SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 55, 2007 — LINGUISTICA BRUNENSIA

KLÁRA OSOLSOBĚ

POPIS GRAMATICKÝCH VÝZNAMŮ (HODNOT) JEDNODUCHÝCH SLOVESNÝCH TVARŮ V ANOTACÍCH ČESKÝCH (SLOVENSKÝCH) KORPUSŮ

V tomto článku se budeme zabývat problematikou morfologických anotací jednoduchých slovesných tvarů v jazykových korpusech. Srovnáme tagsety použité pro značkování slovesných tvarů v českých korpusech (SYN2000/SYN2005, DESAM, KSK) a návrh tagsetu pro značkování SNK.

Poznámka: Korpusy SYN2000 a SYN2005 jsou vyvážené a reprezentativní korpusy psaného jazyka (viz http://ucnk.ff.cuni.cz). Jejich rozsah je řádově 100 milionů tvarů. DESAM je korpus psaného jazyka publicistických textů, jehož rozsah činí řádově 1 milion slovních tvarů (srv. např. Pala, Rychlý, Smrž, 1997). KSK je korpus soukromé korespondence zachycující ručně psanou soukromou korespondenci 2000 pisatelů. Jeho rozsah je řádově 0,5 milionu slovních tvarů (srv. Hladká a kol., 2005). SNK (Slovenský národný korpus) je projekt vybudování reprezentativního a vyváženého korpusu současné slovenštiny (srv. http://korpus.juls.savba.sk).

Pokusíme se navrhnout směr, jímž by se měly ubírat úpravy systémů morfologického značkování tak, aby lépe odpovídaly jak lingvistické teorii, tak potřebám rozličně orientovaných uživatelů jazykových korpusů.

Úvod

Anotované korpusy, tedy korpusy, v nichž je jednotkám textu přiřazena lingvistická interpretace ve formě značek charakterizujících příslušné jednotky na jazykových rovinách (fonetické, morfologické, syntaktické, sémantické), mohou výrazně pomoci badatelům a rozšiřují možnosti dojít k zajímavějším výsledkům v lingvistické práci. Morfologické značkování jazykových korpusů je nejrozšířenějším typem anotací a provádí se většinou automaticky s následnou desambiguací.

Poznámka: Automatická morfologická analýza je obecně nejednoznačná (ambiguitní). Proto na automatickou morfologickou analýzu navazují různé způsoby desambiguace. Desambiguace je volba jednoznačné (kontextově nejpřijatelnější) interpretace. Při automatické desambiguaci (ať už používá statistických metod (srv. Hajič, 2004) nebo je řízena pravidly (srv. např. Oliva et al., 2000, Petkevič, 2001) se zpravidla volí pouze ze značek vygenerovaných automatickou morfologickou analýzou. Při ruční desambiguaci zaručí týž postup možnost lépe dohledat případné chyby anotáto-

rů. Obecně vzato platí, že čím je počet interpretací (značek), které nabízí automatická morfologická analýza, vyšší, tím je následná desambiguace (jednoznačná volba jedné z nabízených interpretací) složitější, ať už počítáme s automatickou nebo ruční desambiguací.

Jedním z důvodů rozšířenosti morfologického tagování je i to, že morfologická analýza je prvním krokem k automatické analýze textu na vyšších rovinách popisu jazyka. K automatickému značkování českých korpusů se v současné době používá různých automatických analyzátorů. Ty přiřazují jednotkám textu (slovním tvarům) tagy, vycházející z klasifikace slovních druhů a gramatických kategorií.

Poznámka: Vzhledem k tomu, že se klasifikace gramatických kategorií a jejich významů v jednotlivých systémech docela nekryjí s žádnou stávající teoretickou lingvistickou koncepcí, budeme nadále v textu používat termíny běžné v literatuře pojednávající o morfologických anotacích. Gramatické vlastnosti odpovídající v podstatě gramatickým kategoriím (slovní druh, osoba, číslo, čas, …) se v anotačních schématech nazývají atributy a jsou zachyceny různým způsobem. Pro gramatické vlastnosti odpovídající gramatickým významům (sloveso, první osoba, singulár, prézens, …) se běžně užívá termín hodnota.

Systémy značek (tzv. tagsety) byly budovány spolu s morfologickými analyzátory a mnohdy modifikovány na základě zkušeností s konkrétní prací s korpusy.

Poznámka: Morfologický analyzátor Jana Hajiče – dále *HA* (srv. Hajič, 1994, 2004) byl použit k anotacím SYN2000/SYN2005 a morfologický analyzátor Radka Sedláčka - *ajka* (srv. Sedláček, 2005, Hlaváčková, Sedláček, 2006) k anotacím DESAM, KSK.

Přesto se ale ukazuje, že i v sebelépe navrženém systému značek se mohou objevit „mezery", tj. korpus může obsahovat takové jazykové jevy, které nelze uspokojivě popsat systémem značek, kterým disponuje automatický morfologický analyzátor. Cílem našeho příspěvku bude poukázat na shody a odlišnosti obou tagsetů (tagsetu *HA* a tagsetu *ajky*) používaných ke značkování českých korpusů a návrhu tagsetu, který vznikl v rámci projektu Slovenského národního korpusu (srv. Garabík, Gianitsová, Horák, Šimková, 2004, Giantisová, 2005), a zamyslet se nad teoretickými i praktickými aspekty jejich předností i omezení.

Značkování korpusu a lingvistická teorie

Gramatické anotace morfologického typu se v korpusové literatuře nejčastěji nazývají tagy. Hovoří se o gramatickém tagování spíše než o gramatických anotacích. Dalším termínem, s nímž se můžeme setkat, je POS – slovnědruhové značkování. Prvním stupněm morfologického značkování je lemmatizace, tedy přiřazení základního tvaru slova jeho textové realizaci.

Každá anotace, a tedy i anotace na úrovni lemmatizace, určení hodnot na úrovni interpretace slovních druhů a hodnot (významů) dalších atributů (morfologických kategorií), tedy morfologické značkování, je obecně vzato interpretace, která je nevyhnutelně ovlivněna lingvistickou teorií. Teoretikové korpusové lingvistiky však často upozorňují na úskalí, která s sebou nese přílišná vázanost anotace korpusových dat k určité teoretické interpretaci (Leech, 1993).

V obecném korpusu, který má sloužit co nejširšímu okruhu uživatelů z řad badatelů i laiků zajímajících se o přirozený jazyk, by anotace měla obsahovat

pouze jasně formulované interpretace přístupné každému konečnému uživateli korpusu. Příčin subjektivní interpretace vtělené do značek může být ovšem více. Nicméně je třeba přiznat, na které úrovni případně sporné interpretace vznikají. Chybné interpretace mohou vnést do značkování korpusů jak automatické programy (značkování velkých - řádově stamilionových korpusů - je třeba provádět automatizovaně), jednak školení anotátoři (při ručním značkování menších, ať už trénovacích nebo speciálních korpusů atd.), při jejichž práci může hrát roli nepozornost, únava, neschopnost konzistentně řešit případy vyžadující postup podle několika pravidel atp., další okruh chyb vzniká na úrovni desambiguace. V tomto článku se chceme věnovat výhradně tomu typu sporných interpretací, které vnáší do tagovaných korpusu používání automatických morfologických analyzátorů.

