

Korpus jako zdroj dat pro opravy chyb automatické morfologické analýzy

Klára Osolobě

1. Úvod

Cílem našeho příspěvku je ukázat, jak mohou korpusová data přispět k doplnění popisů gramatických jevů zachycených ve slovnících a gramatikách na straně jedné a v lingvistických bázích automatických morfologických analyzátorů na straně druhé.

Ke značkování řádově stamilionových korpusů se běžně používají automatické nástroje. V prostředí českých korpusů jsou to především dva analyzátory (taggery) pro morfologické značkování. Morfologický analyzátor Jana Hajiče – dále *HA* (Hajič, 1995, 2004, <http://ucnk.ff.cuni.cz/bonito/index.html>) a morfologický analyzátor Radka Sedláčka – *Ajka* (Sedláček, 2005, <http://nlp.fi.muni.cz/projekty/wwwajka>). Lingvistickou bází těchto analyzátorů byly mluvnické a slovníky, o něž se opíraly algoritmické popisy flexe (Hajič 1995, Osolobě, 1996). 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. Vyšlo najevo, že s ohledem na zkušenosti z konkrétní praxe, je třeba některé interpretace zpětně verifikovat.

V našem příspěvku si všimneme tří jevů: 1) syntetické futurum, 2) stupňování adjektiv a adverbíí a 3) slovnědruhové přechody vybraných slov.

2. Syntetické futurum

2.1. Slovníky a tištěné zdroje

Termínem syntetické futurum se v českých gramatikách označují tvary několika desítek českých nedokonavých sloves tvořených od tvarů indikativu přítomnosti aktíva prefixem *po-* (*pů-*) většinou (kromě *jet*, *jít*) paralelně ke tvarům opisného futura. Tato slovesa jednotlivé mluvnické 1) významově vymezují a 2) uvádějí jejich seznamy. Ve slovnících jsou zachyceny nejednotně. SSJČ uvádí celkem 50 sloves. SSČ pouze 21, z toho je 18 těch, které uvádí i SSJČ a 3, která v SSJČ uvedena nejsou. Další 5 sloves navíc uvádí Čermák (Čermák, 1990, s. 106), 2 slovesa navíc uvádí Šmilauer (Šmilauer 1972, s. 223). V uvedených tištěných pracích je doloženo celkem 60 sloves, která mohou v češtině tvořit tvary futura synteticky prefixem *po-* (viz příloha).

2.2. Automatická morfologická analýza

Automatické analyzátory zachycují sledovaný jev takto: *HA* značkuje a lemmatizuje (lemma bez prefixu *po-*, ve značce na pozici 9 – slovesný čas je uvedeno F - futurum) správně pouze tvary 12 sloves v SYN2000 a 14 sloves v SYN2005. *Ajka* tvary syntetického futura lemmatizuje nesprávně (příslušným tvarům s prefixem *po-* uvádí lemma s tímto prefixem).

2.3. Korpusy

Analýzou korpusových dat jsme zjistili, že v dostupných korpusech (SYN2000, SYN2005, SYN2006PUB, BMK, PMK) se vyskytují další tvary sloves (celkem 20 lemmat) dokládající tvoření syntetického futura v češtině nezaznamenané ve zkoumaných tištěných příručkách (viz příloha). Další doklady jsme našli na internetu.

V tabulce uvádíme výsledky analýzy.

tištěné zdroje	60 lemmat	72,29 %
pouze korpusy	20 lemmat	24,1 %
korpusy i tištěné zdroje	37 lemmat	44,58 %
pouze tištěné zdroje	23 lemmat	27,71 %
pouze internet	3 lemmata	3,61 %
Celkem tištěné zdroje + korpusy + internet	83 lemmat	100 %

