Princípy a problémy spracovania neohybných slovných druhov v ručne anotovanom podkorpuse Slovenského národného korpusu

Mária Šimková

1. Lingvistická anotácia korpusu

Jazykové korpusy sú v súčasnosti nielen čistými textovými databázami poskytujúcimi používateľom obraz reálneho fungovania slova (tvaru slova) v kontexte, ale v čoraz širšom rozsahu prinášajú aj informácie o jeho gramatických, lexikálnych, sémantických a i. vlastnostiach. Textové databázy neobsahujúce lingvistické informácie síce takisto umožňujú skúmanie jazykových javov, a to bez zaťaženia jednou z mnohých gramatických teórií a bez chybných výstupov spôsobených chybnými vstupmi, pri flektívnych jazykoch však nie je možné získať z nich všetky údaje napr. o tvaroch slov tak, ako ich potrebujú rôzne špecifické jazykovedné výskumy. Lingvisticky anotované korpusy disponujú niekoľkými úrovňami lingvistických informácií o každom slove v korpuse. Najzákladnejšia je lematizácia (zvyčajne plne automatizované priradenie základného, slovníkového tvaru), dôležitá najmä pri ohybných slovách, ktoré sa pomocou nej dajú vyhľadať vo všetkých tvaroch jedným príkazom, pričom sa eliminuje časť tvarovej homonymie (napr. *mier* môže byť N al. A sg. maskulína *mier*, G pl. feminína *miera*, imperatív verba *mieriť*; pri dobrej lematizácii by sa každý tvar mal umiestniť k svojej lexéme). Bežná je aj morfologická úroveň, kde sa anotácia (východiskovo ručná, postupne automatizovaná) obligatórne začína zaradením slova do slovného druhu, resp. zaradením textovej jednotky do istej (slovnej) triedy, a pokračuje ďalšími morfologickými kategóriami podľa slovného druhu a typu jazyka (o druhoch a spôsoboch morfologickej anotácie porov. bližšie napr. Leech, 2000; Forróová – Horák, 2004). Na ňu v niektorých korpusoch nadväzuje syntaktická anotácia, ktorá je však oveľa náročnejšia na výber teórie a takisto na čas a financie. Korpusy sú technicky spracované tak, že lingvistické anotácie sú akoby v pozadí, nemusia byť „zapnuté“ a výskum sa môže realizovať na čistom súbore slov napr. pomocou regulárnych výrazov (http://korpus.juls.savba.sk/usage/index.sk.html).

Pri príprave tokenizácie, lematizácie a morfologickej anotácie textov Slovenského národného korpusu (Forróová – Garabík – Gianitsová – Horák – Šimková, 2003; ďalej SNK) sa vychádzalo zo skúseností korpusových pracovísk v okolitých krajinách, najmä v Českej

republike a Poľskej republike. Hlavnými princípmi sa stali formálno-morfologický prístup a kombinácia atribútového a pozičného zápisu morfologických značiek: značia sa iba hodnoty relevantné pre daný tvar, počet znakov je teda variabilný, ale ich poradie v značke je záväzné. A hoci bolo zrejmé, že jednotlivé rozhodnutia treba robiť s vedomím definitívnosti, pretože akékoľvek opravy sú v prípade rozsiahlych databáz náročné, ak nie nemožné, a hoci návrh pravidiel a súboru značiek na morfologickú anotáciu (tagset) prešiel interným oponentským konaním i diskusiou a hodnotením na medzinárodnej konferencii[1], predsa boli po spustení segmentácie a po ručnej anotácii viacerých textov nevyhnutné dve revízie a čiastkové modifikácie tagsetu (Garabík – Gianitsová – Horák – Šimková, 2004; Garabík – Gianitsová – Karčová – Šimková, 2006).

