zaplatil váš táta - - " <b>Řekl jsem</b> /být/VB-S---1P-AA---I <b>podplatil</b> . " - - a pověřil vaše tiskové
! Nebyl přítomen když příkaz vydán . Poslán - <b>byl jsem</b> /být/VB-S---1P-AA---I <b>byl</b> jsem poslán - na pochůzku . Vrátil jsem se
poněkud vybledlého obrázku . Nemohla jsem to pochopit . <b>Myslela jsem</b> /být/VB-S---1P-AA---I <b>shodla</b> jsem se na tom i s kluky , že
ve hmotě sním tucha mě obsedá že nejsem in <b>Uviděl jsem</b> /být/VB-S---1P-AA---I <b>Uviděl</b> jsem že nevidím uslyšel že jsem naskrz hluchý užasl
a zbytek cesty do nemocnice jsem došla pěšky . <b>Řekla jsem</b> /být/VB-S---1P-AA---I <b>došla</b> ? Protože cesta trvala deset minut , ale na
, davy . Nic jiného ho nezajímalo . Jako <b>Řek jsem</b> /být/VB-S---1P-AA---I <b>byl</b> až úzkostlivě čistotný . Bez mučení přiznávám , že
jsem buňič a lítala tam jako tvrdá punkerka . <b>Byla jsem</b> /být/VB-S---1P-AA---I <b>byla</b> trnem v oku a ve škole mě nesnášeli !
se mě žena jak se cítíš zralé k ukřížování <b>řekl jsem</b> /být/VB-S---1P-AA---I <b>stočila</b> se do klubička a přitiskla se mi do klína

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].*"]`.

to změnili , a vy byste se neustále <b>divili</b> " <b>Jsmo</b> /být/VB-P---1P-AA---I dluhy obtíženi , ba víc , teď všeho jsme zbaveni
obrovský labyrint , do něhož jsem kdysi dávno <b>zabloudil</b> . <b>Nejsem</b> /být/VB-S---1P-NA---I si ale svojí rezistencí dopředu nikdy jistý , a proto
promiň , já zapoměla , že ses úplně <b>změnila</b> a <b>jsi</b> /být/VB-S---2P-AA---I pilná jako včelička ( škoda jen , že jako včelička
jednoduše za duševně nemocné To vše jsem si <b>uvědomil</b> když <b>jsem</b> /být/VB-S---1P-AA---I se pokoušel odhalit smysl dvou příběhů které přesně dokazují obecnou
Marlborough byl jsem mu štítem abych se ti <b>libil</b> Myšlenkou <b>jsem</b> /být/VB-S---1P-AA---I ho denně zabíjel rukama bránil žes to poručila a chrabrost
rozloučíte s nižšími ročníky , jak jsme se <b>domluvili</b> . <b>Jsem</b> /být/VB-S---1P-AA---I ráda , že nejste takoví , abyste se opičili po
mld. dostali a 93,68 mld. jsme do Bruselu <b>zaplatili</b> , <b>jsmo</b> /být/VB-P---1P-AA---I tedy zřejmě 0,47 mld. korun v plusu . ( Odkazy
. Nebyla to přitom žádná legrace , <b>věděl</b> už <b>jsem</b> /být/VB-S---1P-AA---I , co bude následovat , ale nemohl jsem zase jen
Když jsem ale později vyhledal svého psychoanalytika , <b>dozvěděl</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I se , že jsem několik let vyšetřoval fiktivního pacienta .
nebo šest let potom , co jsem nepřetřítě <b>šmidlal</b> , <b>jsem</b> /být/VB-S---1P-AA---I najednou začal hrát . Bylo to prý docela dobré .

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 přičestí 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 přegenerování, a teprve poté je pomocí negativního filtru odstranit.)

Marlborough byl jsem mu štítem abych se ti <b>libil</b> Myšlenkou <b>jsem</b> /být/VB-S---1P-AA---I	ho denně zabíjel rukama bránil žes to poručila a chrabrost
. . Nebyla to přitom žádná legrace , <b>věděl</b> už <b>jsem</b> /být/VB-S---1P-AA---I	, co bude následovat , ale nemohl jsem zase jen
Když jsem ale později vyhledal svého psychoanalytika , <b>dozvěděl</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	se , že jsem několik let vyšetřoval fiktivního pacienta .
oblečení , jež na sobě měly ženy . <b>Naslouchal</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	se zlostí a závistí . Žena jako by nad každým
projde se po pokoji Jenom jsem ti to <b>neříkal</b> Chodil <b>jsem</b> /být/VB-S---1P-AA---I	pro poštu a pokaždé jsem je prostě zahazoval zmlkne MLADÁ
jich určitě alespoň deset Jenom jsem ti to <b>neříkal</b> Dostával <b>jsem</b> /být/VB-S---1P-AA---I	dopisy a pak jsem je zahazoval MLADÁ ŽENA tak docela
slova pronesl , kluci pochopili , že je <b>vedl</b> opravdu <b>sem</b> /být/VB-S---1P-AA---6I	. Vašek se zastavil a pak opatrně našlápl na místo
Pak dal mou ruku přes svou paži a <b>řekl</b> já <b>jsem</b> /být/VB-S---1P-AA---I	teď tvoje myš , a začal jezdit podél klávesnice .
profese . Prodávát obrázky . " " A <b>měli</b> už <b>jste</b> /být/VB-P---2P-AA---I	zákazníky , kteří si koupili obraz a předem se na
vzdouvají ty šedivé vodní valy , má duše <b>křičela</b> PROČ <b>JSEM</b> /být/VB-S---1P-AA---I	NEZŮSTAL DOMA ! ? Ale bylo už pozdě . Když
pochopili že smíme beztrestně vše abychom plni zoufalství <b>pochopili</b> že <b>jsem</b> /být/VB-P---1P-AA---I	úplně splynuli s Bohem který si bude sám sobě nepřijemný
vzhůru nohama . Než jste se objevil , <b>měl</b> už <b>jsem</b> /být/VB-S---1P-AA---I	ji krásně rozehranou . " Jenomže Ranleigh ho neposlouchal .
CHLAPEČ Tak to tedy byla ona koho jsem <b>viděl</b> Projel <b>jsem</b> /být/VB-S---1P-AA---I	kolem ní Šla dolů po cestě Já jsem jel nahoru
, pane , já myslel , že madame <b>odešla</b> Myslel <b>jsem</b> /být/VB-S---1P-AA---I	" Ale větu nedokončil . Filip po něm skočil a

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 oprít o spolehlivá data.

Pokusíme se hromadně odstranit ty případy, kdy vyhledaný tvar l-ového přičestí 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.

Když jsem ale později vyhledal svého psychoanalytika , <b>dozvěděl</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	se , že jsem několik let vyšetřoval fiktivního pacienta .
oblečení , jež na sobě měly ženy . <b>Naslouchal</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	se zlostí a závistí . Žena jako by nad každým
projde se po pokoji Jenom jsem ti to <b>neříkal</b> Chodil <b>jsem</b> /být/VB-S---1P-AA---I	pro poštu a pokaždé jsem je prostě zahazoval zmlkne MLADÁ
jich určitě alespoň deset Jenom jsem ti to <b>neříkal</b> Dostával <b>jsem</b> /být/VB-S---1P-AA---I	dopisy a pak jsem je zahazoval MLADÁ ŽENA tak docela
CHLAPEČ Tak to tedy byla ona koho jsem <b>viděl</b> Projel <b>jsem</b> /být/VB-S---1P-AA---I	kolem ní Šla dolů po cestě Já jsem jel nahoru
, pane , já myslel , že madame <b>odešla</b> Myslel <b>jsem</b> /být/VB-S---1P-AA---I	" Ale větu nedokončil . Filip po něm skočil a
Stal se ze mě úplný pan domácí a <b>dělal</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	si na plynovém vařiči čaj , kdy se mi zachtělo
nás vytrhla z pohádkových časů v nichž jsme <b>panovali zšedli</b> <b>jsem</b> /být/VB-P---1P-AA---I	na chodníku sami osiřeli do všední doby Smrt pejska tvůj
nohou na zemi ale abych si srdce řádně <b>obrnila ztratila</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	se mezi tóny Mezi tímhle prvním hitem a vyletem do
větvích Borsalino Představoval jsem si že jsi nečekaně <b>přišla Představoval</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	si že jsi mne políbila a já ucítil meruňky a
ulici ve Vršovicích kde bydlela jsme se dlouho <b>loučili</b> Šel <b>jsem</b> /být/VB-S---1P-AA---I	domů s radostí jak se říká v dobrých básních s
. X. 1974 Na druhý den jsem se <b>probudil vzal</b> jsem <b>jsem</b> /být/VB-S---1P-AA---I	Juliinu mrtvolu shodil jí o tři patra na dvůr a

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.\*"].



Marlbourough byl jsem mu štítem abych se ti <b>libil</b> Myšlenkou <b>jsem</b> /být/VB-S---1P-AA---I	ho denně zabíjel rukama bránil žes to poručila a chrabrost
. . Nebyla to přitom žádná legrace, <b>věděl už</b> <b>jsem</b> /být/VB-S---1P-AA---I	, co bude následovat, ale nemohl jsem zase jen
slova pronesl, kluci pochopili, že je <b>vedl</b> opravdu <b>sem</b> /být/VB-S---1P-AA--6I	. Vašek se zastavil a pak opatrně našlápl na místo
Pak dal mou ruku přes svou paži a <b>řekl já</b> <b>jsem</b> /být/VB-S---1P-AA---I	teď tvoje myš, a začal jezdit podél klávesnice .
profese . Prodávát obrázky . " " A <b>měli už</b> <b>jste</b> /být/VB-P---2P-AA---I	zákazníky , kteří si koupili obraz a předem se na
vzdouvají ty šedivé vodní valy , má duše <b>křičela</b> PROČ <b>JSEM</b> /být/VB-S---1P-AA---I	NEZŮSTAL DOMA ! ? Ale bylo už pozdě . Když
pochopili že smíme beztrestně vše abychom plni zoufalství <b>pochopili</b> že <b>jsem</b> /být/VB-P---1P-AA---I	úplně splynuli s Bohem který si bude sám sobě nepřijemný
vzhůru nohama . Než jste se objevil , <b>měl už</b> <b>jsem</b> /být/VB-S---1P-AA---I	ji krásně rozehranou . " Jenomže Ranleigh ho neposlouchal .
jít . Ale když jsem vzal míč , <b>měl už</b> <b>jsem</b> /být/VB-S---1P-AA---I	jasno , jak to kopnu , " dodal . Mourinho
mladě ženě To je hloupě že jsem ho <b>vzbudila</b> To <b>jsem</b> /být/VB-S---1P-AA---I	nechtěla to je hloupé Posadí se na pohovku K otci
jsi mi upila A cizím hlasem jsi mě <b>tupila</b> Ty <b>jsi</b> /být/VB-S---2P-AA---I	můj fosgen tabun yperit Vidím tě všude Nemám chvíli klid
abych se nezbáznil ale pil jsem protože jsem <b>šilel</b> pak <b>jsem</b> /být/VB-S---1P-AA---I	nepil vůbec teď se zas dostávám do fáze zase mi

Ze seznamu zbylých dokladů je patrné, že mezi tvarem l-ového přičestí 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 přičestí a tvarem slovesa *být* stojí lemma *já|ty|my|vy|už|prý*, jsou hledané složené tvary minulého času a b) jak vypadají ostatní případy.

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

. . Nebyla to přitom žádná legrace, <b>věděl už</b> <b>jsem</b> /být/VB-S---1P-AA---I	, co bude následovat, ale nemohl jsem zase jen
Pak dal mou ruku přes svou paži a <b>řekl já</b> <b>jsem</b> /být/VB-S---1P-AA---I	teď tvoje myš, a začal jezdit podél klávesnice .
profese . Prodávát obrázky . " " A <b>měli už</b> <b>jste</b> /být/VB-P---2P-AA---I	zákazníky , kteří si koupili obraz a předem se na
vzhůru nohama . Než jste se objevil , <b>měl už</b> <b>jsem</b> /být/VB-S---1P-AA---I	ji krásně rozehranou . " Jenomže Ranleigh ho neposlouchal .
jít . Ale když jsem vzal míč , <b>měl už</b> <b>jsem</b> /být/VB-S---1P-AA---I	jasno , jak to kopnu , " dodal . Mourinho
jsi mi upila A cizím hlasem jsi mě <b>tupila</b> Ty <b>jsi</b> /být/VB-S---2P-AA---I	můj fosgen tabun yperit Vidím tě všude Nemám chvíli klid
, aby mě zase někdo nepřepadl , a <b>nedokázala už</b> <b>jsem</b> /být/VB-S---1P-AA---I	se tam cítit bezpečně . Jednou večer jsme s manželi
odcházíme jak štvanci a žebráci leč odcházíme abychom <b>kvetli</b> My <b>jsem</b> /být/VB-P---1P-AA---I	tou setbou k národům hořičičným zrnem a my jsme těmi
na kabát : " Poslouchejte , Wranka , <b>nepracoval vy</b> <b>ste</b> /být/VB-P---2P-AA--6I	před letama letoucima na jedný pile ve Šwiecie ? Vona
mnou mohl soukromě pohovořit . Po jistém zdráhání <b>souhlasila já</b> <b>jsem</b> /být/VB-S---1P-AA---I	strávila příjemnou půlhodinu ve společnosti muže , jenž byl ,
mi útulek , o který žádám . ' <b>Neměl už</b> <b>jsem</b> /být/VB-S---1P-AA---I	odvahu s tím člověkem déle mluvit . Když jsem se
a vášnivý , ale plný křesťanské pokory . <b>Zvolila prý</b> <b>jsem</b> /být/VB-S---1P-AA---I	lepší stránku a že zasluhuji největší pochvaly za svůj milosrdný
něj není . Je z toho vyděšený . <b>Měla už</b> <b>jsem</b> /být/VB-S---1P-AA---I	psa , ale ten mi utekl . Tak jsem si

Výsledkem jsou správné doklady analytického minulého času.

- b) Vrátime se zpět a budeme postupovat tak, že zvolíme filtr (negativní), rozsah hledání  $\langle -1, -1 \rangle$ , **Typ dotazu lemma** a do dotazovacího řádku napíšeme dotaz ve formě *já|ty|my|vy|už|prý*.

Marlborough byl jsem mu štítem abych se ti <b>libil</b> Myšlenkou <b>jsem</b> /být/VB-S---1P-AA---I	ho denně zabíjel rukama bránil žes to poručila a chrabrost
slova pronesl , kluci pochopili , že je <b>vedl</b> opravdu <b>sem</b> /být/VB-S---1P-AA--6I	. Vašek se zastavil a pak opatrně našlápl na místo
vzdouvají ty šedivé vodní valy , má duše <b>křičela</b> PROČ <b>JSEM</b> /být/VB-S---1P-AA---I	NEZŮSTAL DOMA ! ? Ale bylo už pozdě . Když
pochopili že smíme beztrestně vše abychom plni zoufalství <b>pochopili</b> že <b>jsem</b> /být/VB-P---1P-AA---I	úplně splynuli s Bohem který si bude sám sobě nepříjemný
mladé ženě To je hloupé že jsem ho <b>vzbudila</b> To <b>jsem</b> /být/VB-S---1P-AA---I	nechtěla to je hloupé Posadí se na pohovku K otci
abych se nebláznil ale pil jsem protože jsem <b>šil</b> pak <b>jsem</b> /být/VB-S---1P-AA---I	nepil vůbec teď se zas dostávám do fáze zase mi
zahříván zvířaty Když jsi se k betlémským jesličkám <b>skláněla</b> stoup <b>jsem</b> /být/VB-S---1P-AA---I	si za tebe a s tváří anděla stáh jsem ti
na některé . Poté co jsme si ještě <b>zahráli</b> škatulata-batulata <b>jsem</b> /být/VB-P---1P-AA---I	se vydali dále do Mojžíšův pramen . Po cestě jsme
1983 ) ZÁCHOD I . Jak jsem sem <b>vešel</b> řek <b>jsem</b> /být/VB-S---1P-AA---I	" Ach " když známý ovanul mne pach záchodu hospodského
abych poklábosil lámanou španělštinou pot mi po tváři <b>tekli</b> jak <b>jsem</b> /být/VB-S---1P-AA---I	utíkal na vlak mé sešlapané sandály zaprášené jako španělské silnice
tak jsem se roztesknil že jsem až zapomněl <b>zapomněl</b> kde <b>jsem</b> /být/VB-S---1P-AA---I	byl a byl jsem na Rychtě a mrznul u topení
-- DOSVĚDČÍ SAMA TUŽKA ANIŽ BYCH JÍ <b>PSAL</b> že <b>jsem</b> /být/VB-S---1P-AA---I	tě miloval svíčka i plamen v naftákách jak jsem měl

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* označovaných 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].\*["].

docela vstřícně a vlídně , abych ji <b>nepoplašila</b> . Teď <b>jsem</b> /být/VB-S---1P-AA---I	byla na koni . Ona byla vyděšená , mohla jsem
jsem si pomoci . Jak bych také <b>mohl</b> , když <b>jsem</b> /být/VB-S---1P-AA---I	s vámi ? Neberte si to k srdci , netrapte
z toho usnula ! Ani jsem si <b>nevšimla</b> , kdy <b>jsi</b> /být/VB-S---2P-AA---I	tam přišel ! Představ si , nechtěli mě pustit ven
zoologické zahradě , blízko níž jsme kdysi <b>bydleli</b> . Zavřel <b>jsem</b> /být/VB-S---1P-AA---I	oči . Už je to dávno a vybavuji si to
. Ovšem tam jste si ho sotva <b>povšiml</b> , zajímal <b>jste</b> /být/VB-P---2P-AA---I	se o muže , a ne o děti . Ale
stanu s knihami . Když jsem tam <b>dorazil</b> , všiml <b>jsem</b> /být/VB-S---1P-AA---I	si ani ne dvacetileté ženy , která si tam prohlížela
Poslal jste mi ho , abych ho <b>prozkoušel</b> . Vyzkoušel <b>jsem</b> /být/VB-S---1P-AA---I	ho ve všem možném . Je taky jediný , kterého
nespojvalo s postelí , kde jsem teď <b>ležel</b> . Nemohl <b>jsem</b> /být/VB-S---1P-AA---I	se dočkat , až maminka odejde a já se ponořím
na oranžovou , abych to do restaurace <b>stihla</b> . Zatímco <b>jsem</b> /být/VB-S---1P-AA---I	čekala na světlech , která jsem už nestačila projet ,
mrhnutí venku , ale teď jsem to <b>vetoval</b> . Cítil <b>jsem</b> /být/VB-S---1P-AA---I	se Saulem , kterého jsem v duchu viděl , jak
svalnaté hrudi . " Já zase skoro <b>zapomněl</b> , jaká <b>jste</b> /být/VB-P---2P-AA---I	maličká a křehká , nicméně že jste divoká a vášnivá
nebo zpívat , zrovna když jsem si <b>myslela</b> , že <b>jsem</b> /být/VB-S---1P-AA---I	si vysloužila chvilku klidu . " Prosim , maminko
, kdy tě čekám . Tys ji <b>neprozdila</b> , kde <b>jsi</b> /být/VB-S---2P-AA---I	, vidí ? Řeklas , že jedeš za mnou ,

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.

docela vstřícně a vlídně, abych ji **nepoplašila**. Ted' **jsem** /být/VB-S---1P-AA---I byla na koni. Ona byla vyděšená, mohla jsem jsem si pomoci. Jak bych také **mohl**, když **jsem** /být/VB-S---1P-AA---I s vámi? Neberte si to k srdci, netrapte z toho usnula! Ani jsem si **nevšimla**, kdy **jsi** /být/VB-S---2P-AA---I tam přišel! Představ si, nechtěli mě pustit ven zoologické zahradě, blízko níž jsme kdysi **bydleli**. Zavřel **jsem** /být/VB-S---1P-AA---I oči. Už je to dávno a vybavuji si to. Ovšem tam jste si ho sotva **povšiml**, zajímal **jste** /být/VB-P---2P-AA---I se o muže, a ne o děti. Ale stanu s knihami. Když jsem tam **dorazil**, všiml **jsem** /být/VB-S---1P-AA---I si ani ne dvacetileté ženy, která si tam prohlížela. Poslal jste mi ho, abych ho **prozkoušel**. Vyzkoušel **jsem** /být/VB-S---1P-AA---I ho ve všem možném. Je taky jediný, kterého nespojovalo s postelí, kde jsem teď **ležel**. Nemohl **jsem** /být/VB-S---1P-AA---I se dočkat, až maminka odejde a já se ponořím na oranžovou, abych do restaurace **stihla**. Zatímco **jsem** /být/VB-S---1P-AA---I čekala na světlech, která jsem už nestačila projet, mrhnutí venku, ale teď jsem to **vetoval**. Cítil **jsem** /být/VB-S---1P-AA---I se Saulem, kterého jsem v duchu viděl, jak svalnaté hrudi. " Já zase skoro **zapomněl**, jaká **jste** /být/VB-P---2P-AA---I maličká a křehká, nicméně že jste divoká a vášnivá

Vrátíme se o krok zpět, zvolíme filtr (negativní), 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.

tatinkova zoufalství jako před domem kde se **zabila** sestra Vždycky **jsem** /být/VB-S---1P-AA---I věděl jak je těžké v životě obstát A přece trvá přistála u hospody ve Škodějově Právě jsem **vypíral** václavky které **jsem** /být/VB-S---1P-AA---I našel v lese kam choďovala moje maminka s mým synem co bys chtěl vědět Vypočítává Kde jsi **byla** S kým **jsi** /být/VB-S---2P-AA---I byla a tak dále MLADÝ MUŽ Něco bys mi snad znát pravdu, tys ji našel, **byla** taková jakou **jsi** /být/VB-S---2P-AA---I chtěl. Měla tři chatrné nožky. Když k večeru seděla u klaviatury kdy ji naposledy někdo **vystříkal** spodek Byl **jsem** /být/VB-S---1P-AA---I mocně zasažen oblidností nechutného vtipu Znovu jsem se vrátil do k výuce, poté co jsem ji **opustil** Poté co **jsem** /být/VB-S---1P-AA---I asi deset let pobýval v ústraní o samotě, dospěl vu! Hrdost, které bychom se **měli** učit Měl **jsem** /být/VB-S---1P-AA---I to štěstí, že jsem se před léty dostal konkursem neslyšen slyšen Mlčel jsem a mluvil jsem **zpíval** jsem nenáviděl **jsem** /být/VB-S---1P-AA---I tě pohrdal jsem tebou styděl jsem se za tebe trpěl to co jsem dostal to co jsem **vyhrál** to co **jsem** /být/VB-S---1P-AA---I propil to co jsem ztratil po cestě to co jsem víme proč to děláme jak jsme se **stali** tím co **jsme** /být/VB-P---1P-AA---I Svět nás má v železech a vleče nás k bídnému Jo Tak takhle jsem si tě nepředstavovala **schovala** fotku kterou **jsem** /být/VB-S---1P-AA---I jí poslal už je to dávno pamatuju si jen její je nepostradatelný. O čem jsme si **povídali** Druhý den **jsme** /být/VB-P---1P-AA---I se vypravili na procházku Hyde parkem ( zrovna totiž nepřišlo

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

tatinkova zoufalství jako před domem kde se **zabila** sestra Vždycky **jsem** /být/VB-S---1P-AA---I věděl jak je těžké v životě obstát A přece trvá přistála u hospody ve Škodějově Právě jsem **vypíral** václavky které **jsem** /být/VB-S---1P-AA---I našel v lese kam choďovala moje maminka s mým synem co bys chtěl vědět Vypočítává Kde jsi **byla** S kým **jsi** /být/VB-S---2P-AA---I byla a tak dále MLADÝ MUŽ Něco bys mi snad znát pravdu, tys ji našel, **byla** taková jakou **jsi** /být/VB-S---2P-AA---I chtěl. Měla tři chatrné nožky. Když k večeru k výuce, poté co jsem ji **opustil** Poté co **jsem** /být/VB-S---1P-AA---I asi deset let pobýval v ústraní o samotě, dospěl to co jsem dostal to co jsem **vyhrál** to co **jsem** /být/VB-S---1P-AA---I propil to co jsem ztratil po cestě to co jsem víme proč to děláme jak jsme se **stali** tím co **jsme** /být/VB-P---1P-AA---I Svět nás má v železech a vleče nás k bídnému Jo Tak takhle jsem si tě nepředstavovala **schovala** fotku kterou **jsem** /být/VB-S---1P-AA---I jí poslal už je to dávno pamatuju si jen její je nepostradatelný. O čem jsme si **povídali** Druhý den **jsme** /být/VB-P---1P-AA---I se vypravili na procházku Hyde parkem ( zrovna totiž nepřišlo Spadla jsem z oblak, kde jsem **měla** hlavu Vždycky **jsem** /být/VB-S---1P-AA---I si myslel, že mexická mafiánská corrida jsou jenom lidové ošetřující lékařku a ta mne před půlnocí **vyzvedla** Po Evě **jsem** /být/VB-S---1P-AA---I strašně toužil a obcházel jsem proto místo jejího bydliště Tak

Projdeme ručně seznam zbylých dokladů. Ručním tříděním zjistíme, že všechny nalezené doklady jsou případy přegenerování. 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 přičestí, 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<sup>35</sup>). Chyby se ovšem vyskytují i v případech, kdy l-ové přičestí 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 přičestí), 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. Vrátime 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].\*"].

období dočkali zeleného světla . - Během ledna a února <b>jsem</b> /být/VB-P---1P-AA---I <b>museli</b> zrušit patnáct pracovních dní , od března však už docela vstřícně a vřídě , abych ji nepoplašila . Teď <b>jsem</b> /být/VB-S---1P-AA---I <b>byla</b> na koni . Ona byla vyděšená , mohla jsem jediného záběru , kdy jsme potřebovali natočit detail hole , <b>jsem</b> /být/VB-S---1P-AA---I <b>vycházel</b> ze dveří studia a ona mi uvízla v kanále , a v pravém horním rúžku byl diamant . Léta <b>jsem</b> /být/VB-P---1P-AA---I <b>nevěděli</b> , kdo je ta žena ani co ten diamant Čadová . Škodovka si vypomáhá přibíráním absolventů . " Loni <b>jsem</b> /být/VB-P---1P-AA---I <b>přijali</b> celkem 467 lidí , z toho 84 absolventů škol Když se mi v létě pero trochu zatoulalo , nejspíš <b>jsem</b> /být/VB-S---1P-AA---I <b>postřehla</b> pár prostředků , jak dosáhnout svého . Seděla jsem okrají skupiny , ale ne jako její členka . Občas <b>jsem</b> /být/VB-S---1P-AA---I <b>měla</b> nějaké kamarádky . Snila jsem o dnu , kdy , a trůn byl jeho . " " A ty <b>jsi</b> /být/VB-S---2P-AA---I <b>ztratil</b> přítele , " zamumlala Brittany se srdcem těžkým lítostí kávu ? " " Ne , děkuju . Před chvílkou <b>jsem</b> /být/VB-S---1P-AA---I <b>obědvala</b> . " " Fiona se nejspíš za chvíli vrátí minu , říkali jsme jim pekáče . Jen koutkem oka <b>jsem</b> /být/VB-S---1P-AA---I <b>stačila</b> zahlídnout děti stojící opodál , plamen a Thanasakise , toho bylo páte místo . To mě uklidnilo . Už <b>jsem</b> /být/VB-S---1P-AA---I <b>nemyslel</b> dál , jak to dopadne . Hleděl jsem si , takže to prošlo bez komplikací . A teď už <b>jsem</b> /být/VB-S---1P-AA---I <b>měl</b> jen obavy , abychom se s panem profesorem našli dovolenou má samozřejmě i své nevýhody . " Z průzkumů <b>jsem</b> /být/VB-P---1P-AA---I <b>zjistili</b> , že klientům například vadí , když musejí složit
---

<sup>35</sup> 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/tuže, horší – zlý/špatný*).