Poznámka: Tak jako školský jazykový rozbor (určování gramatických kategorií slov/ slovních tvarů) neslouží sám sobě, ale má vést k hlubšímu poznání zákonitostí jazyka, tak i značkování velkých korpusů není samoúčelné. Má především sloužit uživatelům korpusů k tomu, aby mohli efektivněji využívat vyhledávání v jazykovém materiálu (textech čítajících stamilióny slovních tvarů). Uživatelé z řad badatelů-lingvistů hledají objektivní materiál pro ověření nejrůznějších jazykových teorií, tvorbu slovníků, gramatik, učebnic atd., odborníci na strojové zpracování přirozeného jazyka (NLP) se snaží, vycházejíce z jazykových korpusů, tvořit modely jazyka pro nejrůznější aplikace.

Mají-li korpusy budované jako obecné korpusy (ČNK, SNK) skutečně sloužit co nejširšímu okruhu badatelů i dalších zájemců, je třeba, aby bylo pro všechny potencionální uživatele korpusů, jednoznačně vysvětleno, co je obsahem tagů.

V následujícím textu se tedy na základě více než desetiletých zkušeností s používáním automatické morfologické analýzy při anotování českých korpusů budeme zabývat problémy, které souvisejí s lingvistickými interpretacemi slovesných tvarů vyjadřovaných konkrétními tagy.

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 jak řetězec znaků dané abecedy oddělený z obou stran oddělovači (mezery, některé znaky). Takto technicky omezená definice slovního tvaru má při další interpretaci (značkování) slovesných tvarů automatickou morfologickou analýzou své důsledky na všech úrovních (srv. níže).

Automatická morfologická analýza

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

Poznámka: Morfologické analyzátory pracují nad databází slovních tvarů a jejich možných interpretací. Tyto databáze byly zpracovány na základě algoritmických popisů flexe (srv. Hajič 1994, 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.

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.

Lemmatizace a značkování slovesných tvarů

Segmentovaným jednotkám (t**oken)** jsou přiřazovány další interpretace. Těmi je u morfologické analýzy **lemma** a značka (**tag**). Konkrétně se lemmatizací rozumí přiřazení slovníkového tvaru textovému tvaru. Vzhledem k tomu, že se slovníkový tvar generuje automatickou morfologickou analýzou, je výsledek lemmatizace podmíněn konkrétním analyzátorem. U slovesa je tímto tvarem zpravidla infinitiv jednoduchého slovesného tvaru, což přísně vzato neodpovídá lingvistickému přístupu. (Například ve větě „*Už **jsem** se **nechala vyfotit** …*“ budou tři samostatné slovesné tvary, jimž budou odpovídat tři lemmata: tvaru ***jsem*** bude přiřazeno lemma ***být***, tvaru ***nechala*** lemma ***nechat*** a tvaru ***vyfotit*** lemma ***vyfotit.***)

Poznámka: Tento „zjednodušený přístup“ není v oblasti strojového zpracování přirozeného jazyka nikterak výjimečný. Svědectvím jsou například údaje o frekvenci slovesa ***být*** ve všech frekvenčních slovnících češtiny (srv. Jelínek, Bečka, Těšitelová, 1961, Králík, Těšitelová, 1986, Čermák, Křen, 2004). Podobně je tomu i v obdobných lexikografických dílech slovenských (srv. např. Mistrík, 1976). Také v korpusech anglických, které mají v oblasti značkování o něco delší tradici, se při značkování složených slovesných tvarů (připomeňme jen, že angličtina jich má mnohem více než čeština) postupuje analogicky.

Obecně vzato můžeme říci, že „kompletace složek složeného slovesného tvaru“ bývá v praxi v rámci morfologické analýzy odsunuta a řeší se až na úrovni syntaktické analýzy a potažmo syntaktické anotace (srv. k tomu řešení v PZK-PDT, Žáčková 1999).

Poznámka: Rozlišování rovin jazyka (fonetické, morfologické, syntaktické, …) je vedeno snahou o jeho systematický popis. To, že jevy popisované na jednotlivých rovinách spolu úzce souvisejí a že je někdy nelze od sebe mechanicky oddělovat, není v lingvistice ničím novým. Na druhé straně je ve školské praxi běžné určování gramatických kategorií slovesa odlišné od toho, co je, jak jsme naznačili výše, běžnou praxí v morfologicky tagovaných korpusech. (K možnostem vyhledávání složených slovesných tvarů ve značkovaných korpusech srv. např. Osolsobě, 1999.)

Lemmatizace tvarů slovesa

Při srovnání tří systémů se z hlediska lemmatizace jeví problematické dva případy: 1) Lemmatizace tvarů *bych, bys, by, ..., abych, ..., kdybych, ...* 2) Lemmatizace sloves s prefixem *ne-* vyjadřujících negaci.

Poznámka: 1) V mluvnicích jsou tvary *-by-* sloužící ke tvoření složených tvarů kondicionálu interpretovány jako tvary slovesa *být*. Tvary *a-/kdy-by-* se probírají v souvislosti se svojí syntaktickou funkcí spojek. Školní praxe ovšem vyžaduje při analýze na úrovni rozboru slovních druhů a následné interpretace gramatických významů interpretovat slovesný tvar ve vedlejší větě připojené spojkou *aby, kdyby, ...* jako kondicionál, a pokud je zadán požadavek vypsat z takového typu věty vedlejší přísudkové sloveso, je za správnou odpověď pokládán tvar *aby/kdyby/ ...* + l-ové participium slovesa významového, a dále l-ové participium pomocného slovesa *být* v případě kondicionálu minulého. Řešení v jednotlivých tagsetech se rozcházejí. 2) Jde o to, zda tvaru se záporkou *ne-* přiřadit lemma infinitiv a) s nebo b) bez ní. Chápání záporky *ne-* jakožto slovotvorného modifikačního prefixu hovoří pro řešení a), praxe slovníků pro řešení b).

Atributy slovesného tvaru a jejich hodnoty

Slovesa se funkčně i formálně liší od jmen. Po formální stránce mají jména (až na výjimky – plurália tantum, názvy jednotlivin atd.) paralelní paradigma singuláru a plurálu vyjadřující gramatické kategorie rodu, čísla a pádu. Oproti tomu slovesa disponují několika samostatnými tvarovými paradigmaty (srv. např. Romportl, 1970, Karlík, Nekula, Rusínová, 1995). Tvary různých subparadigmat jsou nositeli různých významů souborů gramatických kategorií. Z toho plyne i odlišnost v přístupu k morfologickému značkování sloves.

Tagy, jak bylo řečeno výše, mají být dostatečně jednoduché a teoreticky co nejméně zatížené. Porovnejme nejprve přístupy v tagsetech *HA* (SYN2000/ SYN2005), *ajky* (DESAM, KSK) a v návrhu tagsetu pro SNK.

Slovesný tvar

Nejdříve se pokusíme shrnout to, co mají jednotlivé tagsety společné.

