KORPUS SOUKROMÉ KORESPONDENCE (KSK) Z HLEDISKA MORFOLOGICKÉHO ZNAČKOVÁNÍ

Klára Osolsobě

(Studie byla připravena v rámci projektu GA ČR č. 405/03/0248)

Cílem tohoto článku je poukázat na problémy morfologického značkování jazykových korpusů s vysokým procentem výskytu jazykově nestandardních jevů a ukázat možnosti jejich řešení na základě zkušeností získaných při morfologickém značkování Korpusu soukromé korespondence (KSK; Hladká a kol., 2005). Budeme se zabývat nástroji automatické morfologické analýzy z hlediska jejich použitelnosti pro anotace nespisovných či jiných nestandardních jevů, které se vyskytly v KSK, dále problémy ruční disambiguace automaticky anotovaného korpusu a následného doznačkování slovních tvarů, které automatický analyzátor buď neoznačil, nebo označkoval nesprávně, a které byly tudíž při ruční disambiguaci z nejrůznějších důvodů ponechány stranou.

1. Úvod

Morfologické značkování korpusů psaného jazyka je běžné v oblasti budování obecných korpusů, které mají sloužit lingvistickému výzkumu. Pro potřeby značkování se vytvářejí automatické nástroje – morfologické analyzátory, tedy počítačové programy provádějící segmentaci textu na jednotky odpovídající slovním tvarům, jimž pak přiřazují informace ve tvaru morfologických značek (tzv. tagů). Výsledkem automatické morfologické analýzy je lemmatizace (slovnímu tvaru v textu je automaticky přiřazen příslušný základní tvar – lemma) a morfologická anotace (danému slovnímu tvaru v textu jsou automaticky přiřazeny slovnědruhové a morfologické interpretace - tagy). Automatické morfologické analýze předchází tokenizace, tj. segmentace textu na jednotky, které v ideálním případě odpovídají textovým slovům, v podstatě jde však o zjednodušení lingvistického přístupu v tom smyslu, že slovní tvar se chápe formálně jako řetězec znaků mezi mezerami, popř. jinými oddělovači, jimiž mohou být např. interpunkční znaménka. Automatická morfologická analýza přiřazuje jednotkám textu (textovým slovům, token) všechny výše zmíněné kontextově nezávislé interpretace (lemmata a tagy). Morfologická analýza je obecně nejednoznačná. Příčinou nejednoznačnosti je v jazycích typu češtiny vysoká míra homonymie způsobená tvarovou

homonymií uvnitř paradigmatu jednoho systémového slova, homonymií úplnou nebo částečnou (překrytí všech, či několika tvarů) dvou různých lexikálních jednotek, homonymií vyvolanou funkčními slovnědruhovými transpozicemi mezi jednotlivými (především neohebnými) slovními druhy. Míra koncovkové homonymie uvnitř paradigmatu jednoho slova podstatně vzrůstá, jsou-li do automatické analýzy zařazeny možné substandardní tvary.

2. Automatická analýza formální morfologie spisovného jazyka a možnosti její modifikace

2.1. Morfologický analyzátor ajka

Ajka (http://nlp.fi.muni.cz/projekty/ajka/) je automatický morfologicky analyzátor používaný na FI MU a FF MU primárně pro automatické morfologické tagování korpusů psaného jazyka. Analyzátor ***ajka*** vychází z algoritmického popisu české formální morfologie (Osolsobě, 1996) a ze zkušeností s tvorbou a anotacemi korpusů psaného jazyka na FI MU. Na základě analýzy materiálu dat brněnského mluveného korpusu (BMK) byla vytvořena varianta rozšiřující algoritmický popis české formální morfologie o variantní koncovky vyskytující se v mluvených korpusech (srv. více Hlaváčková, 1998, 2002).

2.2. Úprava automatického analyzátoru

Korpus soukromé korespondence (KSK) vznikl v rámci grantového projektu **Současná soukromá korespondence. Vytvoření databáze a zpracování vybraných jevů z pohledu lexikologicko-lexikografického a dialektologického**. KSK byl pracovně rozdělen do tří subkorpusů**:** KSKdopisy (2 000 dopisů)**,** KSKe-maily (1 000 e-mailů)**,** KSKdopisy1 (1 000 dopisů identických s první tisícovkou dopisů v KSKdopisy)**.** Všechny tři jsou lemmatizovány modifikovanou verzí analyzátoru ***ajka***. Poslední z nich byl i morfologicky označkován a disambiguován.

Korpusy soukromé korespondence zahrnují jazyková data v písemné formě, což má za následek eliminaci řady problémů spojených s vytvářením (budováním) korpusů mluveného jazyka (především tvorby a následujícího dodržování pravidel přepisu nahrávek). Na druhé straně je příznačné, že jde o texty s velkým množstvím lingvistických jevů, které se běžně vyskytují v mluveném jazyce (více srv. např. Hladká, Šindlerová, 2004, Hladká, 2005). Pro potřeby značkování korpusů soukromé korespondence s vysokou mírou výskytu jevů

z hlediska spisovného jazyka substandardních bylo tudíž možno zčásti navázat na zkušenosti se zpracováním morfologie mluveného jazyka.

Hlavním cílem bylo vytvořit verzi automatického morfologického analyzátoru, který by „uměl“ interpretovat maximum substandardních jevů vyskytujících se v korpusech soukromé korespondence (KSK).

Prvním krokem při návrhu upravené verze automatického morfologického analyzátoru bylo označkování KSK morfologickým analyzátorem určeným pro spisovnou češtinu. Po vytřídění slov, která zůstala bez morfologické značky, byl vytvořen jejich frekvenční seznam. Z analýzy tohoto seznamu vyplynulo, které substandardní jevy a s jakou frekvencí se v KSK vyskytují. Jejich klasifikace se stala vodítkem pro navržené změny analyzátoru ***ajka***.