2.4. Tištěné zdroje a korpusy

České gramatiky (např. Komárek a kol., 1987: 412, 417; Čechová, 1996: 208; Karlík, Nekula, Rusínová, 1996: 317, Šmilauer, 1972: 223) definují sledovanou skupinu sloves jako slovesa vidu nedokonavého s významem pohybu, který dále vymezují jako pohyb „určený“ nebo „jednosměrný“. Dále se v nich uvádí, že může jít také o slovesa označující „stav“, jeho „trvání“ nebo „změnu“. Tyto charakteristiky jsou dosti vágní. Korpusová data pomohou uvedené definice upřesnit. Kromě pohybu jednosměrného/ v určeném směru (*půjde, pojedje, poplází se, pokráčí, pošine se, povalí, pomaže, ...*), může jít i o pohyb různými směry (*pošíří se, poline se*). Procesy zahrnující pohyb lze blíže klasifikovat jako pohyb a) přírodní (*povane*); b) s neživým nositelem (rostlina *pokvete, poroste*, strom *ponese*, hudba *poline se, povalí*, vůně *pošíří se, poline se*, voda: *poteče, povalí, zpráva pošíří se*); c) mentální (*popálí mu to*); d) s živým nositelem (*poroste, popase se, pták ponese*); e) ve významu fungovat (*pošlape, poklape*); vycházet s někým (*poklape, popeče*); f) ve významu vést/vládnout (*povede, povládne, pokoučuje, potrénuje*).

3. Stupňování adjektiv a adverbii

3.1. Stupňování adjektiv a adverbii v gramatikách a slovnících

Gramatiky řadí sufixální tvoření tvarů 2. stupně a prefixálně-sufixální tvoření tvarů 3. stupně jak do kapitol pojednávajících o formální morfologii, tak o slovo tvorbě. Slovníky mapují tvary 2./3. stupně nejednotně. Samostatná hesla mají tvary „nepravidelně“ tvořených komparativů (superlativů). U ostatních adjektiv a adverbii se o tom, zda a jak je příslušné adjektivum/adverbium schopno tvořit tvar 2./3. stupně příslušnými afixy neříká nic. Mluvnice i další odborná literatura (srv. např. Karlík, Nekula, Pleskalová, 2002; Hladká, Karlík, 2004) se snaží definovat nejrozumnější pravidla omezující možnost derivovat zmíněné tvary.

3.2. Automatická morfologická analýza

Automatické morfologické analyzátory mají v návaznosti na gramatiky vypracována v lingvistických bázích (strojových slovnících) pravidla, podle nichž se adjektiva/adverbia dělí podle toho, zda tvoří/netvoří afixálně tvary 2. a 3. stupně. Lingvistická východiska jsou v obou analyzátoch podobná. HA implicitně vychází z předpokladu, že afixálně nelze stupňovat a) subjektivě posesivní adjektiva tvořená sufixy na -ův/-in; b) deverbativa derivovaná od přechodníků; c) jmenné tvary s výjimkou problematicky slovnědruhově zařaditelných tvarů *rád, raději, nejráději*; U řady adjektiv, u nichž lze uvažovat o formálních i sémantických omezeních afixálního stupňování (srv. Karlík, Nekula, Pleskalová, 2002), uvádí HA (SYN2000, SYN2005) značku pro adjektiva stupňovatelná. Jedná se o tyto případy: a) deverbální účelová adjektiva na -cí (*prací*); b) kompozita, jejichž druhým členem je adjektivizovaný přechodník (*rychleschnoucí*); c) desubstantivní druhově posesivní adjektiva na -í (*psí, medvědí, kuřecí*); d) deadverbiální adjektivum (*domácí, letošní*). V analyzátoru *Ajka* se +/- možnost tvořit afixálně tvary 2./3. stupně řeší na úrovni vzorů založených na derivačních typech adjektiv.

