FREKVENCIA SLOV A TVAROV V SÚČASNEJ SLOVENČINE

Mária Šimková
Slovenský národný korpus Jazykovedného ústavu Ľudovíta Štúra SAV
Panská 26, 813 64 Bratislava, e-mail: marias@korpus.sk

Šimková, M.: Frequency of Lemmas and Word Forms in Contemporary Slovak. In Slovak Speech, 2011, vol. 76, no. 5-6, p. 322-333.
Abstract: In this paper, basic statistical data about the Slovak linguistic units (lemmas, word forms, collocations, parts of speech or rather word classes) are analysed using various language resources. The frequency of lemmas and word forms in contemporary Slovak according to the dictionary Frekvencia slov v slovenčine (Frequency of Words in the Slovak Language) by Jozef Mistrík (1969) is compared with the data obtained from several Slovak National Corpus versions and some of its subcorpora. This provides an overview about the stability and dynamics of the Slovak language system in various areas of its usage during the last fifty-five years. Statistical information summarized in the frequency dictionaries and lists helps to understand better and more objectively the functioning of the linguistic units in communication. It also helps to determine the attributes of both typologically distant and close languages. This contribution serves as a demonstration of the possibilities of statistical analysis and it will be used as a base for preparation of a new frequency dictionary of Slovak based on the Slovak National Corpus material.
Key words: linguistic units, frequency, stability and dynamics of the Slovak language system

Vymedzenie jadra slovnej zásoby na základe kvantitatívnych ukazovateľov o využití jazykových jednotiek patrí už od konca 19. storočia k základným poznatkom o jazyku, s ktorým pracujú nielen jazykovedci, ale aj bádatelia v oblasti pedagogických a didaktických disciplín, literárnej vedy, logopédie, psychológie, neurológie, matematických a štatistických analýz, najnovšie počítačového spracovania prirodzeného jazyka a v ďalších oblastiach. Donedávna sa frekvenčné analýzy robili ručne na obmedzenom materiáli, v súčasnosti sú k dispozícii elektronické korpusové databázy a štatistické nástroje, pomocou ktorých sa dajú získať relevantné výsledky z veľkých rozsahov rozlične štruktúrovaných a diferencovaných písaných textov či zo zvukových záznamov hovorenej reči a ich prepisov.

Pri tvorbe databáz Slovenského národného korpusu, v ktorých sa spracúvajú texty publikované po r. 1955, sa od verzie prim-2.1 (od r. 2006) zverejňujú na webovej stránke korpus.sk aj štatistiky celého korpusu a osobitných podkorpusov, čo umožňuje porovnať frekvenciu jednotlivých jazykových prostriedkov v rôznych obdobiach a v rôznych štýloch. Vďaka profesorovi Jozefovi Mistríkovi môžeme aktuálne a korpusovými metódami získané výsledky porovnávať aj s údajmi z prvých frekvenčných slovníkov slovenčiny, ktorých je autorom: Frekvencia slov v slovenčine (1969), Retrográdny slovník slovenčiny (1976), Frekvencia tvarov a konštrukcií v slovenčine (1985).

Na rozdiel od jednorazovo zhromaždeného a v podstate uzavretého materiálu, z ktorého sa vychádzalo pri tvorbe uvedených slovníkov, sa Slovenský národný korpus (ďalej SNK) buduje ako verejne na internete prístupná (prezerateľná) nekomerčná databáza – zdroj jazykového materiálu a jazykových informácií na najrôznejšie využitia –, ktorá sa dynamicky dopĺňa a aktualizuje. Každá nová verzia korpusu je kvantitatívnym rozšírením predchádzajúcej verzie, ale sprístupnenie nového korpusu zvyčajne sprevádza aj zlepšená lingvistická anotácia – štýlovo-žánrová a morfologická anotácia a lematizácia. V rámci týchto procesov sa každému textu technicky pripravenému na zaradenie do korpusu ručne priraďuje informácia o jeho štýlovo-žánrových charakteristikách a príslušnosti k vecnej oblasti (ekonómia, prírodné vedy, právo a pod.) a každému slovnému tvaru, každej textovej jednotke (okrem slov sú to interpunkčné znamienka, číslice a iné neslovné znaky v texte – porov. aj ďalej výklad pri grafe č. 2) sa automatizovane priraďuje informácia o základnom tvare a gramatických kategóriách, ktoré nadobúda v danom kontexte.