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

, davy . Nic jiného ho nezajímalo . Jako <b>Řek jsem</b> /být/VB-S---1P-AA---I <b>byl</b> až úzkostlivě čistotný . Bez mučení přiznávám , že
nedíží ; v té hrůze , kterou ze mne <b>měla jsem</b> /být/VB-S---1P-AA---I <b>byl</b> i ten chirurg s nožem v ruce , který
jsem buřič a litala tam jako tvrdá punkerka . <b>Byla jsem</b> /být/VB-S---1P-AA---I <b>byla</b> trnem v oku a ve škole mě nesnášeli !
Sbírka : Stromy u cesty ( 1906 ) . <b>Zpíval jsem</b> /být/VB-S---1P-AA---I <b>Zpíval</b> jsem jen na neschůdných cestách , zpíval jsem v
článek pravděpodobně zaplatil váš táta - - " <b>Řekl jsem</b> /být/VB-S---1P-AA---I <b>podplatil</b> . " - - a pověřil vaše tiskové
. Během vánočních svátků , co jsem u ní <b>byla jsem</b> /být/VB-S---1P-AA---I <b>psala</b> každé den dopis Robinovi / po částech / -
ve hmotě sním tucha mě obsedá že nejsem in <b>Uviděl jsem</b> /být/VB-S---1P-AA---I <b>Uviděl</b> jsem že nevidím uslyšel že jsem naskrz hluchý užasl
! Nebyl přítomen když příkaz vydán . Poslán - <b>byl jsem</b> /být/VB-S---1P-AA---I <b>byl</b> jsem poslán - na pochůzku . Vrátil jsem se
co si povídají . Ani mě to nezajímalo . <b>Začal jsem</b> /být/VB-S---1P-AA---I <b>byl</b> zoufalý z následujícího týdne . Copak neexistuje lék ,
a zbytek cesty do nemocnice jsem došla pěšky . <b>Řekla jsem</b> /být/VB-S---1P-AA---I <b>došla</b> ? Protože cesta trvala deset minut , ale na
na malou chvilinku budete vy sami Žila jsem a <b>milovala jsem</b> /být/VB-S---1P-AA---I <b>byla</b> milována Netruchtele příliš Vidím žiju dýchám svého Pána Plačte
poněkud vybledlého obrázku . Nemohla jsem to pochopit . <b>Myslela jsem</b> /být/VB-S---1P-AA---I <b>shodla</b> jsem se na tom i s kluky , že
se mě žena jak se cítíš zralé k ukřížování <b>řekl jsem</b> /být/VB-S---1P-AA---I <b>stočila</b> se do klubička a přitiskla se mi do klína

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].\**

. Byl to teplouš a voblíbil si mě . Když <b>jsem</b> /být/VB-S---1P-AA---I tam <b>přišel</b> , zkraje jsem myl nádobí , pak to
" Svůj gól vám popsat nemůžu . Nevím , jak <b>jsem</b> /být/VB-S---1P-AA---I ho <b>dal</b> . Jen jsem zasáhl míč hlavou . "
chvilí pobýti v kosmu jejího sluchu . . . Pak <b>jsem</b> /být/VB-S---1P-AA---I se <b>ušklíbl</b> : takové hlouposti ! Leč kdo může věděti
částečně to <b>byl</b> ještě zbytek strachu ( tu a tam <b>jsem</b> /být/VB-S---1P-AA---I si <b>připomínal</b> , že součástí Percyho problému bylo to ,
. Všichni mi připadali velmi slibní . " S hrůzou <b>jsem</b> /být/VB-S---1P-AA---I si <b>vzpomněla</b> na poslední přísun zájemců a s díky jsem
: " Něco vám chci sdělit . Už nějakou dobu <b>jsem</b> /být/VB-S---1P-AA---I vám <b>chtěl</b> říci , že ti , kteří přijeli s
se říká normalizace a její " 40 . výročí " <b>jsme</b> /být/VB-P---1P-AA---I si <b>připomněli</b> 21 . srpna . . . Bylo by
se zaujetím premiéry . Jen některé příliš dutě znějící věty <b>jsme</b> /být/VB-P---1P-AA---I později <b>vypustili</b> . Vodku jsme však neošidili nikdy ! Podobně
v Mexiku , kde jsme točili Legendu o Zorrovi , <b>jsme</b> /být/VB-P---1P-AA---I hodně <b>pracovali</b> v noci a vůbec by nebylo možné jim
. Podal mi kouřící šálek z bílého porcelánu . Pomalu <b>jsem</b> /být/VB-S---1P-AA---I se <b>zvedla</b> a on pronesl tato slova : " Jestli
z toho usnula ! Ani jsem si nevsímala , kdy <b>jsi</b> /být/VB-S---2P-AA---I tam <b>přišel</b> ! Představ si , nechtěli mě pustit ven
. Bezděčně jsem se pokusil ho podepřít , ale než <b>jsem</b> /být/VB-S---1P-AA---I ho <b>stačil</b> zachytit , zhroutil se a rozložil se na
dveřmi . Oblouky Jak můžete vidět na fotografiích , já <b>jsem</b> /být/VB-S---1P-AA---I se <b>rozhodla</b> jednotlivé místnosti propojit velkými a širokými oblouky .
navzájem pochopit . Pak si klidně dumejte , jestli to <b>jste</b> /být/VB-P---2P-AA---I tehdy <b>udělali</b> nebo neudělali správně . " Přikývl jsem .

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.

se neměl nikdo z ČSSD ani pokoušet . Vy sama	<b>jste</b> /být/VB-P---2P-AA---I	<b>ale byla</b> na českobudějovické radnici čtyři roky v koalici s
" Věděla jsem , že to všichni slyší a zropačitelá	<b>jsem</b> /být/VB-S---1P-AA---I	<b>. Svésila</b> jsem hlavu , avšak pokračovala jsem v chůzi
na divadle ? Oliva : Ano , hrála . Hrála	<b>jsem</b> /být/VB-S---1P-AA---I	<b>. Hrála</b> jsem hodně . Byla jsem v hudební revyji
ony , ale těžkopádný cenzurní aparát . Tenkrát v Koktěbelu	<b>jsem</b> /být/VB-P---1P-AA---I	<b>však měli</b> komického listonoše a telegrafistu v jedné osobě .
Ovládlo ji tak mocné vzrušení , až vyjekla . Vyrostl	<b>jsem</b> /být/VB-S---1P-AA---I	<b>. Dospěl</b> jsem , Hvězdooká . Tenkrát na poušti jsem
tří řádky . Dál jsem viděl jen mazanici . Zazmatkoval	<b>jsem</b> /být/VB-S---1P-AA---I	<b>. Snažil</b> jsem si vybavit ty písnička , jak jsem
vyprávěl o své dovolené v mé zemi , " řekl	<b>jsem</b> /být/VB-S---1P-AA---I	<b>a snažil</b> jsem se zahlédnout , co je na plánech
" Tak se mnou nikdy nemluvil . " " Sama	<b>jsi</b> /být/VB-S---2P-AA---I	<b>ale řekla</b> , že byl opilý a nemocný . Teď
je obstarat . " " Tak dobře , " souhlasil	<b>jsem</b> /být/VB-S---1P-AA---I	<b>. Šli</b> jsme za roh kolem budovy námořnických odborů .

Dále se snažíme odstranit případy, kdy tvar pomocného slovesa a l-ové přičestí netvoří dohromady hledaný složený slovesný tvar. Pokoušíme se najít možné doklady přegenerování 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].*"]`.

osobní rozhodnutí . Kdybych šel jen po penězích , tak	<b>jsem</b> /být/VB-S---1P-AA---I	<b>býval podepsal</b> jinde a byl bych letos ještě v první
- primitivní , nedomestikovaný . Nejpodobnější náhradní výraz , který	<b>jsem</b> /být/VB-S---1P-AA---I	<b>našel byl</b> , že je milá , což ovšem mělo
Města , laskavě se navrát a zůstávej tam , jak	<b>jsi</b> /být/VB-S---2P-AA---I	<b>byl pravil</b> . " Cože ? Oni se také chtějí
Gaia Probudil jsem se na jiném místě než tam kde	<b>jsem</b> /být/VB-S---1P-AA---I	<b>usnul Ležela</b> vedle mne neznámá známá Nedohlédl jsem konců jejich
a hlavně tiše , " pokračoval Josef . " Nicméně	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byl udělal</b> chybu a nechal se prozradit . Musil jsem
. Maluju tím způsobem , že postupně likviduju , co	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byl namaloval</b> , až mi na obraze zůstane jedna jen
desítek kilometrů . Taky měli rychlou akceleraci , ale já	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byla získala</b> náskok a dřív dosáhla maximální rychlosti . Na
NA SHLEDANOU ZÍTRA Od chvíle co jste odjela S nikým	<b>jsem</b> /být/VB-S---1P-AA---I	<b>nemluvila Byla</b> jsem štěstím bez sebe Když přišel instalatér Rozbila
obul , že jsem připraven . . . vůně dne	<b>jsem</b> /být/VB-S---1P-AA---I	<b>spřádal Protáhl</b> jsem své prokřehlé údy , vstal z roztháného
byl menší kdyby se vám to líbilo " Na inzerát	<b>jsem</b> /být/VB-S---1P-AA---I	<b>neodpověděl Byl</b> jsem přesvědčen že je to jedna z metod
" Georgi , synu . . . Rozhovor , který	<b>jsem</b> /být/VB-P---1P-AA---I	<b>bývali měli</b> spolu mít , už nikdy mít nebudeme .
a moc se neptal Válek se po mně že když	<b>jsem</b> /být/VB-S---1P-AA---I	<b>vstala byla</b> jsem do krve rozdrápaná od strniště Holka stála
psa bývala mou pokladničkou s úsporami pro zlé časy když	<b>jsem</b> /být/VB-S---1P-AA---I	<b>osaměla vybírala</b> jsem krejčárky lásky z tebe spolu nikdy nezchudneme

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

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`.

osobní rozhodnutí . Kdybych šel jen po penězích , tak	<b>jsem</b> /být/VB-S---1P-AA---I	<b>býval podepsal</b> jinde a byl bych letos ještě v první
Města , laskavě se navrát a zůstávej tam , jak	<b>jsi</b> /být/VB-S---2P-AA---I	<b>byl pravil</b> . " Cože ? Oni se také chtějí
a hlavně tiše , " pokračoval Josef . " Nicméně	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byl udělal</b> chybu a nechal se prozradit . Musil jsem
. Maluju tím způsobem , že postupně likviduju , co	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byl namaloval</b> , až mi na obraze zůstane jedna jen
desítek kilometrů . Taky měli rychlou akceleraci , ale já	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byla získala</b> náskok a dřív dosáhla maximální rychlosti . Na
" Georgi , synu . . . Rozhovor , který	<b>jsem</b> /být/VB-P---1P-AA---I	<b>bývali měli</b> spolu mít , už nikdy mít nebudeme .
, neboť záhadně padaly stejně jako dnes . Až později	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byl zjistil</b> , že mě častokrát sledují čísi oči a
její slova a něoženil se podle jejího gusta . Bo	<b>sme</b> /být/VB-P---1P-AA--6I	<b>byli zvykli</b> posluchač švigermutter celý život na slovo a její
větičku , která mne dosud pronásleduje : , To tenkrát	<b>jsem</b> /být/VB-S---1P-AA---I	<b>býval měl</b> být souzen , ne teď . ' Tak
Města , laskavě se navrát a zůstávej tam , jak	<b>jsi</b> /být/VB-S---2P-AA---I	<b>byl pravil</b> . . . " Nu , říkají to
si v Německu našly desetitisíce nadšených čtenářů . " Vždycky	<b>jsem</b> /být/VB-S---1P-AA---I	<b>byl žvanil</b> , vypravěč příběhů , " řekl Kaminer nedávno

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].\**

, kde bydlím , to nemám daleko , a tak <b>jsem</b> /být/VB-S---1P-AA---I na nabídku <b>kývl</b> , " řekl Právu kouč , který
jasný zvuk jako voda nebo zvon . Vzpomínám , kdy <b>jsem</b> /být/VB-S---1P-AA---I se naposledy <b>zasmála</b> . Možná se Sašou nebo Scottem .
svátek , jinými slovy , Susie má narozeniny . Takže <b>jsem</b> /být/VB-S---1P-AA---I tvůj dopis <b>otvírala</b> v očekávání , že na mě vyskočí
. Odvedl mě přes ášram do budovy , v níž <b>jsem</b> /být/VB-S---1P-AA---I ještě nikdy <b>nebyla</b> , odemkl dveře a přešli jsme k
nechci jinak pěkný den končit něčím nepříjemným . Ale souhlasit <b>jsem</b> /být/VB-S---1P-AA---I s ředitelem <b>musel</b> - kousat kamarády se nesmí . Po
, Coro . Bude vám to ale líto , že <b>jste</b> /být/VB-P---2P-AA---I s námi <b>nešla</b> . Budeme stavět překrásného sněhuláka . "
že jste porazily české jedničky Petrovou s Klapalovou ? Už <b>jsem</b> /být/VB-P---1P-AA---I s nima <b>hrály</b> na evropské tour v kvalifikaci a porazily
bych pod jeho pohledem rostla do země . A přitom <b>jsem</b> /být/VB-S---1P-AA---I se už <b>domýšlela</b> , že tihle christoforyčové , siegfriedové ,
jen koukala a koukala . I ještě pak , když <b>jsem</b> /být/VB-S---1P-AA---I ho vždycky <b>viděla</b> na ulici . Chodil do nailské měšťanky
se možná změní v obtížnou upomínku na lehkověrnost , již <b>jsem</b> /být/VB-P---1P-AA---I zcela nepochopitelně <b>podlehli</b> . Vědomí zachráně z prožitého dne jen
jsem měla úmysl složit delegátství Divčí akademie . - Když <b>jsem</b> /být/VB-S---1P-AA---I svůj úmysl <b>sdělila</b> svému muži , neschvaloval to a řekl
. Zelenou bude mít Afrika nebo Asie , protože my <b>jsem</b> /být/VB-P---1P-AA---I si kněze <b>nezasloužili</b> . Mladé učitelce se honily myšlenky ,
se , že byli obviněni z nějakých krádeží . To <b>jsem</b> /být/VB-P---1P-AA---I tehdy dokonce <b>řešili</b> přes televizní Černé ovce , protože postup
" Popravdě , slavily jsme tak velký úspěch , že <b>jsem</b> /být/VB-S---1P-AA---I se nakonec <b>rozhodla</b> pro návrat do Anglie . Jsem odhodlaná

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 větých 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 přegenerování, tak případy opačné.

, " řekla . " Pojdme ven . " Následoval <b>jsem</b> /být/VB-S---1P-AA---I ji . <b>Zůstali</b> jsme spolu stát na slunci venku na
, naléhal , že je nutné zasílat radiogramy . Projeli <b>jsem</b> /být/VB-P---1P-AA---I Wokingem <b>a seděli</b> jsme dlouho mlčky . Vlak samozřejmě projížděl
, jak v rohoži v ložnici poslouchá zprávy . Ztlumil <b>jsem</b> /být/VB-S---1P-AA---I rádio , <b>posadil</b> jsem se na protější postel a svým
z nich pochvalovala . Podle slov soutěžících , se kterými <b>jsem</b> /být/VB-P---1P-AA---I hovořili , <b>byla</b> v Boleticích příležitost skvěle si zastřílet v
svými šedými očima . " Prohlíželi jsme vás , zda <b>nejste</b> /být/VB-P---2P-AA---I ozbrojeni , <b>prošli</b> jste detektory na kov . Jak jste
, že to bude někde v mém mozku , pomyslel <b>jsem</b> /být/VB-S---1P-AA---I si . <b>Přinesl</b> jsem deku s sebou a představoval si
mi přestali brát telefony . Přestal jsem létat a jezdil <b>jsem</b> /být/VB-S---1P-AA---I autem . <b>Byl</b> jsem celý týden pryč od rodiny a
, to nevím , moc jsem to nevnímal . Ohlásili <b>jsem</b> /být/VB-P---1P-AA---I se <b>a byli</b> jsme do budovy vpuštění , " uvedl
žehnej Karlovi II . " Měla samozřejmě pravdu . Chápala <b>jsem</b> /být/VB-S---1P-AA---I to . <b>Nemohla</b> jsem se sobecky oddat svému smutku .
Mabon . " Hebera založila své vojsko , ale to <b>jsem</b> /být/VB-S---1P-AA---I nevěděl . <b>Znal</b> jsem ji jako laskavou dívku , kterou
na soutěžích tvrdá práce . " Vítězství na mistrovství republiky <b>jsem</b> /být/VB-S---1P-AA---I nečekala , <b>byla</b> jsem překvapená . Porotu zřejmě zaujal můj odlišný
obranu a bezpečnost , kde se Kubiceho zpráva probírala , <b>jste</b> /být/VB-P---2P-AA---I <b>však</b> oficiálně <b>řekla</b> , že Jan Kubice zveřejněním zákon neporušil

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 větého celku).

obranu a bezpečnost, kde se Kubiceho zpráva probírala, **jste** /být/VB-P---2P-AA---I **však** oficiálně **řekla**, že Jan Kubice zveřejněním zákon neporušil se otevřeně stranilo lži a pomlouvám. Z průběhu schůze **jsem** /být/VB-S---1P-AA---I **však** dodatečně **pochopil**, že na vypracování hanopisu se podílelo smažené kuře a celý večer vypadala poněkud zmateně. Já **jsem** /být/VB-S---1P-AA---I **však** zmatený **nebyl**. Ve skutečnosti jsem ještě nikdy neměl myslí, že jsem buddhistický mnich. V průběhu večera **jsem** /být/VB-S---1P-AA---I **však** **uvědomil**, že - kdybych Vlastu viděl poprvé kde se ve mně ta odvaha vzala, rázným posunkem **jsem** /být/VB-S---1P-AA---I mu **však** **ukázal**, ať to nedělá, jenže příliš době, kdy jsme se seznámili, mýtus o Antigoně **jsem** /být/VB-S---1P-AA---I **však** **zapamatoval** příliš málo, než abych ho z Nicméně na ní nebylo vidět žádné známky poškození. Když **jsem** /být/VB-P---1P-AA---I se **však** **přiblížili** k Doughty Street, zahlédli jsme dav, nebyla šance získat pro sebe stůl. Toho večera **jsem** /být/VB-S---1P-AA---I se **však** **rozhodla** usednout před bílý ubrus a vychutnat si ostatní a všechny je opět sepal hnědým řemenem; nejprve **jsem** /být/VB-S---1P-AA---I se **však** **ujistil**, že jsou poskládány jedna na druhé potom jsme to ještě potvrdili. " Ve druhé půli **jste** /být/VB-P---2P-AA---I **však** trochu **polevili**, bylo příčinou sebeuspokojení po tříúhlovém vedení možností? Byl jsem ještě idealisticky naivní. Rudé právo **jsem** /být/VB-S---1P-AA---I **však** každodenně **četl**. Jan Amos Nekomenský Mě asistentství pověřeni lidem se zdálo, že přímo klokotá, každý den **jsem** /být/VB-P---1P-AA---I se **však** **dovívali** něco takového, co v nás vyvolávalo jsme si, že to bude mnohem jednodušší. Těžko **jsem** /být/VB-P---1P-AA---I se **však** **vyrovnávali** s absencí Hradila a s úžasnou bojovností

Dále se snažíme odstranit případy, kdy tvar pomocného slovesa a l-ové přičestí netvoří dohromady hledaný složený slovesný tvar. Pokoušíme se najít možné doklady přegenerování 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].\*"]*.

z nich pochvalovala. Podle slov soutěžících, se kterými **jsem** /být/VB-P---1P-AA---I **hovořili**, **byla** v Boleticích příležitost skvěle si zastřílet v Mabon. " Hebera založila své vojsko, ale to **jsem** /být/VB-S---1P-AA---I **nevěděl**. **Znal** jsem ji jako laskavou dívku, kterou na soutěžích tvrdá práce. " Vítězství na mistrovství republiky **jsem** /být/VB-S---1P-AA---I **nečekala**, **byla** jsem překvapená. Porotu zřejmément zaujal můj odlišný zacvičit jógu ( takže kdyby se mě zeptal, co **jsem** /být/VB-S---1P-AA---I **dělala**, **mohla** bych říct, že jógu ). výpravy a společně se z nich těšili. O budoucnosti **jsem** /být/VB-S---1P-AA---I **nemluvila**. **Myslela** jsem, že nemusím, že Maxim Zelená dáma, máte tvary jako mají pozemšťanky. To **jsem** /být/VB-S---1P-AA---I **nečekal**. **Byl** jsem už jednou v jiném světě. já se přidal, když hráli z těch, které **jsem** /být/VB-S---1P-AA---I **znal**. **Koncertovali** jsme málem do půlnoci. Když ohniště zbyty. Když nepomůže jed, pomůže nůž! Úplně **jsem** /být/VB-S---1P-AA---I **zapomněl**! **Chtěl** bys asi vědět proč. Proč toto. Vzal si klobouk a tmavé kordovánové rukavice, které **jsem** /být/VB-S---1P-AA---I **milovala**. **Věděla** jsem, že rukavice znamenají, že loňského léta jsem byl pořád ve Francii. A co **jsem** /být/VB-S---1P-AA---I **přijel**, **neukázal** jsem se tu, protože - protože velkou pomocí. Je to špičkový hráč. S Columbasem **jsem** /být/VB-P---1P-AA---I **jednali**, **mluvil** jsem i s Voráčkem, ale on daňovou povinnost. Ukončení podnikání a kontrola OTÁZKA: Vážně **jsem** /být/VB-S---1P-AA---I **onemocněl** a **skončil** jsem podnikání v oblasti služeb. Zajímalo Pak jsme si s Deanem promluvili, nebo spíš já **jsem** /být/VB-S---1P-AA---I **mluvil**. **Předestřel** jsem před ním všechny svoje hříchy. si, tohle je ten správný způsob života! Když **jsem** /být/VB-P---1P-AA---I **jedli**, **pustil** jsem se do dalšího vymyšleného kriminálního příběhu ODS zastupovali v oblasti životního prostředí zelení. Nic takového **jsem** /být/VB-S---1P-AA---I **nenapsal**. **Vyjádril** jsem údiv nad plytkými názory místopředsedy ODS

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*.



nic nevím, " na to doktor . " Tam **jsem** /být/VB-S---1P-AA---I **nebyl**, **měl** jsem službu - a musím říct ,  
 vyběhla s libereckou rezervou čtveřice privoligových hráčů . " Zaskočený **jsem** /být/VB-S---1P-AA---I **nebyl**, **věděl** jsem ale , že to pro nás  
 obdivovali zvláště když jsem byla ve stavu , v jakém **jsem** /být/VB-S---1P-AA---I **byla** . **Sedávali** jsme s Karlem na lavičkách , děti  
 a co ho nestojí příliš úsilí . Protože i já **jsem** /být/VB-S---1P-AA---I **byla** lenivá **milovala** jsem potěšení , vzniklo mezi námi okamžité  
 v sobě . Váš život zvučí v mém . Dokud **jste** /být/VB-P---2P-AA---I **byli**, **byla** jsem dítětem , jež pohrdá smrtí .  
 ? ( Krátká pauza ) V prvním lágru , kde **jsem** /být/VB-P---1P-AA---I **byli**, **choval** ses zrovna tak divně . V noci  
 - jak trefně říkáte - mizející za zatáčkou . Nadšená **jsem** /být/VB-S---1P-AA---I **nebyla**, **jela** jsem tenkrát do Chebu , na představení  
 ) , které nemají přímou empirickou reprezentaci ? A jak **jsem** /být/VB-S---1P-AA---I již **byla pravila** , v sociologii je takových pojmů mnoho  
 ... Ale zase takový frajer jako bratří Mašinové **jsem** /být/VB-S---1P-AA---I **nebyl**. **Věděl** jsem , že to prostě nejde .  
 vám pomáhej . - V tom rozpoložení , v jakém **jsem** /být/VB-S---1P-AA---I **byla** , **přijala** jsem tento další malér - nevím ,  
 vrhlo vstříc : " Ach dědečku ! Dědečku ! Kde **jste** /být/VB-P---2P-AA---I **byl** ! **Byl** jste tak dlouho pryč . " Děd  
 i větší kameny z podlahy a ze zdí . Nejvíce **jsem** /být/VB-S---1P-AA---I se **byl zaměřil** zajisté na věž , o níž jsem  
 Čáral jsem po ulicích a vybíral popelnice . Takhle hladověj **jsem** /být/VB-S---1P-AA---I **byl** . **Poflakoval** jsem se kolem restaurací a koukal voknama

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

a co ho nestojí příliš úsilí . Protože i já **jsem** /být/VB-S---1P-AA---I **byla** lenivá **milovala** jsem potěšení , vzniklo mezi námi okamžité  
 ) , které nemají přímou empirickou reprezentaci ? A jak **jsem** /být/VB-S---1P-AA---I již **byla pravila** , v sociologii je takových pojmů mnoho  
 i větší kameny z podlahy a ze zdí . Nejvíce **jsem** /být/VB-S---1P-AA---I se **byl zaměřil** zajisté na věž , o níž jsem  
 nástroj volá o pomoc Zimní klavír Největší jsem byla když **jsem** /být/VB-S---1P-AA---I **byla** malá **smýčila** jsem hvězdy z jehličnaté oblohy žebříky mi  
 dělat znamenité obchody . O páté hodině odpolední , když **jsem** /být/VB-P---1P-AA---I **byli** značně **pookřáli** , vydali jsme se na další pochod  
 nejsem MLADÁ ŽENA Ano Pohlédne na něj příkře No tak **jsem** /být/VB-S---1P-AA---I tedy **byla Byla** jsem s Bastem celý večer Byla jsem  
 ! Poslední roky jsem byl všelike ve světě a ať **jsem** /být/VB-S---1P-AA---I **byl** kde **chtěl** , nikde jsem o svým starým kámošovi  
 oboru realistickém , co-staruju v dobré komedii , jak už **jsem** /být/VB-S---1P-AA---I ti **byl sdělil** , a je to zatím zde veliký  
 Dne 13 . dubna asi kolem osmé ranní , když **jsem** /být/VB-P---1P-AA---I se **byli přiblížili** k našemu místu určení , zůstal vlak  
 pokračovat : dálka hloubka výška ) ale také závrat když **jsi** /být/VB-S---2P-AA---I **byl** chlapec **lezls** na stromy a střechy vlastně se bojíš  
 vzlykat . Willi ji hladil po vlasech . " Jak **jsi** /být/VB-S---2P-AA---I sama **byla řekla** , " pravil mýně , " teď  
 Tvé o tom , cože vzešlo ze solí , kteréž **jsem** /být/VB-S---1P-AA---I **byl** Tobě **poslal** . Je to nedobré a jasně značí  
 Chodím , kamaráde , kolem dřev a trámů , když **jsem** /být/VB-S---1P-AA---I **byl** včera **dodělal** toho Charona a nutkají mě ty trámy  
 celých následujících šestnáct let jsem nedokázal dostátí slibu , jenž **jsem** /být/VB-S---1P-AA---I mu **byl dal** . Neobjevil jsem ani vraha , ani

## 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 přičestí tvořící s pomocným slovesem složený tvar nenajdeme. Pokud jde o výskyt l-ového přičestí na dalších pozicích vpravo od pomocného slovesa (varianty „*jsem* <.\*> *přišel*“, „*jsem* <.\*> <.\*> *přišel*“), tak lze konstatovat, že frekvence postupně klesá.

## Formulace dalších otázek vypluvších ze zkoumání daného jevu

Pozorujte chyby (většinou v interpunkci), které brání efektivnímu vyhledávání případů, kdy tvar l-ového přičestí 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 přičestí 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 přičestí, 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.

za data za rok 2003 vrátit . " Na základě tohoto rozsudku **jsme** /být/VB-P---1P-AA---| pak pěti dalším městům včetně Hronova peníze přijaté za počítačová data **vrátili** , které zůstávají , by neměly před námi ukrýt to , co **jsme** /být/VB-P---1P-AA---| se již díky výzkumům těchto velkých krizí o historii našeho světa **dozvěděli**

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 pojít s řadou slov. Kromě tvarů přičestí, 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 přítomných, 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 2007<sup>1</sup>). Volný morfém *-s* se často pojí s tvarem l-ového přičestí slovesa významového (*přišels*). V morfologické značce takovýchto l-ových přičestí 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 zvrátneho *se/si*, pak je z řady důvodů snazší začít vyhledávání právě od tohoto tvaru, a to proto, abychom odlišili případy, kdy l-ové přičestí 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 přičestím, např. *přišels*, nebo poneciálně n-/t-ovým přičestím, 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

označovány 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.*"]`.

z okna .) Ale nezví ho do hospody , <b>kdybys</b> /kdyby/J.-S---2-----	ho mohl setřást nějakým jiným . . . BARNEY
zdůrazňuju , a navíc za to pykám , až by <b>ses</b> /se/P7-X4--2-----	divila . Joanna mi to může dosvědčit . " Fakticky
se Chloe na Sandie . " Vsadím se , že <b>sis</b> /se/P7-X3--2-----	uvědomila , žeš udělala správně . . . když jsi
Koneckonců jste jeden druhého neviděli pěknou řádku let . Možná <b>ses</b> /se/P7-X4--2-----	o to pokusil trochu předčasně . " Jejich cesta se
" " Původně jsme byly přesvědčené o tom , že <b>ses</b> /se/P7-X4--2-----	zamilovala do Bjerna , " nadhodila Michelle . " Cooo
pomodlili ? Tak se pojdte nasnídat . " " Umyl <b>sis</b> /se/P7-X3--2-----	ruce ? Tak mi pomoz udělat kvásek . " "
, kdepak . Tohle bylo pouze první dějství . Těšil <b>ses</b> /se/P7-X4--2-----	na pokračování , na další zvrácené a hrůzné slasti .
" pokračoval Tony . " Od toho okamžiku , kdy <b>ses</b> /se/P7-X4--2-----	vrátila domů a musela jí napravit hlavu výpraskem . Tenkrát
holčičko . Ale podle mě bylo daleko děsivější , když <b>sis</b> /se/P7-X3--2-----	ho brala . " " Tohle vždycky říkal Jack .
to " myslí dobře " . Byl bych rád , <b>kdybys</b> /kdyby/J.-S---2-----	opravdově okřál a pokračoval ve zrozumovací nepolitické politice nejméně do
do hlavy kamenem ? " Whitticombe strnul . " Myslel <b>sis</b> /se/P7-X3--2-----	, že to nevím , co ? Ale byla jsem
radost , že nejsem baletní mistr v Košicích , a <b>abys</b> /aby/J.-S---2-----	to taky některým vybraným oslům , kteří budou mít vztek
, že nedá najevo svůj strach . " Jenže to <b>ses</b> /se/P7-X4--2-----	spletla , " prohodil Zeke , který mezitím přišel až
svatý slyšet v jemném vánku . Tož Ti přejí , <b>abys</b> /aby/J.-S---2-----	v jemném ovzduší májovém , v přírodě i v církvi
To jediný dává smysl , " řekl Cassidy . " <b>Kdybys</b> /kdyby/J.-S---2-----	byl ženská a chtěl se pomstít chlapovi , nestřílel bys
že si tvoju lásku nezasloužím . Zasloužil bych si , <b>abys</b> /aby/J.-S---2-----	mě opustila - jenže to bych ti nikdy nedovolil .

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

### Víceúrovňová frekvenční distribuce

Frekvenční limit:

Úroveň	Atribut	Nerozlišovat velikost?	Pozice	(Node) začít od <sup>2</sup>	
1.	lemma ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	-
2.	lc ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	<input checked="" type="checkbox"/>
3.	pos ▾	<input type="checkbox"/>	KWIC ▾	slova KWIC nejvíce vlevo ▾	<input checked="" type="checkbox"/>

## Frekvenční distribuce

Frekvenční limit:

[Nastavit limit](#)

Celkem: 12 (1 str.)

	lemma	lc	pos	Frekvence	
1.	p/ n se	ses	P	7 661	
2.	p/ n aby	abys	J	5 018	
3.	p/ n se	sis	P	4 290	
4.	p/ n kdyby	kdybys	J	2 240	
5.	p/ n ty	tys	P	2 175	
6.	p/ n že	žes	J	1 397	
7.	p/ n co	cos	P	802	
8.	p/ n co	čemus	P	5	
9.	p/ n co	číms	P	4	
10.	p/ n co	čems	P	4	
11.	p/ n kdyby	dybys	J	3	
12.	p/ n co	čehos	P	2	

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 označovanými jako *X.\** končícími na *s*, protože mezi nimi převažují cizojazyčná slova<sup>36</sup>, kterých je velké množství. Vhodnější postup je tedy „namátková kontrola“ jednotlivých slov.

Spolehlivá je anotace slovesných tvarů 1-ového přičestí (viz obrázek). Zvolíme **Typ dotazu cql** a do dotazovacího řádku napíšeme `[tag="V[pq].....2.*" & word="*.s"]`.