3.3. Korpusová data

Tvary 2./3. stupně jsou v korpusech velmi málo frekventované. Korpusová data nemohou dosud výrazně pomoci pro upřesnění dosavadních observací. Morfologické značkování nebere v úvahu opisné stupňování. Doklad z korpusu SYN2006PUB „*Basilio je <více Beaumarchaisův> než Rossiniho.*“ nelze pokládat za důkaz stupňovatelnosti (byť opisné) posesivních adjektiv na -ův/-in. Rovněž na internetu nalezený doklad „*aby se dodavatelům nenasazovala <psejší> hlava, než jim přísluší*“ těžko vyvrátí, že některé slovtvorné typy syntetické tvary 2./3. stupně zpravidla netvoří. V obou případech jde o doklady toho, že přenos významu (adjektivum relační – hodnotící) může nejrůznější omezení rušit.

3.4. Analýza derivačních typů a morfologické značkování

Zdá se, že sledovaná problematika stupňování adjektiv a adverbíů zůstává i nadále otevřená pro další diskusi. Klíčovou a dobře formálně popsatelnou roli hraje příslušnost k derivačnímu typu. Proto by dle našeho názoru bylo užitečné k ní při úpravách analyzátorů přihlédnout. Jisté řešení nabízí v tomto směru analyzátor *Ajka*.

4. Slovnědruhá vícestupňovost

Jedním z problémů automatické morfologické analýzy je určení hranice v oblasti slovnědruhových přechodů (transpozic). Z hlediska automatické morfologické analýzy jde o zachycení případů, kdy jedna jednotka může mít více interpretací (značek) z toho důvodu, že může v návaznosti na užití (kontext) nabývat různé platnosti slovního druhu a dalších

slovnědruhově závislých významů (je tedy homonymní). Tradiční gramatiky se tohoto problému dotýkají a v některých případech nabízejí i pravidla řešení. Aplikace uvedených pravidel na data obsažená v jazykových korpusech a především jejich převedení do podoby algoritmu použitelného pro automatickou analýzu, vyžaduje často mnohá doplnění a upřesnění, jimž dosud nebyla věnována pozornost. Ve slovnících (SSJČ, SSČ) je u řady hesel uváděna několikerá slovnědruhová platnost. Z hlediska automatické morfologické analýzy není nejzávažnějším problémem rozšíření počtu možných interpretací, ale spíše užitečnost takového kroku při dalším užití automatických nástrojů. Situace je zdánlivě složitější tehdy, tvoří-li tvary, které lze potencionálně interpretovat více způsoby, otevřenou a početnou skupinu (například substantivizace adjektiv), zdánlivě jednodušší je u relativně uzavřených tříd (např. zájmen, zájmenných příslovcí a určitých číslovek) nebo u tříd poměrně méně početných (předložky, spojky, částice), kde je tento jev více zaznamenán ve slovnících i gramatikách. Sporné není, jak upravit automatické analyzátoři, aby rozlišovaly možné slovnědruhově a další interpretace, ale jakými pravidly se bude řídit disambiguace v jednotlivých případech.

4.1. Slovnědruhová platnost tvarů *mnoho, moc, hodně, (nej)víc(e)* ve slovnících

Tvary *mnoho, moc, hodně, (nej)víc(e)* jsou ve slovnících (SSJČ, SSČ) interpretovány následujícím způsobem.

SSČ	SSJČ
hodně/přísl.	hodně/přísl.
mnoho/přísl. ve spoj. s počít. předmětem v plat. čís. neurč. hovor.	mnoho/přísl. a čís. neurč.
moc/přísl. v plat. čís. neurč. hovor.	moc/přísl. a čís. neurč.
víc/více/přísl. v plat. čís. neurč.	víc/více/přísl. a čís. neurč.
nejvíc/nejvíce/přísl.	ODKAZ k více

4.2. Automatická morfologická analýza

Každé značkování je interpretace a každá interpretace má jistá omezení. Proto se na tomto místě budeme zabývat nikoli otázkou slovnědruhově platnosti slov/slovních tvarů, nýbrž otázkou, jak stanovit a následně dodržovat průhledná kritéria přiřazování zvolených interpretací.

V praxi vidíme na posunu mezi značováním uvedených tvarů v SYN2000 a SYN2005, že problém nezůstal v automatické morfologické analýze na okraji zájmu.