Zmeny v tokenizácii (segmentovaní textu na tokeny – textové jednotky zahŕňajúce nielen slová v klasickom lingvistickom ponímaní, ale aj interpunkciu, číslicové zápisy a pod.) sa týkali najmä zjednotenia segmentácie a následne aj lematizácie zložením typu *ako-tak, ten-ktorý* bez ohľadu na to, či sú zapísané správne so spojovníkom alebo nesprávne s pomlčkou, ba niekedy dokonca aj s medzerami. Všetky časti akýchkoľvek viacslovných celkov (viacslovné spojky, druhotné predložky, viacslovné názvy) sa v SNK tokenizujú osobitne (napr. *v závislosti od* obsahuje tri tokeny: *v*, *závislosti*, *od*). Súčasne však aglutinované tvary typu *oňho* sa chápu ako jeden reťazec znakov, čím sa aj na tejto úrovni uplatňuje formálny prístup, na základe ktorého sa token v SNK definuje ako akýkoľvek reťazec znakov medzi dvoma medzerami vrátane tých reťazcov/jednotlivých znakov, okolo ktorých sa medzery robia umelo, aby boli technicky a štatisticky spracovateľné (interpunkcia, kompozitá so spojovníkom a pod.). Toto riešenie nevylučuje možnosť zapojenia logického modulu, ktorý by mohol byť aplikovaný aj ako základ inej lematizácie a značkovania.

Zmeny v morfologickej anotácii sa dotkli predovšetkým hodnoty „neurčený“, ktorá sa v pôvodnom tagsete priraďovala (kongruentným) kategóriám rodu, čísla a pádu v prípadoch, keď neboli z kontextu jednoznačne určiteľné (*muži a ženy tlieskali*). Podľa poslednej verzie tagsetu sa značka *o* „neurčený“ používa len vo výnimočných prípadoch. Zmenil sa aj počet slovných tried. K pôvodným šestnástim: substantíva (S), adjektíva (A), pronominá (P), numeráliá (N), verbá (V), adverbiá (D), prepozície (E), konjunkcie (O), partikuly (T), interjekcie (J), reflexívne *sa/si* (R), kondicionálová morféma *by* (Y), interpunkcia (Z), abreviácie a značky (W), neurčiteľný slovný druh (Q) a neslovný element (#) pribudli ďalšie tri: formálne participiá (G), citátový výraz (%) a číslice (0). Princípy spracovávania neohybných slovných druhov sa v zásade nemenili, s výnimkou druhotných predložiek, kde sa pôvodné jednotné označovanie všetkých častí značkou pre predložku (E) zmenilo na

[1] SLOVKO 2003 – Slovanské jazyky v počítačovom spracovaní (Bratislava 24. – 25. októbra 2003). Materiály z konferencie neboli publikované.

použitie značky E výlučne pre predložkové zložky a substantívne časti sa označujú ako S s priradením príslušných kategórií.

2. Princípy morfologickej anotácie neohybných slovných druhov v SNK

Pri lexikálnych jednotkách z neohybných slovných tried (adverbiá, prepozície, konjunkcie, partikuly, interjekcie) sú lemy identické s ich slovníkovým tvarom, ktorým je často zmeravený tvar ohybného slova (*končiac, koncom, jasné, figu, (do) hroma*). Kladie to zvýšené nároky tak na ručnú, ako aj na automatizovanú lematizáciu a následnú morfologickú anotáciu, ale v tomto prípade iné riešenie nebolo prijateľné. Hoci anotátori majú k dispozícii základné kodifikačné príručky (Krátky slovník slovenského jazyka, Pravidlá slovenského pravopisu) a opisy morfológie slovenského jazyka, predsa musia jednotlivé prípady neraz riešiť individuálne podľa kontextu; anotačný softvér sa takisto „snaží“ riadiť na základe dvoj- a trojslovných zoskupení slov, tzv. bigramov a trigramov. Rozhodovanie sťažuje vysoká homonymia (najmä medzi neohybnými slovnými druhmi navzájom) pri neprítomnosti iných formálnych delimitačných znakov a značná dynamika vzniku nových lexém z tvarov ohybných slovných druhov alebo použitie tvaru v ešte úplne nelexikalizovanej inej funkcii, ktoré gramatická teória ani lexikografické príručky nezachytávajú.