2.2.1 Substandardní jevy navržené pro automatickou identifikaci

Přednostně byly zpracovány frekventované jevy (pokrývající analýzu slovních tvarů, které se v KSK vyskytly pětkrát a více). Lze je rozdělit do následujících skupin:

- tvary lišící se od spisovné normy koncovkou (hláskoslovné a oblastní varianty – *blbý – blbej*, *prosím – prosim, kluky – klukama, námi – nama, chci – chcu, ...*);

- **hláskoslovné varianty spisovných kmenů** (*být – bejt, mít – mět, půjdu – pudu, vezmu – vemu, prý – prej, ...*);

- **varianty s protetickým *v-*** *(okno – vokno, od – vod, ...*);

- **nespisovné lexikální jednotky** (*maturák, anglina, jazykovka, slohovka, zabíračka, vejška, kámoš, kámoška, pařba, vzrůšo, ...*).

Do analyzátoru byly doplněny nespisovné koncovky pojící se se spisovnými základy slov, které jsou obsaženy v morfologické databázi i_par (Veber, 2003). Morfologickým značkám označujícím tyto tvary přibyl atribut signalizující substandardnost (více srv. Hlaváčková, Sedláček, 2004). Do slovníku kmenů byly zařazeny frekventované nespisovné lexikální jednotky.

3. Lemmatizace

Lemmatizací se v oblasti značkování jazykových korpusů míní první stupeň morfologického značkování, a sice přiřazení základního tvaru (lemmatu) slovnímu tvaru. Definice slovního tvaru je technicky zúžena. Při automatické lemmatizaci je slovní tvar definován jako řetězec znaků (písmen dané abecedy) ohraničený znaky - většinou mezerami a/nebo interpunkčními znaky. Nepočítá se s víceslovnými jednotkami na straně jedné a s některými typy

pravopisných spřežek na straně druhé. Tyto případy se řeší různě. U víceslovných jednotek se lemma přiřazuje každé jednotce, takže např. lemmatizace složených slovesných tvarů, tvarů s volným morfémem *se*, víceslovných příslovcí, předložek, spojek, částic a citoslovcí nehledě na frazémy a idiomy je do jisté míry zjednodušena. Různé morfologické analyzátory se snaží také vyrovnat s lemmatizací zájmenných příslovcí (*nač, zač, ... oň, proň, ...*) a spojek, jejichž genetická vazba na kondicionálovou „částici" způsobuje, že vyjadřují osobu přísudkového slovesa (*aby, kdyby, ...*).

3.1 Lemmatizace substandardních jevů v KSK

Jak bylo řečeno výše, v textech KSK se vyskytly nejrůznější typy substandardních jevů. Otázku lemmatizace těchto případů bylo třeba řešit na prvním místě. Na základě analýzy korpusových dat jsme stanovili dvě skupiny:

A. Tvar je substandardní variantou standardní jednotky, a má tudíž lemma podle standardní varianty.

Jde o tyto případy:

1) varianty se substandardní koncovkou
př.: tvar ***klukama*** má lemma ***kluk***, tvar ***ject*** má lemma ***jet***, tvar ***žijó*** má lemma ***žít***, tvar ***bavěj*** má lemma ***bavit,*** tvar ***dobrej*** má lemma ***dobrý,*** tvar ***kterejma*** má lemma ***který***;
2) varianty se substandardní kmenotvornou příponou (u sloves)
př.: tvar ***myslim*** má lemma ***myslet,*** tvar ***vidim*** má lemma ***vidět***;
3) substandardní tvary zájmen
př.: tvar ***ja*** má lemma ***já,*** tvar ***nama*** má lemma ***my,*** tvar ***teho*** má lemma ***ten***;
4) substandardní tvary slovesa ***být***
př.: tvary ***su, seš, sou, sem, sme, bejt, ...*** mají lemma ***být***;
5) substandardní tvary kondicionálů ***by, ...*** *s* variantami ***aby, .., kdyby, ...***
př.: tvar ***bysme*** má lemma ***by***, tvar ***byjsme*** má lemma ***by***;
6) tvary s protetickým *v-*
př. tvar ***vobšťastňovat*** má lemma ***obšťastňovat***, předložka ***vod*** má lemma ***od***;
7) varianty se substandandardními pravopisnými jevy (chybami)
př.: tvar ***být*** *(A nepiš už taťkovi že mě nemá* ***být*** *po hlavě a že budu blbej ...)* má lemma ***bít***.

B. Tvar není substandardní variantou standardní jednotky (slovotvorně substandardní tvar, nespisovná lexikální jednotka).

Lemmatem slovotvorně substandardních tvarů a nespisovných lexikálních jednotek je pravidelně vytvořený tvar nominativu nebo infinitivu.
Například: tvar ***dopendluju*** má lemma ***dopendlovat***, tvar ***foťáku*** má lemma ***foťák***, tvar ***kámoškou*** má lemma ***kámoška,*** tvar ***strejdou*** má lemma ***strejda,*** tvar ***bráchem*** má lemma ***brácha***, tvar ***ahojky*** má lemma ***ahojky*** atd.
KSK (tedy KSKdopisy**,** KSKe-maily**,** KSKdopisy1) byl lemmatizován upravenou verzí morfologického analyzátoru ***ajka***. Tvary rozpoznané analyzátorem mají lemma (lemmata), která nabízí analyzátor, tvary nerozpoznané mají jakožto lemma uveden tvar sám. Pouze KSKdopisy1 byl disambiguován (96,6 % tvarů má jednoznačně přiřazeno lemma).