V mluvnických popisech se explicitně nebo implicitně pracuje s těmito subparadigmaty: 1) soustava tvarů (dále ST) indikativu prézentu (futura) / (formálního prézentu) aktiva, 2) ST imperativu, 3) ST participia minulého (l-ového), 4) ST participia pasivního (n-/t-ového), 5) ST přechodníku přítomného, 6) ST přechodníku minulého a 7) tvarem (tvary) infinitivu. Samostatně se pozornost věnuje tvarům futura slovesa *být* (*budu, ...*) užívaným pro tvoření analytického futura, tvarům *bych, bys, by ...,* jejich genetické spřízněnosti se slovesem *být* i současnému trendu k poklesu na „volný morfém“ slovesného tvaru kondicionálu. Bývá zmíněno anomální tvoření syntetických tvarů futura (prefix *po-* + tvary indikativu prézentu některých nedokonavých sloves pohybu a některých dalších).

Praktická řešení

V tagsetech automatických analyzátorů i v návrhu pro SNK se pracuje s jednotlivými slovesnými subparadigmaty.

Poznámka: V prvním sloupci je uvedena konkrétní podoba značky (znaky „.*“ jsou tzv. regulární výrazy a označují libovolné opakování libovolného počtu znaků). Ve druhém sloupci je uveden slovní popis hodnoty atributu slovní druh (sloveso/deverbativní adjektivum), ve třetím sloupci je uveden slovní popis definující hodnoty atributu pro subklasifikaci soustav slovesných tvarů.

tab. 1 - SYN2000/SYN2005

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, ...*)
AG.*	adjektivum	adj. odvozené od přech. přít.
AM.*	adjektivum	adj. odvozené od přech. min.
J,.*	spojka	spojky podřadicí vč. *a-/kdy-by,...*

tab. 2 – DESAM/KSK*

značka (tag)	slovní druh (atribut k)	slovesný tvar (atribut m)
k5.*mF	sloveso	infinitiv
k5.*mI	sloveso	formální indikativ (préz./fut.)
k5.*mR	sloveso	imperativ
k5.*mA / *k5.*mA.*zS	sloveso	l-ové participium (vč. tvarů s *–s)*
k5.*mN	sloveso	n-/t-ové participium
k5.*mS	sloveso	přechodník přítom.
k5.*mD	sloveso	přechodník minulý
k5.*mB	sloveso	tvary *budu, budeš, bude, ...*
kY.*mC	kondicionálová částice	tvary *bych, bys, by, a-/kdy-bych, ...,*
k2.*	adjektivum	adjektivizovaná part. = adjektiva

Poznámka: *Pouze ve variantě analyzátoru použité pro automatické anotace KSK se pracuje se značkou pro volný morfém *–s* signalizující význam 2. os. sg. (srv. více Hlaváčková, Sedláček, 2006, Hladká a kol., 2005, Osolsobě, 2006).

tab. 3 - SNK(návrh)

značka (tag)	slovní druh (1. pozice)	slovesný tvar (2. pozice)
VI.*	sloveso	infinitív
VK.*	sloveso	formálný prézent (indik.)
VM.*	sloveso	imperatív
VH.*	sloveso	prechodník
VL.*	sloveso	l-ové príčastie
VB.*	sloveso	futúrum *byť*, *po-*

značka (tag)	slovní druh (1. pozice)	slovesný tvar (2. pozice)
Gk.*	particípium	aktívne
Gt.*	particípium	pasívne
Y.* YO.*	kondicionálová morféma *by* *aby, keby*	kondicionálová morféma kondicionálová morféma + spojka

Ze srovnání představených systémů vyplývá, že:

1) V návaznosti na tradiční gramatické popisy se ve všech třech srovnávaných systémech rozlišují slovesná subparadigmata jednoduchých slovesných tvarů.

2) Nejednotně se řeší slovnědruhové značkování tvarů *by, aby, kdyby, ...*

3) Nejednotně se řeší značkování slovesných tvarů futura slovesa *být (budu, budeš, ...)*.

4) Nejednotně se řeší značkování tvarů syntetického futura *(pojedu, ...)*.

5) Nejednotně a ne vždy důsledně se řeší značkování nesamostatného morfému *–s* signalizujícího význam 2. osoby sg.

6) DESAM, KSK opomíjejí značkování okrajových archaických tvarů.

7) Nejednotně se řeší značkování adjektivizovaných přechodníků.

Poznámka: Otázku značkování deverbativ (adjektivizovaných přechodníků) ponecháme nadále stranou.

Společné rysy a rozdíly

Všechny tři tagsety mají shodně samostatně subparadigma infinitivu a imperativu. (K tvaru infinitivu srv. níže odd. **Vid** a **Negace.)** K samostatnému imperativnímu subparadigmatu srv. níže odd. **Slovesný tvar a slovesný způsob**.)

Všechny tři tagsety shodně zachycují subparadigma indikativu prézentu aktiva nedokonavých a indikativu prézentu/futura aktiva (formálního prézentu) dokonavých sloves. Systém používaný ke značkování korpusu SYN2000/SYN2005 má na rozdíl od systému DESAM, KSK a SNK samostatnou značku pro archaické tvary prézentu (futura) indikativu (*jsemť, nebuduť, ...*) a pro archaické tvary příčestí minulého (*zůstalať, ...*).

Všechny tři tagsety shodně zachycují subparadigma tvarů l-ového participia (v popisech značek se drobně odlišují v terminologii – srv. výše tab. 1 – 3). Tagsety korpusů SYN2000/SYN2005 a KSK mají na rozdíl od DESAM ve značce uvedeno, že součástí tvaru je nesamostatný morfém *–s* signalizující význam 2. osoby singuláru. SYN2000/SYN2005 pouze tehdy, je-li *-s* a) součástí slovesného tvaru (např. *musel**s**, bita**s**)*, b) u tvaru osobního zájmena (*ty**s***), c) u tvarů *se**s**/si**s**,* d) u tvarů *kdo**s**, koho**s**, komu**s**, kým**s*** a e) u tvaru *že**s***. V těchto případech je ve značce na pozici pro vyznačení atributu osoby uvedena hodnota *2. osoba* a na pozici pro vyznačení atributu čísla uvedena hodnota singulár (srv. podrobněji níže odd. **Osoba, Číslo**). KSK má u těchto i řady dalších tvarů, kde se nesamostatný morfém *-s* vyskytuje (*se**s**, si**s**, že**s**, ty**s**,* sloveso v l-ovém participiu + *-s*, *kdo**s**, co**s**, kam**s**, kde**s**, ...,* atd.), signalizovánu jeho přítomnost (atribut *Z* s hodnotou *s*) bez explicitního udání dalších hodnot atributů, které by mohla přítomnost *–s* implikovat.

Poznámka: Vychází se z předpokladu, že hodnoty atributu osoba a číslo jsou nutně 2. osoba a singulár. Ostatní hodnoty mohou být ambiguitní a bylo by třeba přesně popsat, jak by se měly určovat (zda by se měly brát v úvahu hodnoty složeného tvaru, jehož je *–s* součástí, čili nic). Z příslušného způsobu popisu by pak vyplynuly důsledky pro desambiguaci. Tento postup je komplikovaný. Dává se tedy přednost jednoduchému řešení (srv. níže odd. **Osoba, Číslo, Čas a slovesný tvar, Rod a slovesný tvar, Negace**).