Odklon od bežnej klasifikácie slovných druhov predstavuje niekoľko osobitných prípadov (o ďalších podrobnejšie Garabík – Gianitsová – Karčová – Šimková, 2006):
a) nerozlišuje sa medzi *sa, si* ako zvratným zámenom a súčasťou slovesného tvaru, príp. časticou, všetky tokeny *sa, si* sú označené ako R, čo ich z formálneho hľadiska začleňuje do pásma neohybných slovných jednotiek;[2]
b) morféma *by* ako súčasť tvaru podmieňovacieho spôsobu má samostatnú značku Y a značka Y sa nachádza ako druhá v poradí aj pri spojkách (O) a časticiach (T), ktoré predstavujú aglutinované tvary s morfémou *by* (*keby, aby, sťaby, kiežby*);
c) vo viacslovných lexikálnych jednotkách obsahujúcich nesamostatné časti nezaraditeľné do niektorého zo slovných druhov sa tieto, zvyčajne prvé časti označujú značkou Q – neurčiteľný slovný druh (*po slovensky, slovensko-český, na bielo* a pod.);
d) pri predložkách sa okrem spojenia s pádom (na 3. mieste) uvádza na 2. mieste informácia o ich vokalizovanosti alebo nevokalizovanosti (napr. v spojení *so mnou* má predložka *so* značku Ev7);

[2] Toto riešenie výrazne zjednodušuje anotáciu a nekomplikuje používateľom vyhľadávanie v prípade chybného zaradenia, keďže morfémy *sa, si* môžu mať viacero interpretácií a spolu so svojimi slovesami takisto vykazujú značnú dynamiku tvarotvorných a slovotvorných procesov. Jednoznačná charakteristika *sa, si* pri morfologickej anotácii ostáva na špecializované výskumy, pri syntaktickej anotácii (na analytickej rovine) sa zvratné zámeno označuje klasicky ako objekt, morfémy *sa, si* ako reflexívum tantum (značka AuxT) alebo zvratné, opisné pasívum (značka AuxR), v ojedinelých prípadoch ako AuxO – nadbytočná, rytmická častica.

e) jednoslovné druhotné predložky sa označujú značkou E ako všetky predložky, dvoj- a viacslovné druhotné predložky sa však lematizujú a anotujú ako spojenie predložky (predložiek) a substantíva – je to istý ústupok kvôli zjednodušeniu automatizovanej anotácie a takisto vzhľadom na dynamiku tvorby sekundárnych predložiek, súčasne to však zodpovedá východiskovému formálno-morfologickému princípu.

Na základe súboru pravidiel, z ktorých sme tu výberovo uviedli niekoľko týkajúcich sa predovšetkým neohybných slovných druhov, sa od začiatku automatizovane tokenizovali všetky texty v SNK (ručná segmentácia pri miliónoch textových jednotiek neprichádza do úvahy). Lematizácia a morfologická anotácia sa v prvej fáze (po verziu prim-2.0 vrátane) robila takisto automatizovane, ale len pomocou českého lematizátora a morfologického anotátora (a dezambiguátora – nástroja na zjednoznačňovanie, jednoznačné zaradenie homonymných tvarov), ktorý pre SNK poskytlo pracovisko Ústavu formální a aplikovaní lingvistiky MFF UK Praha. Český tagset je však robený spôsobom vyhovujúcejším počítačovému spracovaniu prirodzeného jazyka, no náročnejším pre používateľa-lingvistu (Hajič, 2000; Hana – Hanová, 2002), česká a slovenská gramatická teória nie je úplne rovnaká (už základná slovnodruhová klasifikácia neohybných lexikálnych jednotiek je značne odlišná) a ani lematizácia nebola vždy stopercentná (napr. zámená *všetok, náš* boli lematizované ako *všetok, môj*, adjektíva *irácky, islamský* ako **irácky, *islámsky*, substantívum *Štúr* ako *štúrať*...). Preto sa pristúpilo k tvorbe vlastného slovenského tagsetu, z ktorého sme citovali niektoré zásady spracovania neohybných slovných druhov. Tým sa vlastne morfologická anotácia SNK vrátila na začiatok – na to, aby sa mohla realizovať automatizovaná morfologická anotácia, je potrebná ručná anotácia v rozsahu aspoň 1 milión tokenov. Na ručne anotovanom korpuse sa trénuje morfologický anotátor (príp. aj dezambiguátor), ktorým sa potom automatizovane anotujú všetky textové jednotky v korpuse.

3. Základné výsledky morfologickej anotácie SNK

SNK sprístupnil prvú verziu ručne morfologicky anotovaného korpusu r-mak-1.0 v rozsahu 322 600 tokenov začiatkom r. 2006. Na tomto minimálnom podkorpuse sa trénoval morfologický anotátor, pomocou ktorého bol automatizovane anotovaný celý korpus verzie prim-2.1. Aj keď táto anotácia nie je bezchybná, predsa poskytuje štatistické údaje napr. o frekvencii slovných druhov, resp. slovných tried, ktoré vypovedajú o základných vlastnostiach súčasnej slovenčiny a umožňujú porovnávanie podobných údajov s inými jazykmi. (Na porovnanie je ešte stále k dispozícii aj morfologická anotácia pomocou českého softvéru.)