3.1 Problémy nejasné lemmatizace

V korpusech se obecně mohou vyskytnout a také se vyskytují případy, kdy slovní tvar nelze jednoznačně lemmatizovat a anotovat (srv. k tomuto tématu více Bartůšková, Hlaváčková, Ungermannová 2004).

V KSK je počet těchto případů dost velký. Odpověď na otázku jak je řešit není vždy jednoduchá. Obecně je při značkování korpusů vždy třeba mít na zřeteli poměr úsilí, jež je třeba vynaložit na označkování sporných případů (složitá typologizace jednotlivostí a následné náročné a mnohdy pochybnosti budící rozhodování), a užitečnosti, popřípadě použitelnosti výsledku pro uživatele, jimž je korpus primárně určen. Z tohoto důvodu byla při značkování KSK věnována značná pozornost především značkování typických morfologických jevů. Asi 3,4 % slovních tvarů v KSKdopisy1 zůstalo i po ruční disambiguaci a doznačkování bez adekvátního lemmatu a morfologické značky.

4. Morfologické značkování

Morfologické značky (***tagy***) používané pro značkování morfologickým analyzátorem ***ajka*** mají formu dodržující pevně dané pořadí atributů a hodnot, které atributy aktuálně nabývají pro analyzovaný slovní tvar (***word***). Všechny značky povinně obsahují alespoň jeden atribut. Je to buď atribut slovní druh, anebo atribut interpunkce, zkratka, speciální značka. Přesný popis značek (tagset analyzátoru ***ajka***) lze najít ve formátu pdf nebo ps na

http://nlp.fi.muni.cz/projekty/ajka/ (tags.pdf, tags.ps). (Podrobný popis značek použitých v modifikované verzi analyzátoru ***ajka*** lze najít na CD1, které je součástí práce Hladká a kol., 2005.)

4. 1 Kategorie slovního druhu a další slovnědruhově závislé kategorie

Systém značek (***tagset***) rozlišuje 10 slovních druhů odpovídajících v základních rysech klasifikaci slovních druhů v běžných mluvnicích (MČ 2, 1986). Samostatnou značku bez označení slovnědruhové příslušnosti mají slovní tvary sloužící k vyjádření kondicionálu (*bych, bys, by, bychom, byste, abych, ..., kdybybych, ...*). Je tomu tak především proto, aby se zabránilo problémům slovnědruhově nejednoznačně určitelných tvarů typu *aby, kdyby, ...* .

Zvláštní značky mají zkratky a interpunkční znaky (poměrně velké množství nejrůznějších kombinací) a slovní tvary označkované speciálními značkami (viz níže).

Značky jednotlivých slovních druhů zahrnují v daném pořadí další atributy:

podstatných jmen (rod, číslo, pád, fakultativní atributy – viz níže),

přídavných jmen (negace, rod, číslo, pád, stupeň, fakultativní atributy – viz níže),

zájmen (osoba – fakultativně u zájmen, která vyjadřují osobu, rod – fakultativně u zájmen, která vyjadřují rod, číslo, pád, fakultativní atributy – viz níže),

číslovek (rod – fakultativně u číslovek vyjadřujících rod, číslo – fakultativně u základních číslovek ***jeden, dva, tři, čtyři*** a u adjektivně skloňovaných číslovek, pád, fakultativní atributy – viz níže),

sloves (negace, vid, slovesný tvar, osoba – fakultativně podle slovesného tvaru, pokud ji tvar vyjadřuje, rod - fakultativně podle slovesného tvaru, pokud jej tvar vyjadřuje, číslo – fakultativně podle slovesného tvaru, pokud je tvar vyjadřuje, fakultativní atributy – viz níže),

příslovcí (negace, stupeň, fakultativní atributy – viz níže).

U dalších slovních druhů (**předložek, spojek, částic, citoslovcí**) se uvádějí pouze fakultativní atributy – viz níže.

Tvary ***bych, bys, by, bychom, byste, abych, ..., kdybybych, ...*** mají zvláštní značku, v níž se uvádí atribut slovesný tvar s hodnotou kondicionál, osoba, číslo a fakultativní atributy – viz níže.

Poznámka: Atribut slovesný tvar neodpovídá žádné tradiční gramatické kategorii. Hodnoty, které nabývá (infinitiv, indikativ - jednoduché tvary, imperativ, participium l-ové, participium pasivní, přechodník přítomný, přechodník minulý, tvary *budu, budeš, bude,...*, tvary *bych, bys, by,..., abych, ..., kdybych, ...*) odrážejí složitý systém slovesných subparadigmat v češtině a zároveň umožňují ve značce podchytit potencionální významy

gramatických kategorií slovesa, které vyjadřuje buď kombinace významů různých gramatických kategorií (vid + slovesný tvar), anebo kombinace několika slovesných tvarů (složené slovesné tvary).

U adjektiv a příslovcí se vyplňuje atribut **d** – stupeň s hodnotou **1** – pozitiv i u tvarů, které stupňovat nelze. Jsme si vědomi, že jde o kompromis. Atribut **e** – negace (přítomnost/nepřítomnost prefixu ne- vyjadřujícího negaci) se vyplňuje u všech adjektiv, adverbií a sloves, přičemž se opět jedná o kompromis (ne všechna adjektiva, natož pak adverbia, mohou prefixem *ne-* tvořit negaci).

