Čo sa možno dozvedieť zo Slovenského národného korpusu?

Mária Šimková

(Príspevok bude publikovaný v časopise Čeština doma a ve světě, v čísle venovanom 10. výročiu univerzitnej slovakistiky v ČR.)

V jednom z monotematických čísel časopisu Čeština doma a ve světě sa v r. 2001 predstavil Český národný korpus, ktorý mal v tom čase už rok na internete prístupný stomiliónový reprezentatívny korpus písaných textov súčasnej češtiny SYN2000. Na Slovensku, konkrétne v Jazykovednom ústave Ľ. Štúra SAV a na Pedagogickej fakulte Univerzity Komenského v Bratislave, sa síce už na konci 80. rokov minulého storočia niekoľkí priekopníci v tejto oblasti (na čele s prof. J. Horeckým) zaoberali myšlienkou budovania korpusu textov súčasnej slovenčiny, ale od myšlienky k jej naplneniu muselo uplynúť vyše desať rokov, kým sa v r. 2000 začal formulovať širší projekt. V r. 2001 sa hľadalo a získavalo jeho finančné zabezpečenie, čomu napomáhala jednak existencia zákona o štátnom jazyku, jednak vtedy prebiehajúce prístupové rokovania a perspektíva začlenenia Slovenska do Európskej únie. Rok 2002 je rokom vzniku nového oddelenia JÚĽŠ SAV, oddelenia Slovenského národného korpusu, ktorého pracovníci mali od samého začiatku veľkú oporu v Ústave Českého národného korpusu FF UK, ale aj v ďalších ústavoch a centrách, ktoré sa v ČR venujú korpusom a korpusovej lingvistike: v Ústave formálnej a aplikovanej lingvistiky MFF UK, v Centre komputačnej lingvistiky MFF UK, v Ústave teoretickej a komputačnej lingvistiky FF UK, v Katedre informačných technológií Fakulty informatiky MU v Brne. Na Slovensku sa budovaniu korpusu a korpusovej lingvistike systematicky venuje iba toto jedno špecializované oddelenie JÚĽŠ SAV (8 pracovných miest), pričom príslušný odbor sa nevyučuje na žiadnej vysokej škole.

Hoci sa v JÚĽŠ SAV zhruba od r. 1993 do r. 2002 budoval korpus textov slovenského jazyka a lexikálna báza dát (bližšie Šimková, 2004a, 2004b, resp. http://korpus.juls.savba.sk/korpus/bibliotéka), SNK nemal na vstupe k dispozícii nijaké texty, pretože k predchádzajúcim zmluvy s poskytovateľmi buď neexistovali, alebo neobsahovali možnosť začlenenia textov do korpusu prístupného na internete. Rovnako technika ich spracovania nezodpovedala aktuálnym štandardom – korpus textov slovenského jazyka sa indexoval (bez lematizácie a akejkoľvek anotácie okrem základných bibliografických údajov) a spravoval pod operačným systémom MS DOS programom WordCruncher, ktorý prejavoval výrazné kapacitné limity už pri 200 tisícoch jednotlivých výskytov slov a celkovo na hranici 20 miliónov slov. Necelé dva roky intenzívnej práce oddelenia SNK (od začiatku r. 2003) prinášali postupne výsledky v podobe menších testovacích verzií prim0.1 a prim0.2, ktoré sa

hneď dávali k dispozícii používateľom aj na internete, aby sa overila výkonnosť a spoľahlivosť technického vybavenia pracoviska, vhodnosť softvéru pre potreby používateľov korpusu, najmä lexikografov, ale i spôsob spracovania textov (segmentácia, anotácie). V súčasnosti je na internete prístupný takmer 200 miliónový korpus ***prim1*** – primárny, všeobecný korpus písaných textov súčasnej slovenčiny pozostávajúci v prevažnej miere z publicistických textov posledného desaťročia. Postupne sa spracúvajú staršie texty neexistujúce v elektronickej podobe a plánuje sa vytvorenie vyváženého korpusu pokrývajúceho slovnú zásobu rokov 1955 – 2005. Na prehliadanie sa používa korpusový manažér Manatee s klientom Bonito (z FI MU Brno), ale programátori SNK vyvíjajú vlastný korpusový manažér ***korman***, ktorého testovaciu verziu môžu záujemcovia o prácu s korpusom používať bez registrácie. SNK je lematizovaný (ku každému slovu, resp. tvaru sa priraďuje jeho základná podoba), podrobne bibliograficky a štylisticky-žánrovo anotovaný, pracuje sa na morfologickej anotácii, perspektívne sa časť textov bude anotovať aj syntakticky. V rámci získavania aktuálnych poznatkov z diania v odbore odznelo na pôde SNK doteraz 30 prednášok, zväčša českých autorov (abstrakty i celé handauty sú na stránke SNK, do konca r. 2004 vyjde zborník rozšírených prednášok v anglickom jazyku).