4.2.1. Interpretace v SYN2000 a SYN2005

Následující tabulka zachycuje posun mezi značováním SYN2000 (obyčejné písmo a tučná kurzíva) a SYN2005 (obyčejné písmo a kurzíva).

hodně/hodně/Dg.....1.*	<i>víc(e)/více/Dg.....2.*</i>	<i>nejvíc(e)/více/Dg.....3.*</i>
moc/moc/Db.....-.*		
<i>mnoho/mnoho/Db.....-.*</i>		

	<i>víc(e)/více/Dg.....2.*</i>	<i>nejvíc(e)/více/Dg.....3.*</i>
mnoho/mnoho/Ca..(1 4)....-.*		
mnoha/mnoho/Ca..(2 3 6 7)....-.*	<i>víc(e)/víc/více/ Ca..(1 2 3 4 6 7).*</i>	
moc/moc/Ca.....-.*		

Interpretace v SYN2000:

- 1) V souladu se SSJČ/SSČ ignoruje hojně v korpusech doložené užití tvaru *hodně* v platnosti neurčité číslovky (*hodně lidí, bylo nás hodně*) a hodnotí jej výhradně jako adverbium.
- 2) V rozporu se SSJČ/SSČ ignoruje v korpusech doložené užití tvaru *mnoho* v platnosti adverbia (*pracoval mnoho, mnoho platný*);
- 3) Prakticky řeší problematickou lemmatizaci tvarů *víc(e)/nejvíc(e)* a zároveň ignoruje užití těchto tvarů ve funkci neurčité číslovky (*více lidí, peněz, ...*).

Interpretace v SYN2005:

- 1) Nadále v souladu se SSJČ ignoruje *hodně* synonymní s *moc/mnoho* v platnosti neurčité číslovky (*hodně/mnoho/moc lidí, bylo nás hodně/mnoho/moc*);
- 2) S tím souvisí, že rozlišuje adverbia (*mnoho/moc*), která hodnotí jako nestupňovatelná (*Db.**), a stupňovatelné (*Dg.**) adverbium *hodně* – 1. stupeň, *více* – 2. stupeň, *nejvíce* -3. stupeň. Řeší tak problém obtížné lemmatizace: Je lemmatem tvarů *víc(e)/nejvíc(e)* tvar *mnoho, moc* nebo *hodně*?
- 3) Neignoruje transponované užití tvarů *víc(e)* (nikoli *nejvíc(e)*) v roli číslovky (*více lidí, peněz, ...*). Současně lemmatizuje tyto tvary důsledně tvarem samým.

4.2.3. Interpretace analyzátoru *Ajka*

Problematické řešení v analyzátoru *Ajka* je uvedeno v tabulce.

hodně/hodně/k6.*d1	<i>víc(e)/ hodně/k6.*d2</i>	<i>nejvíc(e)/ hodně/k6.*d3</i>
moc/moc/k6.*d1	<i>víc(e)/ moc/k6.*d2</i>	<i>nejvíc(e)/ moc/k6.*d3</i>
mnoho/mnoho/k6.*d1	<i>víc(e)/ mnoho/k6.*d2</i>	<i>nejvíc(e)/ mnoho/k6.*d3</i>
		<i>nejvíc(e)/nejvíc/k6.*d1</i>

Interpretace *Ajky*:

- 1) Neumožňuje rozumnou lemmatizaci tvarů *víc(e), nejvíc(e)* (není možné rozhodnout, které z lemmat *moc, hodně, mnoho* se má přiřadit tvarům *víc(e)/nejvíc(e)*).
- 2) Ignoruje číslovkovou platnost tvarů.