4.2 Fakultativní atributy

4.2.1 S-atribut

V textech soukromé korespondence se vyskytuje poměrně frekventovaně nesamostatný morfém „*-s*“, který signalizuje 2. osobu singuláru (při tvoření analytických tvarů minulého času a u tvarů kondicionálu s reflexivním formantem *se/si*). Například nejčastěji ***ses, sis*** *(kam by* ***ses*** *chtěla dostat),* l-ové participium významového slovesa *(****mělas*** *mi říct)*, tázací zájmena *(****cos*** *mi napsala)*, příslovce *(****kdes*** *zrovna poletovala)*, spojky *(* ***žes*** *počkala),* ... atd. Tvary s nesamostatným morfémem *s-* patří sice do repertoáru spisovného jazyka, nicméně se v obecných korpusech (např. SYN2000) vyskytují zřetelně méně frekventovaně než v KSK. Je to dáno dialogickým charakterem dopisu, z něhož plyne časté užití tvarů 2. osoby. Ze srovnání korpusu KSK a SYN2000 vyplývá, že v KSK se tvary s nesamostatným morfémem *s-* vyskytují v průměru patnáctkrát častěji než v SYN2000 (srv. tab.).

	SYN2000 – počet výskytů	KSK – počet výskytů	SYN2000 - % z celkového počtu pozic	KSK - % z celkového počtu pozic
ses	2 710	262	0,0022 %	0,027 %
sis	1 475	176	0,0012 %	0,018 %
žes	580	79	0,00047 %	0,0084 %

Verze analyzátoru ***ajka*** určená primárně pro značkování obecných korpusů psaného jazyka tyto tvary neanalyzovala, což patřilo k jedné ze slabin tohoto analyzátoru. Modifikovaná verze ***ajky*** nabízí u tvarů zakončených písmenem *-s*, které po odtržení tohoto *-s* jsou slovními tvary nalezenými v morfologické databázi, příslušné lemma a značku, jejíž součástí je atribut signalizující přítomnost nesamostatného morfému *-s*. Atribut je označen písmenem **z** a má hodnotu **S**.

Například: Tvar ***muselas*** má značku [tag="k5eAaImAgFnS**zS**"], tvar ***žes*** má značku [tag="k8**zS**"].

4.2.2 Atribut „stylistický příznak“

Ve verzi analyzátoru ***ajka*** používané pro značkování spisovných textů mají značky u všech slovních druhů uveden atribut **stylistický příznak** označovaný písmenem **w**. (Tagset používaný pro značkování Českého národního korpusu i Pražského závislostního korpusu má pro postižení stylistické příznakovosti vyhrazenu 15. pozici.) Morfologických charakteristik jazyka se však tento příznak přísně vzato týká pouze v některých případech (srv. k problematice příznakovosti v morfologii Krčmová, 2005). Z tohoto důvodu jsme při značkování KSK přistoupily k jistým změnám.

Poznámka: Hodnoty atributu **w** v lexikální databázi, nad níž pracuje morfologický analyzátor ***ajka***, se kryjí s hodnotami stylových charakteristik uváděnými ve Slovníku spisovného jazyka českého (SSJČ). Strojový slovník českých kmenů (srv. Pala, 1992, Osolsobě, 1996), který byl vytvořen na Ústavu českého jazyka FF MU v letech 1988-96, zahrnul slovní zásobu vycházející z hesláře SSJČ. S tímto slovníkem pracoval automatický analyzátor ***lemma*** (Ševeček, 1995) a později, s řadou úprav a oprav, automatický analyzátor ***ajka***. Analyzátor ***lemma*** a později ***ajka*** sloužil a slouží k automatickému tagování korpusů budovaných na FI MU a ve spolupráci s FI MU na FF MU. Zásadní význam pro značné opravy v lexikální databázi – slovníku kmenů – mělo vytvoření elektronické verze SSJČ (srv. více Smrž, Pala, 2001).

Značkování stylistického příznaku v KSK představuje vzhledem k velmi složitému teoretickému pozadí celé problematiky jistý kompromis. Atribut stylistický příznak je ve značkách KSK pouze fakultativní a má jen jedinou hodnotu („příznakovost“).

Atribut mají vyplněny: 1) varianty se substandardními koncovkami, 2) varianty s protetickým *v*-, 3) chybné užití zájmenných tvarů (*mě/mně, ji/jí, ...*), l-ových participií, tvarů kondicionálu, 4) nekodifikované slovotvorné inovace a nespisovné lexikální jednotky, 5) pravopisné chyby. Vyjmenované "anomálie" jsou signalizovány přítomností atributu **w**, který v těchto případech nabývá hodnoty **H**.

Například: Tvar ***brachem*** má značku [tag=“k1gMnSc7**wH**”], tvar ***kterej*** má podle kontextu např. značku [tag=“k3gMnSc1**wH**”], tvar ***vo*** má značku [tag=“k7**wH**”], tvar ***kámoškou*** má značku [tag=“k1gFnSc7**wH**”]**,** chybně napsaný tvar ***jí*** v kontextu *„ta **jí** poprosila“* má značku [tag=“k3p3gFnSc4**wH**”], chybně napsaný tvar ***být*** v kontextu *„... být po hlavě ...“* má značku [tag=“k5eAaImF**wH**“] atd.

4.3 Zkratky a interpunkce

Zkratky a **interpunkce** nemají atribut slovní druh, značka říká, že jde o zkratku nebo interpunkci. Vzhledem k tomu, že KSK obsahuje velké množství zkratek individuálních a velké množství jevů, které jsou homonymní se zkratkami (jednotlivá „osamocená“ písmena),

byl do modifikované verze analyzátoru ***ajka*** zařazen jen malý počet zkratek. Vysoká míra homonymie totiž do značné míry zatěžuje ruční anotátory a celkově zpomaluje disambiguaci. Řada zkratek víceméně individuálních byla doznačkována ručně a má speciální značku (srv. níže odd. speciální značky).

V KSK se vyskytlo poměrně mnoho případů individuálního použití interpunkce. Šlo především o různé množství několika teček, pomlček, vykřičníků nebo otazníků (tři, čtyři, pět až *x* teček, ...) použitých pisateli s nejrůznějšími záměry. Automatická morfologická analýza těmto „řetězcům" nepřiřadila žádnou značku. Byly doznačkovány automaticky až po ruční disambiguaci ve fázi doznačkovávání neanalyzovaných tvarů.