Slovenský národný korpus poskytuje informácie o systéme slovenského jazyka a prehľad reálneho fungovania jazykových prostriedkov v písaných textoch pre všetkých záujemcov, ktorých môžeme rozdeliť na bežných záujemcov (učitelia, študenti, redaktori a pod.), lingvistov a odborníkov v oblasti NLP (natural language processing – počítačové spracovanie prirodzeného jazyka). V súčasnosti pracuje s prim1 takmer 150 registrovaných používateľov, medzi ktorými sú aj lingvisti z ČR a iných krajín (Bielorusko, Fínsko, Juhoslávia, Maďarsko, Nemecko, Poľsko, Rakúsko, Slovinsko, Srbsko a Čierna Hora). Český používateľ môže byť prekvapený, keď v slovenskom korpuse nájde viacero českých slov – druhé najfrekventovanejšie české slovko *se* sa v prim1 nachádza 7 363 ráz. Vysvetľuje to prítomnosť českých textov v slovenskej publicistike, keďže najmä v prvej polovici 90. rokov prinášali hlavné slovenské denníky pomerne rozsiahle výpisy z českej tlače, ktoré sa v tejto verzii korpusu spracúvajú spolu so slovenskými (ale aj anglickými) textami. Neskôr sa budú súvislé cudzojazyčné texty automaticky selektovať. Predstava neprimerane vysokého podielu českých slov v slovenskom korpuse však môže byť ovplyvnená aj vysokou mierou lexikálnej podobnosti blízkopríbuzných jazykov. Slová ako *abeceda, adresa, bard, babička, bahno, banka, barbar, jeden, jednak, navždy, obsah, odmlada, odnož, odpor, podstata, postup, posun, posyp, pošta, povel, stránka, však, zároveň* a mnoho ďalších vyzerajú rovnako v slovenčine i v češtine. Patria medzi ne aj frekventované slová základnej slovnej zásoby: *a, i, v, na, je, že, s, z, o, do; žena, matka, muž, chlap, chlapec, strom, ráno, večer, včera, zima, ruka, noha, hlava,*

život a pod. Slová typu *dnes, otec* sa líšia len výslovnosťou, resp., ako je to napr. pri slove *nebo*, aj sémantikou (sl. obloha, raj – čes. *nebe*; čes. spojka – sl. *alebo*).

Textové korpusy môžu dobre poslúžiť pri zisťovaní systémových zhôd a rozdielov medzi dvoma jazykmi. Keďže slovenský prim1 je lematizovaný a sú k dispozícii frekvencie základných tvarov slov, porovnáme ich s výsledkami SYN2000, ako boli publikované v spomínanom dvojčísle časopisu Čeština doma a ve světě (Kopřivová – Křen, 2001, s. 98 – 120).

Český korpus SYN2000 obsahuje 100 miliónov textových slov (tokenov), z nich je 1 763 818 rôznych slovných tvarov. Slovenský korpus prim1-snk-sane (prístupný iba oddeleniu SNK, na internete je o niečo menšia verzia) obsahuje 190 351 993 tokenov, z nich je 1 813 096 rôznych slovných tvarov. Rozdiel necelých 50 tisíc rôznych slovných tvarov oproti rozdielu vyše 90 miliónov tokenov medzi českým a slovenským korpusom neznamená, že slovenčina je výrazne chudobnejšia na rôzne slová a tvary. Ide o jav charakteristický pre korpusy všetkých jazykov: kým je korpus menší, pomer všetkých a unikátnych slov, resp. tokenov je väčší. So zväčšovaním korpusu sa pôvodne ostrá krivka rôznych tvarov stále viac zarovnáva, počet výskytov nových typov jazykových prostriedkov v texte postupne klesá, hoci frekvencia nových typov je vždy vyššia v textových segmentoch naprieč textami ako vo vzorke jedného textu. Korpusové štatistiky uvádzajú, že v korpuse, ktorý obsahuje 100 miliónov textových jednotiek, sa 8 tisíc jednotiek nachádza v 95 percentách textu a zvyšných 5 percent reprezentuje 500 tisíc jednotiek. J. Mistrík, autor prvého frekvenčného slovníka slovenčiny, uvádza, že prvých 10 slov vyčerpáva 26,67 % textu, ďalších 10 slov 8,47 % textu, prvých 20 slov spolu 35,14 % textu. Ďalších 10 slov už pokrýva iba 5,52 %, prvých 30 slov spolu 40,66 %. Prvých 100 najfrekventovanejších slov pokrýva 56,13 % všetkých slov (Frekvencia slov v slovenčine; ďalej FSS, 1969, s. 97). Autor ručne pracoval s „korpusom“ v rozsahu 1 milión slov.