V tagsetech korpusů DESAM, KSK je na úrovni atributu pro subklasifikaci ST hodnota subparadigma futura slovesa *být (budu, budeš, ...)*. Tagset navržený pro SNK počítá s využitím téže kombinace značek pro tvary syntetického futura slovesných tvarů tvořených prefixem *po- (pojedem, pojdem, ...*). Totéž pojetí měl v návrhu i jedna z dřívějších variant tagsetu *ajky*.

V tagsetu použitém pro značkování SYN2000/SYN2005 není subparadigma futura zachyceno samostatnou značkou. Význam tvarů *budu, budeš, ...* a tvarů syntetického futura (*pojedu, ...*) je signalizován na úrovni hodnoty atributu čas (srv. níže odd. **Čas a slovesný tvar**).

Tagsety korpusů užitých pro značkování korpusů DESAM, KSK a návrh pro SNK mají tvary kondicionálového *bych, ..., a-/kdy-bych, ...* označeny jako samostatný slovní druh (DESAM/KSK : **kY**; SNK : **Y** pro tvar *by***, YO** pro tvary *a-/ke-by*). DESAM a KSK má v rámci značky u tvarů *by, aby, kdyby, ...* vyznačeny další hodnoty: u atributu slovesný tvar hodnotu kondicionál a příslušné hodnoty atributů osoba a číslo (srv. níže odd. **Způsob a slovesný tvar)**.

Tagset používaný pro SYN2000/SYN2005 má u tvarů *bych, bys, by, bychom, byste* hodnotu atributu slovní druh „sloveso“ (*V*), hodnotu atributu ST „kondicionál slovesa *být“* (*c*). Tvary *a-/kdy-bych, -bys ,-by, ...,* mají hodnotu atributu slovní druh konjunkce (***J***), hodnotu atributu ST (kondicionálová platnost) není na úrovni morfologické značky nijak zachycena. Ve značce jsou ovšem uvedeny hodnoty atributů osoba a číslo (podrobněji viz níže odd. **Osoba, Číslo**).

Značkování dalších hodnot slovesných tvarů

Poznámka: Pro každý atribut jsme sestavili tabulku, do níž zanášíme, zda vůbec a jak jej vymezují srovnávané tagsety (sloupce) u jednotlivých slovesných tvarů (řádky). „+“ značí, že se atribut u slovesného tvaru určuje. „(+)“ je uvedeno, pokud v příslušném tagsetu chybí hodnota příslušného atribut slovesný tvar. „0“ značí, že slovesný tvar se systémově liší a je značkován na jiné úrovni. „-“ značí, že se hodnota atributu neurčuje.

Osoba

Všechny zkoumané tagsety rozlišují v souladu s tradičními gramatikami tři gramatické hodnoty atributu osoba.

Poznámka: Automatické morfologické analyzátory *HA* i *ajka* vycházejíce z tabulkových údajů uvedených v naprosté většině českých mluvnic analyzovaly tvary *by*, *aby, kdyby* výhradně jako tvary s hodnotou 3. osoby. Nebraly v úvahu zvratné tvary sloves, u nichž je *–s* pravidelně ve druhé osobě sg. připojeno k zvratnému *se/si* (*nasmál by ses/dal by sis*). Tento postup je patrný v anotacích SYN2000 a v SYN2005 je patrný pokus tuto chybu eliminovat (viz níže).

SLOVESNÝ TVAR	SYN2000** /SYN2005	DESAM**/KSK**	SNK
infinitiv	-	-	-
imperativ	+	+	+
formální indikativ	+	+	+
l-ové part.	-/+****	-	+
n-/t-ové part.	-/+****	-	0***
přechodník přítomný	-	-	-
přechodník minulý	-	-	0
budu, ...	(+)	+	+
po-,	(+)	(+)	+
by, ...	+/-*	+**	+
aby, kdyby ...	+/-*	+**	+

*Poznámka: V SYN2005 se hodnota osoby určuje pouze u tvarů *(a-/kdy-)- bych, -bys, -bychom, byste*. Tvary *(a-/kdy-)- by* mají vyplněnu hodnotu „neurčuje se“. (Jde o pokus odstranit chybu, jíž je výlučná interpretace tvarů *by, a-/kdy-by* jakožto tvarů 3. osoby. Toto řešení je problematické, bylo by tudíž vhodné na ně upozornit v manuálech uživatelů korpusů a výhledově je odstranit.

** Poznámka: U tvarů *by, aby, kdyby* nabízí automatická morfologická analýza (*ajka*) pouze interpretaci 3. osoba. Případy typu *zasmála by ses/dala by sis* automatický analyzátor ignoruje. Na úrovni ruční desambiguace korpusů DESAM a KSK nebyla tato chyba odstraněna (srv. Osolsobě, 2006). Jde o chybu, která by se měla odstranit.

***Poznámka: Slovenskými ekvivalenty jsou v tagsetu „formálné participia“.

****Poznámka: Pokud je součástí tvaru příslušného příčestí volný morfém *-s*, pak je u atributu osoba uvedena hodnota 2. osoba.

Tagsety použité pro SYN2000/SYN2005, DESAM a KSK na rozdíl od návrhu tagsetu SNK kategorii osoby vyznačují nejen u sloves, ale i u osobních zájmen. Tagset SNK navrhuje naopak rozlišování kategorie osoby u tvarů l-ových příčestí ve všech složených slovesných tvarech minulého času a kondicionálu, což by zajisté znesnadnilo desambiguaci.

Tagset SYN2000/SYN2005 rozlišuje kategorii osoby u některých (srv. výše) tvarů, jejichž součástí je nesamostatný morfém *-s* signalizující hodnotu 2. osoby sg. (*musels, tys, žes, …*).

Poznámka: Problematické ovšem je, že jde pouze o některé (frekventované) případy. V SYN2000/SYN2005 jsou tvary *kdos, kohos, komus, kýms* označkovány jinak než analogické vary *cos, čehos, čemus, čím*s nebo tvary *kdes, kdys, jaks, tehdys, ...* atd.

Značka tvarů *ty/tys* na rozdíl od značek tvarů *ses, sis, žes, kohos, ...* zahrnuje další hodnoty slovesných atributů interpretovatelných do značky z přítomnosti volného morfému *-s* (srv. odd. **Čas a slovesný tvar, Rod a slovesný tvar, Negace**). Toto řešení není úplně transparentní.

Číslo

Všechny zkoumané tagsety rozlišují v souladu s tradičními gramatikami dvě hodnoty atributu číslo u sloves.

SLOVESNÝ TVAR	SYN2000/SYN2005	DESAM/KSK	SNK
infinitiv	-	-	-
imperativ	+	+	+
formální indikativ	+	+	+
l-ové part.	+	+	+
n-/t-ové part.	+	+	0
přechodník přítomný	+	+	-
přechodník minulý	+	+	0
budu, ...	(+)	+	+
po-,	(+)	(+)	+
by, ...	+/-*	+	+
aby, kdyby ...	+ /-*	+	+

*Poznámka: V SYN2005 se hodnota čísla nevyplňuje u tvarů *by, aby, kdyby* (viz výše).

Tagset SYN2000/SYN2005 rozlišuje kategorii čísla u (srv. výše) tvarů, jejichž součástí je nesamostatný morfém *-s* signalizující hodnotu 2. osoby sg. (*žes, ...*).

