ČESKÁ MORFOLOGIE A KORPUSY

KLÁRA **OSOLSOBĚ**

KAROLINUM

Česká morfologie a korpusy

doc. PhDr. Klára Osolsobě, Ph.D.

———

Recenzovali:
PhDr. Josef Šimandl, Ph.D.
doc. RNDr. Vladimír Petkevič, CSc.

Vydala Univerzita Karlova v Praze
Nakladatelství Karolinum
www.cupress.cuni.cz
Grafická úprava Jan Šerých
Sazba DTP Nakladatelství Karolinum
Vydání první

ISBN 978-80-246-2562-1
ISBN 978-80-246-2593-5 (online : pdf)

Univerzita Karlova v Praze
Nakladatelství Karolinum 2014

http://www.cupress.cuni.cz

OBSAH

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 bohemistiky, řadu návodů, jak lze vyhledávat v jazykových korpusech relevantní data, jejichž prostřednictvím mohou pozorovat svou mateřštinu, a také řadu postupů, jimiž 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 vděčni 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 jednodušími i složitějšími otázkami z české morfologie a slovotvorby.

Při práci s korpusy se pro přístup k elektronicky uloženým jazykovým datům používají programy, tzv. korpusové manažery, které umožňují data vyhledávat, zobrazovat, třídit, 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.

V kapitole **Jazykové korpusy z hlediska lemmatizace a morfologického značkování** podáváme komentovaný přehled značek použitých pro značkování korpusů řady SYN (synchronní korpusy psané češtiny budované v ÚČNK[1]). K podrobnějšímu proniknutí do tematiky, které přesahuje rámec potřebný k výkladu v této knize, odkazujeme k další literatuře.

Jádro knihy tvoří kapitola **Korpusová cvičení k české morfologii**, která zahrnuje oddíly týkající se jednotlivých slovních druhů, oddíl s přesahem do slovotvorby 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 slovotvorby.
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 vyplynuvších 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 korpu-

1 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.

sové lingvistiky doporučujeme z kratších česky psaných příruček *Český národní korpus – úvod a příručka uživatele* (Kocek – 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). Aktuální informace lze čerpat z příručky 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 přítulné 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 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[2]

Ve druhém kroku je každé z takto definovaných jednotek (token) přiřazena interpretace.[3]