Najfrekventovanejšie slovné tvary

SYN2000		prim1		FSS	
a	2 690 157	byť	4 253 781	a	35 273
se	1 997 092	v	4 161 871	byť	30 316
v	1 836 848	a	3 815 048	sa	21 401
na	1 532 219	sa	3 526 153	v	21 332

je	950 984	na	2 921 201	na	18 419
že	858 341	ten	1 697 806	on	16 741
s	711 281	ktorý	1 409 842	ten	16 576
z	654 966	s	1 366 095	že	9 045
o	618 915	z	1 302 380	z	8 675
do	589 261	že	1 295 564	ako	8 150

Percentuálne pokrytie príslušného textu frekvenčne prvými tvarmi: v SYN2000 predstavuje *a* 2,69 %, v prim1 *byť* 2,23 %, *a* „len“ 2,00 %. Vo FSS sa *a* nachádza v 3,52 % textu, čo môže byť ovplyvnené podielom štýlov a žánrov vo FSS tak, ako sa v tom čase považovali za relevantné: dialógy 10,53 %, umelecká próza 30,17%, poézia 13,22 %, žurnalistika 14,58 %, náučná próza 31,50 %. SYN2000 má zloženie: umelecká literatúra a literatúra faktu 15 %, noviny a časopisy 60 %, odborná literatúra 25 %. Nevyvážený prim1 obsahuje 95 % publicistiky, 3,5 % umeleckej literatúry, 1,5 % odbornej a populárno-náučnej literatúry. V budúcnosti bude zaujímavé porovnanie terajších frekvencií s frekvenciami vo vyváženom slovenskom korpuse.

Interpunkčné znamienka (FSS ich neuvádza)

SYN2000		prim1	
čiarka	7 647 971	bodka	12 797 784
bodka	7 201 853	čiarka	12 239 155
úvodzovky	1 530 104	úvodzovky	2 145 071
pomlčka	1 220 801	pomlčka	2 124 837
dvojbodka	841 067	dvojbodka	1 710 614
zátvorka)	800 972	zátvorka)	1 519 614
zátvorka (	769 185	zátvorka (	1 508 698
otáznik	260 904	otáznik	338 764
výkričník	134 301	lomka	224 006
lomka	127 555	bodkočiarka	138 975
		výkričník	118 693

Prvé dve najfrekventovanejšie znamienka sa vyskytujú v slovenčine v opačnom poradí v porovnaní s češtinou, ale celkový vzťah medzi čiarkou a bodkou je pri takýchto vysokých číslach dosť tesný – obe znamienka nasledujú v absolútnej frekvencii všetkých tokenov hneď za sebou. Na prvý pohľad medzijazykový rozdiel môže byť v tomto prípade spôsobený aj rozdielmi v tokenizácii (segmentácii textu). Ďalších 5 znamienok sa za prvými dvoma nachádza s rozdielom „triedy“, podobný rádový rozdiel je medzi 7. a 8. miestom. O slovenčine by sme na základe tohto prehľadu mohli povedať, že je menej imperatívna ako čeština – výkričník má v slovenskom korpuse percentuálne viac ako dvojnásobne nižšie zastúpenie ako v českom (0,06 % : 0,13 %).

Najčastejšie predložky

SYN2000		prim1		FSS	
v/ve	2 525 070	v/vo	4 161 871	v/vo	21 332
na	1 664 860	na	2 921 201	na	18 419
s/se	886 521	s/so	1 366 095	z/zo	8 675
z/ze	861 027	z/zo	1 302 380	s/so	7 005
o	660 813	o	1 124 378	do	6 864
do	621 814	do	896 814	o	5 009
k/ke	523 095	za	576 892	za	4 257
pro	401 812	po	566 948	po	4 064
za	391 733	pre	494 822	k/ku	3 871
po	319 788	k/ku	466 866	od	2 404

Na prvých šiestich miestach sú v obidvoch veľkých korpusoch rovnaké predložky, FSS sa odlišuje len v ich poradí zrejme opäť vzhľadom na štýlovo-žánrové rozvrstvenie textov. Na 7. a 8. mieste sa poradie predložiek v prim1 a FSS zhoduje. Príznačné je 7. miesto českej predložky *k* v porovnaní s nižším zastúpením tejto predložky v slovenských textoch vzhľadom na širšiu významovú distribúciu *k/ke* v českom jazykovom systéme. Celkovo je podiel slovenských predložiek od 5. miesta nadol v prim1 percentuálne nižší ako českých v SYN2000 – stále porovnávame korpusy, ktorých veľkosť je v pomere takmer 2 : 1, pričom slovenský korpus je extrémne nevyvážený v prospech publicistiky s jej predpokladane vysokým využívaním menných (predložkových) konštrukcií.

Najčastejšie spojky

SYN2000		prim1		FSS	
a	2 831 331	a	3 815 048	a	35 273
že	867 283	že	1 295 564	že	9 045
i	543 845	aj	957 181	ako	8 150
ale	389 016	ako	772 130	čo	6 519
aby	205 966	ale	472 654	aj	5 619
nebo	201 500	čo	432 547	ale	5 448
když	193 081	však	381 944	keď	4 261
však	190 568	keď	345 037	i	3 051
až	162 687	i	290 175	ani	2 766
než	153 767	aby	286 064	aby	2 166