Frekvencia vymedzených slovných tried v troch verziách SNK

r-mak-1.0	prim-2.1-public-all	prim-2.1-vyv
322 600 tokenov	294 087 581 tokenov	54 357 894 tokenov
S: 71035	S: 71672080	S: 12977758
Z: 62434	Z: 53583544	Z: 10092391
V: 45303	V: 34928614	V: 6578560
P: 27063	E: 24771176	E: 4522345
E: 25180	A: 23254199	A: 4246719
A: 21986	P: 19669576	P: 3809665
O: 20030	O: 15602682	O: 2985752
T: 10862	0: 10731655	0: 1885610
D: 10384	T: 8264014	T: 1587311
R: 9984	D: 7527068	D: 1449534
N: 3822	R: 6379524	R: 1220839
%: 3797	%: 4665643	W: 771410
0: 3398	W: 4520302	%: 653189
G: 3277	N: 3494257	N: 645309
W: 1781	G: 3187448	G: 584085
Y: 878	Y: 847828	Y: 159563
Q: 798	Q: 541042	Q: 99682
J: 313	#: 301556	#: 57506
#: 170	J: 145309	J: 30660
?: 1	?: 64	?: 6
57,9 % umelecké,	60,6 % publicistické,	60 % publicistické,
41,8 % publicistické,	17,5 % umelecké,	20 % umelecké,
0,2 % odborné texty	11,6 % odborné,	20 % odborné texty
	10,3 % iné texty	

V prvom stĺpci sú údaje z ručne anotovaného podkorpusu, ktorý obsahuje 322 600 tokenov, v druhom z celého verejne prístupného korpusu (takmer 300 mil. tokenov), v treťom zo štýlovo vyváženého korpusu v rozsahu necelých 55 mil. tokenov. Obidva veľké korpusy boli morfologicky anotované automatizovane. Na konci každého stĺpca sú informácie o zastúpení hlavných štýlov.

Prvé tri miesta obsadili vo všetkých troch prípadoch podstatné mená, interpunkcia a slovesá. Možno predpokladať, že toto rozloženie sa už nebude veľmi meniť ani so zväčšovaním korpusu, ani so zlepšovaním morfologickej anotácie. Na ďalších miestach má ručne anotovaný podkorpus najskôr zámená, potom predložky a adjektíva, veľké korpusy najskôr predložky a adjektíva, až potom zámená. Rozdiel môže súvisieť s nízkym podielom odborných textov v ručne anotovanom podkorpuse. Všeobecne zhodné je na ďalšom mieste postavenie spojok, no rozdielna štýlovo-žánrová štruktúra zrejme spôsobila, že na 8. mieste sú vo veľkých korpusoch číslice, ktoré sa v podkorpuse r-mak nachádzajú až na 13. mieste. Za

spojkami v r-mak-u tak ako za číslicami v ďalších dvoch verziách nasledujú častice, príslovky a polyfunkčné slová-morfémy *sa, si*. Ich pozícia v strede tabuľky môže byť stabilná aj v ďalších verziách. Dolná časť tabuľky je už značne diferencovaná – prejavujú sa tu jednak rozdiely medzi ručnou (spoľahlivejšou) a automatizovanou (ešte málo spoľahlivou) anotáciou, jednak rozdiely na základe zastúpenia textov podľa štýlov.