<sup>36</sup> Např. plurálové formy anglických slov.

nebyla jsem Edmundem . Ten byl skoro tvým bratrem a	<b>znals</b> /znát/VpMS---2R-AA---I	ho možná líp než sebe . Mě si vlastně nijak
Sáhl po láhvi a znovu si štědře nalil . "	<b>Požádals</b> /požádat/VpMS---2R-AA---P	mě o pomoc , takže ji dostaneš . " Deckard
. Nebo jsi vyvedl soud můj a při mou ,	<b>posadils</b> /posadit/VpMS---2R-AA---P	se na stoličce , soudce spravedlivý . Ohromil jsi národy
zlatem . Nůšemi zlata , " řekla Ilona . "	<b>Slyšelals</b> /slyšet/VpFS---2R-AA---I	pohádku o soli nad zlato ? Nejsou letadla nad zlato
oblečku rozpatlala kusy zmizliny a dortu . Ach jo ,	<b>bylals</b> /být/VpFS---2R-AA---I	strašně zlomyslné děcko . " A změnila jsem se
On se hýbe ! - Na mě se usmál ,	<b>viděls</b> /vidět/VpMS---2R-AA---I	to ? - Vy jste génius , pane doktore !
toho , že ses nechala svést jedním brutálním necitou ,	<b>neprovedlals</b> /provést/VpFS---2R-NA---P	nic hrozného . " Pohlédnu na něho a mám pocit
se tomu na jejím místě taky bránila . . .	<b>neměls</b> /mit/VpMS---2R-NA---I	jet na turnaj ? " " Jo , to měl
Higginbothama , až se ho dotýčný pán zeptal : "	<b>Potkals</b> /potkat/VpMS---2R-AA---P	nějakýho ducha ? " Martin se vzpamatoval a podíval se
Doma Na známých místech známých věcí drůzy , kde prvně	<b>vzhlédls</b> /vzhlédnout/VpMS---2R-AA---IP	v opiovou výš , kovový šum tě kolébá v tvé
a nedovolil , aby se k tobě kdokoliv přiblížil .	<b>Říkalals</b> /říkat/VpMS---2R-AA---I	, že se jmenuješ Gning Gning , a my jsme
bylo moc chytrý . " řekl doktor uštěpačně . "	<b>Měls</b> /mit/VpMS---2R-AA---I	tu takových případů hodně ? " " Jediná vražda ,

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

Následující konkordance ukazují, že v případě homonymie tvaru neurčitého adverbium a tvaru vztažného adverbium 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 zájmeného 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.

## Frekvenční distribuce

Frekvenční limit:  Nastavit limit

**Celkem:** 1 (1 str.)

1.	<u>p/ n</u> kdes	328
<b>lemma</b> <b>Frekvence</b>		

jsem unavenej . V noci jsem ti dvakrát volal . <b>Kdes</b> /kdes/Db-----	byla ? " " To se ptám já tebe !
) Ty seš směšné . Myslíš , že nevím , <b>kdes</b> /kdes/Db-----	byl až do rána ? V Rosicích na pivě .
hřáli u kamen . Drápal zavolal : " Veruno , <b>kdes</b> /kdes/Db-----	koupila ten svetr s těma boulema ? " Byl to
vesluje vrána , z mlází se kouří , a nahoře <b>kdes</b> /kdes/Db-----	hnula se zlehka veveřiči hnízda . . . Někdo jde
Vůbec by mu nedělalo problém být jí věrný . " <b>Kdes</b> /kdes/Db-----	byla celej můj život ? " zeptal se . "
cestovala . Právě jsem se vrátila . " " A <b>kdes</b> /kdes/Db-----	byla ? " " V Naře a v Aomori .
manželem . " Tome , o čem to mluví ? <b>Kdes</b> /kdes/Db-----	ji našel ? " " Pořád to nechápeš ? "
se jich maminka aspoň zeptala : " Kam jdeš ? <b>Kdes</b> /kdes/Db-----	byl ? " Tady se jich nikdo na nic neptal
dělníkú se hemží v hloubi lomů , a parní stroj <b>kdes</b> /kdes/Db-----	v temnu štoly vzdychá . Jak vše tu bílo v
mnoha věcmi - " " Počkej chvíli , Suli . <b>Kdes</b> /kdes/Db-----	tohle všechno sebral ? " " V jednom z pokojů

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.\*"]*.

dobrej ? Jeho syn teď asi brečí . To , <b>cos</b> /cos/NN-----g- mu řekl , Martine , bylo hnusný . Jeho syn
To je báječné ! Ty můj drahej dacane ! A <b>cos</b> /cos/NN-----g- dělal ? JAKUB : Chechtal jsem se . PÁN :
písek ? " dobírala si ji Andrea . " A <b>cos</b> /cos/NN-----g- celou dobu dělala ? Potkala jsi tam nějaké děti ?
to připisovala vinu , omlouvám se . To , o <b>cos</b> /cos/NN-----g- požádal , je ale špatné . Je to špatné pro
( uvažujeme - li diferenciál d ( sin ) = <b>cos</b> /cos/NN-----g- d ) Je zřejmé , že pro jednoznačnou statickou charakteristiku
" To jsem si taky pomyslel . " " A <b>cos</b> /cos/NN-----g- udělal ? Asi sis dal drink . " " Ano
ve všech novinách a v televizi . " " A <b>cos</b> /cos/NN-----g- mu na to řekl ? " zeptal se Christian .
hledí mu na paty . Však vždycky přijde čas a <b>cos</b> /cos/NN-----g- ho vábí zpět , k ženě se vrací zas ,
doufat obmyslem byla věrnost - Od té doby však , <b>cos</b> /cos/NN-----g- ho spatřil , a řekl - proč bych se modlil
do své montážní jámy a vylov toho nejhezčího cucína , <b>cos</b> /cos/NN-----g- v tomhle moly prožraném bazáru kdy měl . "
ten chlap Rettig se mnou odmítá komunikovat . Jenže <b>cos</b> /cos/NN-----g- taky čekal , ty naivko ? položil si Dunlap v

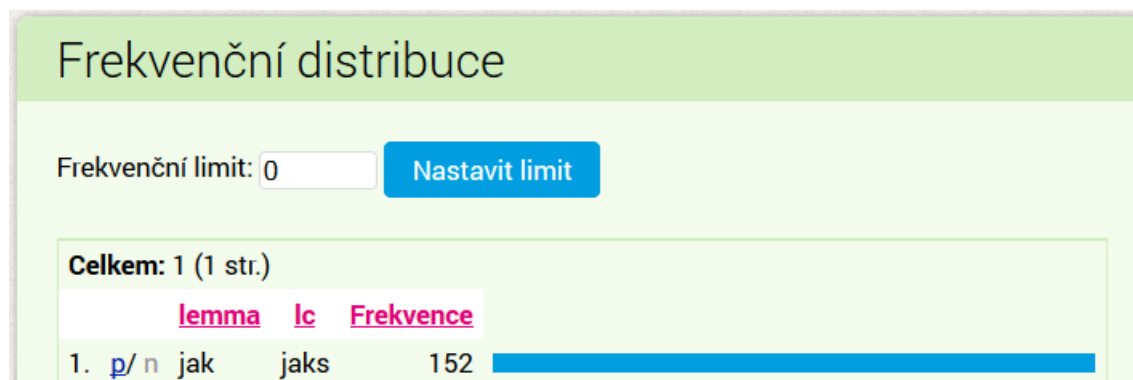
Zvolíme filtr (negativní), rozsah hledání <-1,-1>, **Typ dotazu tag** a do dotazovacího řádku napíšeme *[ZJ].\**. Tímto způsobem odstraníme jasné případy chybné desambiguace vztažného *cos* a můžeme sledovat 75 zbylých řádků.

to připisovala vinu , omlouvám se . To , o <b>cos</b> /cos/NN-----g- požádal , je ale špatné . Je to špatné pro
jsem ovšem nemohl vědět , že ? " Tak <b>cos</b> /cos/NN-----g- tam dělal ? " " To víš , tak jsem
" " Ale vždyť tady nejsi první den . Tak <b>cos</b> /cos/NN-----g- dělal ? " " Plaval , jedl a pil .
a vpadli do kouzelného Melatina . Potůček jím protékající bublal <b>cos</b> /cos/NN-----g- o věčné změně , a než jsme se nadáli ,
stejná jako ta v rozpadu neutronu , tedy s faktorem <b>cos</b> /cos/NN-----g- , zatímco pro vrchol c - d - W je
do noci Prší venku do noci tak tichá v stromech <b>cos</b> /cos/NN-----g- ' jen šelestí a vzdychá ; to jich listy tichounce
zahrad ) Mařenka Paní Kamile Vodičkové K hrobu u zdí <b>cos</b> /cos/NN-----g- mne zvalo na zahradě hřbitova ; nad ním deska bez
" Protože jsem ti loni v zimě dovolila , o <b>cos</b> /cos/NN-----g- žádal . Kdybych se ti nedala tak lacinou , mluvil
se , uměla . Však v téhle chvíli pojednou zašumělo <b>cos</b> /cos/NN-----g- nade mnou , a já jsem ve tmě uviděl vzlétnout
mně vítr v šatech po kapsách slídě zda jsem schoval <b>cos</b> /cos/NN-----g- trvanlivějšího z čeho duše i tam by mohla němě pít
tatík s matkou v pozdní chvíli . V tom obrovského <b>cos</b> /cos/NN-----g- se nad grunt chýlí a vleče přes šindely střechy holé
něj , aby je udržel rovně . " Proč jsi <b>cos</b> /cos/NN-----g- udělal , zvorals to ? " zavrčel . " Koupil
frajerka . Taky že jsem . Ach bože . Máš <b>cos</b> /cos/NN-----g- chtěla . Teddymu bláto evidentně nevádí . Hopsá si v

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-----*. \*

ten hovor už stojí " Když si pomyslím , <b>jaks</b> /jak/Db-----	na toho hochu celou válku čekala - já jen ,
" O čem máme lhát ? " Tak , <b>jaks</b> /jak/Db-----	to už dělal , napřed na severu v okrese Purdom
buď naklepej , nebo namluv jen stráněčku o tom , <b>jaks</b> /jak/Db-----	za první války světové sbíral vajgly , dělal z nich
bych , že ses nakazil od ty pouliční děvky , <b>jaks</b> /jak/Db-----	byl posledně v Baton Rouge . " Nikdy jsem nebyl
nevolky válejí v hlavince . Z toho hrubého obrysu , <b>jaks</b> /jak/Db-----	naznačil , mi čím dál více připadá , že bychom
hochu , řekl si . Musels to udělat tak , <b>jaks</b> /jak/Db-----	to udělal . " Henry , " zeptal se .
! Podívej se , prosím tě , jak vypadáš a <b>jaks</b> /jak/Db-----	' to tu zaprasila ! " Basket strhl ze stolu
podívat do pracovny , jestli je tam všechno tak , <b>jaks</b> /jak/Db-----	to včera opustila ? " Odpoledne jsem pár hodin přebírala
zapalovat pánům cigarety ? A takovým zapalovačem , McNeilová ; <b>jaks</b> /jak/Db-----	k němu přišla ? Je odporný . " A skutečně
lidí , se kterejma se tolik oháníš , dozvědí , <b>jaks</b> /jak/Db-----	zasabotoval Delaeroixovu popravu - " " Zasabotoval - "



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.

### 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.



## **Adverbia a další neohebné slovní druhy**

O adverbiiích jsme se zmínili již v souvislosti se značkováním a tvořením komparativu a superlativu adjektiv. V této kapitole nás bude zajímat to, jak lze na základě dat získaných z jazykových korpusů sledovat proces adverbializace při vzniku tzv. příslovečných spřežek.

### ***Příslovečné spřežky***

#### **Motivační úvod**

V *Internetové jazykové příručce* se v oddíle 2 *Nejde o spřažený výraz (časté chyby)* uvádí toto: „Jelikož se v následujících případech velmi často chybuje, připomínáme, že jen zvlášť píšeme zpodstatnělá přídavná jména, která jsou užívána pouze či převážně s předložkou *na* a pojí se 4. pádem. Jsou to spojení jako: *na shledanou, na viděnou, na slyšenou, na rozloučenou, na odchodnou, na uvítanou, na zotavenou, na posilněnou, na pováženou, na uváženou, na rozmyšlenou, na vysvětlenou, na srozuměnou, na upřesněnou, na vybranou, (jako) na zavolanou*. Dalším slovním spojením, které je často mylně chápáno, a proto psáno jako spřažený (jednoslovný) výraz, je *na rozdíl*. Připomínáme, že i v tomto případě je nutné psát uvedené spojení jako slova dvě.“

#### **Nastínění problému**

V rámci systému slovních druhů působí dynamické vztahy, tzv. slovnědruhovú transpozice, které mají za následek přechod slovnědruhovú forem od jednoho slovního druhu ke druhému. Slovtvorné slovnědruhovú transpozice se vyznačují ukončeným přechodem jednotky A patřící ke slovnímu druhu X do jednotky B patřící slovnímu druhu Y. V živém jazyce je však mnohdy tento typ slovtvorného procesu „v pohybu“. Tato skutečnost způsobuje, že slovní druhy fungují jako otevřené třídy s pevným jádrem a proměnlivou vrstvou obklopující toto pevné jádro. Typickým příkladem mohou být adverbia vznikající v procesu adverbializace prostúch a předložkovúch pádú substantiv (u předložkovúch pádú je v psaném jazyce signálem ukončení procesu adverbializace užívání a případná kodifikace příslovečné spřežky).

Má-li automatická morfologická analýza tento jev postihnout, je třeba, aby byl zachycen i ve slovníku automatického morfologického analyzátoru.

V jazykových korpusech můžeme hledat případy, kdy proces adverbializace je graficky završen vznikem kodifikované příslovečné spřežky. V korpusech můžeme najít i doklady, které jsou v rozporu s platnou kodifikací.

Problém korpusových dokladů tkví v tom, že (zejména tam, kde jde o doklady s nízkou frekvencí) nelze spolehlivě rozlišit, zda jde o záměr autora textu (ať už je důvod takového úzu jakýkoli), či o překlep (chybu v grafické realizaci mezery mezi slovy). Korpusové texty (i když se jedná o texty psaného jazyka a u velkých korpusů řady SYN v naprosté většině i o texty, které prošly nějakou korekturou) obsahují chyby (nezáměrně nekorigované překlepy), které se v korpusech při tzv. konverzi textů (převedení do společného formátu strojově čitelné podoby) úmyslně ponechávají tak, jak jsou v originálních textech (neopravují se). Pro jakoukoli interpretaci, něku-li kodifikaci, opřenou o úzus (korpuse) je ale takové rozlišení naprosto zásadní.

Předešleme tedy, že při interpretaci korpusových dokladů pro účely argumentace spojené s kodifikačními zásahy je třeba postupovat obezřetně.

## Otázky

Přesto je možné užívat korpuse jako relevantní zdroje informací o sledovaném jevu. Zaměříme se na to, zda se v korpusu (SYN2010) vyskytují chyby, na něž upozorňuje *Internetová jazyková příručka*, a jak je možné s korpusovými doklady takových chyb dále pracovat.

## Formulace dotazu pro získání dat z korpusů

Data pro pozorování získáme jednoduše. Zvolíme **Typ dotazu fráze** a do dotazovacího řádku zapíšeme postupně *na shledanou, na viděnou, na slyšenou, na rozloučenou, na odchodnou, na uvítanou, na zotavenou, na posilněnou, na pováženou, na uváženou, na rozmyšlenou, na vysvětlenou, na srozuměnou, na upřesněnou, na vybranou, na zavalanou* a zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku zapíšeme postupně *nashledanou, naviděnou, naslyšenou, narozloučenou, naodchodnou, nauvítanou, nazotavenou, naposilněnou, napováženou, nauváženou, narozmyšlenou, navysvětlenou, nasrozuměnou, naupřesněnou, navybranou, nazavalanou*.

A děkuju. " Není zač. Tak - na /na/RR--4----- shledanou /shledaná/NNFS4-----A----- " Kouknul po mně. " Na shledanou, mi nikdy nezavdáš příčinu to zopakovat. Buď pozdraven, na /na/RR--4----- shledanou /shledaná/NNFS4-----A-----, Artuš, králi Stříbrného trůnu. " Společníci se rozhlížející se s přirozenou ladností a půvabem. " Tak na /na/RR--4----- shledanou /shledaná/NNFS4-----A-----, Roddy, " řekl jsem. " Určitě se toho pondělního před Samanthařím domem jsme nijak nepokročili. Bizy na /na/RR--4----- shledanou /shledaná/NNFS4-----A-----, řekla a já řekl že ano. Prý se začíná starostlivě mračit. Mávám na něj. " Bizy na /na/RR--4----- shledanou /shledaná/NNFS4-----A----- ! " Vyznívá to jako rozkaz a on mizí za pořád telefonoval, tak jsem odešel. Nečeká mi ani na /na/RR--4----- shledanou /shledaná/NNFS4-----A----- . Připadal jsem si jako ve filmu Byl jsem při tři sta a přecházím na Středisko Boston. Díky a na /na/RR--4----- shledanou /shledaná/NNFS4-----A----- píšete. " Záběr se stříhem změnil a letadlo teď vás, až budu mít nějaké podstatné informace. Zatím na /na/RR--4----- shledanou /shledaná/NNFS4-----A----- . " Udělal na Donalda gesto, jako by si cítil zavázaný podat dříve kufří mezi dveře a říci jí na /na/RR--4----- shledanou /shledaná/NNFS4-----A-----, ačkoliv na žádné shledání nepomyslel, a zamávat jí poušti - výcvikové středisko jihlavské SPŠ Pouště. Za rok na /na/RR--4----- shledanou /shledaná/NNFS4-----A----- - Mezinárodní policejní závody v Českých Budějovicích. Dvě strany se dozvíš, jak říct " ahoj " a na /na/RR--4----- shledanou /shledaná/NNFS4-----A----- " v deseti různých jazycích. Nejdůležitější věc, kterou něj sestoupil. Byl opět svobodným člověkem. " Takže na /na/RR--4----- shledanou /shledaná/NNFS4-----A-----, pane Poirote. Neodhalil jste v mém domě,
--

tady z našeho gymnázia sedm . S Pánem Bohem a **nashledanou** /nashledanou/TT-----6- ! V Římě 9 . dubna 1940 Milí hoši !  
z vozu . Usmála se a řekla , " Zítřka **nashledanou** /nashledanou/TT-----6- , pane mlíkaři . " Poganowska odpověděl : " Zítřka  
" Oskar se málem zakuckal . " Není zač . **Nashledanou** /nashledanou/TT-----6- , " rozloučila se ukonejšená Němcová . " Mějte se  
pojede snad v smrti klín ! Lituji chyb , a **nashledanou** /nashledanou/TT-----6- ! Svět mizí v dáli - nevrátím se . P.  
" Děkuju vám , paní , nashledanou . " " **Nashledanou** /nashledanou/TT-----6- ! " Obchodnice otevřela dokořán dveře do vyhřátého vzduchu .  
- v červenci a srpnu - bude Klub uzavřen . **Nashledanou** /nashledanou/TT-----6- s vámi se budeme těšit v září . Přejeme vám  
není , to je jen naše věc . Díky a **nashledanou** /nashledanou/TT-----6- , " řekla nám . Zda je Preissův dobrovolný odchod  
vyměnil pizzy . PIZZAMANN Riziko podnikání . Tak tedy : **nashledanou** /nashledanou/TT-----6- ! A nechte si tady i to pít ! (   
. A ještě jednou děkuju . Přju dobré zdraví a **nashledanou** /nashledanou/TT-----6- . " Z restaurace vyšel Hadrava s dobrým pocitem ,  
děkuji , " uvedl František Trhlík . " Za rok **nashledanou** /nashledanou/TT-----6- v Kostelci nad Orlicí , " neslo se náměstím těsně  
na Třeboň těším , tak domluveno , zítra po deváté **nashledanou** /nashledanou/TT-----6- . " " Nashle . " Major Křelina si po  
stůl . Voda mi brzy kapala na boty . " **Nashledanou** /nashledanou/TT-----6- , " slyšel jsem její tichý úsměvný hlas . Nejspíš

## Třídění a pozorování dat získaných z korpusů

Vytvoříme tabulku a zapíšeme počty tvarů

<i>na shledanou</i>	200	<i>nashledanou</i>	73
<i>na viděnou</i>	31	<i>naviděnou</i>	2
<i>na slyšenou</i>	4	<i>naslyšenou</i>	0
<i>na rozloučenou</i>	818	<i>narozloučenou</i>	0
<i>na odchodnou</i>	17	<i>naodchodnou</i>	0
<i>na uvítanou</i>	169	<i>nauvítanou</i>	0
<i>na zotavenou</i>	61	<i>nazotavenou</i>	0
<i>na posilněnou</i>	12	<i>naposilněnou</i>	0
<i>na pováženou</i>	115	<i>napováženou</i>	0
<i>na uváženou</i>	4	<i>nauváženou</i>	0
<i>na rozmyšlenou</i>	246	<i>narozmyšlenou</i>	0
<i>na vysvětlenou</i>	165	<i>navysvětlenou</i>	0
<i>na srozuměnou</i>	165	<i>nasrozuměnou</i>	1
<i>na upřesněnou</i>	0	<i>naupřesněnou</i>	0
<i>na vybranou</i>	802	<i>navybranou</i>	0
<i>na zavolanou</i>	64	<i>nazavolanou</i>	0

## Formulace závěrů

Za relevantní závěr lze, jak je patrné, pokládat data v případě, kdy je *na shledanou* psáno dohromady (*nashledanou*). Zvolíme-li **Možnost zobrazení**, **Reference** např. opus.genre (žánr), můžeme sledovat, v jakých typech textů se tato chyba proti stávající kodifikaci vyskytuje.

Formulace dalších otázek vyplynuvších ze zkoumání daného jevu

Zajímavá je otázka, jak jsou substandardní tvary značkovány. Pokud budeme chtít v korpusech vyhledat podklady jazykových dat, která jsou z hlediska kodifikace v pohybu, nebo jsou s ní přímo v rozporu, pak je nejlépe vyhledávat podle slovního tvaru a nespolehat se na výsledky automatické morfologické analýzy a desambiguace. Pro příklad můžeme sledovat posun ve značkování nekodifikovaného tvaru *nashledanou* v jednotlivých korpusech řady SYN.

<i>nashledanou</i>	tag	počet tvarů
SYN2000	XX-----	75
SYN2005	NNFS4-----A-----	49
	VB-P---3P-AA----	27
	AAFS4----1A-----	10
	AAFS7----1A-----	3
SYN2006PUB	TT-----6-	73
SYN2009PUB	TT-----6-	284
SYN2010	TT-----6-	73

Výsledky značkování v korpuse SYN2005 je třeba přičíst na vrub okolnosti, že na tomto korpuse byly testovány tzv. guessery (více Hlaváčová 2001, Jelínek 2008).

### **Zadání cvičení, v nichž lze uplatnit analogické postupy**

Najděte v pravidlech pravopisu případy, kdy je dovoleno psát adverbializované pády jmen zvlášť i dohromady, a sledujte zastoupení variant v korpusech.

## Značkování neohebných slovních druhů

V této kapitole se budeme zabývat otázkou lemmatizace a značkování neohebných slovních druhů a naznačíme problémy, spojené s existencí slovnědruhových přechodů na rovině značkování a desambiguace.

### *Ještě několik slov ke značkování neohebných slovních druhů s ohledem na slovnědruhovú transpozice*

#### **Motivační úvod**

Jednou z oblastí slovnědruhových přechodů je tzv. prepozicionalizace, při níž se z adverbia (mnohdy vzniklého ustrnutím pádu jména) stane nevlastní předložka. Mějme určit slovnědruhovou platnost tvaru *místo* v následující větě:

*<Místo> vašich schůzek sis měl lépe vybírat.*

Tato věta není z korpusů, ale demonstrujeme na ní řídký případ, kdy slovnědruhovú homonymie na úrovni jednotlivého tvaru způsobí dvojnou interpretaci na úrovni celé věty. Z korpusu je následující doklad: *První muž se ho pokusil obelstít na trajektu Star, což bylo <místo> dohodnuté schůzky.* V tomto případě vyloučíme asi možnost, že jde o předložku. K vyloučení interpretace je třeba analyzovat celé souvětí vč. koreference.

#### **Nastínění problému**

Homonymie může nastat tam, kde je na základě kontextu možné obojí čtení. Kontext je dvojný, jazykový a mimojazykový. Z hlediska jazykového kontextu je pro interpretaci tvaru jako předložky třeba splnit podmínku, že za tímto tvarem následuje tvar jména / tvary jmen (jmenná skupina) v tom pádě, se kterým se předložka pojí. To je podmínka nutná, nikoli postačující.

#### **Otázky**

Podívejme se na to, jak jsou v korpusech značkovány nepůvodní předložky (seznam najdete např. v mluvnicích). Pokusme se najít případy, kdy tvarová homonymie na úrovni slovního druhu je důvodem chybné desambiguace.

#### **Formulace dotazu pro získání dat z korpusů**

Zvolíme **Typ dotazu slovní tvar** a do dotazovacího řádku zapíšeme *kolem*. Zvolíme filtr (pozitivní), rozsah hledání <-1,-1>, **Typ dotazu cql** a do dotazovacího řádku zapíšeme

[tag=" [APC].N.7.\*"]. Chceme tak získat řádky, na nichž se může vyskytnout substantivum *kolo* rozvíte shodným přívlastkem (adjektivem, zájmenem nebo číslovkou v 7. pádě středního rodu).

### Filtr konkordancí

Filtr:  pozitivní  negativní

Vybraný token:  první  poslední

Rozsah hledání: od -1 do -1  včetně KWIC

Typ dotazu: CQL

CQL:  vložit tag vložit "within" klávesnice

Implicitní atribut: word [Popis morfologických značek](#)

Poté ponecháme filtr (pozitivní), zadáme rozsah hledání <1,1>, **Typ dotazu cql** a do dotazovacího řádku zapíšeme [tag=" ....2.\*"]. Chceme získat pouze ty případy, kdy za tvarem *kolem* následuje jméno ve 2. pádě (předložka *kolem* se pojí se 2. pádem).

### Hledat v korpusu

Korpus:

Typ dotazu: CQL

CQL:  vložit tag vložit "within" klávesnice

Implicitní atribut: word [Popis morfologických značek](#)

Týž postup zopakujeme pro vyhledání tvaru *místo*, přičemž omezíme výběr konkordančních řádků na ty případy, kdy bezprostředně vpravo (<1,1>) za tímto tvarem stojí tvar označovaný jako 2. pád (...2.\*) a bezprostředně vlevo (<-1,-1>) před tímto tvarem stojí tvar označovaný jako adjektivum, zájmeno nebo číslovka středního rodu a jednotného čísla v 1. nebo 4. pádě ([APC].NS[14].\*).

### Třídění a pozorování dat získaných z korpusů

Nyní si pozorně prohlédneme konkordance a zjistíme, že v desambiguaci se vyskytují chyby.

čekat nedokážu . " Stál tam s bílým ručníkem přehozeným kolem /kolo/NNNS7-----A----- krku a rukama svíral jeho konce . Pustila košili zpátky

případných smrtících látek . " Přenesla pohled ke knihám rozloženým kolem /kolo/NNNS7-----A----- sebe . " Zdá se , že jsi měla napilno

květinou , kde nebyla růží . S copem třikrát **pleteným kolem** /kolo/NNNS7-----A----- **čela** , hořekala veselím a radostněla zpěvem . Říkávala :

Chybnou desambiguaci pozorujeme i v případě konkordancí tvaru *místo*.

zná název Khao Lak . V únoru ves jako **nejpostiženější místo** /místo/RR--2----- **Thajska** navštívili američtí prezidenti George Bush starší a Bill Clinton

nenáviděnému " iotistovi " Jan Nejedlý a prosadil na **ono místo** /místo/RR--2----- **svého** chráněnce . Když se v téže době uvolnilo místo

a u něj parkoviště . Od vlakového nádraží je **toto místo** /místo/RR--2----- **vzdálené** asi 500 metrů . Přímou proti parkovišti , už

? Jak to kterýkoliv muž může vůbec vědět ? **Ono místo** /místo/RR--2----- **ženské** odezvy je temným lesem . Jak může nějaký muž

### Formulace závěrů

Je patrné, že slovnědruhové přechody, které mají za následek vícero interpretací na úrovni lemmatizace, slovního druhu i dalších slovnědruhově závislých kategorií, značně ovlivňují obtíž a chyby na úrovni desambiguace. S výsledky automatické analýzy je tudíž třeba pracovat opatrně s vědomím toho, že je třeba údaje vždy zkontrolovat.

### Formulace dalších otázek vyplynulých ze zkoumání daného jevu

Ve výběru konkordancí, na nichž jsme zkoumali úspěšnost desambiguace, jsme vycházeli z toho, že ve zvolených případech by mohlo dojít k problémům při aplikaci desambiguačních pravidel založených na lingvistických předpokladech. Nezkoumali jsme úspěšnost všech konkordančních řádků. Vybrali jsme pouze ty, pro které platilo, že tvar lze na základě lingvistické analýzy bezprostředního kontextu interpretovat obojím způsobem. Pravidlová desambiguace byla použita pro odstranění některých chyb desambiguace stochastické. Zajímavé by bylo porovnat chybovost v jednotlivých korpusech řady SYN a v nereferenčním korpusu SYN (více Jelínek 2008, Skoumalová 2011).

### Zadání cvičení, v nichž lze uplatnit analogické postupy

Sledujte, jak je provedena desambiguace u tvarů *během*, *bokem*, *úderem*, *stranou*, ...  
Postupujte podobným způsobem, jak bylo naznačeno.

## Korpus a některé pravidelné derivace

Některé slootovorné kategorie jsou v rámci automatické morfologické analýzy zahrnuty do morfologického značkování slovních tvarů (Hajič 2004, Osolsobě 1996). Prostudujeme-li Popis morfologických značek v manuálu k programu *Bonito* na www stránkách Ústavu ČNK (viz elektronické zdroje), zjistíme, že na pozici 2 morfologické značky se řada hodnot vztahuje ke slootovorným třídám a typům. Subklasifikace adjektiv zahrnuje značku *AU.\** (adjektivum přivlastňovací na *-ův* i *-in*), značku *AG.\** (přídavné jméno odvozené od slovesného tvaru přítomného přechodníku), *AM.\** (přídavné jméno odvozené od slovesného tvaru minulého přechodníku). Subklasifikace zájmen a číslovek koresponduje s jednotlivými druhy zájmen a číslovek, přičemž u číslovek lze říci, že jednotlivé druhy číslovek se dosti pravidelně derivují od číslovek základních (Osolsobě 1995), takže i tyto charakteristiky můžeme použít pro sledování slootovorných vztahů uvnitř této kategorie. Do oblasti slootovorby se ve většině lingvistických prací řadí stupňování (viz výše).