Vid

Tagset používaný pro značkování SYN2000 nemá ve značkách zachyceno určení atributu vid. V tagsetu SYN2005, DESAM, KSK a návrhu pro SNK se rozlišují shodně tři hodnoty pro označení atributu vid (dokonavost, nedokonavost, obouvidovost). Atribut vid se určuje u všech slovesných tvarů. V *HA* a *ajce* jeho hodnoty odpovídají klasifikacím přiřazeným tvarům na úrovni slovníku automatického analyzátoru. V praxi to znamená, že jak *HA*, tak *ajka* přiřazují tvarům *(ne)-napíš-u, -eš, .., napiš, ..., napsal, ... napsán* lemma *napsat* a tvaru *(ne)-píš-u, -eš…, piš, ..., psal, ..., psán, ...* lemma *psát.* V návrhu tagsetu pro značkování SNK je patrný rozdíl od tohoto pojetí. Stojí zde, že lemmatem tvarů *píšem* i *napíšem* má být *písať*, přičemž se bude vycházet z toho, co uvádí Krátky slovník slovenského jazyka (srv. Giantisová, 2005). U přídavných jmen odvozených od přechodníků (SYN2000/SYN2005) a u „gerundív“ v SNK se hodnota vidu neurčuje.

Rod a slovesný tvar

Samostatnou značku pro určení kategorie slovesného rodu má pouze tagset používaný pro značkování SYN2000/SYN2005.

SLOVESNÝ TVAR	SYN2000/SYN2005	DESAM/KSK	SNK
infinitiv	-	-	-
imperativ	-	-	-
formální indikativ	+	-	-
l-ové part.	+	-	-
n-/t-ové part.	+	-	0

SLOVESNÝ TVAR	SYN2000/SYN2005	DESAM/KSK	SNK
přechodník přítomný	-	-	-
přechodník minulý	-	-	0
budu, ...	(+)	-	-
po-,	(+)	-	-
by, ...	-	-	-
aby, kdyby	-	-	-

Poznámka: Hodnoty tohoto atributu jsou v tagsetu SYN2000/SYN2005 charakterizovány takto: 1) neurčuje se (-), 2) aktívum nebo nikoliv 'pasívum' (***A***), 3) pasívum (***P***).

Klademe si otázku, k čemu takovéto značkování slouží. Sondou do SYN2000/SYN2005 totiž zjistíme, že vyznačení aktívum nebo nikoliv pasívum mají jednak slovesné tvary, které nejsou slovesnými tvary pasivního příčestí (všechny tvary, které mají ve značce uvedeno, že jde o slovní druh sloveso a přitom nejde o tvar pasivního příčestí (n-/t-ového participia) a navíc všechny tvary zájmena *ty*, jehož součástí je nesamostatný morfém *-s* signalizující hodnotu 2. osoby - *tys* (*tys/ty/PP-S1–-2P-AA-*).

Poznámka: Pokud bychom ovšem chtěli interpretovat hodnotu atributu slovesného rodu volného *–s* důsledně, pak bychom museli řešit problém, jakou hodnotu bude mít u *tys* v případech jako *…tys/ty/PP-S1–-2P-AA- od kořene do vršku změněna/změnit/VsQW---XX-AP---…, … tys/ty/PP-S1–-2P-AA- rozprostřen/rozprostřít/VsYS---XX-AP---* … (SYN2000).

Tvary *ses, sis, žes, kohos, …* mají u hodnoty kategorie aktívum/pasívum vyznačenu hodnotu „neurčuje se". Proč tomu tak je, není dosti transparentní.

Atribut pro hodnotu *pasívum* (***P***) mají vyplněn pouze tvary pasivního příčestí (n-/t-ového participia včetně tvarů s volným morfémem *–s* signalizujícím hodnotu 2. osoby sg.). Vzhledem k tomu, že již na úrovni atributu ST (detailní určení slovního druhu – 2. pozice) je uvedeno, že jde o tvar pasivního participia, jde dle našeho názoru o veskrze nadbytečnou informaci.

Způsob a slovesný tvar

Žádný ze sledovaných tagsetů nemá samostatný atribut pro hodnotu slovesného způsobu. Je tomu tak nejspíše proto, že hodnoty lze do jisté míry vyčíst z hodnot atributu ST. Z lingvistického hlediska by bylo zajisté zajímavé mít k dispozici ve značkování nástroj, který by pomohl uživateli rozsáhlých elektronických korpusů vyhledat např. potencionální tvary zvratného pasíva (označkování tranzitiv, reflexiv tantum atd.).

Neproblémově se jeví značkování tvarů imperativu.

Poznámka: Upozorňujeme ovšem, že pokud se bude věnovat pozornost slovesům, která mohou tvořit syntetické futurum pomocí prefixu *po-* (srv. níže odd. Čas a slovesný tvar), pak by se nemělo zapomenout na příslušné imperativní tvary a případná omezení jejich tvoření (*pojď, pojeď, poleť, … ?popal, ?poveď, …*).

Jak bylo řečeno výše, nejproblematičtěji se jeví značkování kondicionálového morfému. Eklekticky působí řešení v tagsetu použitém v SYN2000/SYN2005.

Jednoduše je problém vyřešen v návrhu SNK, zde je ovšem třeba přihlédnout i k odlišné jazykové situaci (ve slovenštině jde o morfém signalizující pouze právě kondicionál, kategorie osoby a čísla nejsou na tvaru formálně signalizovány). Řešení *ajky* se jeví přijatelně za předpokladu, že budou z automatické analýzy odstraněny chybné interpretace tvarů *by, aby, kdyby* jakožto tvarů, které mají u atributu osoba výhradně hodnotu 3. osoba.

Čas a slovesný tvar

Čas se jako samostatný atribut označuje pouze v tagsetu použitém pro SYN2000/SYN2005. Tagset pro DESAM, KSK i návrh pro SNK rezignují na určování hodnot času na úrovni značek omezených na jednotlivé jednoduché tvary (srv. výše odd. **Lematizace a značkování slovesných tvarů**).

SLOVESNÝ TVAR	SYN2000/SYN2005	DESAM/KSK	SNK
infinitiv	-	-	-
imperativ	-	-	-
formální indikativ	+*	-	-
l-ové part.	+**	-	-
n-/t-ové part.	+***	-	0
přechodník přítomný	-	-	-
přechodník minulý	-	-	0
budu, ...	0	-	-
po-,	0	-	-
by, ...	-	-	-
aby, kdyby	-	-	-

Poznámka: Hodnoty času jsou v tagsetu SYN2000 (SYN2005) charakterizovány takto: 1) neurčuje se (-), 2) futurum (budoucí čas) - (*F*), 3) minulost nebo přítomnost - (*H*), 4) prézens (přítomný čas) (*P*), 5) minulý čas (*R*) a 6) libovolný čas (*X*).

*V tagsetu používaném pro anotaci SYN2000/SYN2005 mají budoucí čas (*F*) vyznačeny tvary budoucího času slovesa *být* používané též pro tvoření analytických tvarů futura (*budu, budeš, budeš, ...*) a tvary některých nedokonavých sloves, které mohou tvary s prefixem *po-* + indikativ prézentu aktiva tvořit tzv. syntetické futurum.

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

Problematické se nám zdá, že v rozsáhlém materiálu SYN2000/SYN2005 je poněkud nejednoznačně vyřešeno značkování celé řady tvarů syntetického futura.