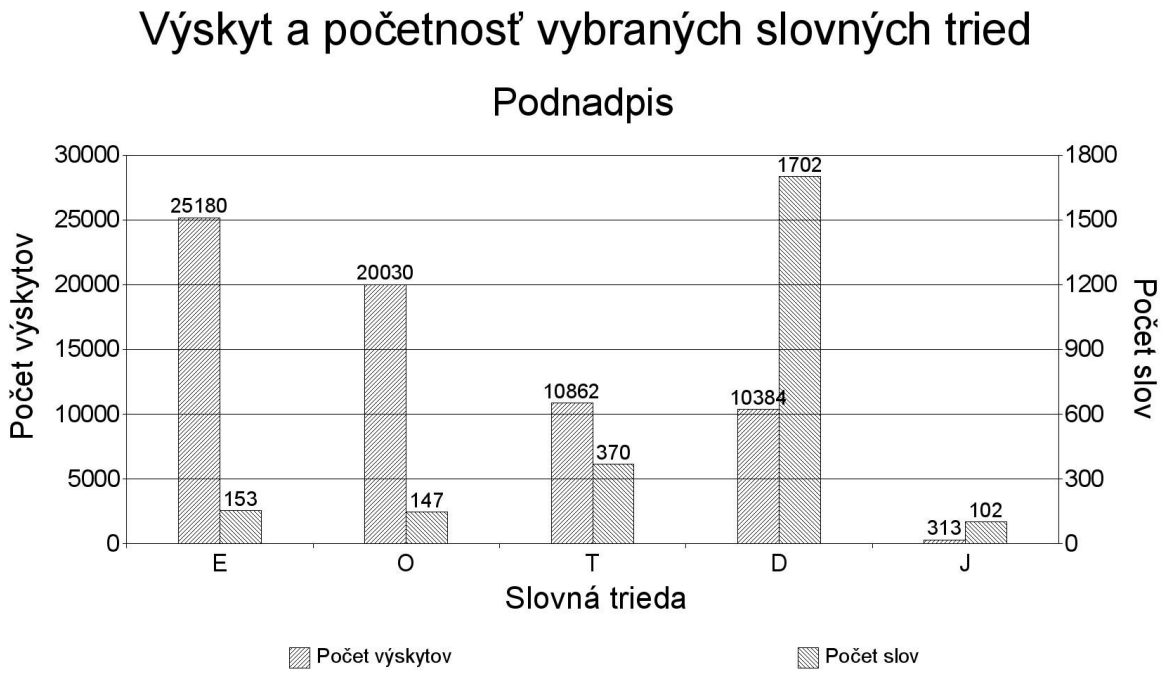
Z hľadiska podielu ohybných a neohybných slovných druhov a iných textových jednotiek vidíme, že na prvom mieste sú substantíva, za nimi interpunkcia a verbá, hoci vo frekvencii tvarov a lem sú na prvých miestach stabilne bodky a čiarky, potom spojka *a* a predložky *v, na* a prvé substantíva a plnovýznamové verbá sa nachádzajú zvyčajne až hlboko v druhej päťdesiatke či stovke (porov. http://korpus.juls.savba.sk/stats). Inventár interpunkčných znamienok, spojok a predložiek je však oproti podstatným menám a slovesám značne obmedzený, takže takéto poradie frekvencie celých slovných tried je pochopiteľné. Predložky a spojky sú napokon najfrekventovanejšie spomedzi neohybných slovných druhov, čo súvisí s vyjadrovaním zložitejších vzťahov medzi slovami a vetami, nasledujú častice, v tesnom závese príslovky a s celkom malým podielom na poslednom mieste (nielen v rámci neohybných slovných druhov) citoslovcia. Pomerne nízka frekvencia prísloviek môže byť trochu prekvapujúcou vzhľadom na potencionalitu derivácie od adjektív, no v podstate len potvrdzuje doterajšie zistenia (napr. aj lexikografov koncipujúcich nový výkladový slovník slovenského jazyka, ktorí priebežne signalizovali nedostatok dokladov na príslovky, o tvaroch komparatívu a superlatívu ani nehovoriac). Malý rozdiel medzi časticami a príslovkami sa môže pri lepšej anotácii zmeniť v prospech prísloviek, ale môže aj posilniť postavenie častíc – v tomto bode nepovažujeme anotáciu (ani ručnú) za bezchybnú a ukončenú. (V aktuálne sprístupňovanej novej verzii ručne anotovaného podkorpusu r-mak-2.0 v rozsahu 511 534 tokenov sa častice opäť nachádzajú tesne pred príslovkami.) Nízka frekvencia citosloviec je do značnej miery ovplyvnená formou textov v korpuse (písané, v prevažnej väčšine oficiálne texty) a ich počet nezvyšuje ani zastúpenie umeleckých textov. O problematike citosloviec podrobnejšie v ďalšej časti.