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-
```

2 K historii automatické morfologické analýzy češtiny srov. též Osolsobě 2007b, Jelínek 2008.

3 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 potenciální (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 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. Většině jednotek je přiřazena více než jedna interpretace.

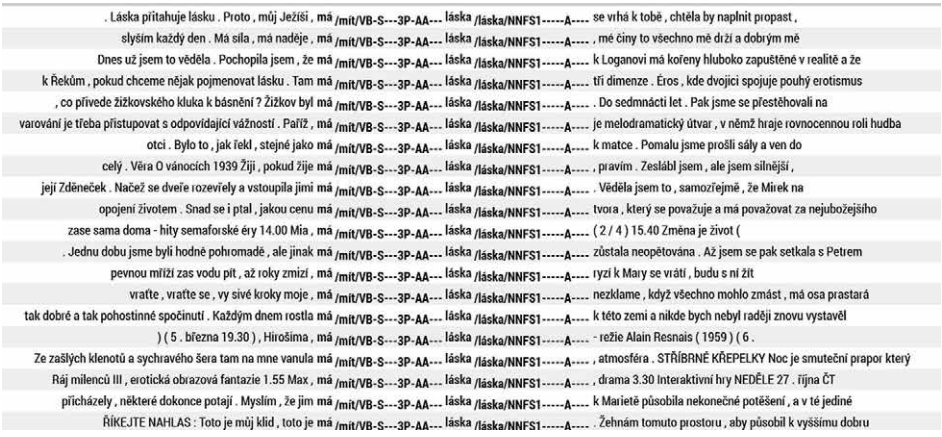
Druhým krokem je desambiguace[4] (disambiguace, zjednoznačnění). Desambiguaci je opět možno provádět buď ručně, nebo pomocí automatických nástrojů. Pokud je automatizována, rozlišujeme různé metody, které se pro zjednoznačnění používají.

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

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 prézentu aktiva (*VB-S---3P-AA---*, resp. *VB-S---3P-AA---I*).[5]



5 Jak číst morfologickou značku srov. níže.

. Leč emoce mají neméně výraznou odvrácenou tvář . Jako má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- svůj rub v nenávisti , v kterou se snadno zvrtne
poklidné krajině . Žiji zde sám v nebeském vzduchu , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- , má píseň blaží mě . I zde se jedná
eso . Je jím v pořadí čtvrtý koncert Provence - má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- , kvůli němuž přijede z Prahy povědomý pár Milan Dvořák
něco nezajímavého a možná i nepříjemného , za což nás má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- kompenzovat . Jenže ona to za nás neudělá . Udělá
. . . - Nenápadně , do ří ídkých vánků má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- vyvát jak dým ; v hebkých vlnách , polehounku a
jako by byla obsažena v tom svém stížení . A má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- , má touha , byly v té váze ; z
svém věku snad již víš , slečinko , že ačkoli má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- nejedno jméno , nakonec vždy muže se ženou vzájemně přitahuje
(zasměje se) . * Jste čerstvě zamilovaná , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- vliv na váš šatník ? Ano , jsem opravdu hodně
se podíváte , narazíte na psa nebo kočku . Odtud má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- ke zvířatům . Kam se dají nacpat , jsou knihy
ano . David , přemítala horečně Claudia . David , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- . David , který mne naučil milovat lásku , skutečnou
řekl po svém , že věcí dechu . Tak jako má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- k bližnímu a Bohu , smíří - li se oba
Prostějovských dnů hudby . Komponovaný pořad Provence - má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- klavíristy Milana Dvořáka a zpěvačky Evy Kríz - Lifkové začíná
pobřeží (1958) , významná Resnaisova díla Hirošima , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- (1958) a Loni v Marienbadu (1961)
v PŘEPADENÍ 13 . OKRSKU , zjistíme , jakou cenu má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- v romantické komedii DOKONALÁ PARTIE , pokusíme se zachránit největší
jsem mluvit , tuším o filmu Alaina Resnaise Hirošima , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- . Později , když už jsme se spřátelili , jedna
v kinu Mír stane dokument z roku 1992 Japonsko , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- . Do tajů východoasijské kultury návštěvníky zasvětí japanolog Martin Vačkář
sebezdokonalování ! Může vás spasit , může vám pomoci - má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- ! Ona je spásná ! Moje láska ! OLGA :
rám ? Řekni mi , svatební noci , zda mě má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- v své moci , nebo jen touha a klam ?
nemohu ani usilovat . Má druhá a drahá polovička , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- , je dána stále stejnou představou , nebo podobou ,
doby se Karel stal pro můj život tak důležitým a má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- k němu tak silným citem , jaký jsem nikdy k