Poznámka: Značku tag=“VB……F.*“ mají v SYN2000/SYN2005 pouze tvary následujících sloves: *být/bude, jít/půjde, jet/pojede, nést/ponese, běžet/poběží, letět/poletí, téct/poteče, vézt/poveze, hrnout/pohrne (se), lézt/poleze, plout/popluje, cestovat/pocestuje, stěhovat/postěhuje, řítit/pořítí (se),* a v SYN2005 navíc tvary *trvat/potrvá* a *plazit/poplazí (se)*. Značku tag=“Vi……-.*“ a lemma bez *po-* mají v SYN2000 a SYN2005 tvary sloves *jít/pojď, slyšet/poslyš, jet/pojeď, lézt/polez, běžet/poběž, letět/poleť*, v SYN2005 navíc i tvary *nést/pones, stěhovat/postěhuj*.

Ze sond do materiálu SYN2000/SYN2005 je patrné, že by sem měla patřit ještě řada dalších slovesných tvarů. Následující tabulka mapuje tvary sloves na *po-* + koncovky ind. préz. akt. / imp. v obou sledovaných korpusech. V pravém sloupci jsou uvedena slovesa, u kterých jsou tvary s prefixem *po-* doložené v sledovaných korpusech výhradně tvary syntetického futura. V levém sloupci jsou slovesa, u nichž tvary na *po-* mohou být podle kontextu buď tvary syntetického futura, nebo tvary odvozených (prefigovaných) sloves většinou s distributivním významem. Čísla v závorkách jsou orientační. Uvádějí počty nalezených dokladů v SYN2000/SYN2005. Ve druhém sloupci se uvádějí pouze celkové součty.

tvary *po-* jsou výhradně tvary synt. futura	**tvary *po-* nejsou výhradně tvary syntetického futura**
putovat/putoval/putuje/poputuje (bude putovat) (171/212)	vést/vedl/vede/povede (bude vést) (se) : povést/povedl/povede (se) (3544/25+2957*)
plynout/plynul/plyne/poplyne (bude plynout) (70/47)	trvat/trval/trvá/potrvá (bude trvat): ?potrvat/ ?potrval/?potrvá (3242/1849*)
plavat/plaval/plave/poplave (poplav) (bude plavat) (51/41)	růst/rostl/roste/poroste (bude růst) : porůst/ porostl/poroste (802/712)
kvést/kvetl/kvete/pokvete (bude kvést) (24/41)	hnát/hnal/žene/požene (bude hnát) (se) : pohnat/pohnal/požene (93/140)
vandrovat/vandroval/vandruje/povandruje (bude vandrovat) (8/4)	valit/valil/valí/povalí/(bude valit) (se) : povalit/ povalil/povalí (29/37)
vládnout/vládl/vládne/povládne (bude vládnout) (6/11)	mazat/mazal/maže/pomaže/ (bude mazat): pomazat/pomazal/pomaže (23/89)
mašírovat/mašíroval/mašíruje/pomašíruje (bude mašírovat) (4/7)	šlapat/šlapal/šlape/pošlape (bude šlapat): pošlapat/pošlapal/pošlape (16/29)
vanout (vát)/vál(vanul)/vane/povane (bude vanout) (1/6*)	vléct/vlekl/vleče/povleče (bude vléct) (se): povléct/povlékl se/povleče (se) (22/38)
pádit/pádil/pádí/popádí (bude pádit) (0/3*)	šinout/šinul/šine/pošine (bude šinout) (se)/(si to): pošinout/pošinul/pošine (si) (2/4)
hasit/hasil/hasí/pohasí (bude hasit) (si to) (0/3*)	pást/pásl/pase/popase (bude pást) (se) : popást/ popásl/popase (se) (1/4)
plavit/plavil/plaví/poplaví (bude plavit) (se) (0/3*)	
řinout/řinul/řine/pořine (bude řinout) (se) (0/3*)	
krátčet/krátčel/krátčí/pokrátčí (bude krátčet) (2/0)	
kutálet/kutálel/kutálí/pokutálí (bude kutálet)(se) (1/1)	
fičet/fičel/fičí/pofičí (bude fičet) (1/1*)	
pelášit/pelášil/peláší/popeláší (bude pelášit) (0/1*)	
plížit/plížil/plíží/poplíží (bude plížit) (se) (0/1*)	

*Mezi korpusy SYN2000 a SYN2005 je ve značkování některých tvarů patrný jistý posun. Týká se značkování tvarů *po-+ved-u, -eš…, po-trv-ám, -áš, ….* V korpusu SYN2005 jsou na úrovni lemmatu rozlišeny tvary od *vést/vést(se)* (25 tvarů) a *povést se* (2957 tvarů). Na první pohled je ale velké procento tvarů desambiguováno chybně. Další nedůslednost spočívá v tom, že tvary, které mají jako lemma uveden tvar *vést* jsou označeny jako tvary s hodnotou prézens (*povede/vést/VB-S---3**P**-AA---I*), což neodpovídá praxi u slovesných tvarů *ponese, poběží*, … Ta mají ve značce na příslušné pozici uvedenu hodnotu futurum (*ponese/nést/VB-S---3**F**-AA---I*). Správně je ve značce

odlišeno, že se jedná o slovesné tvary různého vidu (tvary s lemmatem *vést* vidu nedokonavého, tvary s lemmatem *povést [se]* vidu dokonavého). Jiná situace je u tvarů *potrvá/trvat/VB-S---3F-AA---I*, které na úrovni lemmatu jsou výhradně řazeny k lemmatu *trvat* (1849 tvarů). V SYN2005 lze ovšem nalézt celkem 9 tvarů *potrval/potrvat* lemmatizovaných tvarem *potrvat*. Tento malý počet dokladů pochází z různých zdrojů a vyvolává otázku, zda i mezi tvary *potrvám ,.., potrvají* nejsou takové, které by mohly patřit k lemmatu *potrvat,* nebo zda jsou tvary *potrval/potrvat* nekorektně utvořenými, nicméně v korpusu doloženými tvary. V SYN2000 jsou tvary *povane, pokutálí, pofičím* označkovány jako neznámý slovní druh. V SYN2005 jsou tvary *povane* označkovány jako zájmeno, částice, popř. slovesný tvar od lemmatu *povát*. Tvar *pokulí* jako tvar slovesa *pokulit.* Tvary *popádí, pohasí, poplaví, pořine, popeláší, pofičí* jsou v SYN2005 označkovány jako adjektiva, substantiva, adverbia, lemmatem je tvar sám. Někde je patrné, že jde o chybu způsobenou homonymií (*popádí - povodí Pádu*), v jiných případech není ovšem jasné, na které úrovni chyba vzniká.

Přítomný čas (*P*) mají kromě tvarů, které mají na úrovni slovesného tvaru vyplněno, že jde o slovesný tvar přítomného nebo budoucího času (*VB.**) nebo o archaické slovesné tvary přítomného a budoucího času (zakončení *-ť*) (*Vt.**) vyplněny tvary *tys*. Tvary *ses, sis, žes, kohos, ...* mají u atributu čas vyznačenu hodnotu „neurčuje se“.

