## 6. 3. Různé korpusy a rozdíly v anotačních schématech (tokenizace, lemmatizace, tagging, disambiguace, tagset).

**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 vnětextové 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*[1]**

Ve druhém kroku je každé z takto definovaných jednotek (token) přiřazena interpretace.[2]

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-
```

[1] K historii automatické morfologické analýzy češtiny srov. též Osolsobě 2007[2], Jelínek 2008.

[2] Morfologické analyzátory 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 jsou uloženy potencionální (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.

```
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-
```

Podobně u tvarů *je, má, spíš.*

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. Většině jednotek je přiřazena více než jedna interpretace.

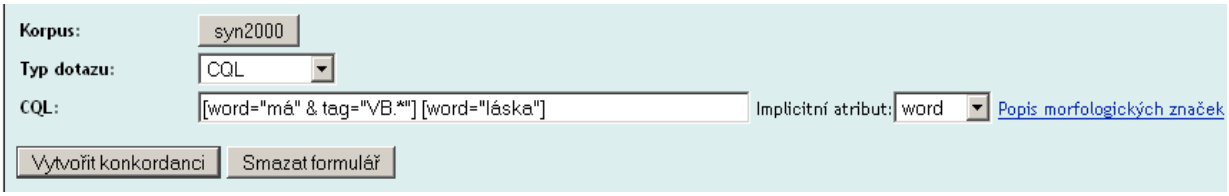
Druhým krokem je desambiguace[3] (disambiguace, zjednoznačnění). Desambiguaci je opět možno provádět buď ručně, nebo pomocí automatických nástrojů. Pokud je automatizována, rozlišujeme různé metody, které se pro zjednoznačnění používají. Rozšířené a užívané jsou metody matematické statistiky. Na opačném 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.

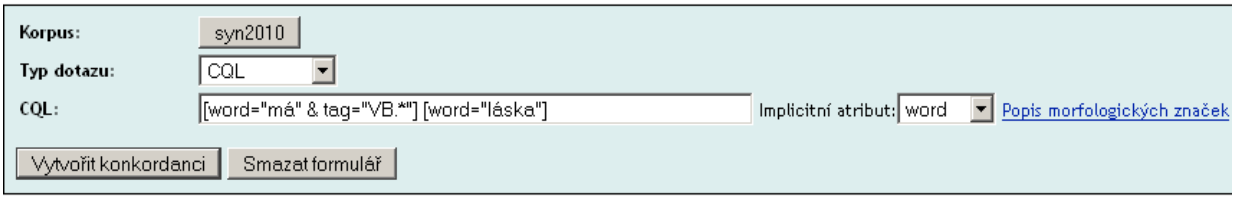
Chybnou desambiguaci vidíme na následujících příkladech z korpusů ČNK, a sice SYN2000 a SYN2010. Vidíme, jak je tvar <*má*> v kontextu <*má*> *láska* ve všech zobrazených vyhledaných dokladech mylně

[3] K problematice desambiguace korpusů ČNK srov. Hajič 2004, Petkevič 2006, Spoustová et al. 2007, Jelínek 2008, Skoumalová 2011.

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 prézentu aktiva (*VB-S---3P-AA---*, resp. *VB-S---3P-AA---I*).[4]








[4] Jak číst morfologickou značku srov. níže.


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 označkovány. 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 všímat 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*** (**p**art **o**f **s**peech), 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.