Grafické znázornenie zastúpenia vymedzených slovných tried v morfologicky ručne anotovanom podkorpuse SNK

4. Špecifiká spracovania a frekvencie neohybných slovných druhov v ručne anotovanom podkorpuse SNK

Základným problémom pri akomkoľvek spracovaní neohybných slovných druhov je okrem ich tvarovej nediferencovanosti aj vysoká homonymia (porov. napr. Šimková, 2001, 2003) a vysoká frekvencia práve tých lexém, ktoré sú homonymné. Vzhľadom na neúplnosť ich spracovania v teórii a dynamiku tvorenia vznikajú problémy pri ich určovaní aj v rámci ručnej morfologickej anotácie. Automatizovaná anotácia sa môže „opierať“ o slovník v prípade existencie morfologického analyzátora a generátora tvarov (pre slovenčinu bol nedávno vyvinutý na pracovisku SNK a začína sa aktuálne používať), aj to však len pri jednoznačných, nehomonymných lexémach, alebo, možno trochu paradoxne, môže na základe „svojho“ prístupu (vyhodnocovanie bigramov a trigramov) ponúknuť materiál na formalizáciu slovnodruhového zaraďovania týchto lexikálnych jednotiek. Napriek tomu, že anotáciu neohybných slovných druhov nepovažujeme za bezchybnú, predsa poskytuje v niektorých ukazovateľoch preukazné výsledky alebo minimálne námety na ďalší výskum.

Grafické znázornenie frekvencie neohybných slovných druhov v morfologicky ručne anotovanom podkorpuse SNK podľa jednotlivých výskytov a celkového počtu výskytov



Ľavé stĺpce zobrazujú celkový počet výskytov členený podľa stupnice s dielmi v hodnote 5000 jednotiek, pravé stĺpce (členené na stupnici, kde jeden diel znamená 300 jednotiek) zobrazujú jednotlivé, unikátne výskyty. Pri predložkách a spojkách sa malý počet slov (153 a 147) podieľa na značnej frekvencii slovného druhu, t. j. jedno slovo z týchto tried sa v ručne anotovanom podkorpuse vyskytuje priemerne 164,6 x (predložky), resp. 136,3 x (spojky). K takémuto priemeru prispievajú najmä vysokofrekventované predložky a spojky, no aj v rámci týchto slovných druhov sa nachádzajú lexémy s výskytom 1.

Prehľad najviac a najmenej frekventovaných predložiek a spojok v ručne anotovanom podkorpuse SNK

na	4591	a	7741
v	3898	že	2461
s	1775	ako	1567
do	1649	keď	820
z	1639	ale	785
o	1242	aby	617
za	905	či	518
k	841	alebo	401
po	759	A	392
V	686	i	325
nehovoriac, Naproti, Sponad, počnúc, pomocou, Nedbajúc, vplyvom, skrze, nevynímajúc,		PRETO, ALE, kde, tomu, Čože, Žeby, Ibaže, takisto, ALEBO, načo, KEĎ, Síce,	

doprostred, naspodok, naproti, dolu, zásluhou, navzdory | hoc, nuž, akonáhle

Naopak, pri príslovkách a citoslovciach sa na celkovej frekvencii slovného druhu, ktorá je najmä pri citoslovciach veľmi nízka, podieľa značné množstvo jednotiek. Jedna príslovka sa tak vyskytuje priemerne 6,1 x, jedno citoslovce dokonca iba 3,1 x. Častice sú uprostred, ale nedosahujú také vysoké frekvencie ako predložky a spojky a s priemerným výskytom jednej jednotky 29,4 x sú bližšie k príslovkám a citoslovciam.

Prehľad 10 najfrekventovanejších častíc, prísloviek a citosloviec v ručne anotovanom podkorpuse SNK

aj	887	veľmi	295	Ach	49
len	750	potom	264	KLAP	21
už	700	už	196	ha	14
nie	510	teraz	178	Mhm	13
A	423	opäť	151	Hej	10
ešte	384	stále	143	Aha	10
však	383	dobre	116	Pozor	10
tak	305	celkom	112	Áu	8
ani	298	treba	102	Hop	7
až	294	dnes	97	Ó	7