s výjimkou Lisbeth Cookeové - kterou často nazývá Lisbeth , má /mít/VB-S---3P-AA---I láska /láska/NNFS1-----A----- . " " Takže žádný záznam o jiné ženě ?
dobrý život . Pak mi to jednou večer Betty , má /mít/VB-S---3P-AA---I láska /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 (např. 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 je 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; 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 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 jsou založeny na statistických metodách) 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:	##
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
bodliná̋č	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
superdříč	Db----------8-	1
zaražeč	Db----------	1
tutáč	Db----------	1
bobkáč	Db----------	1
čajpíč	Db----------	1
neženáč	Db----------	1
pruháč	Db----------8-	1
širokoklobouká	Db----------8-	1
odbíječ	Db----------8-	1
pobízeč	Db----------	1
propouštěč	Db----------8-	1
agent-hráč	Db----------	1
doprovazeč	Db----------	1
pojížděč	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	svážeč	/svážeč/Db----------	musí brzdit " tlapkami " neboli postranními dřevěnými pákami .
přes něj se položí další bun a s touto soupravou	svážeč	/svážeč/Db----------	vyrazí rychlou jízdou do údolí . Na mírných , pomalejších
vystřelovala jako bitva a pálila jako oheň . Kolem projeli	svážeči	/svážeč/NNMP1-----A-----	sena , s nimiž se večer scházela sestra (proti

Jenže v tu chvíli se dveře s rachotem otevřely a	šikmookáč	/šikmookáč/Db----------8-	řekl : " Dejte ruce nahoru ! " " Jak
. místnosti ? " " Mluvte tišeji ! " upozornil	šikmookáč	/šikmookáč/Db----------8-	skoro šeptem . No ano , vždyť za jeho zády
" " Berte a buďte rád ! " okřikl ho	šikmookáč	/šikmookáč/Db----------8-	. A tak Volodin prvně v životě začal šít .
okamžitě tvrdě usnul . Netrvalo však ani dvě minuty a	šikmookáč	/šikmookáč/NNFS4-----A-----	znova s rámusem vrazil do dveří a zvolal : "
cementovou podlahu velké studené umývárny , dveře byly zamčené a	šikmookáč	/šikmookáč/Db----------	ho nechal na pokoji . Snad je to taky člověk

pocestný , že bigošů je bál . Vzali blembák ,	maskáč	/maskáč/NNXXX-----A---8-	, polní , k tomu kanady a kvér , vyrazili
to Němec , nebo Rus , jestli se mu ošoupal	maskáč	/maskáč/Db----------8-	na loktech . . . A teď jde o to
štreku jako nic , já na tvých březích sušíval svůj	maskáč	/maskáč/NNIS4-----A-----	zmáčený , když houkal vlak od Dolních Kralovic . Nad
mnohý z nás udělal nedobrou zkušenost , protože pod "	maskáč	/maskáč/NNIS4-----A-----	" trampa se může schovat leckdo . Nedejme se však
obecná má krásné oči . Ropuchu zelenou chrání dokonalý "	maskáč	/maskáč/Db----------8-	" . Program SAPARD Setkal jsem se s pojmem "

necháme ještě chvíli odležet . Podáváme s knedlíky . Králík **cibuláč** /cibuláč/Db----------8- Rozkrájenou cibuli osmažíme na rozpuštěné slanině , na ni vložíme
10 minut vaříme . Knedlíčky jsou výborné k zajíci " **cibuláči** /cibuláč/NNMP1-----A----- " nebo samotné se zelným salátem či špenátem . Špekový
. Podáváme k ní bramborový nebo houskový knedlík . Zajíc **cibuláč** /cibuláč/Db----------8- Zajíce protáhneme špekem , ale nesolíme , pak jej husí

vstoje v unikajícím teple telefonních budek , nevysněná úzkost , **spoluspáč** /spoluspáč/NNXXX-----A---8- strach . Vymýšlím nějaké účastné slovo , lucerničku v pralese
, vzbuzovaly důvěru . " Jsem John Werner , váš **spoluspáč** /spoluspáč/Db---------- , uvidíte mě jen večer . Pěkně vás vítám !
zkoušely nosnost jeho žeber , ale s tím si již **spoluspáč** /spoluspáč/Db---------- uměl poradit : štípl vždy (podle přátelské dohody z
noci , když ho volali k chroptícímu panu Aronovi , **spoluspáč** /spoluspáč/Db----------8- se jen obrátil na boku , ale oči neotevřel .

Nedostatek toalet hodnotí českobudějovická **zastupitelkyně** /zastupitel/NNFS1-----A---- za stranu Důchodci za životní jistoty

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 doklad 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.

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.