Podobných jevů je celá řada. Problematické jsou zejména případy slovnědruhových přechodů mezi neohebnými slovními druhy (např. adverbii a částicemi, viz výše tvar *spíš*, též prepozicionalizace *místo, kolem, …*). 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 označkovaná 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 označkovaných v korpusu SYN2005 jako adverbia (D). Uvádíme jejich seznam:

| lemma: | tag: ## | |
|---|---|---|
| Šikmooká**č** | Db---------- | 6 |
| p**ř**eklápě**č** | Db----------8- | 4 |
| šikmooká**č** | Db----------8- | 3 |
| maská**č** | Db----------8- | 2 |
| sváže**č** | Db---------- | 2 |
| cibulá**č** | Db----------8- | 2 |
| spoluspá**č** | Db---------- | 2 |
| skupiná**č** | Db----------8- | 2 |
| spoluspá**č** | Db----------8- | 1 |
| P**ř**eklápě**č** | Db---------- | 1 |
| Cecká**č** | Db---------- | 1 |
| procháze**č** | Db----------8- | 1 |
| šikmooká**č** | Db---------- | 1 |
| Rychlovyvíje**č** | Db---------- | 1 |
| skupiná**č** | Db---------- | 1 |
| hrobník-kopá**č** | Db---------- | 1 |
| sedmispá**č** | Db---------- | 1 |
| doprovaze**č** | Db----------8- | 1 |
| autor-vyprav**ěč** | Db---------- | 1 |
| básník-vyprá**věč** | Db---------- | 1 |
| bodliná**č** | Db---------- | 1 |
| mrkvá**č** | Db---------- | 1 |
| inženýr-sváře**č** | Db---------- | 1 |
| gambá**č** | Db----------8- | 1 |
| **ř**emená**č** | Db---------- | 1 |
| záviná**č** | Db---------- | 1 |
| kuchá**č** | Db----------8- | 1 |
| cecká**č** | Db---------- | 1 |
| on-hrá**č** | Db----------8- | 1 |
| superdří**č** | Db----------8- | 1 |
| zaraže**č** | Db---------- | 1 |
| tutá**č** | Db---------- | 1 |
| bobká**č** | Db---------- | 1 |
| **č**ajpí**č** | Db---------- | 1 |
| nežená**č** | Db---------- | 1 |
| pruhá**č** | Db----------8- | 1 |
| širokoklobouká**č** | Db----------8- | 1 |
| odbíje**č** | Db----------8- | 1 |
| pobíze**č** | Db---------- | 1 |
| propoušt**ěč** | Db----------8- | 1 |
| agent-hrá**č** | Db---------- | 1 |
| doprovaze**č** | Db---------- | 1 |
| pojížd**ěč** | Db----------8- | 1 |
| rozjížd**ěč** | Db----------8- | 1 |
| vegá**č** | Db---------- | 1 |

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

| | | | |
|---|---|---|---|
| naviják pevně přišroubován , odklopte | překlápěč | /překlápěč/Db----------8- | , abyste mohli začít házet |
| . 4 Když jste uvolnili | překlápěč | /překlápěč/Db----------8- | , držte pevně vlasec . |
| obtížnější . 3 S odklopeným | překlápěčem | /překlápěč/NNMS7-----A----- | vytahujte vlasec z navijáku a |
| EE Hleďte , ať je | překlápěč | /překlápěč/NNIS1-----A----- | při provlékání vlasce očky odklopený |
| a pak levou rukou odklopte | překlápěč | /překlápěč/Db----------8- | . 2 Ještě s očima |
| a na vlasec . XX | Překlápěč | /Překlápěč/NNXXX-----A---8- | 1 Držte prut pravou rukou |
| Držte prut pravou rukou . | Překlápěč | /Překlápěč/NNIS1-----A----- | je přiklopen . 2 Zvedněte |
| Druhou rukou začněte odklápět oblouk | překlápěče | /překlápěč/NNFS2-----A----- | . 3 Úplně odklopte oblouk |
| . 3 Úplně odklopte oblouk | překlápěče | /překlápěč/NNFS2-----A----- | . Uslyšíte pravděpodobně klapnutí , |