Kým najfrekventovanejšie predložky a spojky dosahujú v 300-tisícovom podkorpuse frekvencie v tisíckach jednotiek a s výskytom 1 ich je iba do dvoch desiatok, najfrekventovanejšie príslovky sa pohybujú v stovkách a citoslovcia iba v desiatkach výskytov. Najväčšie množstvo lexikálnych jednotiek sa v týchto dvoch slovných triedach nachádza s výskytom 1. Pri citoslovciach je aj veľmi zreteľné výraznejšie zastúpenie jedného textu alebo štýlu autora (porov. citoslovce, resp. druhú časť citoslovca *KLAP – KLIPY-KLAP, KLAPY-KLAP*), ktoré osobitne vystupuje do popredia pri malom korpuse, akým ručne anotovaný podkorpus je. Vo veľkom korpuse (prim-2.1) by bolo poradie citosloviec úplne iné, napr. zložke *KLAP* už nepribudol ani jeden ďalší výskyt, takže by sa nachádzala v poslednej tretine frekvenčnej tabuľky a predbehli by ju také lexémy ako *Hej* (4960 x), *Aha* (4449 x), *Ó* (3504 x), *Ahoj* (2932 x), ale aj *Božemôj* (670 x), *Fíha* (370 x) a pod. Ani veľký korpus však nie je celkom mimo vplyvu príznakových textov a špecialitiek nejakého autora, napr. značkou citoslovca bolo 48 x označené slovko *kvá*, z toho 41 x sa nachádzalo u jedného autora (L. Ťažký), ktorý ho však používal aj miesto vulgarizmu *k...*

Aj keď sú viaceré lexémy medzi neohybnými slovnými druhmi značne homonymné a náročne delimitovateľné, predsa sa väčšina z nich (najmä pri takých rozsahom malých

slovných triedach ako predložky, spojky a častice, kde sa počet jednotiek v jednom slovnom druhu pohybuje okolo 400) dá taxatívne vymenovať. Netýka sa to však veľkej časti citosloviec, ktoré sa svojou nesystémovosťou (hláskovou a následne aj grafickou), variabilitou formy (hromadenie a predlžovanie hlások, písanie spolu, osobitne, so spojovníkmi, s pomlčkami) a originalitou vymykajú z bežného postupu inventarizácie jednotiek. Svojou neuchopiteľnosťou sa približujú napr. k vlastným podstatným menám a skratkám, ktoré takisto otvárajú špecifické otázky pri automatizovanej anotácii korpusu a zvyšujú alebo udržiavajú percento jej neúspešnosti.

Osobitné problémy pri anotovaní citosloviec predstavuje:
a) slovnodruhová homonymia – niektoré citoslovcia sú homonymné nielen s časticami alebo príslovkami (*hej, dočerta*), ale aj s tvarmi podstatných mien alebo slovies (*Bože, vitajte*);
b) homonymia formy v širokom zmysle slova – keďže korpus obsahuje množstvo textov z rôznych oblastí vrátane cudzojazyčných prvkov, dochádza k často prekvapujúcim zhodám formy citoslovca s apelatívom (*heš* – hokejista Jiří Heš, *hop* – skupina Hop Trop, *pche* – kláštor Pche Čjaun, *á/Á* – Á. ako iniciála krstného mena, *BUM* – nápoj BUBLI BUM a pod.) alebo s inými výrazmi (*ha/Ha* ako citoslovce, resp. časť citoslovca na napodobnenie smiechu – ha ako značka pre hektár i ako koncovka v rómčine, čínštine a pod., podobne je vo vietnamčine a čínštine Ha, ktoré je aj na začiatku názvu izraelskej hymny Ha Tikva; *hop* ako citoslovce vyzývajúce na skákanie alebo rytmizujúce skákanie – hop ako súčasť anglických pomenovaní hop table, hip-hop a pod.);
c) stanovenie hranice medzi významom a použitím slova ako substantíva alebo verba a ako citoslovca v prípadoch typu *pozor, sláva, horí* – nie každé slovo či zmeravený tvar slova zo základných slovných druhov sa má zaradiť medzi citoslovcia pri emocionálne vypätom použití (najčastejšie v jednočlennej výpovedi); typ *pozor/pozór/pozóór..., horí/hó-rí/horí-í-í* je zrejmý, hoci iba lexéma *pozor* je v doterajších lexikografických príručkách spracovaná aj ako citoslovce, ale výrazy/vyjadrenia ako napr. *prasknéé, peknéé* a pod. sa pri ručnej anotácii hodnotia ako defektné zápisy a priraďujú sa im hodnoty ako slovesnému tvaru *praskne* a adjektívu *pekné*.