Okrem prvých dvoch spojok všetky ostatné vykazujú značné rozdiely nielen medzi slovenčinou a češtinou, ale aj v obidvoch slovenských zdrojoch. Okrem veľkosti a štýlovo-žánrového štruktúrovania korpusov tu svoju úlohu zohráva aj slovnodruhová homonymia a zastúpenie tých istých foriem v inom slovnom druhu, resp. eliminácia tejto skutočnosti zjednoznačnením konkrétnych výrazov pri morfologickej anotácii korpusu. Prim1 zatiaľ nie je morfologicky anotovaný a autor FSS sa podľa našich vedomostí tejto problematike nevenoval, preto napr. frekvencia slov *ako, čo, však* zahŕňa nielen spojky, ale aj zámená, resp. častice. Naproti tomu sme medzi spojky v tomto prehľade nezaradili slovo *ktorý*, hoci sa v prim1 nachádza medzi 10 najfrekventovanejšími slovami – v slovenských slovníkoch a gramatických príručkách sa uvádza ako opytovacie zámeno a len v jeho 3. a 4. význame sa konštatuje, že uvádza vetu, teda ide o vzťažné zámeno vo funkcii spojky. Nazdávame sa, že výraz *ktorý* je frekventovanejší vo funkcii spojky ako v pozícii zámena, čo by po dôkladnejšom zhodnotení korpusového materiálu mohlo priniesť rozšírenie jeho slovnodruhovej príslušnosti a následne zmenu jeho zastúpenia pri posudzovaní frekvencie lexém, teda aj na základe rozlíšenia slovných druhov.

Porovnanie 10 najfrekventovanejších predložiek a spojok z hľadiska celkového a percentuálneho podielu na ploche korpusu

	SYN2000	prim1	FSS

predložky	8 856 533	8,86 %	13 878 267	7,29 %	81 900	8,19 %
spojky	5 739 044	5,74 %	9 048 344	4,75 %	82 298	8,23 %
pomer	1,54 : 1		1,53 : 1		0,99 : 1	

Vo veľkých elektronických korpusoch má prvých 10 predložiek o polovicu vyšší podiel v texte ako prvých 10 spojok, vo FSS je ich pomer vyrovnaný, spojky dokonca nepatrne prevládajú. Vysvetlenie prevahy predložiek môže byť v rozsiahlych menných skupinách súčasnej náučnej literatúry a publicistiky, ktorá má v SYN2000 a najmä v prim1 výrazný podiel. Ak sa podrobnejšie pozrieme na texty analyzované vo FSS, tak okrem časového rozdielu (prevažne 50. a 60. roky 20. storočia, čo bol iný štýl aj v žurnalistike, nieto ešte v iných žánroch) zistíme, že v skupine umeleckej prózy, ale aj odbornej literatúry sú významne zastúpené detské časopisy a učebnice: Včielka, Ohník, Čítanka pre 1. a 2. roč. základnej školy (!), Vlastiveda pre 4. roč. ZŠ, Prírodopis pre 9. roč. ZŠ, Botanika pre 10. roč. SVŠ, Prehľad stredoškolskej matematiky. Navyše FSS obsahuje aj 10 % dialógov (scenáre divadelných hier) a 13 % poézie. Na porovnanie jazyka 50. – 60. rokov s dnešným by bolo potrebné vytvoriť podobne štruktúrovaný korpus a použiť rovnakú metódu analýzy, aká bola uplatnená vo FSS. Schodnejšia cesta však bude vytvoriť v rámci projektu SNK súčasnými metódami porovnateľné podkorpusy pre jednotlivé etapy (dekády) súčasného jazyka, teda aj pre 50. a 60. roky. V tomto prehľade sa potvrdilo, čo sme avizovali vyššie, že prim1 obsahuje percentuálne menej predložiek, ba aj spojok než SYN2000. Na vysvetlenie tohto zistenia by bola potrebná podrobnejšia analýza, obsahujúca napr. prehľad a porovnanie výskytu všetkých predložiek a spojok.

Najčastejšie podstatné mená

SYN2000		prim1		FSS	
rok	406 834	rok	647 188	človek	2 142
člověk	183 509	hodina	240 953	číslo	1 859
Praha	131 958	Slovensko	234 818	rok	1 724
strana	130 835	Bratislava	201 361	čas	1 422
doba	111 053	strana	200 636	deň	1 416
den	110 289	deň	187 69	ruka	1 357

			3		
země	103 622	ľudia	172 452	oko	1 283
společnost	90 657	čas	171 583	svet	1 276
svět	85 503	vláda	171 367	strana	1 204
firma	84 285	mesto	156 803	voda	1 201