| zatočte kličkou navijáku abyste překlopili | překlápěč | /překlápěč/Db----------8- | a jste připraveni k chytání |
| středem očka . Nechte oblouk | překlápěče | /překlápěč/NNFS2-----A----- | odklopený , aby se vlasec |
| za překážkou apod . ) | Překlápěč | /Překlápěč/Db---------- | Součást smekacího navijáku ( obvykle |

| | | | |
|---|---|---|---|
| jako oheň . Kolem projeli | svážeči | /svážeč/NNMP1-----A----- | sena , s nimiž se |
| však nestačí , a tak | svážeč | /svážeč/Db---------- | musí brzdit " tlapkami " |
| bun a s touto soupravou | svážeč | /svážeč/Db---------- | vyrazí rychlou jízdou do údolí |

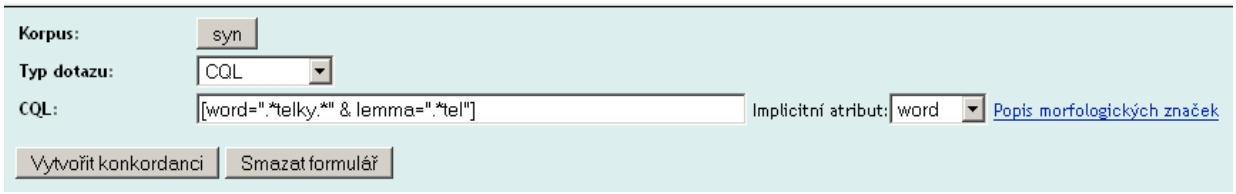
| | | | |
|---|---|---|---|
| se otevřely naprosto tiše . | Šikmookáč | /Šikmookáč/Db---------- | na něho mlčky hleděl . |
| Mluvte tišeji ! " upozornil | šikmookáč | /šikmookáč/Db----------8- | skoro šeptem . No ano |
| mi trochu vody . " | Šikmookáč | /Šikmookáč/Db---------- | přikývl a tiše zamkl . |
| . Státním radou . " | Šikmookáč | /Šikmookáč/Db---------- | soucitně pokýval hlavou a řekl |
| " Za nic . " | Šikmookáč | /Šikmookáč/Db---------- | soucitně přikývl . " To |
| se dveře znovu otevřely . | Šikmookáč | /Šikmookáč/Db---------- | předal Volodinovi z náruče do |
| dveře s rachotem otevřely a | šikmookáč | /šikmookáč/Db----------8- | řekl : " Dejte ruce |
| však ani dvě minuty a | šikmookáč | /šikmookáč/NNFS4-----A----- | znova s rámusem vrazil do |
| , dveře byly zamčené a | šikmookáč | /šikmookáč/Db---------- | ho nechal na pokoji . |
| chvíli se znovu otevřely . | Šikmookáč | /Šikmookáč/Db---------- | podával Volodinovi jehlu a asi |
| rád ! " okřikl ho | šikmookáč | /šikmookáč/Db----------8- | . A tak Volodin prvně |

| | | | |
|---|---|---|---|
| bál . Vzali blembák , | maskáč | /maskáč/NNXXX-----A---8- | , polní , k tomu |
| zkušenost , protože pod " | maskáč | /maskáč/NNIS4-----A----- | " trampa se může schovat |
| na tvých březích sušíval svůj | maskáč | /maskáč/NNIS4-----A----- | zmáčený , když houkal vlak |
| Ropuchu zelenou chrání dokonalý " | maskáč | /maskáč/Db----------8- | " . Program SAPARD Setkal |
| , jestli se mu ošoupal | maskáč | /maskáč/Db----------8- | na loktech . . . |

| | | | |
|---|---|---|---|
| a podáváme s plackami . | Cibuláč | /Cibuláč/NNFS1-----A----- | s vejci 500 g libového |
| jsou výborné k zajíci " | cibuláči | /cibuláč/NNMP1-----A----- | " nebo samotné se zelným |
| Podáváme s knedlíky . Králík | cibuláč | /cibuláč/Db----------8- | Rozkrájenou cibuli osmažíme na rozpuštěné |
| nebo houskový knedlík . Zajíc | cibuláč | /cibuláč/Db----------8- | Zajíce protáhneme špekem , ale |

| | | | |
|---|---|---|---|
| Jsem John Werner , váš | spoluspáč | /spoluspáč/Db---------- | , uvidíte mě jen večer |
| ale s tím si již | spoluspáč | /spoluspáč/Db---------- | uměl poradit : štípl vždy |
| k chroptícímu panu Aronovi , | spoluspáč | /spoluspáč/Db----------8- | se jen obrátil na boku |
| budek , nevysněná úzkost , | spoluspáč | /spoluspáč/NNXXX-----A---8- | strach . Vymýšlím nějaké účastné |

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 generovaní pravidelných derivací při výstavbě slovníku automatického morfologického analyzátoru.