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

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

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

Poznámka: Výše popsaná řešení zůstávají do jisté míry na půli cesty a nejsou průhledná. a) Mají-li všechny tvary l-ového příčestí s volným *–s* (*přišels*) na úrovni atributu čas hodnotu minulost, b) tvary pasivního participia s volným *–s* hodnotu přítomnost nebo minulý čas (*minuto**s*** *rozumem = minuto* ***jsi*** *rozumem*, *minuto**s*** ***bylo*** *rozumem = minuto* ***jsi bylo*** *rozumem*) a c) všechny tvary *ty* s volným *–s* (*tys*) hodnotu prézens, pak by mělo být explicitně řečeno, proč se u všech

ostatních tvarů s volným *–s* význam (hodnota) čas „nevyplňuje“, ačkoliv bychom mohli předpokládat, přinejmenším ambiguitní hodnotu minulost nebo přítomnost (*...žes vyhoštěn ze sedadel knížecích…,* ?*…to by ses před ní ukázal v pěkném světle…*, *?… přesto by ses byl se mnou vyspal…* srv. SYN2000). Pokud by platilo pouze a) a b), pak by se nabízela tato odpověď: Protože ostatní tvary nejsou tvary, které mají na první pozici ve značce vyznačenu hodnotu slovní druh *V* (sloveso). Vzhledem k tomu, že platí také c), je patrné, že se hodnoty slovesných atributů neurčují vždy jen u tvarů, které mají hodnotu atributu slovní druh sloveso. Z toho plyne, že v některých případech se u tvarů s volným *–s* v hodnotách uvedených na úrovni morfologické značky odrazí hodnoty implikované jeho přítomností a v jiných ne. Hodnota atributu čas je u tvarů *tys* vždy prézens, ačkoliv bychom opět očekávali přinejmenším ambiguitní hodnotu minulost nebo přítomnost (*…tys přišel …, ...tys tam byl přijat…*, *... tys nejlepší filozof …, ... tys ta jediná …,* SYN2000). Nemýlíme-li se, nejedná se o chybu statistické desambiguace, ale o chybu způsobenou již na úrovni automatické morfologické analýzy.

V analyzátoru *ajka* se na úrovni morfologické značky signalizuje pouze přítomnost/nepřítomnost volného *–s* (srv. více Hladká a kol., 2005, Osolsobě, 2006). Toto řešení lze doporučit (viz výše diskuse k redundanci a ambiguitě hodnot atributu čas v *HA*).

Negace

Všechny tagsety uvádějí atribut negace. U tohoto atributu mají hodnotu signalizující přítomnost/nepřítomnost prefixu *ne-* vyplněny všechny tvary, které mají na úrovni automatické morfologické analýzy ve strojovém slovníku příslušného analyzátoru vyznačenu možnost tvořit tvar s prefixem *ne-*. Konkrétní značkování některých negativ tantum v SYN2000/SYN2005 ukazuje, že toto technické řešení má svá úskalí na úrovni lemmatizace (srv. *HA* generované lemmatizace *nezbytnost/zbytnost, nezbytný/zbytný, nezbytně/zbytně, ...* na straně jedné a neanalyzované tvary *nenenáviděl* na straně druhé (viz SYN2000/SYN2005)), jimiž se ovšem na tomto místě nebudeme zabývat.

Poznámka: Návrh tagsetu pro SNK se snaží omezení dané tokenizací překonat návrhem, aby i tvary slovesa *je*, kdy *nie* stojí samostatně (případ *nie* ***je***), měly ve značce signalizován hodnotu „negace“.

V SYN2000 SYN2005), DESAM, KSK je lemmatem tvarů sloves s prefixem *ne-* tvar bez *ne-*, což odpovídá praxi tištěných slovníků (slovníky slovesa tvořená negativními prefixy jako samostatná heslová slova běžně neuvádějí). Návrh tagsetu pro SNK počítá s tím, že slovesné tvary s prefixem *ne-* budou mít lemma s prefixem *ne-*.

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

Závěr

Z porovnání dvou tagsetů (*HA* a *ajky*) a tagsetu navrhovaného pro SNK plyne, že ačkoliv neexistují zásadní rozdíly, jednotlivá řešení se a) liší, b) obsahují chyby. Tagset *HA* zachycuje na úrovni značky více informací, nicméně jednotli-

vá řešení nejsou úplně transparentní. Tagset *ajky* zachycuje některé jevy konzistentněji, řadu jednotlivostí ponechává ovšem stranou. Návrh tagsetu SNK nabízí některá jednoduchá a přijatelná řešení (viz výše), srovnáváme ovšem pouze návrh, nikoli praktickou aplikaci. Je totiž patrné, že každý korpus je širší než sebelépe navržený tagset. Porovnání jednotlivých řešení může ovšem posloužit k optimalizacím automatického morfologického značkování na rovině teoretické i praktické.

Pozornost by se při případné úpravě tagsetů měla zaměřit na řešení následujících problémů: 1) konzistentní řešení lemmatizace a značkování tvarů s *–by–,* 2) odstranění netransparentních řešení označování jednotlivých gramatických kategorií slovesných u tvarů s volným morfémem *–s*, 3) zjednodušení a zprůhlednění označování hodnot času a slovesného rodu a 4) oprava lemmatizace a značkování tvarů syntetického futura.

Na tomto místě bychom rádi zdůraznili, že první tři body spadají mezi jevy obecnější povahy, které představují v automatické morfologické analýze známý problém. Zjednodušeně můžeme hovořit o „případech, kdy jedna nerovná se jedné". Sem se řadí velmi různorodé jevy, které bychom mohli opět metaforicky označit jako případy, kdy „jedna se chápe jako více než jedna, ale v automatické morfologické analýze je tomu jinak". (Na obdobné nedostatky, s nimiž se setkáváme při značkování tvarů *–by–* a tvarů s *–s*, narážíme u značek přidělených automatickou morfologickou analýzou *HA* zájmenným spřežkám typu *oň, ..., oč,...*, a půjdeme-li ještě dále, narazíme na velmi rozsáhlou oblast zkratek všeho druhu.) V rámci tohoto článku nám šlo především o to, upozornit na nejednotné zachycení analogických jevů. Konkrétním problémům u značkování hodnot atributů ST u tvarů *a-/kdy-by,* ... jsme se podrobněji věnovali v odd. **Společné rysy a rozdíly.**

Druhou stranou téže mince je otázka lingvisticky adekvátnějšího značkování a disambiguace víceslovných jednotek (budeme-li se držet naší metafory, půjde o případy, kdy „více než jedna se běžně chápe jako jedna, ale v automatické morfologické analýze je tomu jinak"). Na tomto místě nelze probírat širokou škálu problémů, které bývají v souvislosti s anotováním korpusů označovány jako multiword expressions (MWE). V našem textu jsme na ni poprvé upozornili v odd. **Lemmatizace a značkování slovesných tvarů** v souvislosti s chápáním „slovního tvaru" (token/atribut *word*) obecně a s dopadem takového pojetí na morfologické značkování tvarů sloves konkrétně. Má-li být toto zjednodušující technické řešení zároveň transparentní, mělo by být buď jednoduché, nebo by se mělo přesně deklarovat, co se čím míní, a to i na úrovni široce používaných manuálů značkovaných korpusů. Detailní rozbor na jedné straně redundantního a na druhé straně vágního značení hodnot času a slovesného rodu v tagsetu i praxi *HA,* které jsme podrobně popsali v oddílech **Čas a slovesný tvar** a **Rod a slovesný tvar,** těmto požadavkům neodpovídá. Pokud totiž budeme usilovat o to, abychom na úrovni morfologické značky jednotlivých složek zachytili hodnoty, které interpretujeme u složeného slovesného tvaru až na základě jeho „kompletace", pak bude třeba přesně určit, která složka nese odpovědnost za tu kterou hodnotu, a poté vyznačovat hodnotu u příslušných složek jednotně a závazně.