4.3. Návrh na zjednodušení a zprůhlednění dosavadní interpretace v *HA* i *Ajce*

hodně/hodně/Da.....1.* k6.*d1		
moc/moc/Da.....-.* k6.*d1		
mnoho/mnoho/Da.....-.* k6.*d1		
	<i>víc(e)/více/Da.....2.* k6.*d2</i>	<i>nejvíc(e)/více/Da.....3.* k6.*d3</i>
mnoho/mnoho/Ca..(1 4)....-.* k4.*		
mnoha/mnoho/Ca..(2 3 6 7)....-.* k4.*		

Argumentace pro navrhované řešení:

1. Pouze *mnoho* rozlišuje užití ve funkci číslovky formálně, a to nikoli tvary původní jmenné flexe (dochované v ustrnulých adverbializovaných tvarech *mnohem*, *namnoze*), nýbrž paradigmatickým složeným ze dvou tvarů, které odpovídá typu skloňování číslovek od 5 výše (srv. Komárek, 2006). Proto je vhodné pouze tvary *mnoho/mnoha* analyzovat jako tvary se značkou slovní druh číslovka, přičemž tvar *mnoho* je třeba disambiguovat (může mít též platnost adverbia a odpovídající značku).

2. Značkování tvarů *víc(e)*, *nejvíc(e)* jako slovní druh adverbium, detailní určení slovního druhu číslovka neurčitá/adverbium míry a lemmatizace tvarů *víc(e)*, *nejvíc(e)* lemmatem *víc* se řeší a) problémy disambiguace (výsledky ve značkových korpusech jsou dosti nepřesné), b) případ „sporné lemmatizace“ (viz výše) a c) z lingvistického hlediska je obhajitelné poukazem na to, že chápeme-li tvoření tvarů komparativu a superlativu jako slovtvornou modifikaci, pak je přijetí lemmatizace tvarů komparativu a superlativu tvarem pozitivu pouze přijetím zavedené lexikografické konvence. Tuto konvenci je pak možné z dobrých důvodů porušit. „Nezávislost“ tvarů komparativu/superlativu na tvaru pozitivu není totiž jen nezávislostí formy (nejde o tvoření afixální, ale o tvary, které se interpretují jako nepravidelné tvoření od supletivních kořenů), ale i nezávislostí na „jediném“ tvaru pozitivu (ambiguita určení tvaru nahrazujícího komparativ/superlativ při transformaci do pozitivní konstrukce).

3. Značkování tvarů *hodně*, *moc* (lze uvažovat i o dalších např. *tuze*, *velmi*, *velice*) jako slovní druh adverbium detailní určení slovního druhu číslovka neurčitá/adverbium míry by mohlo pomoci systematictějšímu a průhlednějšímu zachycování zkoumaného problému na úrovni automatické morfologické analýzy.

5. Závěr

V našem příspěvku jsme se snažili ukázat, jak mohou korpusová data přispět k upřesnění a doplnění údajů uvedených v gramatikách, slovnících a v neposlední míře i v lingvistických bázích automatických morfologických analyzátorů, které byly na základě dostupných příruček budovány. Dílčí závěry jsou uvedeny v kapitolách 2.4., 3.4. a 4.3., proto je na tomto místě pouze stručně shrneme.

Na základě analýzy dostupných reprezentativních korpusů psaného jazyka jsme dokázali zhruba o čtvrtinu rozšířit seznam sloves, která mohou v češtině vyjadřovat futurum synteticky, a pokusili jsme se přesněji definovat jejich významy. Stranou ponecháváme stylistickou klasifikaci jednotlivých dokladů. Z porovnání popisu zkoumaného jevu ve starších slovnících s jeho výskytem v synchronních reprezentativních korpusech vyplývá, že se jedná o skupinu sloves, již lze rozdělit na stabilní jádro (frekventovaná slovesa stojící v centru slovní zásoby doložená jak v tištěných příručkách, tak v korpusech) a proměnlivou

periferii, kam lze zahrnout lemmata zaznamenaná v tištěných příručkách a nedoložená v korpusech ani na internetu a lemmata, jejichž výskyt dosvědčují výhradně korpusey a internet.