Vidíme, že kým v súčasných korpusoch s vysokým podielom publicistiky prevládajú časové a miestne, príp. spoločensko-politické a ekonomické informácie, v minulosti sa ľudia viac venovali bežným ľudským potrebám (*voda*) a sebe (*ruka, oko*). Vysoká absolútna frekvencia podstatného mena *číslo* vo FSS môže súvisieť so spomínanými časopismi alebo Prehľadom stredoškolskej matematiky – slovo *číslo* má podstatne nižšiu relatívnu frekvenciu, ktorá ho podľa tohto kritéria zaraďuje vo FSS ďaleko za prvých 10 podstatných mien. V elektronických korpusoch je nepopierateľné vysoké prvenstvo podstatného mena *rok* zrejme opäť vďaka publicistike, hoci reprezentatívny SYN2000 obsahuje podstatne menej textov tohto štýlu ako prim1 – na frekvencii slova *rok* sa to však neprejavuje. 2. miesto podstatného mena *člověk* v SYN2000 má v prim1 na prvý pohľad pendant v tvare *ľudia*, ktorý je až na 7. mieste. Pravdepodobne však nejde o rozdiel vo vnímaní individuálneho postavenia človeka, ale o rozdiel v lematizácii – prim1 má osobitnú lemu *ľudia* a osobitnú *človek* s frekvenciou 63 245. Ak by sme obidve čísla spočítali, dostaneme výsledok 235 697 výskytov a spoločná lema *človek* by sa dostala v prim1 na 3. miesto s neveľkým odstupom od druhej *hodiny*.

Najčastejšie slovesá a prídavné mená (v citovanom prehľade vybraných frekvencií SYN2000 sa neuvádzajú)

prim1		FSS	
byť	4 253 781	byť	30 316
mať	976 824	mať	7 119
môcť	426 071	môcť	3 285
povedať	221 680	vedieť	2 920
musieť	216 036	ísť	2 392
chcieť	213 731	povedať	2 348
ísť	188 835	chcieť	2 283
vedieť	166 058	musieť	2 022
dať	162 732	dať	1 891

dostať	146 226	vidieť	1 773

Presvedčivé je postavenie prvých troch slovies aj s veľkými rozdielmi medzi nimi. Na 4. – 7. mieste je skupina slovies s malými rozostupmi, v skupine sa striedajú 3 rovnaké slovesá (vôľové, pohybové a sloveso hovorenia), štvrté z nich (*vedieť*, *musieť*) sa striedajú v skupine a v tesnom závese za ňou. 9. miesto patrí rovnakému slovesu *dať*. K poslednému riadku len lakonicky: lepšie je dostať, ako vidieť ..., pravda, ak ide o niečo dobré, užitočné. Tieto dve slovesá sú aj jediné odlišné, ostatných 9 najfrekventovanejších slovies je rovnakých vo FSS i v prim1 napriek veľkým rozdielom v čase a zložení textov.

prim1		FSS	
slovenský	279 559	veľký	1 845
nový	277 555	celý	1 740
veľký	263 347	nový	1 352
ďalší	197 316	slovenský	1 104
dobrý	189 253	dobrý	1 017
celý	144 761	starý	1 012
vysoký	125 500	malý	889
posledný	124 199	mladý	584
malý	111 923	vysoký	576
štátny	105 053	vlastný	441

V prvej desiatke najfrekventovanejších prídavných mien sa aj s rozdielom 40 rokov opakuje 7 prídavných mien, čo považujeme takisto za dosť vysoké percento aj vzhľadom na to, že v celkovom poradí sa posledné slovo tejto skupiny (*štátny*) nachádza v prim1 blízko miesta 200, teda dosť hlboko v slovnej zásobe. Namiesto prídavných mien *starý, mladý, vlastný* z prvej desiatky FSS sa v prim1 nachádzajú slová *ďalší, posledný, štátny* – a to v 60. rokoch už takmer neexistovalo súkromné vlastníctvo a naopak v 90. rokoch sme sa usilovali vymaniť z područia štátneho ... Za nimi sú na nasledujúcich pozíciách prídavné mená *domáci, politický, starý, český* (89 474 výskytov), *európsky* (82 954, ale to môže byť aj príslovka),

americký (81 518). Hoci slovo *americký* sa medzi prídavnými menami nachádza na pomerne vysokom 16. mieste, nevidíme medzi najfrekventovanejšími výrazmi ani v tejto skupine, ani v predchádzajúcich skupinách nijaké z „obávaných“ všade a najmä do publicistiky prenikajúcich anglicizmov či amerikanizmov. Ak by sme len za anglicizmy nepovažovali slová *prezident* (151. miesto podľa absolútnej frekvencie v prim1), *percento* (164.), *situácia* (202.), *firma* (233.), *banka* (237.) a pod., čo sú však slová latinského, resp. posledné dve talianskeho pôvodu. Ako prvé z angličtiny prevzaté slovo novšieho dáta nachádzame slovo *tím* na 297. mieste so 62 260 výskytmi, čo je však skreslené prítomnosťou českých textov a, ako sme zistili, nesprávnou lematizáciou tvaru zámena *ten* v 7. páde.