Popis slootovorby jako celku není na úrovni morfologického značkování zahrnut. Přesto lze z jazykových korpusů poměrně snadno získávat data pro výzkum v oblasti tvoření slov.

V následujících kapitolách ukážeme, jak lze aplikovat znalosti o gramatických vlastnostech slovesných tvarů při vyhledávání některých typů deverbativ. Dále se zaměříme na desubstantivní adjektiva v konstrukcích tzv. hebrejského superlativu, formálně připomínající procesuální adjektiva. Ukážeme, jak mohou korpusy přispět k charakteristice tohoto okrajového derivačního typu. Na závěr se zaměříme na jeden typ hláskových alternací (vokalické alternace v českých prefixech). Ukážeme, jak lze z jazykových korpusů získat relevantní data pro pozorování příslušného jevu.

### ***Deverbativa od slovesného kmene a jejich vyhledávání v korpusech***

#### **Motivační úvod**

Výzkum slootovorby se orientuje jak na výzkum fungování slov utvořených, tak na vznik nových pojmenování různými slootovornými postupy. V korpusech vidíme slova v přirozených kontextech jejich užití, korpusy zahrnují velké množství textů novějších, než jsou ty, ze kterých vyšly dostupné tištěné výkladové slovníky. Je tudíž užitečné vědět, jak lze korpusy použít pro vyhledávání a třídění dokladů pro výzkum slootovorby. Slova utvořená nějakým slootovorným prostředkem (prefix, sufix) lze v korpusech poměrně snadno vyhledávat. Už v kapitole věnované adjektivům jsme ukázali, jak lze využít morfologické



vlastnosti slovesa při hledání kandidátů na účelová adjektiva. V této kapitole budeme používat analogické postupy.

### **Nastínění problému**

O některých deverbativech se v mluvnicích tvrdí, že se tvoří paradigmaticky (lze je utvořit od každého slovesa bez omezení). Platí to o dějových jménech tvořených od tvaru shodného s potenciálním tvarem pasivního přičestí. Zatímco adjektiva z přechodníků mají ve značkování použitém pro anotace korpusů řady SYN zvláštní značky, substantiva na *-ní/-tí* jsou označena jako všechna ostatní substantiva. Tato praxe neodpovídá běžné praxi českých výkladových slovníků. Ty totiž pravidelně tvořená dějová jména nezaznamenávají jako samostatná heslová slova. Deverbální substantiva na *-ní/-tí* si ovšem uchovávají řadu gramatických vlastností fundujícího slovesa (vid, fakultativně zvrtnost signalizovanou morfémem *se/si*, do určité míry valenční doplnění). Z tohoto hlediska představují slovtvorný typ, jehož zkoumání může být z mnoha aspektů zajímavé.

### **Otázky**

Slovtvorným základem pro odvozování jmen (substantiv a adjektiv) od sloves může být buď kořen (základ) slovesa (např. *mluv-i-t > mluv-čí*), nebo jeden ze slovesných kmenů (kořen rozšířený o kmenotvornou příponu např. *uč-i-t > uč-i-tel*), nebo slovesný tvar (např. *kut-i-t > kuti-i-l/Vp.\* > kut-i-l/N.\**). V mluvnicích se v oddílech věnovaných slovtvorbě hovoří o derivaci od kořene a od kmene.

Které typy deverbativ jsou derivovány od slovesného tvaru, které od kmene a které od kořene? Většinou se v mluvnicích uvádí, že odvozovací kmen je kmen minulý. Pod tímto termínem se ale rozumí jistá abstrakce, protože tvary od kmene minulého (někdy též infinitivního) mohou být až tři<sup>37</sup> (infinitiv – *brát*, l-ové přičestí + od něj tvořený přechodník minulý *bral/nabrav*, pasivní přičestí + od něj tvořené verbální adjektivum a substantivum *brán, braný, braní*). Je patrné, že kmenotvorná přípona (v našem případě *a/á*) alternuje, má více variant. Je vždy potřeba zvážit, která z variant je spolufornantem při odvozování od slovesného kmene.

### **Formulace dotazu pro získání dat z korpusů**

---

<sup>37</sup> Více Osolsobě 2011<sup>1</sup>.

Pokusme se vyhledat v korpusu všechna substantiva slovesná na *-ní/-tí*. Uvedli jsme, že tato substantiva se tvoří od tvaru shodného s potenciálním tvarem přičestí trpného. Mohli bychom tedy vyjít ze znalosti morfemické stavby pasivního přičestí u jednotlivých slovesných tříd a formulovat dotaz pro korpusový manažer tak, abychom zachytili pouze ta substantiva, která by mohla být dějovými jmény. Můžeme ale také zvolit postup kombinovaný. Budeme si všimnout korpusových dat a porovnávat výsledky jejich pozorování se znalostmi morfologie slovesného tvaru.

Dějové jméno se musí skloňovat podle vzoru *stavení* a musí končit na řetězec *[nt]i*. Zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme `[lemma=".*[nt]i" & tag="NNN.*"]` (hledáme všechna substantivní neutra, jejichž lemma končí na řetězec *-ní/tí*). Podívejme se na frekvenční distribuci lemmat.

### Hledat v korpusu

**Korpus:**

**Typ dotazu:**

**CQL:**

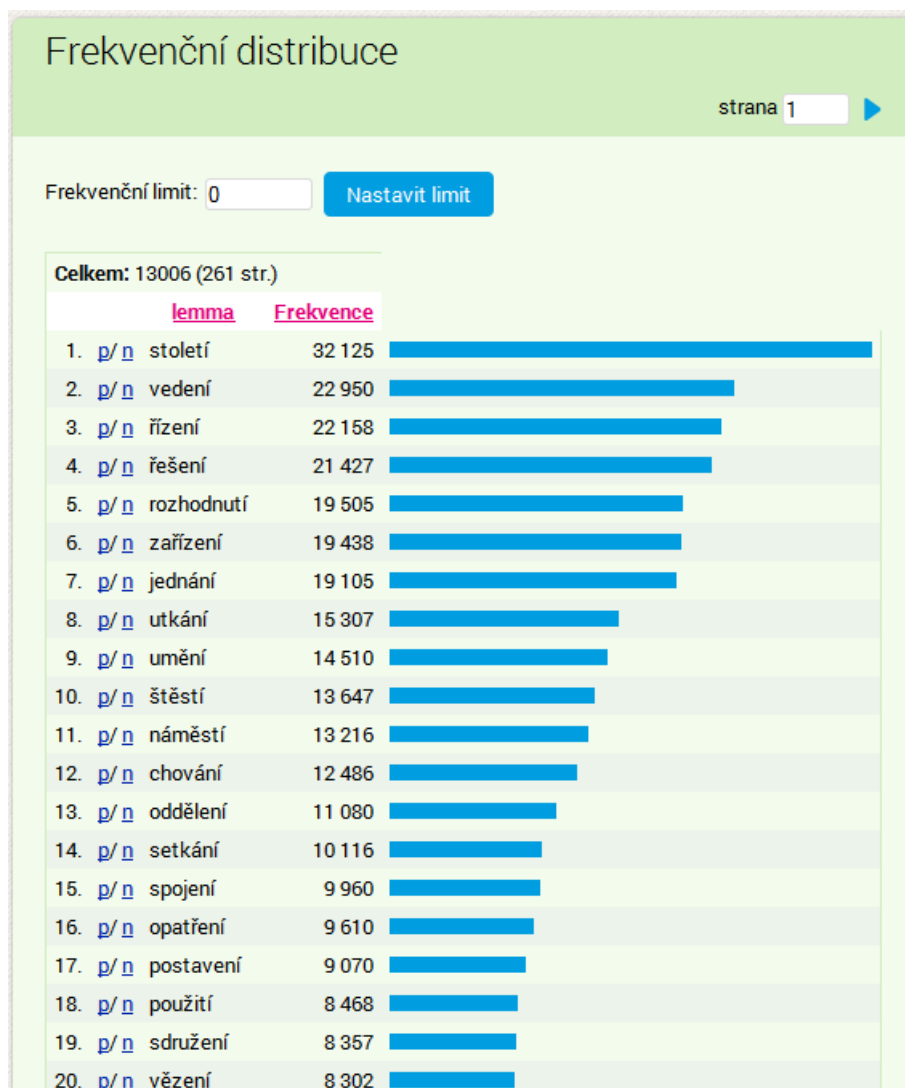
Implicitní atribut:  [Popis morfologických značek](#)

zákon o registračních pokladnách vyvolává podivení i obavy zákon o právní jistoty , a to vzhledem k tomu , že	<b>vydávání</b> /vydávání/NNNS6-----A-----	dokladů za částky vyšší než 50 Kč . Nepochopitelnou se
do roku 1374 . Nepravidelný půdorys západního dvoulodí ambitu sleduje	<b>řízení</b> /řízení/NNNS1-----A-----	podle čl. 234 SES ve vztahu k otázkám neplatnosti nestanovuje
matka neuměla ty krásně prožité červenové dny dvašedesátého roku pojmenovat	<b>zakřivení</b> /zakřivení/NNNS4-----A-----	Štupartské ulice ( dodržení uliční čáry bylo pro středověkou zástavbu
nebo informací potřebných k jeho vyřešení - stereotypy , neschopnost	<b>přijímání</b> /přijímání/NNNS7-----A-----	, na to jsem myslel a kráčel již dávno asfaltovou
že mé ženě hrozí nebezpečí . " Měl jsem	<b>izolování</b> /izolování/NNNS2-----A-----	problému , neschopnost vidět problém z jiné strany .
a nároky . To proto , abychom s výběrem a	<b>vidění</b> /vidění/NNNS2-----A-----	, pane Thorne . . . " Povídá jste
. Computer Press 2007 , 208 s . , 249 Kč	<b>umístění</b> /umístění/NNNS7-----A-----	hydromasážní vany byli spokojeni i za dva , pět ,
doma tkané látky a ozbrojených ručnicemi . Když dojeli k	<b>řešení</b> /řešení/NNNS1-----A-----	manažerských problémů pomocí nástrojů dynamické simulace je moderní a efektivní
nebo na taneční hodinu , což by mohlo ohrozit vaše	<b>úpatí</b> /úpatí/NNNS3-----A-----	skaliska , zastavili se ke krátké poradě . " Prameny
vnějšího prostředí ) na vylučování hormonů působí . Vysvětlíme to	<b>předsevzetí</b> /předsevzetí/NNNS4-----A-----	pustit se do cvičení . Při cvičení doma můžete číhat
perzekuce Srbů a Romů ze strany albánských šovinistů ) po	<b>propojení</b> /propojení/NNNS7-----A-----	mezi nervovou soustavou a hormonální soustavou prostřednictvím podvěsku mozkového (
, že Británie válku prohrává . Oficiální zůvodnění pro Hessovo	<b>omezení</b> /omezení/NNNS6-----A-----	? lidskoprávního ? zájmu jen na Tibet , Kubu či
Pešková MŠVinařská Největší trest promě za zlobení by byl zákaz	<b>přemístění</b> /přemístění/NNNS4-----A-----	je však mnohem bizarnější . V květnu 1942 zřejmě nastaly
motivace , proč jsem v politice , je nějaké vnitřní	<b>koukání</b> /koukání/NNNS2-----A-----	na televizi . Filip Hartl MŠVinařská Kdybych hodně zlobil ,
poúnorové kulturní politiky představovaly vlastně již čistky realizované krátce po	<b>přesvědčení</b> /přesvědčení/NNNS1-----A-----	, že je možné si představit změnu k lepšímu a
	<b>uchopení</b> /uchopení/NNNS6-----A-----	moci . Klasické stalinské heslo o tom , že "

To, že lemma substantiva neutra končí na *-ní/tí*, je podmínka nutná, nikoli postačující pro to, aby slovo bylo dějovým jménem, jak o tom svědčí slova jako *století*, *štěstí*, *náměstí*, *zápěstí*, *ústí*, ... Podíváme-li se na tato slova blíže, vidíme, že některá mají společnou vlastnost, kterou nemá ani jedno ze slov ze zobrazené části seznamu, která za dějová jména pokládat lze. O

jakou vlastnost se jedná? U substantiv *šťěstí* a *náměstí* stojí před *-tí* souhláska. Všechny ostatní doklady (*rozhodnutí, užití, napětí, ...*) mají před *-tí* samohlásku. Také všechna zobrazená neutra na *-ní*, která lze pokládat za dějová jména, mají před *-ní* samohlásku. Můžeme si položit další otázky. Může před *[nt]í*, má-li jít o dějové jméno, stát souhláska? Může před *[nt]í*, má-li jít o dějové jméno, stát libovolná samohláska? Odpovědi budeme hledat nejdříve v korpusu.

Podívejme se na frekvenční distribuci lemmat.



### Třídění a pozorování dat získaných z korpusů

Zvolíme filtr (pozitivní), **Typ dotazu lemma** a do dotazovacího řádku zapíšeme *.\*[bcčdd'fghjklmnňpqrřsšt'vwxyzž][nt]í*. Podívejme se na frekvenční distribuci lemmat. Projdeme-li seznam 56 lemmat, zjistíme, že dějová jména se nevyskytují. To znamená, že ve zvoleném korpusu není doloženo žádné dějové jméno sledovaných vlastností. Vrátime se o

krok zpět, zvolíme filtr (negativní) a do dotazovacího řádku zapíšeme *.\*[bcčdd'fghjklmnňpqrřsšt'vwxxž][nt]í*.

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od  do   včetně KWIC

**Typ dotazu:**  ▼

**Dotaz:**  klávesnice

### Frekvenční distribuce

strana 1 ▶

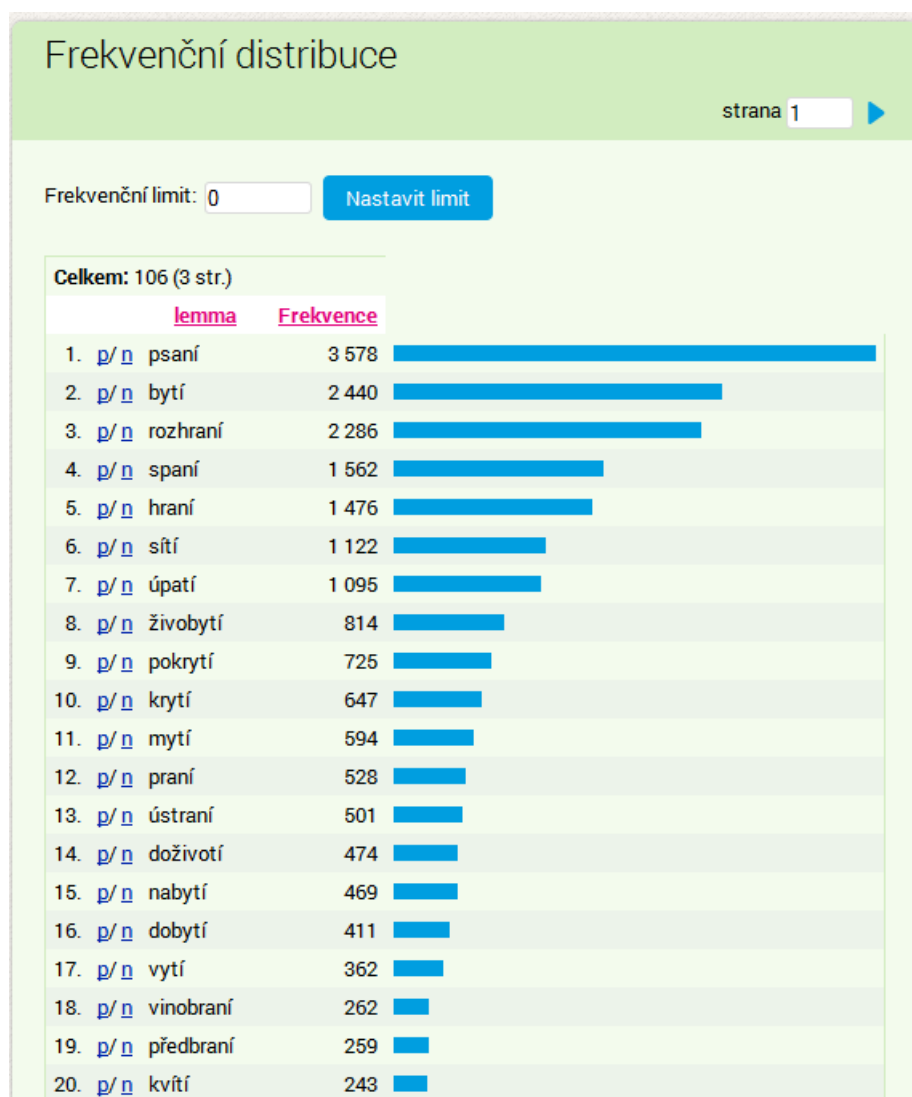
Frekvenční limit:

**Celkem: 56 (2 str.)**

	lemma	Frekvence
1.	p/ů štěstí	13 647
2.	p/ů náměstí	13 216
3.	p/ů Ústí	4 989
4.	p/ů neštěstí	2 981
5.	p/ů úmrtí	2 227
6.	p/ů listí	2 039
7.	p/ů zápěstí	1 893
8.	p/ů předměstí	1 776
9.	p/ů ústí	1 368
10.	p/ů předloktí	747
11.	p/ů rozcestí	393
12.	p/ů houští	380
13.	p/ů zmí	299
14.	p/ů Pobaltí	292
15.	p/ů tmí	272
16.	p/ů předmostí	146
17.	p/ů scestí	139
18.	p/ů roští	105
19.	p/ů klestí	69
20.	p/ů nadloktí	64

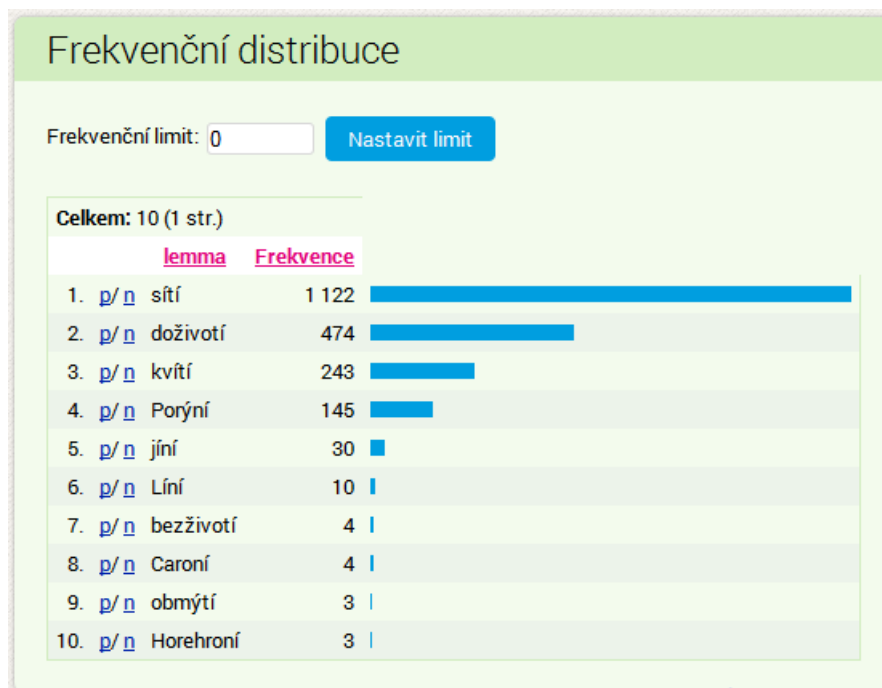
Dále budeme pozorovat, které samohlásky se vyskytují před *[nt]í* u substantiv, která jsou prokazatelně dějovými jmény. Mezi substantivy na první zobrazené stránce frekvenčního seznamu nacházíme taková, že před *[nt]í* předchází *[áeěiu]*. Všimneme si, že kromě *á* jde ve všech ostatních případech o krátký vokál. Nejprve se můžeme zaměřit na ty vokály, které chybějí, tj. *[aéioóůýý]*.

Vrátíme se k původní konkordanci, zvolíme filtr (pozitivní), **Typ dotazu lemma** a do dotazovacího řádku zapíšeme *.\*[aéioóůýý][nt]í*. Podívejme se na frekvenční distribuci lemmat. Pozorujeme, že předcházejí-li další krátké vokály *[aiy]*, pak jde až na jedinou výjimku (*úpatí*) o dějové jméno (nebo kompozitum, jehož druhým členem je dějové jméno). Naopak předcházejí-li dlouhý vokál, pak nejde v obou případech o dějové jméno (*sítí, kvítí*). Nemáme k dispozici doklady o jednom krátkém a dalších (kromě *i*) čtyřech dlouhých vokálech *[oéioóůýý]*.



Vrátíme se zpět. Zvolíme filtr (pozitivní), **Typ dotazu lemma** a do dotazovacího řádku zapíšeme *.\*[oéioóůýý][nt]í*. Podívejme se na frekvenční distribuci lemmat. Pozorujeme, že mezi

10 nalezenými lemmaty není ani jedno dějové jméno. Některé hledané kombinace nejsou doloženy vůbec.



Jde v případě lemmatu *obmýti* o překlep? Ve slovníku slovo *obmýti* existuje.

**ssjc** Slovník spisovného jazyka českého

## obmýti

-í s. les. doba od vzniku lesního porostu do jeho zmýcení; obmýtní doba

Vrátíme se o krok zpět, zvolíme filtr (negativní) a do dotazovacího řádku zapíšeme *\*[oéíóúý][nt]í*.

### Formulace závěrů

Na základě pozorování dat získaných z korpusů můžeme tvrdit, že ve sledovaném korpusu jsou doložena pouze dějová jména taková, že před *[nt]í* předchází *[aeěiuýá]*. Jsou doložena neutra skloňovaná podle vzoru *stavení*, která nejsou dějovými jmény, u nichž před *[nt]í* předchází buď souhláska, nebo *[oiý]*. Nejsou doložena žádná slova taková, u nichž před *[nt]í* předchází *[éóů]*.

Ve sledovaném korpusu jsou doložena lemmata neuter na *[ea][nt]í*, která nejsou dějovými jmény (*století, úpatí, ...*).

Závěrem lze říci, že podmínkou nutnou, nikoli postačující k tomu, aby mohlo jít o dějové jméno, je zakončení lemmatu neutra na [aeěiuyá][nt]í.

Empirická pozorování korpusových dat můžeme doplnit na základě znalosti gramatiky. Kmenotvorná přípona, která je spolufornantem pro tvoření dějových jmen v češtině, může být pouze [aá], a to u sloves I. třídy vzorů *brát, mazat* (srov. *brání, mazání*), III. třídy vzoru *kupovat* (*kupování*) a V. třídy vzoru *dělat* (*tkání, dělání*), dále [e], a to u sloves I. třídy vzoru *umřít* (srov. *tření*), [eě] II. třídy vzoru *začít* (srov. *početí, stětí*), [eě], a to u sloves IV. třídy vzorů *trpět, sázet* (srov. *trpění, sázení*) a [nm]u, a to u sloves II. třídy vzorů *tisknout, minout* a *začít* (srov. *stisknutí, pominutí, vzednutí*). Po KmV 0 následuje [e] u sloves I. třídy vzorů *nést, péci* (srov. *nesení, pečení*), [eě] u sloves II. třídy vzoru *tisknout* (srov. *tištění, nadšení*) a [e] u sloves IV. třídy vzoru *prosit* (srov. *vyprošení*)<sup>38</sup>. Zakončení [nt]í stojí bezprostředně za kořenovým vokálem a) u sloves III. třídy vzoru *krýt*, která v češtině, pokud nejde o KoV [aá] (např. *hraní, sání*), mají u těchto derivátů vždy KoV krátký – všechny krátké vokály kromě o lze doložit, neboť seznam sloves je úplný, protože jde o uzavřenou třídu, (srov. *spění, vypletí, bití, obutí, umytí*) a b) u atematických/nepravidelných sloves (*jedení, chtění, bytí, ...*).

Výsledky empirie jsou takto opřeny o pravidla gramatiky.

### Formulace dalších otázek vyplynulých ze zkoumání daného jevu

Najděte v korpusu lemma *obmýtí*.

Lze jednoduše (jedním dotazem) vybrat dějová jména, která jsou zaručeně odvozena od sloves určité třídy? Lze totéž udělat pro určitý vzor?

Ručně vyberte frekventovaná slova, která splňují formální podmínky zakončení dějových jmen, ale dějovými jmény nejsou (*století, úpatí, ...*). Je všechna z nich třeba vyhledávat ručně? Lze najít nějaký postup, který by pomohl při ručním vyhledávání?

<sup>38</sup> Toto rozdělení vychází z Komárkových prací a podrobně je popsáno ve druhém svazku *Mluvnice češtiny* (Komárek a kol. 1987 : 427–495). Odpověď na otázku, proč je e/ě u sloves některých tříd a vzorů interpretováno jako kmenotvorná přípona a u jiných jako část tvarotvorné přípony pasivního participia, je možné synchronně zdůvodnit na základě a) analogie v rámci tvarů od kmene minulého a b) shod a neshod na úrovni hláskových alternací finálního konsonantu kořene (před kmenotvornou příponou, již je 0 nebo e/ě). Proč je e/ě u sloves I. třídy vzorů *nést, péci*, II. třídy vzoru *tisknout* a IV. třídy vzoru *prosit* součástí tvarotvorné přípony a u sloves I. třídy vzoru *umřít* a IV. třídy vzorů *trpět, sázet* je chápáno jako kmenotvorná přípona? V případě sloves I. třídy vzorů *nést, péci* a II. třídy vzoru *tisknout* jde o systém tvarů od kmene minulého. Všechny tvary od kmene minulého mají u sloves I. třídy vzorů *nést* a *péci* kmenotvornou příponu 0. Tvary od kmene minulého mohou mít u sloves vzoru *tisknout* kmenotvornou příponu buď *nu/nou*, nebo 0. V případě sloves IV. třídy je důvod jiný. Kmenotvorná přípona 0 u sloves IV. třídy vzoru *prosit* nezpůsobuje alternaci finál [bmv] na švu kořen/kmenotvorná přípona (*rozzloben, vykrmen, potupen, zastaven, pohřben, ...*), ale fakultativně způsobuje alternace jiných finál (např. *s>š prošen, z>ž zaražen, t>c zaplacen, d'>z zrazen, ...*). Kmenotvorná přípona e/ě u sloves 4. třídy vzorů *trpět* a *sázet* způsobuje alternaci finál [bmv] na švu kořen/kmenotvorná přípona (*vyráběn, srozuměn, protrpěn, zastavěn, ...*), ale nezpůsobuje jiné alternace, a to zejména u sloves vzoru *sázet* (např. *zasazen, obletěn, předváděn, ...*). Pokud by se v obou případech počítalo s kmenotvornou příponou e/ě, museli bychom předpokládat dvě různá e v kmenotvorné příponě sloves téže třídy.

## Zadání cvičení, v nichž lze uplatnit analogické postupy

Pokuste se aplikovat podobné postupy pro vyhledání kandidátů na deverbativa (maskulina životná) na *-tel*, *-č* (jména činitelská).

### Ještě jednou k adjektivům na *-cí*

#### Motivační úvod

V jedné z předchozích kapitol jsme ponechali stranou skupinu adjektiv derivovaných od podstatných jmen sufixem *-oucí*. Jan Gebauer (srov. Gebauer 1920 : 76) píše, že „v příponě *-oucí* bývá zároveň význam vzmocňovací; v tom významu se přenáší i ke kmenům jmenným, např. div *divoucí*, pravda *pravdoucí*, bílý – *běloucí* ...“ František Trávníček (srov. Trávníček 1951 : 354) upozorňuje na fakt, že adjektiva typu *divoucí* vyjadřují nejvyšší míru vlastnosti. Vladimír Šmilauer (Šmilauer 1971 : 124) zmiňuje „silně citová adjektiva na *-oucí* v etymologické figuře *pravda pravdoucí*“. V *Mluvnici češtiny I* (Dokulil a kol. 1987 : 378) se uvádí, že „zcela okrajově patří k jakostním desubstantivním adjektivům příležitostná expresivní adjektiva s příponou *-ouc(i)* – typ *tmoucí*. Převádějí obsah určovaného substantiva ve vlastnost a zároveň tuto vlastnost zesilují.“

V českých slovnících jsou zaznamenána tato adjektiva:

	PSJČ	SSJČ	SSČ	ČSFI
<i>bídoucí (bída)</i>	+	+	–	–
<i>divoucí (div)</i>	+	+	+	+
<i>hanboucí (hanba)</i>	+	+	–	–
<i>hrůzoucí (hrůza)</i>	+	+	+	+
<i>krásoucí (krása)</i>	+	+	+	+
<i>křivdoucí (křivda)</i>	+	+	–	+
<i>letoucí (léta/leta)</i>	+	+	+	+
<i>nicoucí (nic)</i>	+*	+	–	–
<i>pravdoucí (pravda)</i>	+	+	–	+
<i>smrt'oucí (smrt)</i>	+	+	+	+
<i>smůloucí/smoloucí (smůla)</i>	+	+	–	–
<i>strachoucí (strach)</i>	+	+	–	–
<i>světoucí (svět)</i>	+	+	–	–
<i>tmoucí (tma)</i>	+	+	+	+
<i>večeroucí (večer)</i>	+	+	–	–
<i>věkoucí (věky)</i>	+	+	–	–
<i>zlatoucí (zlato)</i>	+**	–**	–	–

\* Spojení *nic nicoucí* uvádí jen SSJČ. PSJČ uvádí pouze význam nicotný.

\*\* Spojení *zlato zlatoucí* uvádí jen PSJČ v rámci výkladu adj. *zlatoucí* synonymního ke *zlatavý/zlatistý* (označuje expr. příjemně do zlatova zbarvený). V SSJČ struktura *zlato zlatoucí* zmíněna není.



SSJČ i PSJČ uvádějí dále např. adjektivum *krvoucí* ve významu *zakrvácený*.