Na základě analýzy popisu afixálního tvoření tvarů 2./3. stupně v lingvistické literatuře a automatických morfologických analyzátoch upozorňujeme na možnost hlouběji využít analýzu derivačních typů při dalších úpravách existujících automatických analyzátorů.

Na základě zkušeností s problémy slovnědruhovými víceznačností tvarů *hodně*, *moc*, *mnoho*, (*nej*)*víc(e)* upozorňujeme na některá sporná řešení a navrhuje jistá zjednodušení, která by nebyla v rozporu s lingvistickou teorií na straně jedné a přispěla by k průhlednějšímu řešení na úrovni automatické morfologické analýzy na straně druhé.

Příloha: Seznam lemmat sloves, která mohou tvořit synteticky tvary futura

Tvary uvedené ,*‘ jsou homonymní s tvary sloves tvořených prefixem *po-*. ,*?’ uvádíme u slovesa *potrvá/trvat*. V korpusech SYN2006PUB jsme našli 9 dokladů tvaru *potrval(-a/-y)* od předpokládaného slovesa *potrvat*.

půjde/jít	#1	*pocupe/cupat	&0	postěhuje se/stěhovat se	#1
*pojede/jet	#1	pohasí si to /hasit si to	\$1	pocestuje/cestovat	#1
poběží/běžet	#1	*pojechá/jechat	&1	*?potrvá/trvat	#3
*poletí/letět	#1	poklape/klapat	-1	poputuje/putovat	\$1
poleze/lézt	#1	poklíčí/klíčit	&1	povane/vanout	\$5
popluje/plout	#1	pokulhá/kulhat	&1	*pomaže/mazat	\$5
poplave/plavat	\$1	pokulí se/kulit se	&1	*poválí (se)/valit (se)	\$5
*potáhne/táhnout	\$1	pokvačí/kvačit	&1	pokráčí/kráčet	\$5
*poteče/téct	#1	poline se/línout se	&1	*popase se/pást se	\$5
*povleče/vléct	\$1	poloudá se/loudat se	&1	*pošine se/šínout se	\$5
*požene/hnát	\$1	popeláší/pelášit	\$1	posviští/svištět	\$5
*ponese/nést	#1	poplaví se/plavit se	\$1	*potáhne (se)/táhnout (es)	\$5
*povede/vést	?\$1	poplází se/plazit se	#1	*potrénuje/trénovat	\$5
poveze/vézt	#1	poplízí se/plížit se	\$1	*popeče s kým/péci s kým-6	
*poroste/růst	\$1	porázuje/rázovat	&1	pokluše/klusat	@1
pokvete/kvést	\$1	pořine se/řínout se	\$1	pošupe/šupat	@7
povládne/vládnout	\$1	poštráduje si to/štrádovat	&1	poprodí se/brodit se	@7
povrávorá/vrávorat	\$1	pověje/vát	&1	pocrčí/crčet	@7
*pošlape/šlapat	\$1	povine se/vínout se	&1	pošupajdí/šupajdit	@7
pomašřuje/mašřovat	\$1	povisí/víset	&1	poharcuje/harcovat	@7
pofičí/fičet	\$1	*poleje/polije/lít	&1	pokoučuje/koučovat	@7
?pokutálí se/ kutálet se	\$1	podere se/drát se	&1	*povesluje/veslovat	@8
pohrne se/hrnout se	#1	*porazí/razit	&2	poprýští/prýštit	@8
pochvátá/chvátat	&1	pošífí se/šfít se	&2	pošupačí/šupačit	@8
pokvapí/kvapit	&1	*poprší/pršet	&2	potryská/tryskat	%9
poplyne/plynout	\$1	*popálí/pálit	&2	potrtá/trtat	%9
pocupitá/cupítat	#1	porajzuje/rajzovat	&2	pozdrhá/zdrhat	%9
pocape/capat	&0	pořítí se/řítit se	#1		