5. Disambiguace

Automatické morfologické analyzátory vykazují pro jazyky s vysokou mírou tvarové homonymie, k nimž patří čeština, značné procento nejednoznačně označkovaných slovních tvarů. U některých slovních druhů (adjektiv) nabízí automatická morfologická analýza u jednoho tvaru i více než dvacet interpretací. Pro zjednoznačnění – výběr kontextově správné interpretace (disambiguaci) se používá různých metod. Na jedné straně stojí metody strojové (automatické), na druhé ruční disambiguace, kdy rozhodnutí provádí školený anotátor. Pro potřeby značkování korpusů menšího rozsahu (řádově statisíce slovních tvarů) lze s ohledem na časové i finanční náklady použít ruční disambiguaci. Ani vysoká odborná fundovanost a odpovědnost anotátorů není ovšem vždy zárukou bezchybné anotace. Přesto lze oprávněně předpokládat, že u korpusů s vysokým procentem substandardních tvarů je alespoň zatím spolehlivější než metody založené na stochastických přístupech, či pravidlech. Navíc je ruční disambiguace pro použití těchto metod východiskem a zdrojem zkušeností. Chyby vzniklé při ruční disambiguaci jsou způsobeny především nepozorností a únavou anotátora. Nicméně i ručně anotovaný korpus lze podrobit následné strojové kontrole, a tím počet případných chyb snížit.

Na FI MU byl vytvořen pro ruční disambiguaci program **CED** s dávkou **desam** (Veber, 2003). Zkušenosti s ruční disambiguací korpusů psaného spisovného jazyka daly vznik rukopisnému manuálu (Bartůšková, Hlaváčková, Ungermannová, 2004) používanému při ruční disambiguaci korpusů budovaných na FI MU (např. korpus DESAM, srv. Pala, Rychlý, Smrž, 1997). Na základě těchto zkušeností s přihlédnutím ke specifikům textů soukromé korespondence byla stanovena pravidla pro ruční anotátory KSK. Řada problémů vyšla však najevo až při práci samé a jejich řešení bylo třeba teprve hledat (srv. níže).

5.1 Rozlišování neohebných slovních druhů

Kromě vysoké míry koncovkové homonymie ohebných tvarů jsou k disambiguaci nejčastěji nabízena synsémantika. Homonymie je dána funkčními slovnědruhovými transpozicemi bez formální signalizace.

Například: Slovnímu tvaru *tak* přiřadí ***ajka*** čtyři možné interpretace (adverbium, spojka, částice nebo citoslovce - [tag="k6"] |[tag="k8"]|[tag="k9"]|[tag="k0"]).
Na základě zkušeností jak anotátorů, tak uživatelů jsme se rozhodli tyto případy ponechat nedisambiguované, tzn. že je možné je vyhledat podle všech potencionálních značek. Uživatel pak může dále pracovat s takto získanými daty pomocí „filtrů".

Jedná se celkem o 170 slovních tvarů, které se ovšem vyskytují značně frekventovaně.

5.2 Nárůst počtu interpretací u tvarů nabízených k disambiguaci

Modifikace morfologického analyzátoru, která umožnila identifikaci nestandardních koncovkových podob ohebných tvarů jmen a sloves, měla za následek nárůst počtu možných interpretací nabízených k disambiguaci.

Příklad: Nárůst tvarové homonymie tvrdých adjektiv:

tvar adjektiva	počet značek	standardní	substandardní
blbý	23	5	18
blbé	18	13	5

V celém KSK - dopisy se vyskytlo přes devět tisíc tvarů na *-ý* a kolem sedmi tisíc tvarů na *–é*, které jsou tvary tvrdých adjektiv, dále zájmen a číslovek skloňovaných jako tvrdá adjektiva. Morfologický analyzátor ***ajka*** tyto tvary analyzuje a nabízí u každého z nich k ruční disambiguaci namísto původních pěti/třinácti třiadvacet/osmnáct možných značek, z nichž musí anotátor vybrat podle kontextu značku jednu. (Pro zajímavost uveďme, že počet těchto tvarů představuje 20 % všech slovních tvarů v KSK. Předběžné výsledky ukazují, že zatímco mezi tvary na –ý je kolem 45 % tvarů susbtandardních, mezi tvary na –é je jich necelé 1 %.).

5.3 Problémy disambiguace

Ruční disambiguace je časově i finančně náročná. U řady víceznačných jednotek školený anotátor vybere podle kontextu bez problému jednu z interpretací. Řešení některých případů naráží ovšem při ruční disambiguaci na problémy. Při práci na disambiguaci KSK byla

stanovena zásada, podle níž mají být nejasnosti ponechány nedisambiguovány, aby se tak zabránilo případným inkonzistentním, či chybným řešením.

Poznámka: KSK byl disambiguován více anotátory (studenty FF MU). Ze zkušeností s ruční disambiguací je známo, že inkonzistentní řešení se vyskytují i v práci jednoho anotátora nehledě na jeho odborné kvality a odpovědnost. Proto se někdy volí metoda, při níž je týž text disambiguován alespoň dvěma různými anotátory. Výsledky jejich práce se strojově porovnají a odlišnosti jsou následně překontrolovány. Toto řešení nebylo při práci na disambiguaci KSK z časových a finančních důvodů možné.

5.3.1 Pravopis - morfologie

Při ruční disambiguaci byly nejčastěji ponechávány bez disambiguce následující případy: a) tvar zájmena *mě* ve 3. pádě, b) tvar zájmena *jí* ve 4. pádě c) tvary *by jsme, by sme, by jste*, d) chyby ve shodě v l-ovém participiu. Automatická analýza k těmto slovním tvarům nabízí značky, z nichž ani jedna nebyla v příslušném kontextu správná. Tyto případy se přednostně řešily při následujícím ručním doznačkování.