Frekvencia názvov mesiacov

SYN2000		prim1		FSS	
září	18 891	september	31 026	máj	225
leden	18 541	január	30 990	apríl	104
květen	18 197	máj	30 409	jún	95
listopad	17 812	marec	29 088	september	76
červen	17 341	jún	29 049	október	70
říjen	16 604	november	27 058	január	67
duben	15 380	apríl	26 130	august	66
březen	15 190	júl	25 823	júl	65
srpen	14 806	december	25 332	november	56
prosinec	14 383	august	24 897	marec	50
červenec	14 048	október	24 454	február	49
únor	13 663	február	23 197	december	43

Vo veľkých korpusoch je poradie mesiacov napoly totožné – na 1., 2., 3., 5., 7. a 10. mieste sú v SYN2000 aj v prim1 rovnaké názvy mesiacov. Mesiace *september* a *január* na

prvých miestach zrejme odrážajú zvýšenú všeobecnú i mediálnu aktivitu na začiatku školského a kalendárneho roka. Na posledných miestach sa zas pohybujú mesiace letných a zimných prázdnin a mesiace nasledujúce v kalendári po exponovaných mesiacoch (*október, február*). FSS pracuje s malým rozsahom heterogénnych textov, kde je frekvencia špecifickejších pomenovaní typu názvy mesiacov a dní značne determinovaná obsahom konkrétnych textov.

Frekvencia názvov dní

SYN2000		prim1		FSS	
sobota	21 592	sobota	74 042	nedeľa	110
neděle	19 855	nedeľa	54 157	sobota	60
pátek	13 580	streda	33 795	piatok	50
pondělí	12 800	piatok	31 340	streda	30
středa	12 505	pondelok	26 390	pondelok	28
čtvrtek	10 429	štvrtok	25 905	štvrtok	15
úterý	10 329	utorok	24 039	utorok	15

Poradie názvov dní neopakuje rozloženie názvov mesiacov podľa pracovnej vyťaženosti, ba je práve opačné. Na posledných miestach sú *štvrtok* a *utorok*, ktoré by mali byť najvyťaženejšími pracovnými dňami, na prvých miestach sú víkendové dni a predvíkendový *piatok*, ktorý v prim1 predbehla futbalová *streda*. Na rozdiel od predchádzajúcich prehľadov frekvencií, ktoré nevyžadovali zásadnejšie korekcie (okrem slovnodruhovej homonymie pri spojkách a čiastočne predložkách), názvy dní sa objavujú aj ako vlastné mená ľudí a obcí, čo kladie zvýšené nároky na jednoznačnú lematizáciu. Pritom nestačí automaticky selektovať slová začínajúce sa malým a veľkým písmenom, pretože aj všeobecné meno môže stáť na začiatku vety a byť napísané s veľkým začiatočným písmenom. Autori príspevku o vybraných frekvenciách SYN2000 podrobili prvé zistené výsledky podrobnejšej analýze a po korekcii homonymných tvarov upravovali niektoré frekvencie od – 327 (*sobota*) až po + 105 (*piatok*). Poradie na prvých miestach sa nezmenilo, ale pôvodne

posledný *štvrtok* sa posunul pred *utorok* (v tabuľke citujeme upravenú frekvenciu). Výskyty názvov dní v prim1 nie sú korigované z hľadiska homonymie. Vieme, že *streda* obsahuje aj množstvo výskytov z pomenovaní *Dolná Streda, Horná Streda, Nitrianska Streda* a najmä z obcí *Dunajská Streda* a *Streda nad Bodrogom* frekventovaných v športovej publicistike. Do frekvencie *štvrtka* sú zahrnuté aj názvy *Štvrtok na Ostrove, Plavecký Štvrtok, Spišský Štvrtok*, do *soboty* priezvisko známeho českého herca, ale aj *Rimavská Sobota* a *Spišská Sobota*. V pondelky, utorky a piatky sa na Slovensku v minulosti asi nekonávali trhy, pretože tieto názvy dní sa v pomenovaniach obcí objavujú len výnimočne (*Pondelok* – časť Hrnčiarskej Vsi); český korpus však zachytáva obce *Úterý* a *Pátek nad Ohří*. *Neděla* ako pomenovanie sviatočného dňa sa v názvoch obcí nezvykne vyskytovať vôbec. V prehľade zo SYN2000 sa uvádza modifikovaná podoba *Neděliště* a podobnou príponou sa utvorilo na Slovensku slovo *Sobotište*, ktoré pomenúva miesto konania trhov.

Najčastejšie mužské mená

SYN2000		prim1		FSS	
Jan	41 570	Ján	65 833	Ondrej	364
Jiří	34 506	Peter	65 262	Ján	338
Václav	28 125	Jozef	52 693	Jozef	274
Petr	26 018	Martin	45 174	Pavol	239
Josef	25 226	Vladimír	32 996	Michal	219
Pavel	21 384	Pavol	32 590	Štefan	185
Karel	19 641	Ivan	31 591	Adam	180
Vladimír	16 650	Milan	28 444	Peter	89
Jaroslav	16 062	Mikuláš	24 435	František	79
Martin	14 736	Štefan	23 123	Juraj	74