V této kapitole bychom se k adjektivům tohoto typu rádi vrátili.

### Nastínění problému

Figury založené na opakování příbuzných slov označuje stylistika (poetika) termíny **polyptoton** (opakování téhož slova v různých flektivních tvarech, např. *věky věků*) a **figura etymologica** (spojení *sloveso* + *objekt* nebo *substantivum* + *atribut* slov téhož základu, např. *dobrý boj jsem bojoval, věky věkoucí*).

Struktury N+N (gen. pl.)<sup>39</sup>, tedy např. *píseň písní, kniha knih, král králů*, které vyjadřují nejvyšší stupeň vlastnosti, se označují termínem **hebrejský superlativ**. O hebrejském superlativu lze dle našeho názoru hovořit i v případě konstrukcí obdobného významu typu N+(A<N). Korpusy dokládají např. výskyt těchto spojení: *div divů, hora hor, hrůza hrůz, chvíle chvíl, kniha knih, král králů, ?konec konců, krása krás, legenda legend, pán pánů, píseň písní, pravda pravd, slovo slov, věky věků, zrada zrad, žena žen, ...* K případům druhého typu je možné řadit konstrukce *div divoucí, hrůza hrůzoucí, pravda pravdoucí, věky věkoucí* atd.<sup>40</sup>

Jazykové korpusy sloužily od dob svého vzniku lexikografickým účelům. Slovníky s různou mírou konzistence zaznamenávají (důsledně/nahodile) některé slovotvorné formace. Tak např. pravidelně tvořená podstatná jména slovesná (na *-ní/-tí*) se pro paradigmaticnost (pravidelnost tvoření) neuvádějí, stejně jako některá deverbální adjektiva (adjektiva tvořená od přechodníků, která se na rozdíl od nich běžně užívají). Podobně jsou na tom i tvary komparativu a superlativu adjektiv tvořené pravidelně (viz výše).

V elektronických slovnících, které tvoří bázi programů automatické morfologické analýzy – tagování korpusů, je třeba, aby řada informací, které si uživatel slovníku – člověk (zejména rodilý mluvčí) domyslí, byla uvedena explicitně.<sup>41</sup>

Jednou z „bolestí“ automatické analýzy přirozeného jazyka je „nekonečnost“ (otevřenost) slovníku. Navíc je dokázáno, že počet jednotek s velmi nízkou frekvencí (řekněme s frekvencí 1 – hapax legomena) s nárůstem rozsahu textu neklesá (je stabilní).<sup>42</sup>

<sup>39</sup> Srov. hebrew4christians [online] . 2012. Cit. 5. 06. 2012, dostupné z:

<[http://www.hebrew4christians.com/Grammar/Unit\\_Five/Comparative\\_Usage/comparative\\_usage.html](http://www.hebrew4christians.com/Grammar/Unit_Five/Comparative_Usage/comparative_usage.html)>.

<sup>40</sup> V lingvistické literatuře se můžeme setkat s dalšími hodnoceními (srov. např. Schejbalová 2011, Boissin 1957, Aitzemüller 1950).

<sup>41</sup> Automatická morfologická analýza (jednou z jejích aplikací je i morfologické značkování jazykových korpusů, ale také např. běžně užívané korektory překlepů) bývá založena na strojovém slovníku, v němž jsou jednotky definovány jako trojice tvar / základní tvar / morfologická/é interpretace.

<sup>42</sup> Pro aplikace automatických analyzátorů zmíněného typu to znamená, že slovník analyzátoru a slovník korpusu nejsou zpravidla totožné. V praxi je důsledkem jistý počet „nerozpoznaných jednotek“, které snižují pokrytí

Jedním ze způsobů, jak snížit počet nerozpoznaných/neznámých slov, je neustálé doplňování slovníku. Tato cesta je do jisté míry pohybem v bludném kruhu (hapax legomena budeme mít stále s sebou).

Pro řadu jazyků, k nimž patří i čeština, je typické tvoření nových pojmenování na základě existujících slov. Slovtvorba, jako jeden ze způsobů obohacování slovníku, je do velké míry pravidelná, a lze ji tudíž popsat formálními pravidly. Tento fakt byl úspěšně využit při budování strojových slovníků pro automatické zpracování přirozeného jazyka (NLP), konkrétně automatickou morfologickou analýzu. Některé pravidelné slovtvorné formace byly zahrnuty na základě týchž postupů, které se používaly pro popis slovních tvarů. Další propojení formální morfologie (tvarosloví) a slovtvorby je výzvou pro automatickou morfologickou analýzu.

Podmínkou pro zahrnutí formálních pravidel slovtvorby na úrovni NLP je ovšem formální popis slovtvorby. Předpokladem je přesná znalost fungování těchto pravidel. Z tohoto hlediska je řada slovtvorných typů dobře popsána v gramatikách i slovnících. O okrajových typech mnohdy chybějí potřebné informace. Ty je ale mnohdy možné získat z masových dat, v nichž je možné předpokládat výskyt okrajových jevů, které nebyly dosud zachyceny.

O uvedeném typu (*hrůza hrůzoucí*) adjektiv si řada uživatelů jazyka asi pomyslí, že jde o cosi okrajového. Podívejme se nyní na to, jak může analýza masových dat obsažených v jazykových korpusech pomoci rozšířit naše znalosti o adjektivech tvořených od substantiv, která stojí jako shodný přívlástek v postpozici za motivujícím substantivem.

## Otázky

Jak lze vyhledat v korpusech tato adjektiva? O čem mohou svědčit doklady nalezené v korpusech?

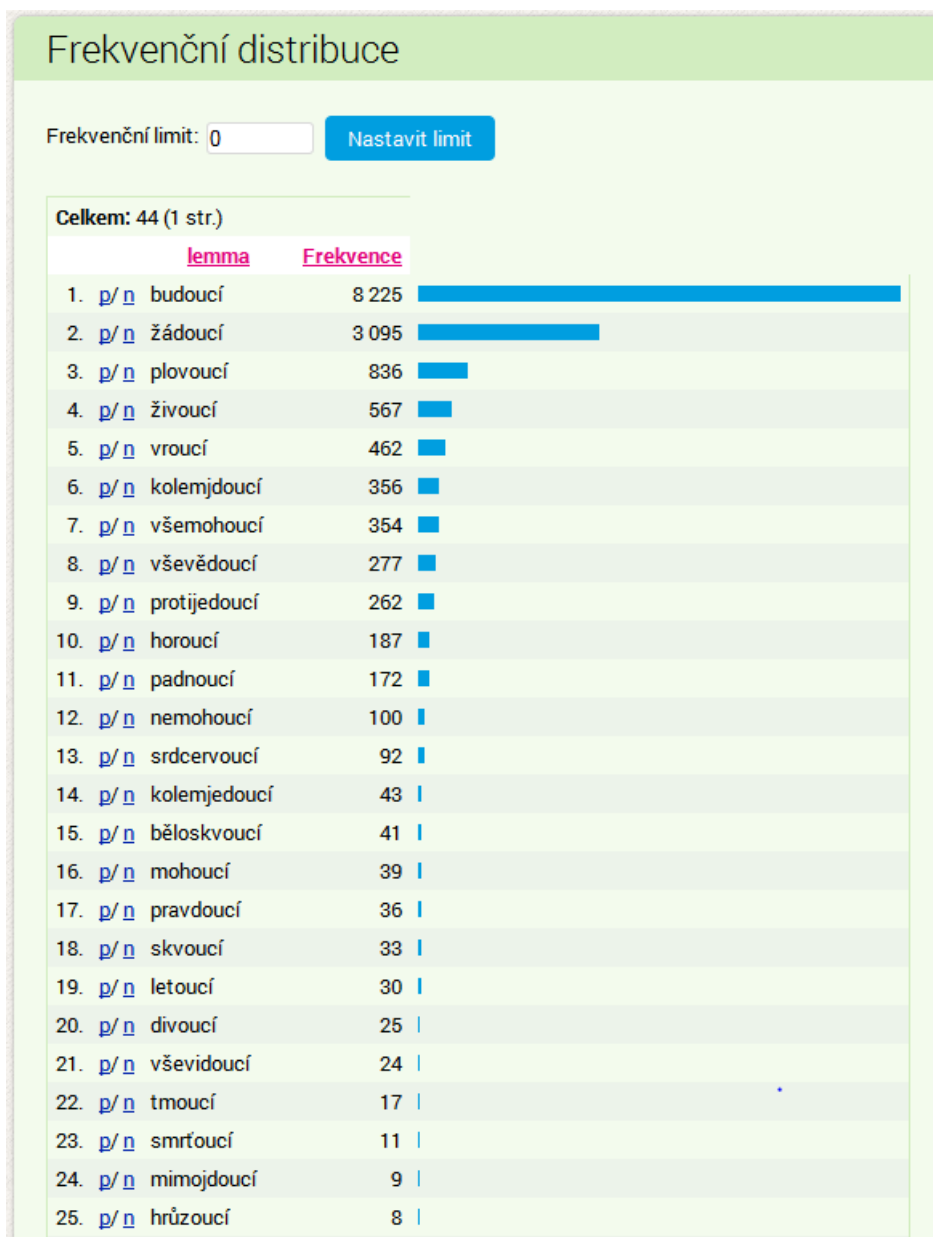
## Formulace dotazu pro získání dat z korpusů

V jedné z předešlých kapitol jsme poukázali na to, že adjektiva na *-oucí* jsou a) procesuální adjektiva označovaná jako *AG.\** (*rostoucí*), b) procesuální adjektiva označovaná jako *AA.\** včetně kompozit z procesuálních adjektiv (*plovoucí, kolemjdoucí*), c) deverbativa podobající se procesuálním adjektivům, která se ale synchronně tvoří jinak (*žádoucí*) a d) desubstantivní adjektiva s významem nejvyšší míry vlastnosti (*pravdoucí, letoucí*).

---

automatické morfologické analýzy. Běžný uživatel korektoru se s touto praxí denně setkává v případě, kdy mu automatický korektor označí „podezřelé“ vlastní jméno či méně běžné slovo (neologismy, archaismy, okazionalismy i termíny), ale i slovo/text v cizím jazyce atd.

Budeme-li chtít z korpusů získat data pro sledování poslední skupiny, budeme postupovat následovně. V menu **Konkordance** zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme `[lemma="*.oucí" & tag="AA.*"]`. Podívejme se na frekvenční distribuci lemat.



### Třídění a pozorování dat získaných z korpusů

Z korpusu SYN2010 získáme seznam 44 lemat. Projdeme jej a vypíšeme adjektiva, o nichž předpokládáme, že mohou být tvořena od substantiv a figurovat ve spojeních tzv. hebrejského superlativu (*pravdoucí, letoucí, divoucí, tmoucí, smrtoucí, hrůzoucí, krásoucí, věkoucí*).

## Formulace závěrů

Vidíme, že v korpusu jsme našli pouze 8 adjektiv ze 17, které jsou uvedena ve sledovaných slovnících (viz tabulka výše). Ze zkušenosti víme, že při sledování okrajových jevů v korpusech je třeba být obezřetný, užíváme-li morfologické značkování. Pokusíme se najít další doklady mezi slovy, která nejsou zachycena ve slovníku (slova označkována jako X.\*).

V menu **Konkordance** zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme `[lemma=".*oucí" & tag="X.*"]`. Podívejme se na frekvenční distribuci lc.



Z korpusu SYN2010 získáme seznam 61 lc.<sup>43</sup> Projdeme jej a vypíšeme adjektiva, o nichž předpokládáme, že mohou být tvořena od substantiv a figurovat ve spojeních tzv. hebrejského superlativu (*hanboucí, tmoucím, pravd'oucí, bohoucí, tmoucího, krvoucí, srdcoucí, dešťoucího, nicoucí, nicoucích, ...*).

Zkontrolujeme, zda naše předpoklady platí. V seznamu Frekvenční distribuce klikneme na příslušném řádku na **p** a sledujeme konkordance. V **Možn. zobrazení** můžeme přes volbu **Atributy, struktury a metainformace ...**, vyznačit **opus.genre, opous.název a opus.autor**, abychom mohli sledovat v jakých textech se příslušná adjektiva vyskytují.

---

. V neděli 11 . září pokřtíme mou knížku Pravda **pravd'oucí** /pravd'oucí/X@----- aneb Pohádky ze středního Polabí . Počtvrté se u vás . Pro děti napsal knihy Pověsti z Kaplanky a Pravda **pravd'oucí** /pravd'oucí/X@----- a letos vyšla jeho publikace S pohádkami do Svatojiříského lesa

---

sochy spadly do zahrady a ještě tam ležej . Hanba **hanboucí** /hanboucí/X@----- , povídám , hanba hanboucí . Budete s tím moct ještě tam ležej . Hanba hanboucí , povídám , hanba **hanboucí** /hanboucí/X@----- . Budete s tím moct něco udělat , mylady ?

---

-- hoře horoucí -- krev krvoucí rve srdce **srdcoucí** /srdcoucí/X@----- . Strašlivá , strašlivá , strašlivá samota prázdnin . Co

---

v nichž nerozpustná tajenka přes všechna vytrásadla má spadeno na **nicoucí** /nicoucí/X@----- . Jen chvíle ještě chybí krokům za zdí , aby

---

místo aby lehounce upravil svou posici . Ó , Bože **bohoucí** /bohoucí/X@----- ! Co s těmi Slovanami , co nemaj moře středozemní

Můžeme konstatovat, že se v korpusu SYN2010 vyskytla adjektiva sledovaného typu, která nejsou zachycena ve výkladových slovnících. Dále můžeme konstatovat, že doložená adjektiva mívají frekvenci 1 (viz výše hapax legomena). Dokladů je málo, nicméně jejich výskyt může svědčit o tom, že a) nejde o uzavřenou třídu a b) v dalších korpusech můžeme najít další doklady. Vzhledem k tomu, že výskyty v korpusech mohou svědčit o modelovém idiomatičtém tvoření a zároveň je patrné, že se jedná o případy řídké, nabízí se možnost hledat další doklady i v dalších korpusech (v korpusech řady SYN, ale třeba i ve velkých korpusech budovaných na FI MU<sup>44</sup>) a za dodržení jistých podmínek i na internetu.<sup>45</sup>

## Formulace dalších otázek vyplynulých ze zkoumání daného jevu

<sup>43</sup> Všimněme si, že mají většinou frekvenci 1 (hapax legomena).

<sup>44</sup> Srov. Pala – Rychlý 2011.

<sup>45</sup> Moderní definice korpusu uvádějí čtyři základní rysy, které musí korpus v moderním slova smyslu splňovat: 1) strojově čitelná podoba, 2) reprezentativnost zaručená zastoupením vzorků nejrůznějších textů (vyváženost), 3) vymezený a omezený rozsah (díky tomu lze opřít o data frekvenční analýzy) a 4) standardní anotace (McEnery – Wilson 1996). Pouze při splnění těchto čtyř požadavků je možné podrobovat data získaná z korpusů kvantitativním analýzám a srovnáním. Přesto je teoreticky dokázáno, že jevy zastoupené okrajově, řídké (hapax legomena) se v textech vyskytují se stabilní frekvencí, tj. s nárůstem rozsahu (počtu textů) neklesá počet nově se vyskytávajících jednotek. Proto se především pro účely lexikografické někdy ustupuje od požadavků vymezeného a omezeného rozsahu korpusu a budují se tzv. monitorovací korpusy, banky textů atd. Podobně lze používat i internet.

Při vyhledávání dokladů na internetu existují dva problémy: a) texty, v nichž jsme našli doklad, mohou být odstraněny a b) internetové vyhledávače nepracují s takovou mírou lingvistické podpory jako korpusové manažery. Je tedy třeba hledat intuitivněji (více Osolobě 2009<sup>1</sup>). Lze v korpusech / na internetu najít i případy, kdy se desubstantivní adjektivum osamostatní a je užito mimo etymologickou figuru?

### **Zadání cvičení, v nichž lze uplatnit analogické postupy**

Zopakujte popsany postup ve všech korpusech řady SYN a sledujte, v jakých typech textů se vyskytují nalezené doklady. Vyhledejte v korpusech (na internetu) adjektiva tvořená od adjektiv označující kvalitu barvy sufixem *-oucí* (*běloucí, zlatoucí, ...*).

### **Hláskové alternace ve slootvorbě**

#### **Motivační úvod**

K hláskovým alternacím (pravidelnému střídání hlásek, na němž je založena alomorfie) dochází při odvozování slov. Lze je tedy interpretovat jako spoluformanty při slootvorbě. Pro formální popis jazyka je důležité explicitně popsat pravidla jejich výskytu. V předchozích kapitolách jsme se věnovali hláskovým alternacím na rovině tvarotvorné. V této kapitole se budeme zabývat alternacemi v prefixu.

#### **Nastínění problému**

Slovesa v češtině nemívají zpravidla v prefixu dlouhou samohlásku. Alternace krátkého vokálu v prefixu (*na>ná, za>zá, do>dů, po>pů, pro>prů, vy>vý, u>ú*) provázejí tvoření jmen (substantiv, adjektiv) od sloves (srov. např. *na-br-a-t>ná-bor, za-vř-í-t>zá-vor-a, do-mysl-e-t>dů-mysl-0, po-hn-a-t>pů-hon, pro-lom-i-t>prů-lom-0, vy-hrá-t>vý-hr-a, u-skoč-i-t>ú-skok-0*). Tyto alternace vykazují různou míru pravidelnosti u různých derivačních typů. Nikdy k nim nedochází u paradigmatického tvoření dějových jmen na *-ní/-tí* a při adjektivizaci participií (Karlík 2007, Scheer 2001, Ziková 2010). Naopak jsou dosti pravidelné u konverzní derivace názvů dějů, přičemž výjimkou jsou prefixy *do-, po-, pro-*, u nichž se setkáváme jak s derivacemi s alternací, tak bez ní, přičemž existují případy, kdy je přítomností/nepřítomností alternace odlišen význam (srov. *průjezd/projezd, důraz/doraz, ...*). Začíná-li sloveso v češtině na řetězec *((ná)|(zá)|(dů)|(pů)|(prů)|(při)|(vý)|(ú)).\**, pak je buď odvozeno od substantiva, adjektiva, zájmena, adverbia s příslušnou variantou zdlouženého prefixu, nebo nejde o prefix, ale pouze o náhodnou formální shodu. (K náhodným shodám počítáme případy, kdy nelze ani na základě etymologie vydělit prefix, např. *zábst.*) Existuje

ovšem několik málo výjimek (na některé upozorňuje Scheer 2001), které obecnou platnost uvedeného pravidla narušují.

### **Otázky**

Kolik takových sloves je? Mohou korpusy přispět k vytvoření úplných seznamů takových sloves?

### **Formulace dotazu pro získání dat z korpusů**

Při vyhledávání sloves, která mohou mít délku v prefixu, budeme postupovat následovně. Zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme `[lemma="((ná)|(zá)|(dů)|(pů)|(prů)|(pří)|(vý)|(ú)).*" & tag="V.*"]` (hledáme všechny tvary obsahující na začátku řetězec, který může být dlouhým prefixem, označované jako *V.\**, tedy slovesné tvary). Podívejme se na frekvenční distribuci lemmat.

## Frekvenční distribuce

strana 1

Frekvenční limit:  [Nastavit limit](#)

Celkem: 67 (2 str.)

	lemma	Frekvence
1.	<a href="#">p/ř</a> působit	22 182
2.	<a href="#">p/ř</a> záležet	11 180
3.	<a href="#">p/ř</a> následovat	11 166
4.	<a href="#">p/ř</a> záviset	4 479
5.	<a href="#">p/ř</a> účastnit	3 974
6.	<a href="#">p/ř</a> půjčit	2 950
7.	<a href="#">p/ř</a> zářit	2 633
8.	<a href="#">p/ř</a> náležet	2 206
9.	<a href="#">p/ř</a> důvěřovat	2 040
10.	<a href="#">p/ř</a> přísahat	1 920
11.	<a href="#">p/ř</a> útočit	1 918
12.	<a href="#">p/ř</a> závidět	1 753
13.	<a href="#">p/ř</a> závodit	1 249
14.	<a href="#">p/ř</a> půjčovat	1 067
15.	<a href="#">p/ř</a> účinkovat	1 006
16.	<a href="#">p/ř</a> přít	998
17.	<a href="#">p/ř</a> příslušet	936
18.	<a href="#">p/ř</a> zápasit	920
19.	<a href="#">p/ř</a> účtovat	819
20.	<a href="#">p/ř</a> zásobovat	806
21.	<a href="#">p/ř</a> ústit	780
22.	<a href="#">p/ř</a> výt	404
23.	<a href="#">p/ř</a> zápolit	363
24.	<a href="#">p/ř</a> násobit	351
25.	<a href="#">p/ř</a> přičít	338
26.	<a href="#">p/ř</a> příst	318

### Třídění a pozorování dat získaných z korpusů

Projdeme seznam kandidátů na slovesa, která je možné interpretovat jako slovesa s délkou v prefixu. Vidíme, že kromě hledaných sloves se vyskytují a) případy, kdy jde o náhodnou shodu počátečního řetězce a hledaného prefixu (*výt, přít*), b) slovesa derivovaná od jiných slovních druhů (substantiv, adjektiv, adverbí, ...), jejichž fundující slova mají v prefixu délku (*účinkovat, zásobovat, ...*), a c) hledané doklady (*záležet, záviset, ...*). Vytvoříme tabulku a budeme zapisovat počty sloves s příklady.

Při vyplňování tabulky si ruční práci zjednodušíme tak, že vytřídíme postupně skupiny s potencionálními sufexy. Zvolíme filtr (pozitivní) **Typ dotazu cql** a do dotazovacího řádku



zapišeme [lemma="ná.\*" & tag="V.\*"]. Podíváme se na frekvenční distribuci lemmat, projdeme seznam a zapišeme výsledky do tabulky. Vrátime se o krok zpět a zopakujeme postup pro tvary s lemmatem začínajícím na další řetězce. (Přihlédneme k interpretacím v etymologickém slovníku.)

lemma	nalezeno	derivace od S,A,P,D	náhodné shody	délka v prefixu	překlepy
<i>ná.*</i>	7	5 ( <i>násobit</i> )	0	2 ( <i>následovat</i> )	0
<i>zá.*</i>	19	9 ( <i>závodit</i> )	4 ( <i>zábst</i> ) 2 ( <i>zápolit</i> )*	3 ( <i>záležet</i> )	1 ( <i>zábránit</i> )
<i>dů.*</i>	1	1 ( <i>důvěřovat</i> )	0	0	0
<i>pů.*</i>	5	2 ( <i>působit</i> )	1 ( <i>půlit</i> )	2 ( <i>půjč[i (ova)]t</i> )	0
<i>prů.*</i>	2	2 ( <i>průměřovat</i> )	0	0	0
<i>pří.*</i>	9	1 ( <i>příštipkařit</i> )	4 ( <i>příst</i> )	3 ( <i>přísahat</i> )	1 ( <i>příspěvat</i> )
<i>vý.*</i>	7	3 ( <i>výletovat</i> )	4 ( <i>výt</i> )	0	0
<i>ú.*</i>	17	11	4 ( <i>ústit</i> )	2 ( <i>útoč[i íva]t</i> )	0
<b>CELKEM</b>	67	34	19	12	2

\* Ke slovesům *zápolit* a *zápasit* srov. motivaci předložkou *za* (viz Rejzek 2001).

### Formulace závěrů

Z celkového počtu 67 lemmat sloves, která mohou být kandidáty na slovesa s délkou v prefixu, jde pouze ve 12 případech o kombinaci prefix s délkou + sloveso. V ostatních případech je délka již ve fundujícím slově (34 případů), nebo jde o náhodnou shodu počátečního řetězce (19 případů). Ve 2 případech je doklad překlep.

Z pozorování korpusových dat je patrné, že hledaná slovesa jsou většinou frekventovaná. Pozorujeme, že se někdy vyskytují ve více variantách (slovesa odvozená od sloves jako např. *přísah[a|áva]t*). To může svědčit o tom, že od některých sloves doložených v korpusu budou existovat varianty, které ve sledovaném korpusu doloženy nejsou (např. méně obvyklá iterativa).

Tento předpoklad můžeme ověřit následovně. Předpokládejme, že od slovesa *závidět* lze vytvořit iterativum *závidívat*. To není doloženo v korpusu SYN2010. Lze je ovšem najít např. v korpusu SYN2006PUB.

málo rozrůzněným co do vkusu i zájmů . Jak jen **závidíval** Rusům Turgeněvovi a Dostojevskému či jiným současníkům ve Francii , se ziskem prodali . V této souvislosti vždy také nezakrytě **závidivá** potomkům Lva Nikolajeviče Tolstého , " ze kterých se na

U některých sloves, která jsme zařadili mezi ta, jež mají délku v prefixu a nejedná se o derivát od jmen, lze připustit dvojí možnou motivaci, tedy motivaci slovesem i dějovým jménem (*utkat (se)/útok>útoč[i]iva)t*).

V případě slovesa *půjč[i]ova)t* lze segmentaci, a tedy i délku v prefixu určit až na základě etymologie a analogie s příbuznými slovanskými jazyky (srov. slovesky *požičat'*)<sup>46</sup>.

« KRÁTKY SLOVNÍK SLOVENSKÉHO JAZYKA

**požičat'** -ia dok.

poskytnúť niekomu (na jeho žiadosť) na dočasné používanie: **požičat'** priateľovi peniaze, knihu; **požičaj** mi ceruzku

● aký požičaj, taký vráť;

nedok. **požičiavat'** -a

// **požičat' si** vyžiadať si od niekoho na istý čas na dočasné používanie, vypožičať si: **požičať si** knihy z knižnice; **požičať si** pero od suseda;

nedok. **požičiavat' si**

### Formulace dalších otázek vypluvších ze zkoumání daného jevu

V jedné z předchozích kapitol jsme se zabývali alternací *o/e* v substantivní flexi. S alternací *e/o* se můžeme setkat i při tvoření tvarů sloves (např. ... *obě baby byly <státy> ...*, ... *získal nad protivníkem převahu a <setnul> mu hlavu ...*), ale také při odvozování sloves prefixy (např. tzv. vokalizace prefixů na švu s kořenem např. *tnout* i *tít* / *setnout* i *stít* ale *stínat*, *dřít/sedřít* ale *sdírat*, *jít/nadejít* ale *nadcházet*). Vyhledejte v korpusu (korpusech) doklady pro sledování naznačených alternací a pokuste se vysledovat pravidla, jimiž se řídí distribuce alternace *e/o* u jednotlivých tvarů.<sup>47</sup>

### Zadání cvičení, v nichž lze uplatnit analogické postupy

Projděte ostatní korpusy a zjistěte výskyt dalších dokladů (např. iterativ).

<sup>46</sup> Srov. elektronická verze slovníků slovenštiny *SLEX99*.

<sup>47</sup> Srv. více Osolsobě 2011<sup>1</sup> : 40n.

## **MWE – gramatika a slovník**

Tzv. Multiword Expressions (víceslovné výrazy) patří k oblasti, která při počítačovém zpracování přirozeného jazyka přináší řadu problémů. MWE je zastřešující pojem pro různé případy od analytických tvarů (sloves, ale třeba i komparativů a superlativů), přes frazeologická spojení až po víceslovné názvy osob, míst, institucí, děl atd., tedy jednotek, které běžně chápeme v jejich celku. Tyto jednotky jsou při automatickém zpracování textů na úrovni tzv. tokenizace (viz výše) rozděleny na části. Nicméně je pro různé potřeby výzkumu chceme vidět jako celek. Něco z této problematiky bylo naznačeno již výše v souvislosti s analytickými tvary sloves. V tomto oddíle poukážeme na problémy, které mohou nastat v případech, kdy jednotky, z nichž je složen MWE, následují bezprostředně za sebou v textu. Všimneme si případů homonymních slovosledných formací, tedy případů, kdy slova, která po formální stránce odpovídají zadání pro vyhledání MWE, žádanou jednotku netvoří. Jako příklad použijeme posloupnosti *(nej)více/(nej)méně* + adjektivum ve tvaru pozitivu, které nemusí nutně být příklady opisného stupňování v češtině. V další podkapitole upozorníme na některé problémy spojené s vyhledáváním slovesných frazémů v korpusech.

### ***Opisné stupňování***

#### **Motivační úvod**

Problematické opisného stupňování se v českých gramatikách nevěnuje příliš mnoho pozornosti. Některé problémy jsou nastíněny v monografickém zpracování (Marvan 2008). Občas se objeví upozornění na to, že formulace jako ... *první možnost je <nejvíce optimální>* ... jsou pleonasmy, protože adjektivum *optimální* je z latinského *optimus*, tedy *nejlepší*. Jak a kdy se nahrazuje tvar syntetický tvarem opisným, je jistě zajímavé. Ještě zajímavější je, že zatímco pro vyjádření vyšší/nejvyšší míry směrem „nahoru“ existují tvary syntetické, pro srovnání „dolů“ máme pouze analytické (opisné) tvary. Korpusy představují zdroje dat pro zkoumání toho, jak se používá těch i oněch tvarů.

#### **Nastínění problému**

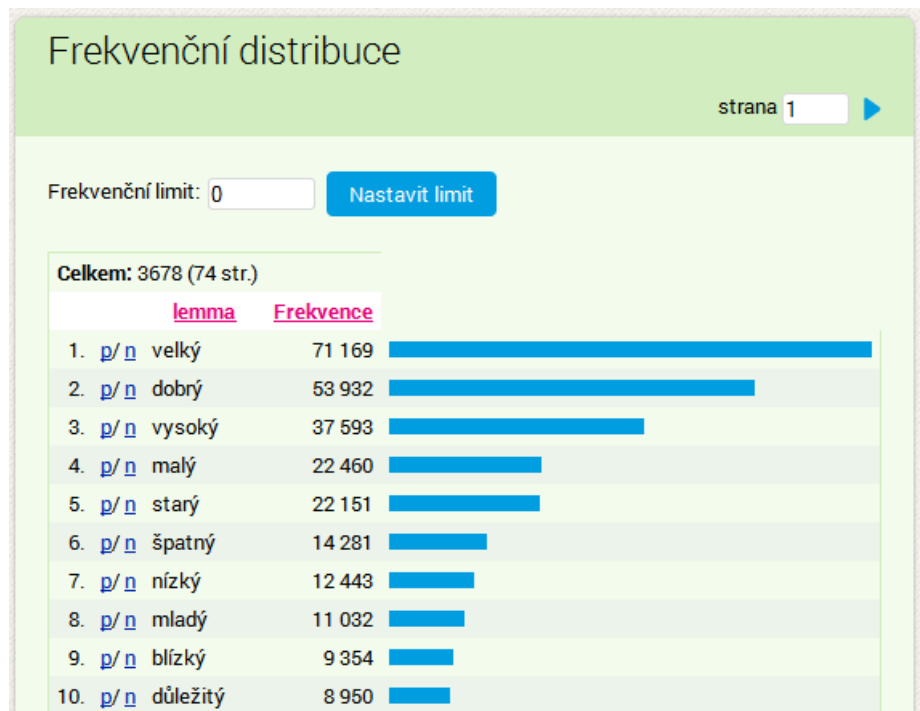
Jak jsme si ukázali v jedné z předchozích kapitol, jsou syntetické tvary komparativu a superlativu minimálně doloženy i ve stamilionových korpusech. Jedno z hledisek, které můžeme sledovat, je tedy hledisko frekvence (např. *způsobilejší* a *více způsobilý*).