5.3.2 Určení kategorií vyjadřujících shodu (rod, číslo, pád) v případech elipsy substantiva

Anotátoři ponechávali nerozhodnuty případy, kdy nebylo možné na základě kontextu určit některý z gramatických významů. K častým případům patřila elipsa substantiva, s nímž by se měl shodovat tvar nabízený k disambiguaci. Mnohdy nebylo možné ani na základě širšího kontextu zjistit, o jaké substantivum jde.

V takovýchto případech se jako nejpřijatelnější jeví řešení, které by ponechávalo hodnoty příslušných atributů (gramatických významů) nevyplněné. To ovšem naráží na omezení dosavadního systému značek (tagsetu automatického morfologického analyzátoru ***ajka***) a zároveň naznačuje směr, kterým by se měla ubírat jeho další modifikace.

5.3.3 Určení kategorie rodu u hypokoristik

V KSK se vyskytlo poměrně velké množství nejrůznějších tvarů hypokoristik (srv. více Osolsobě, 2005). Frekventované domácké podoby vlastních jmen byly zařazeny do modifikované verze programu ***ajka***, méně frekventované byly ponechány k ručnímu doznačkování. Problémy vyvstaly u některých případů při určování gramatické kategorie

rodu, který nebylo možné disambiguovat ani na základě prohledání celého kontextu příslušného dopisu. Vzhledem k tomu, že se jedná o jednotlivosti, byly tyto případy ponechány bez značek.

6. Ruční doznačkování

Anotátoři, kteří prováděli ruční disambiguaci KSK, záměrně vynechávali řešení některých sporných případů. V následující části se budeme zabývat systematizací případů, které při ruční disambiguaci způsobovaly obtíže.

6.1 Označkované tvary vynechané při ruční disambiguaci

Při ruční disambiguaci byly záměrně vynechány případy, kdy a) anotátor si nebyl jistý, kterou z nabízených variant vybrat (srv. výše 5.3.2, 5.3.3), b) žádná z nabízených variant nebyla správná (srv. výše 5.3.1). Tyto slovní tvary byly ponechány bez značky a část z nich byla doznačkována ručně. Přednostně byly vybrány případy, kdy bylo možné jednoznačně doplnit značku.

6.2 Tvary neoznačkované automatickou morfologickou analýzou

Jak již bylo řečeno v kapitole 2.2, při modifikaci automatického morfologického analyzátoru ***ajka*** byly brány v úvahu pouze frekventované jevy. Během práce na značkování KSK se ukázalo, že se v korpusu vyskytuje velké množství substandardních jevů s velmi malou frekvencí, které je ovšem možno označkovat podobným způsobem jako jevy frekventované. Jednalo se především o nejrůznější varianty substandardních podob koncovek frekventovaných slov (především zájmen), nejrůznější varianty slovesa *být*, tvarů *by, aby, kdyby, ...* atd. Těmto slovním tvarům byly ručně doplněny příslušné značky.

6.2.1 Značkování speciálními značkami

Morfologické tagy nemají být konečnou a neměnnou instancí (srv. např. Leech, 1993). Korpusy s vysokou frekvencí substandardních jevů ukazují, že řadu problémů nelze uspokojivě řešit pomocí tagsetů navržených primárně pro značkování korpusů psaného jazyka (každý korpus je z principu bohatší než sebelépe navržený systém značek). Ukázalo se ovšem,

že problematické jevy lze alespoň třídit, a tak o nich získat přehled. Na základě předběžné klasifikace byly navrženy speciální značky. Ty se do budoucna mohou stát inspirací při navrhování nových systémů pro anotaci (tagsetů).

Speciální značky byly navrženy na základě průzkumu materiálu slovních tvarů, k nimž automatický analyzátor nenabídl žádnou značku a jimž nebylo možno přiřadit některou z existujících značek. Primárně jsme se zabývali frekventovanými jevy.

Jedná se o následující případy:

Grafická chyba [tag="<graficka_chyba>"] – do této kategorie řadíme:

- neúplná slova

př.: *ta* místo *tak* – vynechané písmeno [lemma="ta" & tag="<graficka_chyba>"];

- spojení více slov do jednoho řetězce (pozice)

př.: *AhojBlani* místo *Ahoj Blani* – vynechaná mezera mezi slovy [lemma="AhojBlani" & tag="<graficka_chyba>"];

- rozdělené slovo

př.: *říkej me* místo *říkejme* - mezera mezi částmi jednoho slova [lemma="říkej" & tag="<graficka_chyba>"][lemma="me" & tag="<graficka_chyba>"];