Pri spracúvaní, dezambiguovaní niektorých homonymných tvarov v korpuse pomáhajú rôzne formálne znaky. Ukazuje sa však, že pri citoslovciach to nemôžu byť ani veľké začiatočné písmená (okrem spoluvýskytu podobných apelatív aj homonymné častice sa môžu začínať veľkým písmenom), ani postavenie (aj homonymné častice, podstatné mená a slovesá môžu stáť na začiatku vety), ba ani písanie čiarky či iných interpunkčných znamienok, ktoré je rozkolísané tak v norme, ako aj v používaní. Predmetom samostatného lingvistického výskumu a spracovania by mohlo byť postavenie citosloviec vo vete: oproti charakteristike

citosloviec ako samostatných výpovedných jednotiek (ekvivalentov vety) stojí ich reálne používanie v zapojenosti do vetnej štruktúry ako prísudkov, predmetov, podmetov, príslovkových určení. Tieto ich funkcie podstatne prevyšujú pôvodné rečové a teoreticky formulované použitie ako samostatných výpovedí. Napokon, korpus písaných textov neprináša veľa dokladového materiálu na výskum citosloviec a keby v ňom nebolo dostatok prekladov z cudzích jazykov, frekvencia tohto slovného druhu by bola ešte nižšia.

Nové verzie Slovenského národného korpusu sú už automatizovane morfologicky anotované aj na základe natrénovania anotačného nástroja na väčšom ručne anotovanom podkorpuse, aj s pomocou analyzátora tvarov slovenského jazyka, ktorý obsahuje súpisy nehomonymných lexikálnych jednotiek. Postupne sa do skvalitňovania morfologickej anotácie zapojí aj dezambiguátor a syntakticky anotovaný podkorpus, pomocou ktorých sa morfologická anotácia opäť skvalitní. Doterajšie pozorovania a výsledky môžu poslúžiť na signalizáciu citlivých miest a na porovnanie dosiahnutého percenta úspešnosti v jednotlivých verziách anotovaného korpusu. Pre časť neohybných slovných druhov, najmä pre citoslovcia prinesie nový výskumný materiál pripravovaný korpus hovorených komunikátov.

Literatúra

FORRÓOVÁ, Martina – GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Návrh morfologického tagsetu SNK. In: Medzinárodná konferencia SLOVKO 2003 – Slovanské jazyky v počítačovom spracovaní (Bratislava 24. – 25. októbra 2003), rkp. 20 s. http://korpus.juls.savba.sk/publications/block2

FORRÓOVÁ, Martina – HORÁK, Alexander: Morfologická anotácia korpusu. In: Slovenčina na začiatku 21. storočia. Ed.: M. Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 174 – 186.

GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. Bratislava: Slovenský národný korpus JÚĽŠ SAV 2004. Interný materiál, rkp. 24 s. http://korpus.juls.savba.sk/publications/block2

GARABÍK, Radovan – GIANITSOVÁ, Lucia – KARČOVÁ, Agáta – ŠIMKOVÁ, Mária: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. Bratislava: Slovenský národný korpus JÚĽŠ SAV 2006. Interný materiál, rkp. 27 s. http://korpus.juls.savba.sk/publications/block2

HAJIČ, Jan: Popis morfologických značek – poziční systém. Praha: Ústav Českého národného korpusu FF UK 2000. http://ucnk.ff.cuni.cz/manual/znacky.html

HANA, Jiří – HANOVÁ, Hana: Manual for morphological annotation. ÚFAL/CKL Technical report TR-2002-14. Praha: Ústav formální a aplikované lingvistiky MFF UK 2002.

LEECH, Geoffrey: Anotační systémy pro značkovaní korpusu. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae – Philologica 3 – 4. Praha: Karolinum 2000, s. 185 – 197.

ŠIMKOVÁ, Mária: Z gramatiky a lexikografie tzv. malých slovných druhov. In: Lexicographica ′99. Zborník na počesť Kláry Buzássyovej. Ed. S. Ondrejovič – M. Považaj. Bratislava: Veda 2001, s. 189 – 202.

ŠIMKOVÁ, Mária: Tzv. gramatické slová v morfológii a v syntaxi. In: Tradícia a perspektívy gramatického výskumu na Slovensku. Ed. M. Šimková. Bratislava: Veda 2003, s. 233 – 239.