## Otázky

Která adjektiva z těch, jež jsou v korpusech doložena v komparativu/superlativu, jsou doložena též v opisných tvarech? Vyberme nejfrekventovanější lemmata adjektiv doložených ve tvarech syntetického stupňování a porovnejme je s frekvencí a užitím tvarů analytických.

## Formulace dotazu pro získání dat z korpů

Zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme `[tag="AA.....[23].*"]` (vyhledáme všechny tvary adjektiv ve 2. a 3. stupni). Podívejme se na frekvenční distribuci lemmat. (Vidíme lemmata adjektiv, která se v češtině stupňují, seřazená dle frekvence.)



Nyní zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme `[word="((více)|(méně))"]` `[tag="AA.....I.*"]`. Hledáme doklady opisného stupňování.

obrazu je beztretné, zatímco veřejné vyjadřování sympatií neonacismu prostřednictvím samotného klienta. Předexporní úvěr je tak dostupný i pro Číně nebude žádná americká firma. Bude Čína lepší, zaplacené v zahraničí, odpovídající zbývající částem příjmů dosažených za možné, že výsledky na středověchodní a hospodářské frontě budou poměrně více biologickou nežli muž, muž je bytostí poměrně ke kousání (případ našeho Miši), více či zajímavé si povšimnout, že americké chytré bomby GPS mají a sama odezní. Antibiotika: raději méně než. Až na některé velmi vzácné výjimky jsou samice naopak ji - nestalo se nic. Další zkoušku s materiálem je i nadále úžasná. Chileané jsou sice o něco IP systém je snadno rozšiřitelný, protože správa systému je podobného jako "krajinný ráz", ale je ještě	méně /málo/Dg-----2A----- známého /známý/AAIS2----1A----- méně /málo/Dg-----2A----- bonitní /bonitní/AAMP4----1A----- méně /málo/Dg-----2A----- represivní /represivní/AAFS1----1A----- více /více/Ca--4----- zdaňovacích /zdaňovací/AAFP2----1A----- méně /málo/Dg-----2A----- jednoznačné /jednoznačný/AAIP1----1A----- více /hodně/Dg-----2A----- sociologickou /sociologický/AAFS7----1A----- méně /málo/Dg-----2A----- nebezpečné /bezpečný/AAFP1----1N----- méně /málo/Dg-----2A----- přesný /přesný/AAIS4----1A----- více /hodně/Dg-----2A----- alergické /alergický/AAFP1----1A----- méně /málo/Dg-----2A----- náruživé /náruživý/AAFP1----1A----- méně /málo/Dg-----2A----- odolným /odolný/AAAMST----1A----- méně /málo/Dg-----2A----- horkokrevní /horkokrevný/AAAMP1----1A----- méně /málo/Dg-----2A----- nákladný /nákladný/AAFS1----1A----- méně /málo/Dg-----2A----- uchopitelný /uchopitelný/AAIS1----1A-----	symbolu (například organizace Blood & Honour) může být klienty. Na druhou stranu banka obezřetně sleduje způsob vynakládání? Pohne se vpřed? Nemyslím. Čína nemůže být období, na které je uplatněn zápočet. V případech nebude to velký úspěch ani zřejmě katastrofa. V takovém nešli žena" (tamže). Zde studujeme vývoj. Jednou jsem se vyděsila, když mu manžel našel záložní navigační systém, který však nelze rušit, pro reakce se podle doktorky Vermerové nejčastěji vyskytují u penicilinových antibiotik. Jak si již dávno povšiml slavný Hunter, samice provedl Franta a ponořil do potůčku koneček lana. Znovu než jejich argentinskí sousedé, reakce na projíždějící vozy jsou, instalace snadná, připojení jednoduché bez závislosti na místě. Asi nejlepší definice "ducha místa" je to
--	--	--

Nyní si postup rozdělíme do kroků. Zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme `[word="((více)|(méně))"]` `[lemma="velký"]` (hledáme všechny případy, kdy

v korpusu za sebou následuje tvar *více* nebo *méně* a lemma *velký* – možné opisné stupňování). (Do dotazu by bylo možné zařadit též podmínku, že chceme vyhledat pouze tvary pozitivu:  $[word=((vícě)|(méně))][lemma="velký" \& tag="AA.....I.*"]$ . Pokud ji nezařadíme, mohou se v konkordancích objevit také doklady pleonasmů.)

Následující doklad pleonasmu nalezneme v korpusu SYN2006PUB.

nový \* Kachyňa ? Filmaři to budou mít čím dál **více horší** . Kachyňa natočil přes šedesát filmů . To je nějakých

Vrátíme se o krok zpět a postup zopakujeme pro další adjektiva. Dotaz může být formulován i složitěji. Zvolíme **Typ dotazu cql** a do dotazovacího řádku zapíšeme  $[word=((vícě)|(méně))][lemma=((špatný)|(nízký)|(mladý)|(blízký)|(důležitý)|(různý)|(silný)|(široký))]$ .

, než dodal : * Tenhle vlak je určený pro	<b>méně</b> /málo/Dg-----2A-----	<b>důležitý</b> /důležitý/AAFS2----1A-----	osoby . * * Jak - pro méně důležité ?
je tedy třeba vybrat ty nejhodnější ( doporučuje se použít	<b>více</b> /více/Ca--2-----	<b>různých</b> /různý/AAFP2----1A-----	metod a jejich výsledky kombinovat ) a vhodně nastavit jejich
ukázalo , že potřebu statutu , prestiže a funkce jako	<b>méně</b> /málo/Dg-----2A-----	<b>důležitý</b> /důležitý/AAIP1----1A-----	považují obecně zaměstnanci ve výrobních podnicích . Z hlediska vývoje
systemový certifikát pro přístup k více datovým schránkám . Mezi	<b>méně</b> /málo/Dg-----2A-----	<b>důležitý</b> /důležitý/AAANS1----1A-----	, avšak o to více praktické schopnosti patří možnost vytvoření
hongkongský Oasis Air . Stejně jako Flyglobespan nabízí v letadlech	<b>více</b> /více/Ca--2-----	<b>různých</b> /různý/AAFP2----1A-----	tříd včetně pohodlnější obchodní . Ceny jednosměrné letenky začínají na
, což vylučuje zásah sil řádu nadřazeného materií , více	<b>méně</b> /málo/Dg-----2A-----	<b>blízkých</b> /blízký/AAFP2----1A-----	lidské vůli . Descartes nevidí pod iluzorním zdáním smyslových údajů
nikdy znám dopředu . Nabízena je široká škála videokazet s	<b>méně</b> /málo/Dg-----2A-----	<b>širší</b> /široký/AAFS7----2A-----	možností výběru žánru . Filmy , vesměs s kung-fu tematikou
Bude to absolutní pravda , ale taky jenom ta její	<b>méně</b> /málo/Dg-----2A-----	<b>důležitá</b> /důležitý/AAFS1----1A-----	část . Protože my už jsme jeho tajemství odhalili tajemství
korporace uspěly v konkurenci , pak je tato znalost ještě	<b>více</b> /hodně/Dg-----2A-----	<b>důležitá</b> /důležitý/AAFS1----1A-----	pro každého jednotlivce . Američtí experti odhadují , že v
kteřé byly dosud spíš doménou mužů , se dostává stále	<b>více</b> /více/Ca--4-----	<b>mladých</b> /mladý/AAFP2----1A-----	žen , například do zákonodárství , vyšších úřednických míst či
a učí se novým dovednostem . Multi skilling umožňuje zastávat	<b>více</b> /více/Ca--1-----	<b>různých</b> /různý/AAFP2----1A-----	pracovních míst , což je výhodné i pro zaměstnavatele .

## Třídění a pozorování dat získaných z korpusů

Všimli jsme si, že mezi nalezenými doklady jsou případy, kdy tvary *více/méně* nefungují ve významu měrových příslovcí, ale ve významu číslovek. Vyjadřuje se jimi míra neurčitého počtu, čemuž odpovídá i forma jmenné skupiny (počítaný předmět bývá v češtině zpravidla v genitivu – tzv. numerativ). Nejde tedy o hledané doklady. Nesprávné vyhledávky bychom mohli odstranit tak, že bychom z druhé pozice vpravo od KWIC (klíčového slova – tedy v našem případě adjektiva) odstranili ty řádky, na nichž je tvar v genitivu plurálu.

Zvolíme filtr (negativní), rozsah hledání <2,2>, **Typ dotazu cql** a do dotazovacího řádku zapíšeme  $[tag="...P2.*"]$ .

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od 2 do 2  včetně KWIC

**Typ dotazu:** CQL

**CQL:** [tag="...P2.\*"]

Implicitní atribut: word [Popis morfologických značek](#)

, než dodal : " Tenhleten vlak je určený pro ukázalo, že potřebu statutu , prestiže a funkce jako systémový certifikát pro přístup k více datovým schránkám . Mezi , což vylučuje zásah sil řádu nadřazeného materií , více nikdy znám dopředu . Nabízena je široká škála videokazet s Bude to absolutní pravda , ale taky jenom ta její korporace uspěly v konkurenci , pak je tato znalost ještě a trpělivě mi vysvětlovaly , že jsou věci důležité a analytická agentura Aberdeen Group , tedy využití standardizované verze softwaru průser roku byla označena politika v Česku , o něco dopadla i vyjednávání v méně rozsáhlé , ale o nic a novináři pro celou Českou republiku , je přitom logicky jsem mnoho Vánoce ve Skotsku , ale tam bývaly Vánoce hru se hodně soustředím , takže si většinou povídáme o	<b>méně</b> /málo/Dg-----2A----- <b>důležité</b> /důležitý/AAFS2----1A----- <b>méně</b> /málo/Dg-----2A----- <b>důležité</b> /důležitý/AAIP1----1A----- <b>méně</b> /málo/Dg-----2A----- <b>důležité</b> /důležitý/AAANS1----1A----- <b>méně</b> /málo/Dg-----2A----- <b>blízkých</b> /blízký/AAFP2----1A----- <b>méně</b> /málo/Dg-----2A----- <b>širší</b> /široký/AAFS7----2A----- <b>méně</b> /málo/Dg-----2A----- <b>důležitá</b> /důležitý/AAFS1----1A----- <b>více</b> /hodně/Dg-----2A----- <b>důležitá</b> /důležitý/AAFS1----1A----- <b>méně</b> /málo/Dg-----2A----- <b>důležitá</b> /důležitý/AAANS1----1A----- <b>více</b> /více/Ca-7----- <b>různými</b> /různý/AAFP7----1A----- <b>méně</b> /málo/Dg-----2A----- <b>špatní</b> /špatný/AAAMP1----1A----- <b>méně</b> /málo/Dg-----2A----- <b>důležité</b> /důležitý/AAFS6----1A----- <b>více</b> /hodně/Dg-----2A----- <b>špatný</b> /špatný/AAFS2----1A---6- <b>méně</b> /málo/Dg-----2A----- <b>důležité</b> /důležitý/AAFP1----1A----- <b>méně</b> /málo/Dg-----2A----- <b>důležitých</b> /důležitý/AAFP6----1A-----	osoby . " Jak - pro méně důležité ? považují obecně zaměstnanci ve výrobních podnicích . Z hlediska vývoje , avšak o to více praktické schopnosti patří možnost vytvoření lidské vůli . Descartes nevidí pod iluzorním zdáním smyslových údajů možnost výběru žánru . Filmy , vesměs s kung-fu tematikou část . Protože my už jsme jeho tajemství odhalili tajemství pro každého jednotlivce . Američtí experti odhadují , že v že těch nedůležitých je nutno se vzdát , obětovat firmami kvůli maximalizaci zisku poskytovatele , pak jistě nemohl uspět se účastníkům ankety jevil zpěvačka Ewa Farna či duo Ery stěře , vedená Sověty podporovanými " lublinskými Poláky " a než pojem Česko . Nedovedu pochopit , jak je možné než oslavy Nového roku . Neexistovala tam taková atmosféra tlumeného věcech , než jsou velké zakázky , " tvrdí .
--	---	---

Vidíme některé doklady opisného stupňování s *méně*. Podívejme se na řádky s „*více*“. Zvolíme filtr (negativní), vybraný token první, rozsah hledání <0,0>, **Typ dotazu word** a do dotazovacího řádku zapíšeme *méně*.

### Filtr konkordancí

**Filtr:**  pozitivní  negativní

**Vybraný token:**  první  poslední

**Rozsah hledání:** od 0 do 0  včetně KWIC

**Typ dotazu:** Slovní tvar

**Slovní tvar:** méně

Slovní druh: nspecifikováno  Shoda velikosti:

Procházíme konkordance a hledáme případy opisného stupňování. Všimněme si i problematické disambiguace (zachycení rozdílů slovnědruhových přesahů adverbium *Dg* a číslovka *C*).

korporace uspěly v konkurenci , pak je tato znalost ještě	více /hodně/Dg-----2A-----	důležitá /důležitý/AAFS1----1A-----	pro každého jednotlivce . Američtí experti odhadují , že v
analytická agentura Aberdeen Group , tedy využití standardizované verze softwaru	více /vice/Ca-7-----	různými /různý/AAFP7----1A-----	firmami kvůli maximalizaci zisku poskytovatele , pak jistě nemohl uspět
a novináři pro celou Českou republiku , je přitom logicky	více /hodně/Dg-----2A-----	špatný /špatný/AAFS2----1A--6-	než pojem Česko . Nedovedu pochopit , jak je možné
Katka a Dalibor Štruncovi - prošel jedním , možná i	více /hodně/Dg-----2A-----	důležitými /důležitý/AAFP7----1A-----	rozcestími . Nevím , jak se mi bude líbit jejich
, kdy se produkt činnosti jednoho genu funkčně uplatňuje ve	více /hodně/Dg-----2A-----	různých /různý/AAIP6----1A-----	buněčných pochodech , tudíž podmiňuje nebo ovlivňuje fenotyp většího počtu
památou noc strávit přinejmenším s dvěma , ale možná i	více /hodně/Dg-----2A-----	různými /různý/AAFP7----1A-----	besedníky , kteří sice přecházeli jeden v druhého plynule ,
mimo ně . Informace o zákaznících se totiž nacházejí ve	více /hodně/Dg-----2A-----	různých /různý/AAFP6----1A-----	aplikacích , které nejsou vzájemnými podmnožinami , a tudíž se
potřeby jeho webů . Stejný systém najdeme jen málokdy na	více /vice/Ca-6-----	různých /různý/AAIP6----1A-----	webech nebo webech více vydavatelů . Z toho vyplývá ,
složky od sebe odtrhovat a klasifikovat jednu práci dvěma či	více /hodně/Dg-----2A-----	různými /různý/AAFP7----1A-----	známkami za každou složku zvlášť : vždyť každý text vnímá
Dívka , která si točí vlasy na papírky , není	více /hodně/Dg-----2A-----	mladou /mladý/AAFS7----1A-----	bohyní , nýbrž budoucí ženou v domácnosti a tak také
v různých hloubkách . Pomáháme si zde sice zobrazováním ve	více /vice/Ca-6-----	různých /různý/AAFP6----1A-----	projekcích , avšak riziko falešného nálezu či neodhalení anomálie v
" Nebylo na to času . V chvíli tu byly	více /hodně/Dg-----2A-----	důležité /důležitý/AAFS2----1A-----	jiné záležitosti . Lhal bych , pokud bych tvrdil ,
v soutěži málo vhodný . Podstatnými faktory úspěchu byly mnohem	více /hodně/Dg-----2A-----	nízké /nízký/AAIP1----1A-----	mzdové náklady a znalosti ( know-how ) zaměstnanců ve výrobě
, nepůsobí jako zvláštní terno . O naftové šestnáctivostce se	více /hodně/Dg-----2A-----	špatného /špatný/AAANS2----1A-----	prohlásit nedá . Dokonce je s ní možné jezdit za
. Mezi organizacemi založenými po roce 1989 nebylo zaznamenáno významně	více /vice/Ca-1-----	mladých /mladý/AAAMP2----1A-----	než v ostatních organizacích . Zaměřme - li se na
vyrostá jen s jedním z rodičů nebo se dvěma či	více /hodně/Dg-----2A-----	blízkými /blízký/AAFP7----1A-----	osobami . Důležitější je , Děti sice potřebují pevný vztah
že se mu zákon nelíbí , a že proto nebude	více /hodně/Dg-----2A-----	mladé /mladý/AAAMP4----1A-----	soudce jmenovat , mohli soudní adeпти již dva roky praktikovat
zisk než 17 - 27 % . ( Odhady se	více /hodně/Dg-----2A-----	nerůzní /různý/AAAMP1----1N-----	. ) Ještě horší je efekt , který má spalování
ještě bezpečně zajištění stáří pro jinak chudé rodiče . Čím	více /hodně/Dg-----2A-----	mladých /mladý/AAAMP2----1A-----	později obhospodařuje rodnou hroudu , tím lépe se starším vyplatila
kolem vozu dav stovek lidí . Oddání a ještě mnohem	více /hodně/Dg-----2A-----	mladí /mladý/AAAMP1----1A-----	lidé táhli vůz po ploše festivalu dvě hodiny a nechali
" , Henri Matisse . " Někdy více měřitelné vytlačuje	více /hodně/Dg-----2A-----	důležité /důležitý/AAFS2----1A-----	" , Rene Dubois . " Pokud se matematické zákony
sama je přirozeným místem setkávání , což je o to	více /hodně/Dg-----2A-----	důležité /důležitý/AAFS2----1A-----	, že bude jediným místem zcela otevřeným studentům všech technických
tento dojem zrušení času by mohl být způsoben synchronním plynutím	více /vice/Ca-2-----	různých /různý/AAIP2----1A-----	časoměr , což je jedna z podstatných vlastností mé nové
a pouze dodal : Je těžké říci k tomu něco	více /hodně/Dg-----2A-----	blízkého /blízký/AAANS2----2A-----	než to , že se jednalo o chybu , která

Ručně jsme vytřídili řádky , které nebyly hledanými doklady analytického stupňování . Všimli jsme si i případů , u nichž je možná dvojí interpretace , a případů , kde je význam interpretovatelný až na základě širšího kontextu .

## Formulace závěrů

Získali jsme seznam následujících dokladů opisného stupňování z korpusu SYN2010 .

určitou dobu ztratit " a trochu se toulat po < méně> frekventovaných koutech světa  
písečné a kamenité pustině hledal stopy po < méně> zřetelných rozsáhlých tvarech - bud  
čnejíce uprostřed sutin a balvanů zbylých po < méně> odolných susedech , tu v řadách či  
východ . Ve Střední Asii s jejich stále < více> uzavřenými hranicemi se nové tratě  
tato asi dvacetimetrová ruina ve srovnání s < více> známým vrakem americké lodí Liberty z  
. . Vztahy obou mužů se tak staly ještě < více> nepřátelskými , a to do té míry , že  
Slavíka , Oldřicha Králíka , Žofie Hekelové po < méně> známé a dnes už zapomenuté autory .  
sociálním vývojem společnosti . Ve spojení s < více> autoritativním nebo naopak  
hodnotové orientace společnosti , uřčované stále < více> materialistickými projevy života ,  
některé exekuční úkony prováděny společně vůči < více> povinným , exekuční správní orgán  
terapeutické přístupy či jejich modifikace , stále < více> prosazovaným ekvivalentem lege  
navzájem svou činnost hodnotí , a to stále < více> formalizovaným způsobem (   
pracovištěm moderního člověka , ale je stále < více> iniciačním a řídícím střediskem  
Kordony se třemi , čtyřmi , nebo dokonce i < více> svislými rameny vyžadují pracný  
studijních zón postupuje směrem nahoru k tišším a < více> individuálním formám studia až do  
příštích pěti letech skutečně mohli stát < více> konkurenceschopnými a efektivními  
se pro malé a střední podniky stává stále < více> přijatelnějším model SaaS - "  
v sortimentu ( i nadále bude poptávka po < méně> kvalitním zboží za nízkou cenu ,  
nasycený trh Český spotřebitel stává stále < více> náročným a sofistikovaným , což  
změn se tak český spotřebitel stává stále < více> náročným a sofistikovaným , což  
všech , aby byli stále jedinečnějšími , < více> odlišnými a individualizovanými  
které se dají příležitostně zvětšit , ve < více> stísňených kuchyních je možné  
poměrně ojedinělé . Velmi zajímavým , a stále < více> podporovaným druhem učení s  
týče věku i stupně omezení . Při práci s < více> postiženými a dezorientovanými  
reformovali a dali prostor méně ubliženým a < více> racionálním politikům ze svého  
několik cest . Dnešní tip na výlet vede po < méně> známé , ale zajímavé trase  
rozeberou , zda francouzské události jsou < více> spontánním výbu - chem vzteku ,  
proraženou lebkou . Několik brankářů totiž i po < méně> hrůzostrašných střetech se  
Jaroslav Klimek . Těžký terén vyhovoval < více> domácím Prusinovicím , které  
nad Metují se lze vydat na projížďku po < méně> frekventovaných silnicích . Hned  
nahohlkou , mírně slanou , s příchutí po < méně> pražených pistáciích . mandle Ar  
pomocí přístrojů navíc není v nemocnicích po < méně> závažných operacích běžnou praxí  
rozloučil . Stánkaři putují po městě po < méně> výhodných místech kolín /  
velmi láteři , protože při stěhování po < méně> lukrativních místech ztrácí zisky

Vyhledávky svědčí o tom, že opisem se v češtině 2. a 3. stupeň vyjadřuje dosti zřídka (ostatně i výskyt syntetických tvarů je nízký, viz výše). Pokud bychom chtěli zkoumat tento jev, potřebovali bychom více rozsáhlejších<sup>48</sup> korpusů. Zajímavé je, že počet opisů s *více* a *méně* je vcelku vyrovnaný, přestože se česky srovnání směrem dolů vyjadřuje výhradně analyticky (opisem).

### **Formulace dalších otázek vypluvších ze zkoumání daného jevu**

Pomocí adverbii míry *více/méně* lze vyjádřit míru i u dalších slovních druhů, např. míru intenzity děje. V korpusech můžeme podobným postupem hledat doklady jako ... *ženy se potí <méně> než muži ...* nebo ... *přátel se bojí <méně> než neznámých ...*

### **Zadání cvičení, v nichž lze uplatnit analogické postupy**

Opisnou formou *více/méně* + tvar pozitivu lze stupňovat také adverbia. Zopakujte analogický postup pro vyhledávání dokladů typu *více/méně úspěšně* v korpusu SYN2010.

### ***Slovesné frazémy jako typ MWE***

#### **Motivační úvod**

Frazémy a idiomy tvoří v rámci lexika jednotky, které je třeba zkoumat komplexně. Z hlediska NLP tak patří pod zastřešující pojem MWE.<sup>49</sup> Podívejme se, jak je možné postupovat v případě vyhledávání frazémů a idiomů, jejichž části nestojí vždy lineárně za sebou.

#### **Nastínění problému**

Jak upozorňuje přední badatel na poli frazeologie a idiomatiky (autor *SČFI*) a korpusový lingvista (bývalý ředitel ÚČNK) František Čermák, řada slovních tvarů je sporná z hlediska lemmatizace, neboť se vyskytují pouze v ustálených spojeních. Příkladem může být tvar „*holičkách*“ lemmatizovaný v korpusech řady SYN jako tvar 6. pádu plurálu substantiva *holička*.

---

<sup>48</sup> Sic! Nejde o pleonasmus.

<sup>49</sup> Více k této problematice srov. Hnátková 2002, Hnátková, Kopřivová 2013.



Projdeme-li konkordanční řádky, zjistíme, že o substantivum označující kadeřnici v žádném z dokladů nejde. Jen v několika málo případech není v levém bezprostředním kontextu substantiva v 6. pádě plurálu předložka *na* a nejedná se o frazeologické spojení, ale např. o proprium s obdobnou motivovací.

operní pěvkyni Emě Destinnové zůstal na statku u Bublů v **Holičkách** /holička/NNFP6-----A----- nedaleko Stráže nad Nežárkou stručný rukopisný vzkaz : " Ryby

## Otázky

Z toho, co se uvádí v *SČFI*, je patrné, že zatímco spojení *na holičkách* je pevně dáno, slovesná část frazému je „volněji“, existuje více sloves (*být, zůstat, nechat, octnout se, ...*), která tvoří frazém.

**scfis** Slovník české frazeologie a idiomatiky: výrazy slovesné

### být/bejt/zůstat na holičkách

viz nechat někoho na holičkách

### nechat<sup>o</sup> někoho na holičkách

(kol, nepřizn, hanl) **O** ot, pas, imp, préz, 1. sg **Verb** octnout se na h., být/zůstat na h. (Čl. zvl. vůči přátelům v nesnázích ap.) opustit n. zradit někoho a nepomoci mu, nechat někoho bez pomoci. ♦ *Na holičkách*, užívané jen ve frazeologii, souvisí s *holý* (šlo pův. o holé, tj. prázdné ruce). S nepomocí někomu, nechat někoho — v nesnázích, nechat někoho v bryndě/v loži, nepochout pro někoho ani prstem A podat někomu pomocnou ruku, postavit někoho na nohy Cf vysrat se na někoho j. na placatej kámen, obrátit se k někomu zády ♦ A leave so. in the lurch N jm im Stich lassen F laisser tomber qn

### octnout se na holičkách

viz nechat někoho na holičkách

Dokládá užití spojení *na holičkách* v korpusech pouze slovesa uvedená ve slovníku, nebo lze najít i nějaké inovace?

## Formulace dotazu pro získání dat z korpusů

Zvolíme **Typ dotazu fráze** a do dotazovacího řádku zapíšeme *na holičkách*.

To je klub , který vás nenechal po ničivých záplavách	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	a nejen materiálně , ale i morálně vás podpořil ?
jeho intervence tak omezená rozsahem a proč nechal nakonec Španělsko	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	? Nebo intervenoval proto - jak se domnívají katolíci -
zlobí kvůli tomu , jak ji v sobotu večer nechala	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. Sandy se rozhodla , že jí potom zavolá nebo
o pacienty a lékaře zůstává stranou . Kraj nenechal Čáslav	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	Finanční pomoc městu Čáslav na odstranění následků loňské havárie u
. . . Ten Scheidemann . . . Já nenechal armádu	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	, to armáda nechala mě . . . Do Berlína
tam nešel nikdo z rodiny . Charles Doyle nechal mámu	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	a odsoudil ji i děti k nenápadné bídě . Byl
, že už nikdy nikomu nedovolíte , aby vás nechal	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. Na svou tabuli můžete zapsat kohokoli a pak mít
Jenže má to i druhou stránku . Gólmani často zůstávají	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. " Máme body , ale 35 inkasovaných gólů se
" řekl Louis , " nemohli jsme prostě nechat měsíc	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	, ne ? " Opět se zasmál , vytáhl ženu
, která se mizivě tyčí do výše , nechává lidi	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. " Každý objev , " zaznívá v krtkově vyprávění
závislosti , ani pocit , že vás vaši blízcí nechají	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. Především musíte vytrvat v přesvědčení , že zvítězíte .
řekl , " to nemůžeme . Nikdy bych nenechal Andrého	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. " Radši uměš ? zeptal se Astaroth . Na
. Za oknem se mihl stín . " Ponechání výtahem	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	, " mumlal trpce Zafod . Necitil se zrovna ve

Třídění a pozorování dat získaných z korpusů

Zvolíme filtr (negativní), rozsah hledání <-5,5>, **Typ dotazu lemma** a do dotazovacího řádku zapíšeme *(být)|(nechat)|(zůstat)|(octnout)*.

Jenže má to i druhou stránku . Gólmani často zůstávají	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	" Máme body , ale 35 inkasovaných gólů se
, která se mizivě tyčí do výše , nechává lidi	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	" Každý objev , " zaznívá v krtkově vyprávění
. Za oknem se mihl stín . " Ponechání výtahem	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	" mumlal trpce Zafod . Necítil se zrovna ve
Znojma ) : Orli se trápí a svého gólmána nechávají	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	V neděli dokonce po pěti inkasovaných gólech poprvé v
, že je nechalo z toho či onoho důvodu náboženství	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	Neberte má slova jako něco podobného zákonu . Zabývejte
ani láskyplná slova : Je člověk , kterého ostatní nechávají	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	Vienna V úterý , téměř týden a půl po
Neřešil , že ho spoluhráči nechávali v první půli často	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	a obrana nefungovala . " Z brány jsem moc neviděl
jak ošklivě - a jak dlouho - je Loďstvo nechávalo	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	Vzdychl a otočil se na křesle , aby nahlédl
korun . Podle KSČM naopak nechává reforma chudé až příliš	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	" Například lidé nad padesát let těžko shánějí práci
klade čeřeny . Nic mu nepřítakává tak zhola ! Úpí	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	A jízba odkáže to potomnímu . Řekne tedy :
to dokážeš . At vyjde pravda najevo . Nenechávěj mě	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	Ale to si ze mne dělá legraci . Ji
zemědělské politiky . " Chci reformu , protože nehodlám ponechat	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	farmáře , kteří nežádají pomoc , kteří nechtějí žít ze
tuším i znervózňuje . Myslí si , že mě nechává	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	, když z něho takhle vyprchá energie , jenže ono
. Poté chtějí nechat vládcem nad vysílací licencí Ecclestone "	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	" . At už spor dopadne jakkoliv a Grand Prix
, že vás nechává na holičkách - že nechává loď	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	A svým způsobem má pravdu . Nevím proč ,
skončil tady , s krví podlitýma očima , neoholený ,	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	uprostřed neznáma a s myšlenkou na zoufalý čin . Jak
zápasech . * V první třetině vás ale obrana nechávala	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. Na začátku jsme byli trochu poblázněni .
alytické metody automaticky musí , pak svého pacienta nechává	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	. I přesto , že pacient možná hledá to ,
ukousnout její pěstěné ruce . Ubohý kosmolog tak zůstává často	na /na/RR--6-----	holičkách /holička/NNFP6-----A-----	, osířelý jako mlynářovo dítě . Obrací mezi prsty těžkou

Pozorujeme konkordanční seznam a vidíme, že se objevují iterativa *nechávat*, *zůstávat*, ale i varianta *ponechat/ponechávat*.