Medzi najfrekventovanejšími krstnými menami sú síce tradičné (domáce) mená, no prejavujú sa tu isté rozdiely v kultúrnych vzorcoch. Kým čeština viac frekventuje mená *Jiří* (v

slovenčine je meno *Juraj* trochu prekvapujúco až na konci prvej desiatky, aj to iba vo FSS), *Václav, Karel, Vladimír* a *Jaroslav*, slovenskí muži sú častejšie nositeľmi mien *Martin, Ivan, Milan, Mikuláš, Štefan*. Na pomerne vysokú pozíciu mena *Mikuláš* iste vplýva aj to, že jeho nositeľom je premiér Dzurinda, rovnako ako v češtine posunuli meno *Václav* na 3. miesto bývalý prezident Havel a terajší prezident Klaus, ktorí boli v čase budovania českého korpusu jednými z najexponovanejších politikov. Ako uvádzajú autori prehľadu frekvencií SYN2000, problémom pri týchto menách je homonymia ich tvarov s niektorými tvarmi ženských krstných mien (*Jan, Jana, Janu, Jani, Jany*), v slovenskom korpuse ide aj o homonymiu s názvami miest (*Martin*, resp. *Turčiansky sv. Martin, Liptovský Mikuláš, Borský Mikuláš*). Navyše sa tu z už spomínaných dôvodov nachádza pomerne často na prvý pohľad české meno *Jan* (11 964), ktorého frekvencia je však navýšená nesprávnou lematizáciou foriem *JAN* (v športových rubrikách skratka pre január), *Janov* (mesto v Taliansku) a tvaru ženského mena *Jana*. FSS vzhľadom na svoj rozsah a výber textov nemá ani tu veľkú výpovednú hodnotu, no dá sa z neho získať základný prehľad neoficiálnych, familiárnych modifikácií krstných mien (*Jano, Janko, Janíčko; Mišo, Miško; Peťo, Petrík; Paľo, Paľko; Juro, Ďurko; Fero, Ferko*). Frekvencia týchto tvarov je v niektorých prípadoch vyššia ako pri oficiálnych menách, čím sa opäť potvrdzuje rozdiel medzi publicistikou, kde sa používajú neutrálne podoby, a umeleckou literatúrou, kde naopak prevládajú citovo podfarbené zdrobneniny.

Najčastejšie ženské mená

SYN2000		prim1		FSS	
Marie	9 070	Mária	11 478	Eva	168
Jana	7 679	Anna	11 377	Mária	156
Eva	5 839	Eva	10 616	Anna	117
Anna	4 260	Zuzana	8 629	Hedviga	106
Hana	3 934	Jana	7 944	Zuzana	85
Helena	3 343	Katarína	6 560	Helena	84
Věra	3 325	Lucia	5 891	Katarína	76
Petra	3 307	Martina	5 633	Johanka	58
Kateřina	3	Andrea	4 090	Klára	31

	218				
Zuzana	3 121	Marta	3 719	Terézia	21

Pri porovnaní mužských a ženských krstných mien si všimneme dve veci: mužské mená sú 4- až 6-krát frekventovanejšie ako ženské, ženské mená prejavujú väčšiu konzistenciu ako mužské. Na prvých troch miestach sú v zásade rovnaké mená (*Mária, Anna, Eva*, s výnimkou českej *Jany*), v celej skupine je rovnakých 6 mien. Rozdielne mená v češtine a slovenčine sú v druhej polovici tabuľky (čes. *Hana, Helena, Věra, Petra* – sl. *Lucia, Martina, Andrea, Marta*). Vo FSS je z hľadiska zloženia jeho textov príznačné umiestnenie mien *Hedviga* a *Johanka*, ktoré vo veľkom korpuse, resp. medzi nositeľkami nemajú nijako vysoké výskyty. Najbohatšie hniezdo familiárnych tvarov je pri mene *Mária* (*Mariša, Marka, Marienka, Mara, Maňa*), ostatné zdrobneniny uvedené vo FSS sú pomerne štandardné (*Anička, Anča; Katka; Zuza, Zuzka; Evka*).

Frekvencia kontinentov

SYN2000		prim1		FSS	
Evropa	41 121	Európa	64 492	Európa	88
Amerika	10 030	Amerika	17 902	Amerika	58
Afrika	4 199	Austrália	8 796	Afrika	36
Asie	3 787	Afrika	5 878	Ázia	15
Austrálie	3 520	Ázia	5 002	Austrália	14
Antarktída	428	Antarktída	463	Antarktída	3

Najčastejšie krajiny

SYN2000		prim1		FSS	
ČR	46 607	Slovensko	421 871	Slovensko	281
USA	45 576	USA	174 665	Rakúsko	75
Německo	28 744	Česko	50 686	Nemecko	72
Rusko	20 694	Rusko	39 967	Československo	70
Slovensko	20 353	Nemecko	38 719	Rusko	56
Francie	16	Francúzsko	26 177	Taliansko	56

	591				
Polsko	13 037	Maďarsko	24 747	Poľsko	49
Itálie	12 961	Rakúsko	22 969	Anglicko	49
Rakousko	10 725	Poľsko	22 323	Británia	46
Československo	9 788	Taliansko	20 549	Čína	45