LEGENDA:

doloženo v korpusech – lemma bez *po-*/tag=“VB.....F.*“

?# doloženo v korpusech – lemma bez *po-*/tag=“VB.....P.*“

\$ doloženo v korpusech – lemma s *po-*/tag=“VB.....P.*“

@ doloženo v korpusech – lemma **tvar sám (nesmyslný tvar)**/tag!=“VB.....P.*“

- doloženo v korpusech – korpus není lemmatizován a tagován

& není doloženo v korpusech, pouze v tištěných zdrojích

% není doloženo ani v korpusech, ani v tištěných zdrojích, pouze na internetu

- 1 SSJČ/SSČ
- 2 Čermák, 1991
- 3 doloženo jen SYN2000/SYN2005/SYN2006PUB – lemma bez po-/tag="VB.....F.*"
- 4 doloženo jen SYN2005/SYN2006PUB lemma bez po-/tag="VB.....P**"
- 5 doloženo jen SYN2000/SYN2005/SYN2006PUB – lemma s po-/tag="VB.....P.*"
- 6 doloženo jen v BMK
- 7 doloženo jen SYN2000/SYN2005 – lemma tvar sám (nesmyslný tvar)/tag!="VB.....P.*"
- 8 doloženo jen SY2006PUB – lemma tvar sám (nesmyslný tvar)/tag!="VB.....P.*"
- 9 doloženo jen na internetu
- 0 Šmilauer, 1972

Bibliografie

- ČECHOVÁ A KOL. (1996): *Čeština, řeč a jazyk*. Praha : ISVN.
- ČERMÁK, F. (1990): *Syntagmatika a paradigmatika českého slova*. Praha : Karlova Univerzita.
- HAJIČ J. (1994): *Unification Morphology Grammar*. Praha : MFF UK, (disert. práce).
- HAJIČ J. (2004): *Desambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha : Karolinum.
- HAVRÁNEK, B. A KOL. (1989): *Slovník spisovného jazyka českého (SSJČ)*. Praha : Academia.
- FILIPEC, J. A KOL. (2005): *Slovníku spisovné češtiny pro školu a veřejnost (SSČ)*. Praha : Academia.
- KARLÍK, P., NEKULA, M., RUSÍNOVÁ, Z. (1996): *Příruční mluvnice češtiny*. Praha : Nakladatelství Lidové noviny.
- KARLÍK, P., NEKULA, M., PLESKALOVÁ, J. (2002): *Encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny.
- KARLÍK, P., HLADKÁ, Z. (2004): Kam s ním? (problém stupňování adjektiv). In: Karlík, P., Pleskalová, J. (eds.), *Život s morfémy*. Brno : FF MU, 73-93.
- KOMÁREK, M. a kol. (1986): *Mluvnice češtiny II.*, Praha : Academia.
- KOMÁREK, M. (2006): *Příspěvky k české morfologii*. Olomouc : Periplum.
- OSOLSOBĚ, K. (1996): *Algoritmický popis české morfologie a strojový slovník češtiny*. Brno : FF MU (disert. práce).
- OSOLSOBĚ, K. (2007): Syntetické futurum v češtině – gramatiky, slovníky, korpusy. *Přednášky a besedy z XL. běhu LŠSS*. Brno, s. 131-144.
- OSOLSOBĚ, K. (2007): Popis gramatických významů jednoduchých slovesných tvarů v anotacích českých (slovenských) korpusů. *SPFFMU A, 55, v tisku*.
- SEDLÁČEK, R. (2004): *Morphematic analyser for Czech*. Brno : FI MU (disert. práce).
- ŠMILAUER, V. (1972): *Nauka o českém jazyku*. Praha : SPN.
- Český národní korpus - SYN2000, SYN2005, SYN2006PUB, BMK, PMK, ORAL2006. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupné z WWW: <<http://ucnk.ff.cuni.cz>>.
- Morfologický analyzátor Ajka* dostupný z <http://nlp.fi.muni.cz/projekty/ajka/ajkacz.htm>.