Poznámka: Současné automatické zpracování korpusů je založeno na automatické segmentaci textu na tzv. pozice (tokenizaci), která pak musí být pro veškeré další strojové zpracování korpusu zachována beze změn. Jedné pozici odpovídá jedno nebo více lemmat a značek, není ale možné rozdělit pozici a jedné její části přiřadit jednu značku a druhé části jinou, popřípadě spojit dvě pozice do jedné a přiřadit takto vzniklé nové jednotce (pozici) odpovídající značku tak, jak by to odpovídalo ve výše uvedených případech lingvistické intuici. Za jeden případ „rozdělených slov" by bylo možné pokládat morfologicky nesprávně utvořené tvary kondicionálu (hyperkorektní tvary *by jsme, by jste* a analogické tvary *by sme*) , které se v KSK sem tam vyskytly. Jejich ruční doznačkování je kompromisní (automatický analyzátor k těmto tvarům nabízí značky, které jsou pro ruční anotátory z lingvistického hlediska nepřijatelné). Kompromis při ručním doznačkování spočívá v tom, že každá jednotka má vlastní značku, protože z technických důvodů nelze dvěma pozicím přiřadit jednu značku, jak by to odpovídalo lingvisticky přijatelnému řešení. Pokud je izolovaný slovní tvar spisovný (*by, jsme, jste, ...*), značka atribut **w** nemá, pokud je izolovaný slovní tvar nespisovný (*sem, sme, ste, ...*), pak značka má atribut **w** s hodnotou **H** a signalizuje tak susbtandardnost izolovaného slovního tvaru. Substandardní kombinace použitých tvarů není signalizována. Toto řešení není nikterak ideální. Odpovídá však obecnému řešení v dosavadní praxi. V českých korpusech se složené slovesné tvary značkují tak, že každý tvar je lemmatizován a označkován samostatně, tj. bez ohledu na to, že je součástí víceslovné jednotky. Ve značce slovesného tvaru se uvádějí hodnoty příslušných gramatických kategorií nezávisle na hodnotě kategorie, jak ji vyjadřuje složený tvar slovesný jako celek, což je chyba (srv. značkování v případech typu *„To **by ses** před ní ukázal v pěkném*

světle ...“ , slovní tvar *by* je označen jako tvar 3. osoby). Nepatrně odlišnou praxi nalezneme například v návrhu tagsetu pro Slovenský národní korpus, (srv. Garabík, Gianitsová, Horák, Šimková, 2004).
Například: Tvary *by jsem (... už nevím, co* ***by***/by/kYmCp3nS ***jsem***/být/k5eAaImIp1nS *měla o dovolené napsat ...)* jsou označkovány dle formy, tvar *by* má u atributu **p** hodnotu **3** (3. osoba), což neodpovídá lingvisticky správnému řešení. Navíc není označeno, že celý tvar (kombinace ***by+jsme***) je substandardní. Obdobně jsou označkovány tvary *by sem (...předem mého dopisu* ***by***/by/kYmCp3nS ***sem***/být/k5eAaImIp1nSwH *Ti chtěla zdělit ...)*. Pouze u tvaru *sem* je jeho substandardnost vyznačena, a to přítomností atributu **w** s hodnotou **H**.

- neidentifikovatelné slovo

př.: ... *nelze vyjít* ***ze*** *brány knihovny* ... - z kontextu není zřejmé, zda jde o překlep *ze<z*, nebo *ze<za* [lemma=”ze” & tag=”<graficka_chyba>”].

Zkratka [tag=”<zkratka>”]

Tato značka byla ručně přiřazena případům zkratek, které nebyly označkovány automatickou morfologickou analýzou značkou pro zkratky (kA).

Poznámka: Otázka značkování zkratek se na základě dosavadních zkušeností jeví jako nedořešená. Značku („nálepku“) zkratka má řada značně různorodých jevů (grafické zkracování frekventovaných slov, grafické zkracování frekventovaných slovních spojení, normované grafické symboly pro matematické, fyzikální, chemické pojmy, iniciálové zkratky atd.), které by měly být jemněji klasifikovány. Situace zkratek je svým způsobem zrcadlová k situaci víceslovných výrazů (v oblasti strojového zpracování přirozeného jazyka - NLP se hovoří o šířeji pojatých multi word expression - MWE). Tak jako jsou víceslovné výrazy v dosavadní praxi značkovány odděleně, aniž by se jakkoliv signalizovalo, že k sobě patří, tak zkratky, které v řadě případů zastupují víceslovný výraz (grafické zkracování frekventovaných slovních spojení, iniciálové zkratky atd.), mají pouze „nálepku“ zkratka. Jak už jsme několikrát naznačili, značky nejsou a nemají být neměnnou a konečnou instancí, ale pouze pomocným systémem usnadňujícím práci s masovými daty, systémem který může a má být modifikován a optimalizován především na základě zkušeností uživatelů v nejširším slova smyslu.

Cizí slovo

Delší úseky textů v cizích jazycích byly při přepisu z rukopisu do elektronicky čitelné podoby dat zařazeny do ”poznámky”, takže nejsou zpracovávány morfologickou analýzou. Záměrně však byla v textu ponechána jednotlivá cizojazyčná slova a slovní spojení. Makarónský způsob vyjadřování je totiž charakteristickým rysem v dopisech zejména mladých pisatelů.

Vzhledem k tomu, že k označkování těchto slovních tvarů se nehodily stávající značky, přistoupili jsme ke speciálním značkám. Bereme v úvahu především jazyky, které se vyskytly v KSK. Obecně lze ovšem říci, že ve stávajících analyzátorech (resp. příslušných tagsetech) není situace značkování cizojazyčných jednotek v textu uspokojivě řešena. Navrhované řešení je prvním a jistě ne jediným možným pokusem, jak stávající stav změnit.

Některým frekventovaněji užitým anglickým, francouzským, německým, slovenským, ruským aj. slovům v textech jsou přiřazeny následující značky:

[tag=”<anglicky>”]

[tag=”<nemecky>”]

[tag=”<francouzsky>”]

[tag=”<jiny_jazyk>”]

Závěr

Korpusy existují v neanotované (surové) nebo anotované (označkované, taggované) podobě. Obsahují holé texty nebo texty s přídatnou především lingvistickou (morfologickou, slovnědruhovou, syntaktickou, sémantickou) informací. Už surové korpusy mohou výrazně pomoci lingvistovi v jeho bádání, obecně lze ovšem tvrdit, že anotované korpusy výrazně rozšiřují možnosti dojít k zajímavějším výsledkům v lingvistické práci a představují tudíž „lepší korpusy“. Morfologické značkování jazykových korpusů je nejrozšířenějším typem anotací, neboť je prvním krokem k automatické analýze textu na vyšších rovinách popisu jazyka. Značkování KSK přineslo řadu zkušeností, které je možné využít a) při značkování korpusů s vysokým počtem substandardních tvarů (korpusů mluvených, korpusů psaných textů nepodléhajících jazykové korekci atd.), b) při úpravě automatických morfologických analyzátorů pro takové typy korpusů, c) při teoretických úvahách o mezích a možnostech morfologického značkování jazykových korpusů a v neposlední řadě d) při tvorbě nových systémů značek (tagsetů), které by lépe odpovídaly skutečnosti jazyka, již jazykové korpusy reprezentují.