## Formulace závěrů

Zkoumání korpusových dat ukázalo, že slovesné tvary, které jsou součástí slovesných frazémů, mají více variant užití, než je zachyceno ve slovníku. Na tento aspekt je třeba brát zřetel zejména v případě, že jmenná část slovesného frazému bude slovo / slovní tvar, který se bude vyskytovat v různých frazémech či kolokacích.

## Formulace dalších otázek vyplývajících ze zkoumání daného jevu

V *SČFI* se tvar *rukou* objevuje v celé řadě frazeologických spojení s různými významy (např. *být v dobrých rukou*, *sehnat něco pod rukou*, ...). Aplikujte postup naznačený výše k odlišení jednotlivých frazémů.

## Zadání cvičení, v nichž lze uplatnit analogické postupy

Vyhleďte v korpusu tvar *panáka* a sledujte, se kterými slovesy tvoří kolokaci s významem „*dopřát sobě/někomu určitou dávku tvrdého alkoholu*“. Podívejte se do *SČFI* a sledujte, jak je příslušný frazém zachycen. Porovnejte výsledky svého šetření s korpusem.

## Závěr

Cílem textu **Česká morfologie a korpusy** je podnítit studenty lingvistických oborů, bohemisty, korpusové lingvisty i všechny, které zajímá čeština a její gramatika na straně jedné a metody korpusové lingvistiky na straně druhé, aby sami zkoušeli hledat postupy, jimiž lze z jazykových korpusů čerpat data pro řešení lingvistických problémů. Snažíme se názorně ukázat, jak se na poli korpusového výzkumu kombinují přístupy corpus based (hledání dokladů pro lingvistický výzkum v korpusech) a corpus driven (optimalizace dotazů založená na pozorování korpusových dat). Kniha nabízí řadu přesně dokumentovaných postupů řešení úloh z oblasti formálního tvarosloví a tvoření slov. Vedle toho ovšem také klade další otázky, které vyvstanou, když začneme skutečně do hloubky pracovat s korpusovými daty.

Charles J. Fillmore (1992 : 35) napsal: „*I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore ... [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way.*“<sup>50</sup>

V tomto smyslu by kniha měla být výzvou pro další bádání. Každý z oddílů kapitoly, která nese v názvu titul celé knihy, obsahuje stručné formulace dalších možných otázek, jež mohou vyvstat po vyřešení modelového úkolu. Připojuje také otázky, na něž lze hledat odpovědi na základě analogických postupů. Obojí má vést čtenáře k další samostatné práci s jazykovými korpusy.

Součástí knihy je bibliografie. Na řadě míst v textu se odkazuje k publikacím, které mohou doplnit to, co je stručně naznačeno. Mnohde odkazujeme k článkům, které obsahují řešení související s výkladem i dalšími směry bádání naznačenými v „dalších otázkách“.

Za jednu z předností korpusové lingvistiky se pokládá fakt, že výsledky získané na základě pozorování korpusových dat umožňují všem (dalším badatelům, ale i autorovi samému) zpětnou verifikaci. Korpusová data jsou přístupná pro ověření tvrzení, která z nich vycházejí. Proto pokládáme za dobré upozornit na to, co již bylo na základě korpusů o českém jazyce napsáno, protože tím nikterak nemíníme „podvazovat“ budoucí bádání, ani kritizovat minulá opomenutí. Mýliti se je lidské. A existují i omyly, které mohou být tím, co se v Exultet<sup>51</sup> nazývá *šťastnou vinou (felix culpa)*, nebo tím, co básnicky vyjádřil František Halas verši:

---

<sup>50</sup> Nemyslím si, že by mohly existovat sebevětší korpusy, v nichž bych našel všechny informace o slovníku a gramatice angličtiny, které chci zkoumat ... (ale) i ten nejmenší korpus, který jsem měl možnost zkoumat, mě přivedl k věcem, na něž bych jinak nebyl býval přišel.

<sup>51</sup> Exultet – velikonoční liturgický zpěv.

...*Co zvedám/ zavíní čísi pád/ co upouštím/ bude mu vzlétnutím ...*<sup>52</sup> Odkazujeme-li na články a další publikace z oboru korpusové lingvistiky, které řeší některé z nastíněných problémů, vyzýváme tím čtenáře mimo jiné i k tomu, aby se pokusili buď pokračovat, nebo změnit směr. Rozhodně nemají „hodit flintu do žita“ nebo „přešlapovat na místě“, protože „to už někdo udělal“. V mnoha případech se totiž můžeme něčemu naučit i tehdy, když projdeme cestou, kterou už někdo šel, protože si na ní můžeme všimnout věcí, kterých si on nevšiml. Přeji všem, kteří se pustí na pouť s korpusy po češtině, aby jim tato kniha byla dobrým průvodcem. Šťastnou cestu.

---

<sup>52</sup> Halas, F.: A co básník. In: Kundera, L. (ed.), *A co*, Praha : Československý spisovatel, 1983, s. 111.

## Bibliografie

- Aitzenmüller, R.: Ein baltisch-slavisches Elativsuffix und seine Entsprechungen in den übrigen indogermanischen Sprachen; der griechische Superlativ auf *-atos / -tatos*. *Slavistična revija* 3, 1950, 289–296.
- Blatná, R. – Petkevič, V. (eds.): *Jazyky a jazykověda. Sborník k 65. narozeninám prof. Františka Čermáka*. Praha : FF UK Ústav ČNK, 2004.
- Blatná, R. – Čermák, F.: *Manuál lexikografie*. Jinočany : H&H, 1995.
- Boissin, H.: Quelques procédés de renforcement nominal en serbo-croate. In: *Revue des études slaves*, Tome 34, 1957, s. 32–36.
- Caha, P. – Scheer, T.: *Syntax and Phonology of Czech Templatic Morphology*. (Dostupné z: <[http://www.linguistics.stonybrook.edu/fasl/abstracts/H\\_Caha%20Scheer.pdf](http://www.linguistics.stonybrook.edu/fasl/abstracts/H_Caha%20Scheer.pdf)>). Handout pro *FASL 16* 2. 5. – 4. 5. 2007.
- Cvrček, V. a kol.: *Mluvnice současné češtiny*. Praha : Nakladatelství Karolinum, 2010.
- Cvrček, V. – Kovaříková, D.: Možnosti a meze korpusové lingvistiky. *Naše řeč* 3, 94, 2011, s. 113–133.
- Čermák, F.: *Syntagmatika a paradigmatica českého slova*. Praha : Karlova Univerzita, 1990.
- Čechová, M. a kol.: *Čeština, řeč a jazyk*. Praha : ISV Nakladatelství, 2000.
- Čermák, F. – Klímová, J. – Petkevič, V. (eds.): *Studie z korpusové lingvistiky*. Praha : Karolinum. 2000.
- Čermák, F. – Blatná, R.: *Korpusová lingvistika. Stav a modelové přístupy*. Praha : Nakladatelství Lidové noviny, 2006.
- Čermák, F. a kol.: *Slovník české frazeologie a idiomatiky 1–4 (SČFI)*. Praha : LEDA, 2009.
- Čermák, F. – Křen, M.: *Frekvenční slovník češtiny*. Praha : Nakladatelství Lidové noviny, 2004.
- Dokulil, M.: *Tvoření slov v češtině. 1, Teorie odvozování slov*. Praha : Academia, 1967.
- Dokulil, M.: *Tvoření slov v češtině. 2, Odvozování podstatných jmen /zpracoval kolektiv pracovníků Ústavu pro jazyk český ČSAV za redakce Fr. Daneše, M. Dokulila, J. Kuchaře*. Praha : Academia, 1967.
- Dokulil, M. – Komárek, M. a kol.: *Mluvnice češtiny 1, 2*. Praha : Academia, 1987.
- Filipec, J. a kol.: *Slovník spisovné češtiny pro školu a veřejnost (SSČ)*. Praha : Academia, 2005.
- Fillmore, Ch.: “Corpus linguistics” or “Computer-aided armchair linguistics”. In: Svartvik, J. (ed.), *Directions in Corpus Linguistics*, 35–60, Berlin : Mouton de Gruyter, 1992.

- Gebauer, J.: *Gebauerova mluvnice česká pro školy střední a ústavy učitelské. 1. Hláskosloví; Nauka o slově*. Praha : Česká grafická Unie, 1920.
- Hajič J.: *Unification Morphology Grammar*. Praha : MFF UK, 1994. (Disertační práce.)
- Hajič J.: *Desambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha : Karolinum Charles University Press, 2004.
- Hajičová, E. – Panevová, J. – Sgall, P.: Smysl formalismu v teoretické lingvistice a úloha českého výzkumu v ní. *Slovo a slovesnost* 52, s. 175–182.
- Havránek, B. a kol.: *Slovník spisovného jazyka českého (SSJČ)*. Praha : Academia, 1989.
- Hladká, Z. – Osolobě, K.: Příprava elektronických korpusů češtiny. In: *Přednášky a besedy z XXXIV. běhu LŠSS*. Brno : FF MU, 2001, s. 63–70.
- Hladká, Z.: Zkušenosti s tvorbou korpusů češtiny v ÚČJ FF MU v Brně. *SPFFBU A* 53, Brno : FF MU, 2005, s. 115–124.
- Hladká, Z. a kol.: *Čeština v současné soukromé korespondenci. Dopisy, e-maily, SMS*. [CD-ROM]. Brno : Masarykova univerzita, 2005.
- Hlaváčková, D.: Korpus mluvené češtiny z brněnského prostředí a jeho morfologické značkování. *Slovo a slovesnost* 62, 2001, s. 62–70.
- Hlaváčková, D. – Osolobě, K.: Morfologické značkování mluvených korpusů, zkušenosti a otevřené otázky. In: Kopřivová, M. – Waclawičová, M. (eds.), *Čeština v mluveném korpusu*, Praha : Nakladatelství Lidové noviny, 2008, s. 105–114.
- Hlaváčková, D. – Pala, K.: Computer Processing Derivational Relations in Czech. In: Garabík, R. (ed.), *Computer Treatment of Slavic and East European Languages*, Bratislava : Slovak National Corpus, L. Štúr Institute of Linguistics, Slovak Academy of Sciences, 2007, s. 198–208.
- Hlaváčková, D. – Osolobě, K. – Pala, K. – Šmerk, P.: Relations between Formal and Derivational Morphology in Czech. In: Dočekal, M. – Ziková, M. (eds.), *Czech in Formal Grammar*, Mnichov : Lincom, 2009<sup>1</sup>, s. 79–87.
- Hlaváčková, D. – Osolobě, K. – Pala, K. – Šmerk, P.: Exploring Derivational Relations in Czech with the Deriv Tool. In: Levická, J. – Garabík, R. (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research*, Bratislava/Smolenica : Tribun, 2009<sup>2</sup>, s. 152–161.
- Hlaváčková, J.: Morphological Guesser of Czech Words. In: Matoušek, V. (ed.), *Text, Speech and Dialogue*, Berlin : Springer-Verlag, 2001, s. 70–75.
- Hlaváčková, J.: *Formalizace systému české morfologie s ohledem na automatické zpracování českých textů*. Praha : UK, 2009. (Disertační práce.)

Hnátková, M.: Značkování frazémů a idiomů v českém Národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost* 63, 2002, s. 117–126.

**Hnátková, M., Kopřivová, M.:**

Horák, A. – Pala, K. – Rambousek, A. – Rangelova, A.: Nové nástroje pro českou lexikografii – DEB2. In: Štícha, F. – Šimandl, J. (eds.), *Sborník konference Grammar & Corpora 2005*, Praha : Ústav pro jazyk český Akademie věd ČR, 2006, s. 190–196.

Hujer, O. – Smetánka, E. – Weingart, M. – Havránek, B. – Šmilauer, V. – Ziskal, A. (red.): *Příruční slovník jazyka českého (PSJČ)*. Praha : Státní nakladatelství, 1935–1957.

Jelínek, T.: Nové značkování v Českém národním korpusu. *Naše řeč* 91, 2008, s. 13–20.

Jelínek, T. – Petkevič, V.: Systém jazykového značkování korpusů současné psané češtiny. In: Petkevič, V. – Rosen, A. (eds.), *Korpusová lingvistika Praha 2011 – 3. Gramatika a značkování korpusů*, Praha : Nakladatelství Lidové noviny, 2011, s. 154–170.

Karlík, P. – Nekula, M. – Rusínová, Z.: *Příruční mluvnice češtiny*. Praha : Nakladatelství Lidové noviny, 1995.

Karlík, P. – Nekula, M. – Pleskalová, J.: *Encyklopedický slovník češtiny*. Praha : Nakladatelství Lidové noviny, 2002.

Karlík, P. (ed.): *Korpus jako zdroj dat o češtině*. Brno : Masarykova univerzita, 2004.

Karlík, P.: Několik drobných poznámek k dějovým jménům. *Linguistica Brunensia* A 55, Brno : FF MU, 2007, s. 235–246.

Karlík, P. – Pleskalová, J. (eds.): *Život s morfémy*. Brno : Masarykova univerzita, 2004.

Klímová, J. – Štícha, F.: K možnostem a mezím sufixální derivace substantiv. In: Štícha, F. (ed.): *Možnosti a meze české gramatiky*, Praha : Academia, 2006, s. 127–138.

Koček, J. – Kopřivová, M. – Kučera, K.: *Český národní korpus – úvod a příručka uživatele*. Praha : Ústav Českého národního korpusu FF UK, 2000.

Konečná, D.: Ukázka použití statistického zkoumání při přípravě strojové syntézy českého jednoduchého slovesného tvaru. *Slovo a slovesnost* 22, 1961, s. 268–273.

Komárek, M.: Příspěvky k české morfologii. Olomouc : Periplum, 2006.

Kopečný, F.: *Slovesný vid v češtině*. Praha : Nakladatelství československé akademie věd, 1962.

Kosek, P. – Křístek, M. – Osolobě, K. – Vojtová, J. – Ziková, M.: První korpusová mluvnice češtiny. (Recenze práce: Václav Cvrček a kolektiv autorů: *Mluvnice současné češtiny 1 – Jak se píše a jak se mluví*). *Naše řeč* 94, 2011, s. 142–152.

Králík, J. – Těšitelová, M.: *Retrogradní slovník současné češtiny*. Praha : Academia, 1986.

- Leech, G.: Corpus annotation schemes. *Literary and Linguistic Computing* 8 (4), 1993, s. 275–281.
- Marvan, J.: *České stupňování*. Praha : Karolinum, 2008.
- Mistrík, J.: *Retrográdny slovník slovenčiny*. Bratislava : Univerzita Komenského, 1976.
- McEnery, A. – Wilson, A.: *Corpus Linguistics*. Edinburgh : Edinburgh University Press, 1996.
- Němec, I.: Hláskoslovné předpoklady produktivity slovotvorných formantů. *Slovo a slovesnost* 32, 1971, s. 217–224.
- Osolsobě, K. – Pala, K.: Czech Stem Dictionary for IBM PC XT/AT. In: *Conference on Computer Lexicography*, Balatonfüred : Research Institute for Linguistics Hungarian Academy of Sciences, 1991, s. 163–172.
- Osolsobě, K.: Česká formální morfologie na počítači, aneb jak se počítač učil časovat česká pravidelná a nepravidelná slovesa. In: *Přednášky a besedy z XXVII. běhu LŠSS*, Brno : FF MU, 1994, s. 16–31.
- Osolsobě, K.: Automatické rozpoznávání a generování českých určitých číslovek a od nich odvozených číselných pojmenování na počítači. *SPFFMU A 43*, Brno : FF MU, 1995, s. 31–48.
- Osolsobě, K.: *Algoritmický popis české morfologie a strojový slovník češtiny*. Brno : FF MU, 1996. (Disertační práce.)
- Osolsobě, K.: Formale Beschreibung der Tschechischen Morphologie und ihre Anwendung auf dem Gebiet der linguistischen Datenverarbeitung. *SPFFBU A 44*, Brno : FF MU, 1996, s. 59–70.
- Osolsobě, K.: Formale Beschreibung der Tschechischen Morphologie. In: Junghanns, U. – Zybatow, G. (Hrsg.), *Formale slavistik*, Frankfurt am Main : Ververt Verlag, 1997, s. 443–451.
- Osolsobě, K. – Pala, K. – Rychlý, P.: Frekvence vzorů českých substantiv (na materiálu ČNK). *SPFFBU A 46*, Brno : FF MU, 1998, s. 77–94.
- Osolsobě, K. – Pala, K. – Rychlý, P.: Frekvence vzorů českých sloves (na materiálu Českého národního korpusu), *Slovo a slovesnost* 59, 1998, s. 265–277.
- Osolsobě, K.: Morfologické značkování složených slovesných tvarů v korpusu. *SPFFBU A 47*, Brno : FF MU, 1999, s. 33–50.



- Osolsobě, K.: Mluvnice versus korpus, několik poznámek k problémům dubletních a variantních koncovek českých substantiv. In: Hladká, Z. – Karlík, P. (eds.), *Čeština, univerzália a specifika 4*, Praha : Nakladatelství Lidové noviny, 2002, s. 333–336.
- Osolsobě, K. – Pala, K. – Sedláček, R. – Veber, M.: A Procedure for Word Derivational Processes Concerning Lexicon Extension in Highly Inflected Languages. In: Rodriguez, M. G. – Araujo, C. P. (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC*, Las Palmas de Gran Canaria : ELRA, 2002, s. 1254–1259.
- Osolsobě K.: Hypokoristika v korpusu soukromé korespondence KSK. *SPFFBU A 53*, Brno : FF MU, 2005, s. 125–136.
- Osolsobě, K.: Korpus soukromé korespondence z hlediska morfologického značkování. *SPFFBU A 54*, Brno : FF MU, 2006, s. 187–201.
- Osolsobě, K. – Pala, K. – Sedláček, R.: Brněnský atributivní tagset. Brno : NLP FIMU, 2006. (Dostupný z: <<http://nlp.fi.muni.cz/projekty/ajka/tags.pdf>>.)
- Osolsobě, K.: Popis gramatických významů (hodnot) jednoduchých slovesných tvarů v anotacích českých (slovenských) korpusů. *SPFFBU A 55*, Brno : FF MU, 2007<sup>1</sup>, s. 201–218.
- Osolsobě, K.: Matematická lingvistika. In: Pleskalová, J. – Krčmová, M. – Večerka, R. – Karlík, P. (eds.), *Dějiny české jazykovědné bohemistiky*, Praha : Academia, 2007<sup>2</sup>, s. 447–467.
- Osolsobě, K.: Syntetické futurum v češtině – gramatiky, slovníky, korpusy. In: *Přednášky a besedy z XL. běhu LŠSS*, Brno : FF MU, 2007<sup>3</sup>, s. 131–144.
- Osolsobě, K.: Značkování gramatických kategorií v korpusech ČNK a jejich zachycení v gramatice a ve slovníku (syntetické futurum, stupňování adjektiv, neurčité číslovky a příslovce míry). In: Štícha, F. (ed.), *Grammar & Corpora / Gramatika a korpus 2007*, Praha : Academia, 2008<sup>1</sup>, s. 407–416.
- Osolsobě, K.: Propria (příjmení na -č) – problém automatické morfologické analýzy. In: Čornejová, M. – Kosek, P. (eds.), *Jazyk a jeho proměny. Prof. Janě Pleskalové k životnímu jubileu*, Brno : Host, 2008<sup>2</sup>, s. 205–216.
- Osolsobě, K.: Čeho je moc, toho je příliš aneb jaké má čeština komparativy a superlativy? In: *Přednášky a besedy z XLI. běhu LŠSS*, Brno : FF MU, 2008<sup>3</sup>, s. 145–158.
- Osolsobě, K.: Formální popis deverbativ na -č. *Linguistica Brunensia 56*, Brno : FF MU, 2008<sup>4</sup>, s. 121–135.
- Osolsobě, K.: K jednomu typu vyjadřování stupně v češtině. *Bohemica Olomuciensia 3 – Linguistica Juvenilia*, Olomouc : Univerzita Palackého v Olomouci, 2009<sup>1</sup>, s. 123–138.

- Osolsobě, K.: *Deriv* – nástroj pro automatické vyhledávání slovotvorných vztahů. Slovtvorný stroj pro češtinu – sen nebo skutečnost? In: *Přednášky a besedy z XLII. běhu LŠSS*, Brno : FF MU, 2009<sup>2</sup>, s. 132–137.
- Osolsobě, K.: Deriváty na -čí: gramatika, slovník a korpus. *Linguistica Brunensia* 57, Brno : FF MU, 2009<sup>3</sup>, s. 115–127.
- Osolsobě, K.: Kajícny a nevěřicny – adjektiva na -cí/-cný: slovníky, gramatiky, korpusy. In: Hlaváčková, D. – Horák, A. – Osolsobě, K. – Rychlý, P. (eds.), *After Half a Century of Slavonic Natural Language Processing*, Brno : Masarykova univerzita, 2009<sup>4</sup>, s. 173–183.
- Osolsobě, K.: Několik poznámek na okraj derivací od sloves s uzavřeným kmenem minulým. In Bičan, A. – Klaška, J. – Macurová, P. – Zmrzliková, J. (eds.), *Karlík a továrna na lingvistiku. Prof. Petru Karlíkovi k životnímu jubileu*, Brno : Host, 2010, s. 316–330.
- Osolsobě, K.: *Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny*. Brno : Masarykova univerzita, 2011<sup>1</sup>.
- Osolsobě, K.: Korpus jako zdroj dat pro studium slovotvorby. In: Petkevič, V. – Rosen, A. (eds.), *Korpusová lingvistika Praha 2011 – 3. Gramatika a značkování korpusů*, Praha : Nakladatelství Lidové noviny, 2011<sup>2</sup>, s. 10–23.
- Osolsobě, K., Vališová, P.: Using data-driven methods in teaching Czech as a foreign language. In: Thomas, J. – Boulton, A. (eds.), *Input, Process and Product. Development in Teaching and Language Corpora*, Brno : MUP, 2012, s. 183–174.
- Pala, K.: *Počítačové zpracování češtiny*. Brno : FF MU, 1992. (Habilitationní práce.)
- Pala, K. – Všiansky, J.: *Slovník českých synonym*. Praha : Nakladatelství Lidové noviny, 1996.
- Pala, K. – Sedláček, R. – Veber, M.: Vztah mezi tvarotvornými a slovotvornými vzory v češtině. In: Hladká, Z. – Karlík, P. (eds.), *Čeština univerzália a specifika 5*, Praha : Nakladatelství Lidové noviny, 2004, s. 151–162.
- Pala, K. – Sedláček, R. – Veber, M.: Relations between Inflectional and Derivation Patterns. In: Vitas, D. – Erjavec, T. – Tadić, M. (eds.), *Proceedings of the Workshop on Morphological Processing of Slavic Languages*, Budapest : EACL, 2003, s. 1–8.
- Pala, K. – Sedláček, R.: Enriching WordNet with Derivational Subnets. In: Sedláček, R. (ed.), *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing CICLING 2005*, Berlin Heidelberg New York : Springer Verlag, 2005, s. 305–311.
- Pala, K.: Derivational Relations in Slavonic Languages. In: Tadić, M. – Dimitrova-Vulchanova, M. – Koeva, S. (eds.), *FASSBJ6-2008 (Proceedings of the Sixth International*

*Conference on Formal Approaches to South Slavic and Balkan Languages 25–28 September 2008, Dubrovnik, Croatia*), Zagreb : Croatian Language Technologies Society – Faculty of Humanities and Social Sciences, 2008, s. 21–28.

Pala, K. – Hlaváčková, D.: Repräsentace významu sloves (valence a sémantické role). In: *Sborník konference Kognice 2010*, Praha : Kognice, 2010.

Pala, K. – Rychlý, P.: Do We Need Very Large Web Corpora? In: Čermák, F. (ed.), *Korpusová lingvistika Praha 2011 – 2. Výzkum a výstavba korpusů*, Praha : Nakladatelství Lidové noviny, 2011, s. 30–40.

Panevová, J. – Ševčíková, M.: Jak se počítají substantiva v češtině: poznámky ke kategorii čísla. *Slovo a slovesnost* 3, 72 (2011), s. 163–176.

Petkevič, V.: Reliable Morphological Desambiguation of Czech: Rule-Based Approach is Necessary. In: Šimková, M. (ed.), *Insight into the Slovak and Czech Corpus Linguistics*, Bratislava : Veda, 2006, s. 26–44.

Petkevič, V.: Využití vidu ke zkvalitnění automatického značkování češtiny. In: Bičan, A. – Klaška, J. – Macurová, P. – Zmrzlíková, J. (eds.), *Karlík a továrna na lingvistiku. Prof. Petru Karlíkovi k životnímu jubileu*, Brno: Host, 2010, s. 368–387.

Pleskalová, J. – Krčmová, M. – Večerka, R. – Karlík, P.: *Kapitoly z dějin české jazykové bohemistiky*. Praha : Academia, 2007.

Pořízka, P. – Schäfer, M.: Morph-Con. A Software for Conversion of Czech Morphological Tagsets. In: Levická, J. – Garabík, R. (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research*, Bratislava/Smolenica : Tribun, 2009, s. 292–301.

Rejzek, J.: *Český etymologický slovník*. Voznice : LEDA, 2001.

Romportl, S.: *Struktura gramatické složky slovesných tvarů určitých v češtině*. Praha : Academia, 1970.

Rosa, V. J.: *Čechořečnost seu Grammatica linguae Bohemicae*. Praha, 1672.

Rosen, A.: Morphological Tags in Parallel Corpora. In: Čermák, F. – Corness, P. – Klégr, A. (eds.), *InterCorp : Exploring Multilingual Corpus*, Praha : Nakladatelství Lidové noviny, 2009, s. 205–234.

Sag, I. A. – Baldwin, T. – Bond, F. – Copestake, A. A. – Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. F. (ed), *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, London : Springer-Verlag, London, UK, 2002, s. 1–15. (Dostupný z: <<http://lingo.stanford.edu/pubs/WP-2001-03.pdf>>.)

- Scheer, T.: The Rhythmic Law in Czech: Vowel-final Prefixes. In: Zybatow, G. – Junghanns, U. – Mehlhorn, G. – Szucsich, L. (eds.), *Current Issues in Formal Slavic Linguistics*, Frankfurt am Main : Lang, 2001, s. 37–48.
- Scheer, T.: O samohláskové délce při derivaci v češtině. In: Hladká, Z. – Karlík, P. (eds.), *Čeština – univerzália a specifika 5*, Praha : Nakladatelství Lidové noviny, 2004, s. 224–239.
- Schejbalová, Z.: Reduplikace jako slootovorný prostředek v češtině a ve francouzštině. In: *Přednášky a besedy z XLIV. běhu LŠSS*, Brno : FF MU, 2011, s. 192–212.
- Sedláček, R.: *Morphematic analyser for Czech*. Brno : FI MU, 2004. (Disertační práce.)
- Sgall, P.: *Generativní popis jazyka a česká deklinace*. Praha : Academia, 1967.
- Skoumalová, H.: Porovnání úspěšnosti tagování korpusu. In: Petkevič, V. – Rosen, A. (eds.), *Korpusová lingvistika Praha 2011 – 3. Gramatika a značkování korpusů*, Praha : Nakladatelství Lidové noviny, 2011, s. 199–207.
- Slavičková, E.: *Retrográdní morfemický slovník češtiny*. Praha : Academia, 1975.
- Sokolová, M. – Moško, G. – Šimon, F. – Benko, V.: *Morfemický slovník slovenčiny*. Prešov : Náuka Prešov, 1999.
- Spoustová, D. – Hajič, J. – Votrubec, J. – Krbec, P. – Květoň, P.: The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. Praha : ACL, 2007, s. 67–74.
- Stich, A.: Stupňování přídavného jména „vzrušující“. *Naše řeč* 52, 1969, s. 62–64.
- Šlosar, D.: *Slootovorný vývoj českého slovesa*. Brno : Univerzita J. E. Purkyně, 1981.
- Lamprecht, A. – Šlosar, D. – Bauer, J.: *Historická mluvnice češtiny*. Praha : SPN, 1986.
- Šlosar, D. – Rusínová, Z.: Průřez vývojem slootovorné soustavy adjektiv v češtině. *SPFFMU A 15*, Brno : FF MU, 1967, s. 37–64.
- Šlosar, D.: *Česká kompozita diachronně*. Brno : Masarykova univerzita, 1999.
- Šmilauer, V.: *Nauka o českém jazyku*. Praha : Státní pedagogické nakladatelství, 1972.
- Šmilauer, V.: *Novočeské tvoření slov*. Praha : Státní pedagogické nakladatelství, 1971.
- Šmerk, P.: *K počítačové morfologické analýze češtiny*. Brno : FI MU, 2010. (Disertační práce.)
- Šimandl, J.: Číslovky: obecný rámec, číslovky základní a řadové. In: Štícha, F. (ed.), *Kapitoly z české gramatiky*. Praha : Academia 2011, s. 689–727.
- Štícha, F. (ed.): *Možnosti a meze české gramatiky*. Praha : Academia, 2006.
- Štícha, F. (ed.): *Kapitoly z české gramatiky*. Praha : Academia, 2011.
- Tichonov, A. N.: *Slovoobrazovatel'nyj slovar' russkogo jazyka*. Moskva : Russkij jazyk, 1985.