| | | | |
|---|---|---|---|
| vybral slečně Božence - hlavní | zakladatelkyni | /zakladatel/NNFS4-----A----- | a okrase Slovanských zábav - |
| neboť ony jsou nejprvnější jich | učitelkyně | /učitel/NNFS1-----A----- | . . . " Taková |
| . Nedostatek toalet hodnotí českobudějovická | zastupitelkyně | /zastupitel/NNFS1-----A----- | za stranu Důchodci za životní |
| zklamání nad jednáním jedné paní | zastupitelkyně | /zastupitel/NNFS2-----A----- | za ODS zvítězil a já |
| , poprosila jsem manžela paní | zastupitelkyně | /zastupitel/NNFS2-----A----- | , aby mi po svém |
| rozuzlení sporu O mandátu karlovarské | zastupitelkyně | /zastupitel/NNFS2-----A----- | Moniky Makkiehové ( ODS ) |
| . Jedinou kandidátkou byla krajská | zastupitelkyně | /zastupitel/NNFS1-----A----- | a podnikatelka Hana Nováčková-Zelenková , |
| MÍSTEK - Rezignaci na post | zastupitelkyně | /zastupitel/NNFS2-----A----- | Frýdku - Místku oznámila ve |
| hlasy , " sděluje pražská | zastupitelkyně | /zastupitel/NNFS1-----A----- | Olga Sedláčková ( SNK ED |
| jak se bude ve funkci | zastupitelkyně | /zastupitel/NNFS2-----A----- | pro devátou městskou část 9 |
| sedmdesáti let dopravu zdarma . | Zastupitelkyně | /zastupitel/NNFS1-----A----- | Erika Sedláčková ( KSČM ) |
| pověřilo náměstka Petra Keřku a | zastupitelkyni | /zastupitel/NNFS4-----A----- | Moniku Makkiehovou . Ti mají |
| Otázka pro Moniku Makkiehovou , | zastupitelkyni | /zastupitel/NNFS4-----A----- | ( ODS ) Jako náměstkyně |
| dobový průvod Horní Jiřetín - | Zastupitelkyně | /zastupitel/NNFS1-----A----- | Helena Dernerová představila ve středu |
| turnaj fotbalistů , " plánovala | zastupitelkyně | /zastupitel/NNFP4-----A----- | . Podzim by byl zahájen |
| . " Na návrh jedné | zastupitelkyně | /zastupitel/NNFS2-----A----- | bylo upuštěno od veřejného zastřelení |

V praxi se jednalo o vybrané typy paradigmatických derivací jako podstatná jména slovesná tvořená od základů shodných s pasivním příčestím, adjektiva tvořená od těchto základů, adjektiva tvořená od přechodníků, tvary II. a III. stupně adjektiv a adverbií, 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 adverbií 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 zjistitelnou 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***

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 slovotvorných 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 adverbií dle +/- derivace komparativu a superlativu (koresponduje s pozicí 10 Stupeň), f) různé.

Pro lepší přehled uvedeme tabulky pro a) – e[5]).

[5] Podrobné popisy vztahů gramatických značek v různých tagsetech používaných pro lemmatizaci a značkování českých korpusů srov. např. Pořízka – Schäfer 2009, Osolsobě 2007[1]). Vylepšená verze open-source webového

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. tv. (indik.) |
| Vi.* | sloveso | imperativ |
| Vp.* | sloveso | l-ové příčestí (vč. tvarů s *-s*) |
| Vq.* | sloveso | l-ové příčestí (vč. tvarů na *-ť*) |
| Vs.* | sloveso | pasivní příčestí (vč. tvarů s *-s*) |
| Ve.* | sloveso | přechodník přítomný |
| Vm.* | sloveso | přechodník minulý |
| Vc.* | sloveso | kondicionál sl. být (*bych, ...*) |
| J,.* | spojka | spojky podřadicí vč. *aby, ... kdyby, ...* |

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

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

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.

| 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 „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 | *naň* | spřežka předložka+osobní zájmeno *on* |
| P1 | *jehož* | vztažné zájmeno *jehož* |
| P4 | *jaký, který* | tázací zájmeno *čí, čípak, jaký, jakýpak, jakýž, jakýže, který, kterýpak, kterýž, ...* |
| P5 | *něj* | osobní zájmeno *on* tvary po předložce (*n*-) |
| P6 | *sebe* | zvratné zájmeno tvary *sebe, sobě, sebou* |
| P7 | *se, si* | zvratné zájmeno tvary *se, si, ses, sis* |
| P8 | *svůj* | přivlastňovací zvratné zájmeno *svůj* |
| P9 | *něhož* | vztažné zájmeno *jehož* tvary po předložce (*n*-) |
| PD | *tento* | ukazovací zájmena *ten, tento, takový, tenhle, onen, týž, tentýž, takovýto, takovýhle, tenhleten, toť, tamten, taký, tamhleten,* |

| | | |
|---|---|---|
| | | *tadyten, tuhleten* |
| PE | *což* | vztažné zájmeno *což* |
| PH | *mě* | krátké (příklonné) tvary osobních zájmen *mi, mě, ti, tě, ji, je, ...* |
| PJ | *jenž* | vztažné zájmeno *jenž* |