Frekvencia prvých dvoch kontinentov je vo všetkých korpusoch jednoznačná – sme súčasťou Európy, hneď potom je „najbližšia“ Amerika. Tretie miesto Austrálie v prim1 pravdepodobne podmienilo konanie letných olympijských hier v Sydney v r. 2000. Pri krajinách je situácia v prvej desiatke podstatne pestrejšia, ale zdá sa, že pomerne reálne odráža vzťahy a geografické kontakty dvoch susediacich a veľmi blízkych, no predsa rozdielnych štátov. Prvé miesta v SYN2000 i v prim1 patria vlastnému štátu, na druhom sú zhodne USA. Ďalšie miesta hovoria, že v českej tlači sa spomína najmä Nemecko, potom je v pozornosti Rusko a Slovensko. V SR sa po USA najviac uvádza ČR, potom Rusko a Nemecko. Francúzsko je v oboch korpusoch zhodne na 6. mieste, ďalej sa striedajú tri rovnaké krajiny, ktoré v slovenskom korpuse predbehlo Maďarsko a v českom korpuse uzatvára Československo. Toto pomenovanie bývalého spoločného štátu sa v slovenskom korpuse nachádza až na 3366. mieste s 5252 výskytmi zrejme v súvislosti s tým, že takmer všetky publicistické texty v prim1 pochádzajú z obdobia po rozdelení Československa.

Zaujímavosť pred záverom – frekvencia akademických a vedeckých titulov (FSS ich neuvádza)

SYN2000		prim1	
Ing.	11 677	Ing.	21 773
Dr.	9 007	MUDr.	9 404
MUDr.	2 764	Dr.	7 227
prof.	2 554	JUDr.	5 640
JUDr.	2 318	Mgr.	2 926
CSc.	1 633	CSc.	2 472
doc.	1 351	PhDr.	2 402
PhDr.	1 196	prof.	1 966
Mgr.	1 026	doc.	1 964
RNDr.	790	RNDr.	1 373
DrSc.	349	DrSc.	636
MVDr.	221	MVDr.	604
ThDr.	102	PaedDr.	359
PaedDr.	56	PhD.	282

RSDr.	31	ThDr.	155
Bc.	17	PharmDr.	109
PharmDr.	15	RSDr.	80
RCDr.	5	Bc.	51
		Thlic.	3
		RCDr.	0

včera – dnes – zajtra

SYN2000		prim1		FSS	
dnes	79 927	dnes	126 010	dnes	821
včera	60 799	včera	119 865	včera	211
zítra	8 922	zajtra	18 814	zajtra	151

Vo využívaní základných časových orientátorov v podobe prísloviek *dnes, včera, zajtra* sa ukazuje, že najdôležitejšie je to, čo je dnes, prípadne, čo bolo včera, ale zajtrajšok sa spomína výrazne menej. V úvode príspevku sme takisto vychádzali z minulého stavu budovania korpusu textov slovenského jazyka, predovšetkým sme sa však sústredili na súčasný stav a na porovnanie prehľadov frekvencií vybraných skupín jazykových prostriedkov v dvoch blízkopríbuzných jazykoch. Porovnanie však často presiahlo rámec jazykových systémov a týkalo sa širších spoločensko-politických a kultúrno-historických súvislostí, ktoré sa odrážajú aj v takom úzkom výbere, akým je prvých desať najčastejšie používaných slov, resp. zástupcov príslušného druhu slov. Relevantnosť zistených výsledkov vyplýva z dostatočne veľkého rozsahu a v prípade SYN2000 aj dostatočne reprezentatívneho zastúpenia textov, na ktorých sa frekvencie zisťovali. Zajtrajšie plány Slovenského národného korpusu súvisia práve aj s potrebou vybudovania vyváženého korpusu, v ktorom by boli zastúpené texty všetkých štýlov a žánrov rokov 1955 – 2005 tak, ako to predpokladá projekt SNK, čím by sa o. i. významne zlepšila výpovedná hodnota porovnávania s existujúcimi reprezentatívnymi korpusmi typu ČNK. Českým záujemcom o korpusy a korpusovú lingvistiku môžu v blízkej budúcnosti poslúžiť plánované špecifické súčasti Slovenského národného korpusu, ako je napr. budujúci sa paralelný slovensko-český korpus, prínosom budú aj výsledky spolupráce na využití spoločných vlastností češtiny a slovenčiny pri budovaní anotovaných národných jazykových korpusov, ktoré by sa mohli využiť pri automatizovaných prekladoch medzi češtinou a slovenčinou.

Literatúra

KOPŘIVOVÁ, M. – KŘEN, M.: Korpusový test. In: Čeština doma a ve světě, 2001, roč. 9, č. 1 a 2, s. 91 – 120.

MISTRÍK, J.: Frekvencia slov v slovenčine. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1969. 726 s.

ŠIMKOVÁ, M.: Slovenský národný korpus – východiská a plány. In: Slovenčina na začiatku 21. storočia. Ed: M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004 (a), s. 150 – 158.

ŠIMKOVÁ, M.: Možnosti využitia Slovenského národného korpusu na štúdium slovenského jazyka. In: Studia Academica Slovaca. 33. Ed.: J. Mlacek – M. Vojtech. Bratislava: STIMUL 2004 (b), s. 204 – 218.

http://korpus.juls.savba.sk