- Trávníček, F.: *Mluvnice spisovné češtiny I*. Praha : Slovanské nakladatelství, 1951.
- Tušková, J. M.: *Variantní a dubletní tvary v současné deklinaci apelativních feminin*. Brno : Masarykova univerzita, 2006.
- Tušková, J. M.: *Deklinační systém femininních oikonym v češtině. Synchronní pohled na základě Českého národního korpusu*. Praha : Nakladatelství Lidové noviny, 2011.
- Vališová, P.: Výukové materiály založené na korpusu. In: Čermák, F. (ed.), *Korpusová lingvistika Praha 2011 – 2. Výzkum a výstavba korpusů*, Praha : Nakladatelství Lidové noviny/Ústav Českého národního korpusu, 2011, s. 313–323.
- Večerka, R.: *Staroslověnština*. Praha : SPN, 1984.
- Večerka, R. a kol.: *K pramenům slov. Uvedení do etymologie*. Praha : Nakladatelství Lidové noviny, 2006.
- Veselý, V.: Obojí, oboje, obě, všechno. *Naše řeč* 3, 94 (2011), s. 181–193.
- Wagner, R.: Das Auffinden von reflexiven Verbalsubstantiven im tschechischen Nationalkorpus: Grenzen der morphologischen Annotation. In: Štícha, F. – Šimandl, J. (eds.), *Gramatika a korpus 2005*, Praha : Ústav pro jazyk český, 2005, s. 295–304.
- Worth, D. S. – Kozak, A. S. – Johnson, D. B.: *Russian Derivational Dictionary*. New York – London – Amsterdam : American Elsevier Publishing Company, 1970.
- Ziková, M.: Morphological and Phonological Domains: Prefix Lengthening in Czech. Handout pro *FDSL 2010*. (Dostupný z: <<http://www.phil.muni.cz/jazyk/fdsl/abstracts/Zikova.pdf>>.)

### **Elektronické zdroje:**

- Korpus SYN2010: Český národní korpus – SYN2010*. Ústav Českého národního korpusu FF UK, Praha 2010. Cit. 16.01. 2014, dostupný z WWW: <<http://www.korpus.cz>>.
- Korpus SYN2009PUB: Český národní korpus – SYN2009PUB*. Ústav Českého národního korpusu FF UK, Praha 2010. 16.01. 2014, dostupný z WWW: <<http://www.korpus.cz>>.
- Korpus SYN2006PUB: Český národní korpus – SYN2006PUB*. Ústav Českého národního korpusu FF UK, Praha 2006. Cit. 16.01. 2014, dostupný z WWW: <<http://www.korpus.cz>>.
- Korpus SYN2005: Český národní korpus – SYN2005*. Ústav Českého národního korpusu FF UK, Praha 2005. Cit. 16.01. 2014, dostupný z WWW: <<http://www.korpus.cz>>.
- Korpus SYN2000: Český národní korpus – SYN2000*. Ústav Českého národního korpusu FF UK, Praha 2000. Cit. 16.01. 2014, dostupný z WWW: <<http://www.korpus.cz>>.

*Korpus SYN: Český národní korpus – SYN*. Ústav Českého národního korpusu FF UK, Praha. Cit. 16.01. 2014, dostupný z WWW: <<http://www.korpus.cz>>.

Rychlý, P.: *Bonito – grafické uživatelské rozhraní systému Manatee*. Verze 1.49. 1998–2003. (Dostupná z: <<http://ucnk.ff.cuni.cz/bonito/>>A.)

Korpusový manažer BONITO [online]. 2011. Cit. 20. 03. 2011, dostupný z: <<http://ucnk.ff.cuni.cz/bonito/>>.

Korpusový vyhledávač NoSke [online]. 2013. Cit. 16.01. 2014, dostupný z: <<https://korpus.cz/corpora/>>.

Korpusový vyhledávač KonText [online]. 2014. Cit. 16.01. 2014, dostupný z: <[https://kontext.korpus.cz/run.cgi/first\\_form](https://kontext.korpus.cz/run.cgi/first_form)>.

Šmerk, P.: *Dberiv – weové rozhraní*. 2009. (Dostupné z: <<http://deb.fi.muni.cz/deriv.>>)

*Automatický analyzátor ajka* [online]. 2011. Cit. 16.01. 2014, dostupný z: <<http://nlp.fi.muni.cz/projekty/wwwajka>>.

*Internetový vyhledávač Google* [online]. 2011. Cit. 20. 03. 2011, dostupný z: <<http://www.google.com/>>.

*Internetová jazyková příručka* [online]. 2011. Cit. 16.01. 2014, dostupná z: <<http://prirucka.ujc.cas.cz>>.

DEB Dict – *Obecný prohlížeč slovníků* [online]. 2011. Cit. 16.01. 2014, dostupný z: <<chrome://debdict/content/debdict.xul>>.

*SLEX99 – Elektronický lexikón slovenského jazyka*. Forma, s.r.o. Bratislava 1998–2007. Cit. 20. 03. 2011, dostupný z: <<http://www.slex.sk>>.

hebrew4christians [online] . 2012. Cit. 5. 06. 2012, dostupné z: <[http://www.hebrew4christians.com/Grammar/Unit\\_Five/Comparative\\_Usage/comparative\\_usage.html](http://www.hebrew4christians.com/Grammar/Unit_Five/Comparative_Usage/comparative_usage.html)>.

## Algoritmus určování slovesných tříd a vzorů v češtině<sup>53</sup>

© Klára Osolsobě

Algoritmus funguje v případě, že mluvčí je schopen správně tvořit všechny požadované tvary. Řada rodilých mluvčích totiž při postupu občas chybují. Pokud např. pochybíme při tvoření infinitivu, např. máme-li zařadit do třídy/ke vzoru slovesný tvar jiný než infinitiv, a utvoříme infinitiv od příbuzného slovesa, lišícího se většinou videm/způsobem slovesného děje, je toto pochybení důvodem nesprávné odpovědi (neselže algoritmus, ale naše jazyková schopnost).<sup>54</sup> Dále je třeba si dát pozor na případy kolísání mezi třídami/vzory (viz níže).<sup>55</sup>

1. Utvoř infinitiv.<sup>56</sup> Jdi na 2.
2. Infinitiv končí na  $-([aá]t)|([aá]ti)$ ? Ano: Jdi na 3. Ne: Jdi na 6.
3. Utvoř 3. osobu indikativu přítomného aktiva. Jdi na 4.
4. 3. osoba indikativu přítomného aktiva končí na *á*? Ano: Jdi na 5. Ne: Jdi na 17.
5. Jde o sloveso 5. tř. vzoru *dělat*. KONEC
6. Infinitiv končí na  $-([eěi]t)|([eěi]ti)$ ? Ano: Jdi na 7. Ne: Jdi na 17.
7. 3. osoba indikativu přítomného aktiva končí na *í*. Ano: Jdi na 8. Ne: Jdi na 17.
8. Před infinitivní koncovkou  $-(t)|(ti)$  předchází *i*? Ano: Jdi na 9. Ne: Jdi na 10.
9. Jde o sloveso 4. tř. vzoru *prosít*. KONEC
10. Před infinitivní koncovkou  $-(t)|(ti)$  předchází  $[eě]-$ ? Ano: Jdi na 11. Ne: Jdi na 15.
11. Utvoř imperativ (2. os. sg.). Jdi na 12.
12. Tvar končí na  $-[eě]j$ ? Ano: Jdi na 13. Ne: Jdi na 14.
13. Jde o sloveso 4. tř. vzoru *sázet*. KONEC
14. Jde o sloveso 4. tř. vzoru *trpět*. KONEC
15. Utvoř tvar I-ového přičestí. Jdi na 16.
16. Před *-l* předchází *i*? Ano: Jdi na 9. Ne: Jdi na 11.
17. Infinitiv končí na  $-([áeěi]y)|(ova)|(ou)t|([áeěi]y)|(ova)|(ou)ti)$ ? Ano: Jdi na 18. Ne: Jdi na 31.
18. 3. osoba indikativu přítomného aktiva končí na *-je*. Ano: Jdi na 19. Ne: Jdi na 23.
19. Před infinitivní koncovkou  $-(t)|(ti)$  předchází *ova-*. Ano: Jdi na 20. Ne: Jdi na 21.
20. Jde o sloveso 3. tř. vzoru *kupovat*. KONEC
21. Tvar není pouze *je* (*<být>*)? Ano: Jdi na 22. Ne: Jdi na 45 (*být*).
22. Jde o sloveso 3. tř. vzoru *krýt*. KONEC
23. Infinitiv končí na  $-([nm]out)|([nm]outi)$ ? Ano: Jdi na 24. Ne: Jdi na 31.

<sup>53</sup> O přechodnících se všeobecně tvrdí, že se v češtině příliš neužívají. Nicméně se od nich (vcelku pravidelně) tvoří adjektiva (typ *kupující* nebo *dostavivší se*), která (alespoň v prvním případě) frekventovaná a užívaná jsou. Pravidelně se pak chybují při záměnách adjektiv tvořených od přechodníku přítomného (procesuálních *žehlicí, kryjící, ...*) a adjektiv účelových, tvořených od kmene minulého (typ *žehlicí, krycí, ...*). Ve zvukové realizaci adjektiv odvozených od sloves 4. třídy vzoru *prosít* se totiž stírá rozdíl mezi dlouhým a krátkým *i/t*, což je pro češtinu typické i jinde, např. ve tvarech zájmena *ona – jí/jí*. To má pak vliv na pravopisné chyby při grafickém zaznamenání kvantity u příslušných adjektiv. To jsou mimo jiné důvody, pro něž je třeba tvoření přechodníků věnovat patřičnou pozornost.

<sup>54</sup> Uživatelé korpusu si mohou pomoci tak, že tvar slovesa vyhledají v korpusu a zobrazí si lemma.

<sup>55</sup> Kolísání mezi vzory se může odrazit v lemmatizaci tak, že méně obvyklý tvar (archaický) není rozpoznán automatickou morfologickou analýzou (lemma je tvar sám a tag je *X.\**).

<sup>56</sup> První krok je uveden pro případ, že máme za úkol určit třídu a vzor slovesa v libovolném tvaru (textovém). Tento krok není většinou nutný pro určení třídy, je ale nezbytný pro určení vzoru. Usnadní také práci s výjimkami. Pomůckou ke správnému utvoření infinitivu od libovolného tvaru může být následující postup: a) utvoříme větu, v níž bude příslušný tvar tvarem (součástí tvaru) přísudkového slovesa, a b) provedeme substituci konstrukcí s modálním slovesem. Máme např. určit třídu a vzor slovesného tvaru *odpluje*. Studenti chybovali a tvořili nesprávně infinitiv *odplouvat*. Správně je ovšem *odplout*. Jestliže nahradíme přísudkový tvar *Lod' <odpluje>* z *přístavu* tvarem *Lod' <smí odplout>* z *přístavu*, pak se uvedené chybě vyhneme.

24. 3. osoba indikativu přítomného aktiva končí na *-[nm]e*? Ano: Jdi na 25. Ne: Jdi na 46 (*zabřede*).
25. Před infinitivní koncovkou *-(t)|(ti)* předchází *[bcčdfghjkmptsštvxzž]/[nm]ou*-? Ano: Jdi na 26. Ne: Jdi na 30.
26. Utvoř tvar I-ového participia. Jdi na 27.
27. Před *-l* může předcházet jak *(a)*-, tak *([nm]u)*-? Ano: Jdi na 28. Ne: Jdi na 29a.
28. Jde o sloveso 2. tř. vzoru *začít*. KONEC
- 29a. Jde o sloveso *dmout*? Ano: Jdi na 28. Ne: Jdi na 29b.
- 29b. Jde o sloveso 2. tř. vzoru *tisknout*. KONEC
30. Jde o sloveso 2. tř. vzoru *minout*. KONEC
31. Infinitiv končí na *-([szc]t)|([szc]ti)|(ci)|([aái]t)|([aái]ti)*? Ano: Jdi na 32. Ne: Jdi na 45 (*jet*).
32. 3. osoba indikativu přítomného aktiva končí na *-[eě]*? Ano: Jdi na 33. Ne: Jdi na 45 (*stát, bát se, spát, mít, jíst*).
33. Před infinitivní koncovkou *-(t)|(ti)|(i)* předchází *[szc]*-? Ano: Jdi na 34. Ne: Jdi na 38.
34. Před *-e* ve 3. os. sg. ind. prez. akt. předchází *[čž]*-? Ano: Jdi na 35. Ne: Jdi na 36.
35. Jde o sloveso 1. tř. vzoru *péci*. KONEC
36. Před *-e* ve 3. os. sg. ind. prez. akt. nepředchází *n*-? Ano: Jdi na 37. Ne: Jdi na 46 (*řici/říct, stříci se, bříst(i)*).
37. Jde o sloveso 1. tř. vzoru *nést*. KONEC
38. Před infinitivní koncovkou *-(t)|(ti)* předchází *[řl]i*-? Ano: Jdi na 39. Ne: Jdi na 40.
39. Jde o sloveso 1. tř. vzoru *umřít*. KONEC
40. Před infinitivní koncovkou *-(t)|(ti)* předchází *i*- a zároveň před *-e* ve 3. sg. ind. prez. akt. předchází *[nm]*-? Ano: Jdi na 28. Ne: Jdi na 41.
41. Před infinitivní koncovkou *-(t)|(ti)* předchází *[aá]*-? Ano: Jdi na 42. Ne: Jdi na 45 (*jít, chtít*).
42. Před *-[eě]* ve 3. os. sg. ind. prez. akt. předchází měkká souhláska (*[čn<sup>57</sup>řšž|(šl)]*-)? Ano: Jdi na 43. Ne: Jdi na 44.
43. Jde o sloveso 1. tř. vzoru *mazat*. KONEC
44. Jde o sloveso 1. tř. vzoru *brát*. KONEC
45. Jde o nepravidelná slovesa: *být, jet, jíst, mít, jít, chtít, stát, spát, bát se*. KONEC
46. Jde o slovesa kolísající mezi 1. tř. (*nést, péci*) a 2. tř. (*tisknout*) (*řici, stříci se, břísti, ...*). KONEC

Kolísající slovesa:

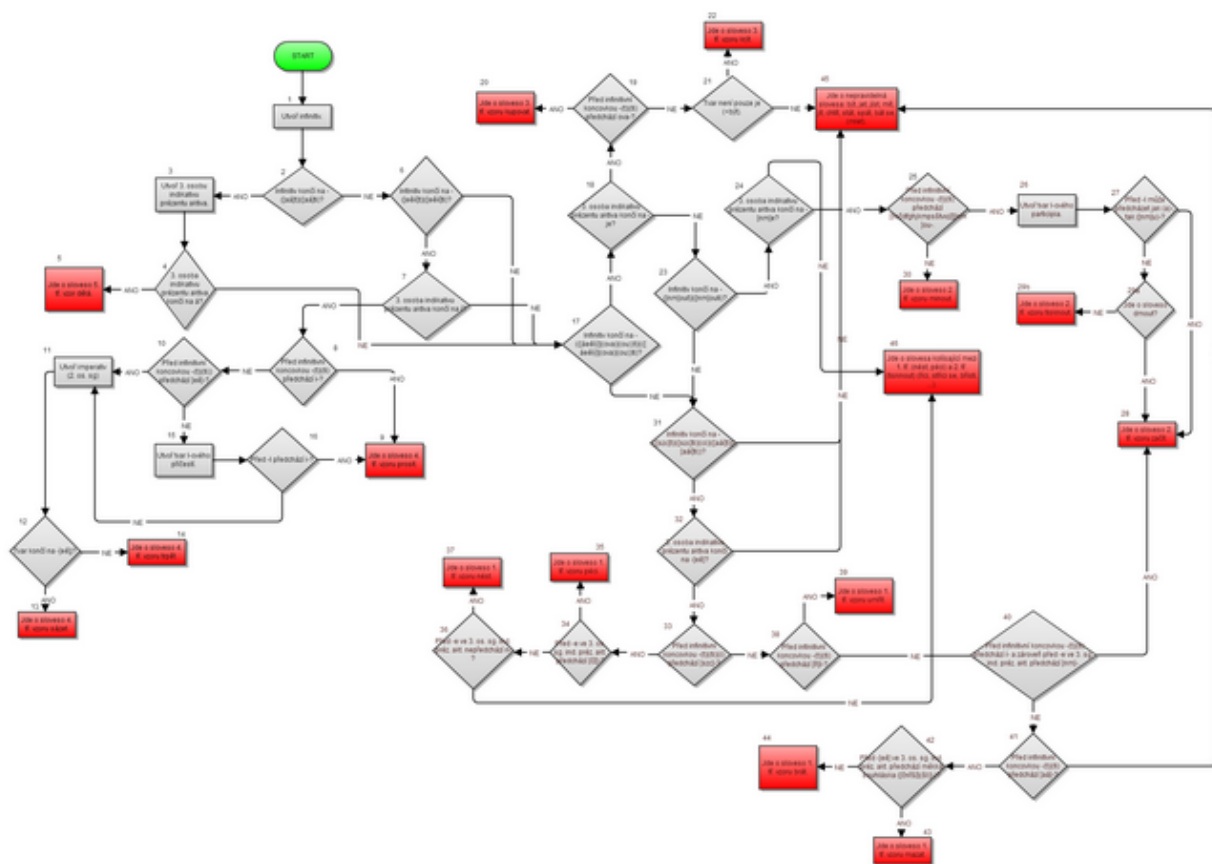
- A. Mezi 1. třídou (pouze vzory *mazat, brát*) a 5. třídou. V mluvnicích přechodný vzor (*kopat*).
- B. Uvnitř 4. třídy mezi vzory *trpět/sázet* (imperativ).
- C. Mezi 1. třídou (pouze vzory *nést, péci*) a 2. třídou (*řici, střechnout se/stříci se, (za)břednout/(za)břísti, ...*).
- D. Během historického vývoje si některá slovesa vytvořila paralelní systém tvarů podle 3. třídy vzoru *kryt* a 2. třídy vzorů *minout, začít*. Dvojice takových sloves mohou/nemusí být synonymní (srov. např. *dout/dmout, ale plout/plynout*).
- E. Mezi 1. třídou (vzor *umřít*) a 4. třídou (vzor *prosit*) srov. tvary slovesa *zřít*.
- F. Mezi 3. třídou (vzor *kryt*) a 4. třídou (vzor *prosit* a *trpět*) srov. např. tvary sloves *dlít, dštít*.<sup>58</sup>

<sup>57</sup> V tom případě se *e* graficky realizuje jako *ě*. Jde, pokud je nám známo, o jediné sloveso *stonat–stůně* a jeho prefigované deriváty.

<sup>58</sup> Srov. např. doklady jako ... *zítra <zetliš> v mých očích ...*, ... *již <tlíte> v zemi, k níž jste láskou věřeli ...*, jakož i chybné značkování tvaru *tlím* (ve skutečnosti jde o překlep ve slově *film*) v korpusu SYN ( ... *nejlepší*



G. Mezi **3. třídou** (vzor *krýt*) a **5. třídou** (vzor *dělat*) srov. např. tvary slovesa *dát* (*se*), *klát*, *plát*.<sup>59</sup>



propagandistický <tlím/tlít/VB-S---IP-AA---I> všech dob ...) versus ... bez vlastní viny pučím, raším, <nesetleji>, ..., ... <tlejou> tam pod vodou, ... <tlejou> a už je jich moc ... Nebo např. doklad z korpusu SYN jako ... anebo tu zášť', <dštím> poněkud kultivovaněji ... versus na internetu doložené ... všichni tady na to <dštijou> síru ...

<sup>59</sup> Sloveso *zdát* (*se*), které etymologicky souvisí se slovesem *dát se* (*děje se*, ...), srov. *udát se*, tvoří synchronně tvary podle vzoru 5. třídy *dělat*. S něčím podobným se setkáváme u slovesa *klát*, které přešlo z 1. třídy (*kolu*, ...) do 5. třídy. V SSJČ i PSJČ se uvádí, že sloveso *proklát* se vyskytuje „jen ve tvarech minulých“, nicméně např. v korpusu SYN jsou doloženy i tvary od kmene přítomného ... a chladnokrevně jej <proklá> oštěpem ... utvořené podle vzoru *dělat* a tamtéž najdeme i doklad tvaru podle vzoru *krýt* ... *Jupiter rozkomihá smrtící blesk a <proklaje> mě* ... Pravidelně se nechová též sloveso *plát*, u něhož jsou řidce doloženy tvary podle vzoru *krýt* (srov. doklad z korpusu SYN ... *tak oko mi <plaje>* ... a ... *zář jiter i západů v jeho krvi nám horoucně <plá>* ...), běžnější jsou ovšem tvary podle vzoru *dělat*. V korpusu SYN je poměr tvarů podle vzoru *krýt*:*dělat* buď 2:80, nebo 3:81. U doložených tvarů 3. pl. ind. prez. akt. nelze totiž rozlišit třídu a vzor (srov. homonymní *pla-j-i* jako *hra-j-i* nebo *pl-aj-i* jako *tk-aj-i*).

## Algoritmus tvoření přechodníků v češtině<sup>60</sup>

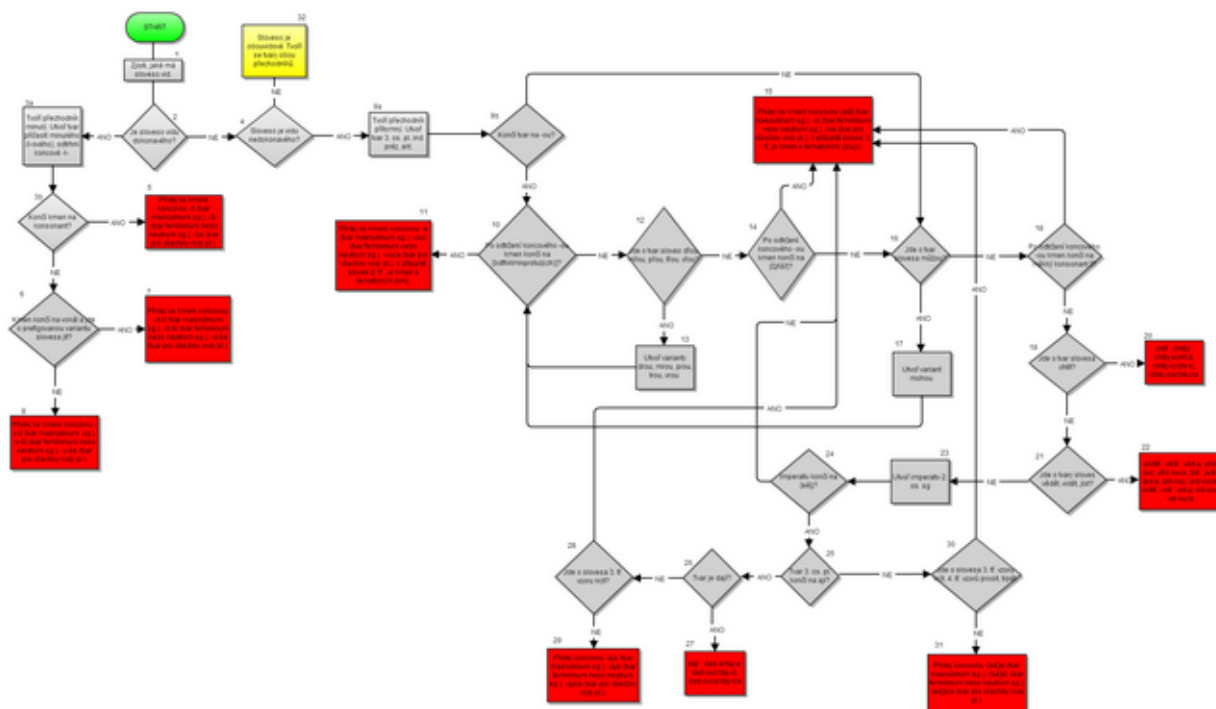
© Klára Osolsobě

1. Zjistí, jaké má sloveso vid. Jdi na 2.
2. Sloveso je vidu dokonavého? Ano: Jdi na 3. Ne: Jdi na 4.
3. Tvoří přechodník minulý. Utvoř tvar příčestí minulého (*l-ového*), odtrhni koncové *-l-*. Po odtržení koncového *-l-* kmen končí na konsonant? Ano: Jdi na 5. Ne: Jdi na 6
4. Sloveso je vidu nedokonavého? Ano: Jdi na 9. Ne: Jdi na 32.
5. Přidej ke kmeni koncovku *-0* (tvar maskulinum sg.), *-ši* (tvar femininum nebo neutrum sg.), *-še* (tvar pro všechny rody pl.). KONEC
6. Po odtržení koncového *-l-* kmen končí na vokál a jde o prefigovanou variantu slovesa *jít*: Ano: Jdi na 7. Ne: Jdi na 8.
7. Přidej ke kmeni koncovku *-d-0* (tvar maskulinum sg.), *-d-ši* (tvar femininum nebo neutrum sg.), *-d-še* (tvar pro všechny rody pl.). KONEC
8. Přidej ke kmeni koncovku *-v-0* (tvar maskulinum sg.), *-v-ši* (tvar femininum nebo neutrum sg.), *-v-še* (tvar pro všechny rody pl.). KONEC
9. Tvoří přechodník přítomný. Utvoř tvar 3. os. pl. ind. prez. akt. Tvar končí na *-ou*? Ano: Jdi na 10. Ne: Jdi na 16.
10. Po odtržení koncového *-ou* kmen končí na *[bdfhklmnpstvz(ch)]*? Ano: Jdi na 11. Ne: 12.
11. Přidej ke kmeni koncovku *-a* (tvar maskulinum sg.), *-ouc* (tvar femininum nebo neutrum sg.), *-ouce* (tvar pro všechny rody pl.). V případě sloves **2. tř.** je kmen s tematickým *[nm]*. KONEC
12. Jde o tvar sloves *dřou, mřou, prou, třou, vřou*. Ano: Jdi na 13. Ne: Jdi na 14.
13. Utvoř varianty *drou, mrou, prou, trou, vrou* a vrať se zpět na 10.
14. Po odtržení koncového *-ou* kmen končí na *[čjňšř]*? Ano: Jdi na 15. Ne: Jdi na 16.
15. Přidej ke kmeni koncovku *-[eě<sup>61</sup>]* (tvar maskulinum sg.), *-íc* (tvar femininum nebo neutrum sg.), *-íce* (tvar pro všechny rody pl.). V případě sloves **3. tř.** je kmen s tematickým *(j)(uj)*. KONEC
16. Jde o tvar slovesa *můžou*. Ano: Jdi na 17. Ne: Jdi na 18.
17. Utvoř variant *mohou* a vrať se zpět na 10.
18. Po odtržení koncového *-ou* kmen končí na měkký konsonant *ž*? Ano: Jdi na 15. Ne: Jdi na 19.
19. Jde o tvar slovesa *chtít*? Ano: Jdi na 20. Ne: Jdi na 21.
20. *chtít : chtějí : chtěj-e/cht-ě, chtěj-íc/cht-íc, chtěj-íce/cht-íce* KONEC
21. Jde o tvary sloves *vědět, vidět, jíst*? Ano: Jdi na 22. Ne: Jdi na 23.
22. *vědět : vědí : věd-a, věd-ouc, věd-ouce, jíst : jedí : jed-a, jed-ouc, jed-ouce, vidět : vidí : vid-a, vid-ouc, vid-ouce*.
23. Utvoř imperativ 2. os. sg a jdi na 24.
24. Imperativ končí na *[eěj]*? Ano: Jdi na 25. Ne: Jdi na 15.
25. Tvar 3. os. pl. končí na *ají*? Ano: Jdi na 26. Ne: Jdi na 30.

<sup>60</sup> O přechodnících se všeobecně tvrdí, že se v češtině příliš neužívají. Nicméně se od nich (vcelku pravidelně) tvoří adjektiva (typ *kupující* nebo *dostavivší se*), která (alespoň v prvním případě) frekventovaná a užívaná jsou. Pravidelně se pak chybí při záměnách adjektiv tvořených od přechodníku přítomného (procesuálních *žehlicí, kryjící, ...*) a adjektiv účelových, tvořených od kmene minulého (typ *žehlicí, krycí, ...*). Ve zvukové realizaci adjektiv odvozených od sloves 4. třídy vzoru *prosít* se totiž stírá rozdíl mezi dlouhým a krátkým *i/í*, což je pro češtinu typické i jinde, např. ve tvarech zájmena *ona – jí/jí*. To má pak vliv na pravopisné chyby při grafickém zaznamenání kvantity u příslušných adjektiv. To jsou mimo jiné důvody, pro něž je třeba tvoření přechodníků věnovat patřičnou pozornost.

<sup>61</sup> Retnice alternují též ve tvarech přechodníku přítomného sloves podle vzoru *prosít* (na rozdíl od tvarů pasivního příčestí a dějových jmen: srov. *nemluvě*, ale *mluvení*), v grafické rovině se po nich tedy píše *ě*, stejně jako se *[d'e], [t'e], [ňe]* graficky realizuje jako *dě, tě, ně*.

26. Tvar je *dají*? Ano: Jdi na 27. Ne: Jdi na 28.  
 27. *dají* : *dad-a/daj-e*, *dad-ouc/daj-íc*, *dad-ouce/daj-ice*. KONEC  
 28. Jde o slovesa **3. tř.** vzoru *krýt*.<sup>62</sup> Ano: Jdi na 15. Ne: Jdi na 29.  
 29. Přidej koncovku *-aje* (tvar maskulinum sg.), *-ajíc* (tvar femininum nebo neutrum sg.), *-ajíce* (tvar pro všechny rody pl.). KONEC  
 30. Jde o slovesa **3. tř.** vzoru *krýt*, **4. tř.** vzorů *prosit*, *trpět*.<sup>63</sup> Ano: Jdi na 15. Ne: Jdi na 31.  
 31. Přidej koncovku *-[eě]je* (tvar maskulinum sg.), *-[eě]jíc* (tvar femininum nebo neutrum sg.), *-[eě]jíce* (tvar pro všechny rody pl.). KONEC  
 32. Sloveso je obouvidové. Tvoří se tvary obou přechodníků.



<sup>62</sup> Všimněme si, že kromě sloves **5. tř.** vzoru *dělat* může ve 3. sg. ind. prez. akt. končit na *ají* a v imperativu výjimečně na *ej* i sloveso **3. tř.** vzoru *krýt* (srv. *tkát/tkají/tkej* ale i *hrát/hrají/hrej* vedle *hraj*).

<sup>63</sup> Všimněme si, že kromě sloves **4. tř.** vzoru *sázet* a **5. tř.** vzoru *dělat* mohou na řetězec *[eě]j* v imperativu končit slovesa **3. tř.** vzoru *krýt* (srv. *sázet/sázej* ale i *pět/pěj*, *spět/spěj*, *smát se/směj se*, *vát/věj*, ...) a **4. tř.** vzoru *prosit* (srv. *sázet/sázej* ale i *vykolejit/vykolej*).