| PK | *kdo* | vztažné/tázací zájmeno *kdo, kdopak, kdožpak, kdož, kdos* |
| PL | *všechen* | zájmena vymezovací (limitativa) *všechno, všecek, sám, samý, veškerý* |
| PP | *ty* | osobní zájmena *já (my), ty (vy), on,* tvar *tys* |
| PQ | *co* | vztažné/tázací zájmeno *co, copak, cožpak, cos, což* |
| PS | *můj* | přivlastňovací osobní zájmena *můj, tvůj, jeho, náš, váš* |
| PW | *nic* | záporná zájmena *nic, žádný, nikdo, pranic, nijaký, pražádný, nižádný* |
| PY | *oč* | spřežka vztažné/tázací zájmeno předložka+*č* (*oč, nač, zač, več, ...*) |
| PZ | *nějaký, něco* | neurčitá zájmena *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*, kdeco, 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, všelikerý* |

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

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

| POS&SUBPOS | tvary – příklady | popis |
|---|---|---|
| C= | *1* | arabské číslice |
| C} | *XIV* | římské číslice |

| | | |
|---|---|---|
| Ca | *mnoho* | tvary „číslovky“ *mnoh-o,-a, ...* |
| Cd | *čtverý* | druhové číslovky *dvojí, obojí, trojí,* a další tvořené sufixem -*erý* |
| Ch | *jedny* | druhová číslovka *jedny* |
| Cj | *čtvero* | úhrnné číslovky *dvé, obé, tré* a další tvořené sufixem -*ero* |
| Ck | *čtvery* | souborové číslovky *dvoje, oboje, troje* a další tvořené sufixem -*ery* |
| Cl | *tři* | základní číslovky *jeden, dva, oba, tři, čtyři* |
| Cn | *pět* | základní číslovky *pět* a výše |
| Co | *tolikrát* | číslovka zájmenná ukazovací násobná *tolikrát* |
| Cr | *druhý* | číslovky řadové |
| Cu | *kolikrát* | číslovka zájmenná tázací násobná *kolikrát* |
| Cv | *sedmkrát* | číslovky určité násobné .*-krát* |
| Cw | *nejeden* | nejeden |
| Cy | *desetina* | číslovky dílové vyjadřující určitý počet .*-ina* |
| Cz | *kolikátý* | číslovka zájmenná tázací/vztažná řadová *kolikátý* |

e) **Detailní určení slovního druhu – klasifikace adverbií**

| POS | Detailní určení slovního druhu (SUBPOS) |
|---|---|
| D | db |

| POS&SUBPOS | tvary – příklady | popis |
|---|---|---|
| Db | *nahoru* | všechna příslovce, která nelze stupňovat |
| Dg | *rychle* | příslovce, která lze stupňovat |

Literatura ke studiu (odkazy):

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.

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.

Hlaváčová, J.: Morphological Guesser of Czech Words. In: Matoušek, V. (ed.), *Text, Speech and Dialogue*, Berlin : Springer-Verlag, 2001, s. 70–75.

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.

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.: 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[1], s. 201–218.

Osolsobě, K.: Syntetické futurum v češtině – gramatiky, slovníky, korpusy. In: *Přednášky a besedy z XL. běhu LŠSS*, Brno : FF MU, 2007[3], 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[1], s. 407–416.
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.
Sedláček, R.: *Morphematic analyser for Czech*. Brno : FI MU, 2004. (Disertační práce.)
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.

**Elektronické zdroje:**

*Korpus SYN2010: Český národní korpus – SYN2010.* Ústav Českého národního korpusu FF UK, Praha 2010. Cit. 20. 03. 2011, dostupný z WWW: <http://www.korpus.cz>.
*Korpus SYN2009PUB: Český národní korpus – SYN2009PUB.* Ústav Českého národního korpusu FF UK, Praha 2010. Cit. 20. 03. 2011, dostupný z WWW: <http://www.korpus.cz>.
*Korpus SYN2006PUB: Český národní korpus – SYN2006PUB.* Ústav Českého národního korpusu FF UK, Praha 2006. Cit. 20. 03. 2011, dostupný z WWW: <http://www.korpus.cz>.
*Korpus SYN2005: Český národní korpus – SYN2005.* Ústav Českého národního korpusu FF UK, Praha 2005. Cit. 20. 03. 2011, dostupný z WWW: <http://www.korpus.cz>.
*Korpus SYN2000: Český národní korpus – SYN2000.* Ústav Českého národního korpusu FF UK, Praha 2000. Cit. 20. 03. 2011, dostupný z WWW: <http://www.korpus.cz>.
*Korpus SYN: Český národní korpus – SYN.* Ústav Českého národního korpusu FF UK, Praha. Cit. 20. 03. 2011, 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/>.
*Internetový vyhledávač Google* [online]. 2011. Cit. 20. 03. 2011, dostupný z: <http://www.google.com/>.
*Internetová jazyková příručka* [online]. 2011. Cit. 20. 03. 2011, dostupná z: <http://prirucka.ujc.cas.cz>.
DEB Dict – *Obecný prohlížeč slovníků* [online]. 2011. Cit. 20. 03. 2011, dostupný z: <chrome://debdict/content/debdict.xul>