MORPHOLOGICAL TAGGING OF KSK (CORPUS OF PRIVATE CORRESPONDENCE)

Corpora may exist in two forms: unannotated (plain text, raw corpus) or annotated (enhanced with various types of linguistic information – lemmatisation, part-of-speech annotation, ...). The utility of the corpus is considerably increased by the provision of annotation.

The experience with annotating the KSK (i.e. the corpus of private correspondence) can be used a) by annotating corpora with a large number of substandard forms (spoken corpora, informal corpora, ...), b) by modification of automatical analyser for such types of corpora,

c) by the theoretical consideration about limits and possibilities of tagging of the corpora and last but not least d) by project of new tagset.

Bibliografie

Bartůšková, D., Hlaváčková, D., Ungermannová, M.: Manuál pro značkování a desambiguaci slovních tvarů v jazykových korpusech, Brno : FI MU 2004. (pdf verze: http://nlp.fi.muni.cz/projekty/desman/)

Hajič J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Praha : Karolinum 2002.

Havránek, B., Jedlička, A.: Česká mluvnice. Praha : SPN 1981.

Hladká, Z.: Korpus soukromé korespondence jako zdroj poznání jazykového úzu. In: M. Šimková (ed.), Tradícia a perspektívy gramatického výskumu na Slovensku. Bratislava : Veda 2003, s. 130-135.

Hladká, Z., Šindlerová, H.: Jakou češtinou si dopisujeme na Moravě. In: J. Fiala (ed.), AUPO, Fac. Phil., Moravica 1. Olomouc: UP 2004, s. 105-114.

Hladká, Z.: Univerbizace – korpusy – slovníky (malá materiálová sonda). In: R. Blatná – V. Petkevič (eds.), Jazyky a jazykověda. Sborník k 65. narozeninám prof. Františka Čermáka. Praha : ÚČNK FF UK 2005, s. 503-514.

Hladká, Z.: Zkušenosti s tvorbou korpusů češtiny v ÚČJ FF MU v Brně, SPFFMU A 53, 2005, s. 115-124.

Hladká, Z. a kol.: Čeština v současné soukromé korespondenci. Dopisy, e-maily, SMS. Brno : Masarykova univerzita 2005.

Hlaváčková, D.: Korpus mluvené češtiny. Dipl. práce na FF MU, (rkp.), Brno 1998.

Hlaváčková, D.: Brněnský mluvený korpus a jeho morfologická analýza. In: Pořízka, P., Polách, V. P.: Vztah langue a parole v perspektivě interaktivního obratu v lingvistickém zkoumání. Sborník příspěvků z 3. mezinárodní konference Setkání mladých lingvistů konané na Filozofické fakultě Univerzity Palackého Olomouc ve dnech 14. a 15. května 2002. Olomouc : VUP 2004, s. 167 – 173.

Hlaváčková, D., Sedláček, R.: Morfologické značkování korpusu soukromé korespondence, XIV. kolokvium mladých jazykovedcov, 8. - 10. 12. 2004, Šintava pri Seredi, Slovenská republika.V tisku.

Krčmová, M.: Příznakovost a její specifika v morfologii. In: Karlík P., Pleskalová, J. (eds.) Život s morfémy. Sborník studií na počest Zdenky Rusínové. Brno : Masarykova univerzita 2005, s. 111-122.

Leech, G.: Corpus annotation schemes, Literary and linguistic Computing 8 (4), 1993, s. 275-281.

Osolsobě, K.: Algoritmický popis české morfologie a strojový slovník češtiny. Disertační práce na FF MU (rkp.), Brno 1996.

Osolsobě, K.: Hypokoristika v korpusu soukromé korespondence KSK, SPFFMU A 53, 2005, s. 125-136.

Pala, K.: Počítačové zpracování češtiny. Habilitační práce na FF MU (rkp.), Brno 1992.

Pala, K., Rychlý, P., Smrž, P.: DESAM - Annotated Corpus for Czech. In Proceedings of SOFSEM 97. Heidelberg : Springer Verlag 1997, s. 523-530.

Petr, J. & kol (eds.).: Mluvnice češtiny II. Praha : Academia 1986.

Sedláček, R., Smrž, P.: A New Czech Morphological Analyser ajka. In Text, Speech and Dialogue, 4th International Conference, TSD 2001. Berlin : Springer-Verlag 2001, s. 100-107.

Sedláček, R.: Morphematic analyser for Czech. Disert. práce na FI MU (rkp.), Brno 2004.

Smrž, P., Pala, K.: Elektronická podoba SSJČ. 2001. MSM 143300003.

Ševeček, P., Morfologický analyzátor a lemmatizátor pro češtinu – implementace v jazyce C, Brno : FI MU 1995.

Veber, M.: Nástroje pro textové korpusy a morfologické databáze. Disert. práce FI MU (rkp.), Brno 2003.

Český národní korpus - SYN2000. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: <http://ucnk.ff.cuni.cz> (http://ucnk.ff.cuni.cz/bonito/)

http://nlp.fi.muni.cz/projekty/ajka/ajkacz.html.

http://ufal.mff.cuni.cz/pdt/ (http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/morph.html).