Tento přístup by v praxi kladl vyšší nároky na desambiguaci a teoreticky by jej nebylo možno vytrhnout z širších souvislostí (viz výše). Druhou variantou je rezignace na snahu zachytit na úrovni značkování automatickým morfologickým analyzátorem ty hodnoty, které lze vždy jednoznačně určit až na vyšší úrovni analýzy. Praxe takového přístupu se uplatňuje v morfologickém analyzátoru *ajka*, figuruje i v návrhu SNK a její přijetí by mohlo být plodné.

Problémy značkování tvarů syntetického futura (viz výše bod 4) řeší jednoduše a přijatelně návrh tagsetu SNK přidáním hodnoty „slovesný tvar futurum (*budu, budeš, ...*) a tvary syntetického futura“ na úrovni značkování slovesných subparadigmat (srv. podrobněji odd. **Praktická řešení**). Vzhledem k tomu, že oba české analyzátory značkují tyto tvary problematicky (*ajka* vůbec ne a *HA* pouze některé viz odd. **Čas a slovesný tvar**), zdá se, že přijetí řešení návrhu SNK spojené s úpravou dat strojových slovníku obou analyzátorů by mohlo být užitečné.

Na závěr zbývá dodat, že úpravy tagsetů automatických morfologických analyzátorů by jednak otevřely cestu k vzájemné převoditelnosti probíraných systémů, jednak by pro širší uživatelskou obec zprůhlednily vazby mezi tagsety (popisy morfologických značek) a konkrétními tagy.

BIBLIOGRAFIE

Čermák, F., Křen, M.: Frekvenční slovník češtiny. Praha: NLN, 2004. + 1 CD-ROM.

Garabík, R., Gianitsová, L., Horák, A., Šimková, M.: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. <http://korpus.juls.savba.sk/publications/index.sk.html>, 2004.

Giantisová, L.: Morphological Analysis of the Slovak National Corpus. In: Insight into Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2005, s. 166 – 178.

Hajič, J.: Unification Morphology Grammar. Praha : MFF UK, (disert. práce), 1994.

Hajič, J.: Desambiguation of Rich Inflection (Computational Morphology of Czech). Praha : Karolinum, Charles University Press, 2004.

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

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

Hlaváčková, D., Sedláček, R.: Morfologické značkování korpusu soukromé korespondence. In Varia XIV. Bratislava : Slovenská jazykovedná spoločnosť pri SAV, 2006, s. 371–379.

Jelínek, J., Bečka, J. V., Těšitelová, M.: Frekvence slov, slovních druhů a tvarů v českém jazyce. Praha : SPN, 1961.

Karlík, P., Nekula, M., Rusínová, Z.: Příruční mluvnice češtiny. Praha : NLN, 1995.

Komárek, M. a kol.: Mluvnice češtiny II., Praha, Academia : 1986.

Králík, J., Těšitelová M.: Retrográdní slovník současné češtiny. Praha: Academia, 1986.

Mistrík, J.: Retrográdny slovník slovenčiny. Bratislava : Univerzita Komenského, 1976.

Mráková (Žáčková), E., Pala, K.: Corpus-Based Rules for Czech Verb Discontinuous Constituents. In Matoušek V., Mautner P., Ocelíková J., Sojka P. (eds.): Text, Speech and Dialogue 1999. Berlín : Springer Verlag, 1999, s. 325–328.

Mráková, E., Sedláček, R.: From Czech Morphology through Partial Parsing to Desambiguation. In Gelbukh, A. (ed.): Computational Linguistics and Intelligent Text Processing. Berlin : Springer-Verlag, 2003, s. 126–135.

Leech, G.: Corpus annotation schemes, Literary and linguistic Computing 8 (4), 1993, s. 275–281.
Oliva, K., Hnátková, M., Květoň, P., Petkevič, V.: The Linguistic Basis of a Rule-Based Tagger of Czech. In: Matoušek, V. et al. (eds.): Text, Speech and Dialogue 2000. Berlín : Springer-Verlag, 2000, s. 3–8.
Osolsobě, K.: Algoritmický popis české morfologie a strojový slovník češtiny, Brno : FF MU, (disert. práce), 1996.
Osolsobě, K.: Morfologické značkování složených slovesných tvarů v korpusu, SPFFBU A 47, 1999, s. 33 – 50.
Osolsobě, K.: Korpus soukromé korespondence (KSK) z hlediska morfologického značkování. SPFFMU A, 54, 2006, s. 187–201.
Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In Jeffery, K. G., Prášil, F. (eds.): Proceedings of SOFSEM 97. Heidelberg : Springer Verlag, 1997, s. 523–530.
Petkevič, V.: Neprojektivní konstrukce v češtině z hlediska automatické morfologické disambiguace českých textů. In: Hladká, Z. – Karlík, P. (eds.): Čeština – univerzália a specifika 3. Brno : Masarykova univerzita, 2001, s. 197–205.
Petkevič, V.: Grammatical Agreement and Automatic Morphological Disambiguation of Inflectional Languages. In: Matoušek, V. et al. (eds.): Text, Speech and Dialogue 2001. Berlín : Springer-Verlag, 2001, s. 47–53.
Romportl, S.: Struktura gramatické složky slovesných tvarů určitých v češtině. Praha : Academia, 1970.
Sedláček, R.: Morphematic analyser for Czech. Brno : FI MU, (disert. práce), 2004.
Žáčková, E.: Parciální syntaktická analýza (češtiny), Brno : FI MU, (disert. práce), 2004.
Český národní korpus – SYN2000/SYN2005. 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.htm.
http://ufal.mff.cuni.cz/pdt/ (http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/morph.html).
http://korpus.juls.savba.sk

TAGGING OF VERB FORMS IN CZECH (SLOVAK) CORPORA (paralels, differencies, defects, possible upgrade)

The confrontation of two tagsets used for tagging of Czech corpora (SYN2000/SYN2005 and DESAM, KSK i.e. the corpus of private correspondence) and a tagset proposal for SNK (i. e. Slovak National Corpus) doesn't show substantial differences. Nevertheless individual cases differ in 1) lemmatisation and tagging of conditional verb-forms (conditional particle) *-by-*, 2) tagging of free-morpheme *–s* (for 2. person singular i.e. ***ses**, **sis**, **žes**, **tys**,...*), 3) tagging of grammatical category of tense and voice and 4) lemmatisation and tagging of synthetic future forms. The comparison of the different solutions and the analysis of mistakes can be the first step to upgraded and reciprocally convertible tagsets.

Klára Osolsobě
Ústav českého jazyka
Filozofická fakulta Masarykovy univerzity
Arna Nováka 1
602 00 Brno