Hlavná, základná databáza písaných textov SNK sa označuje ako ***prim*** (primárna) a používatelia majú v súčasnosti k dispozícii ôsmu verziu označenú ako *prim-5.0*.[1] Z hlavného korpusu sa vyčleňujú podkorpusy v závislosti od licencie na získané texty: prístupné verejne všetkým – *public*, prístupné v rámci Jazykovedného ústavu – *juls* a dostupné iba v rámci oddelenia – *snk*. V podkorpuse *all* sú zhromaždené všetky texty, ktoré sa v danej verzii dostali do výsledného spracovania, podkorpus *sane* je výsledkom očistenia *all* od textov nezodpovedajúcich všetkým kritériám kladeným na štandardnú slovenčinu (napr. texty bez diakritiky, texty od zahraničných Slovákov). Z neho sa potom vyčleňujú ďalšie podkorpusy podľa troch hlavných štýlov: podkorpus publicistických textov – *inf*, podkorpus umeleckých textov – *img*, podkorpus odborných a populárno-náučných textov – *prf*. Samostatné celky tvoria podkorpus pôvodných textov slovenských autorov – *skimg* a podkorpus vyvážený podľa štýlov – *vyv* (po jednej tretine z každého z troch uvedených štýlov). Na porovnávacie analýzy dynamiky jazyka po politicko-spoločenských zmenách v r. 1989 a stavu pred ňou slúži podkorpus *r55az89*, v ktorom sú iba texty z r. 1955 – 1989. Tento podkorpus je na rozdiel od hlavného, základného korpusu s pomerne vysokým podielom publicistických textov – súčasných denníkov a týždenníkov – zložený hlavne z textov umeleckej literatúry. Z oblasti publicistiky v korpuse r55az89 zatiaľ chýbajú vtedajšie denníky (jediným textom tohto druhu je tu Smena z augusta 1968), väčšie zastúpenie v ňom majú týždenníky a mesačníky, z ktorých je však veľká časť skôr literárneho alebo populárno-náučného charakteru. Ako referenčný (a trénovací) korpus s presnou morfologickou anotáciou je k dispozícii ručne morfologicky anotovaný korpus *r-mak*. Okrem korpusov písaných textov sa v rámci komplexu projektov SNK buduje aj databáza

[1] Prvé, pilotné korpusy prim0.1 a prim0.2 boli sprístupnené v r. 2003, každý nasledujúci rok pribudla nová verzia: prim1, prim-2.0, prim-2.1, prim-3.0, v r. 2009 prim-4.0 a v r. 2011 prim-5.0.

zvukových záznamov prehovorov v štandardnej slovenčine a ich prepisov – Slovenský hovorený korpus. Rozsah najnovšej verzie hovoreného korpusu *s-hovor-3.0* (1,65 mil. jednotiek) už dosahuje rozmery bežného korpusu tohto typu a umožňuje aj porovnanie frekvencie jazykových prostriedkov v písanej a hovorenej komunikácii.

Štatistiky z verzií a podkorpusov SNK, zverejňované na internete od verzie prim-2.1, predstavujú jednoduchý zoznam absolútnych frekvencií foriem – konkrétnych slovných tvarov a lem (základný tvar slova, v ktorom sú obsiahnuté všetky výskyty všetkých jeho tvarov), ktorý však aj v takomto „surovom" stave má svoju výpovednú hodnotu a často sa využíva v rôznych počítačových, ako aj lexikografických aplikáciách (čiastkové analýzy porov. Šimková, 2008, 2010, Šimková – Ľos, 2009). Frekvenčný zoznam sa môže využiť pri tvorbe frekvenčného slovníka, v ktorom sú slová a ich frekvencie lingvisticky upravené (spracovanie vlastných mien, homoným, variantov, viacslovných jednotiek a pod.) a poskytujú sa v ňom okrem absolútnej frekvencie aj ďalšie údaje: relatívna frekvencia, využitie/výskyt v jednotlivých štýloch, monokolokabilita a pod. J. Mistrík predložil vo svojich slovníkoch výsledky ručnej analýzy presne štrukturovaného materiálu v rozsahu 1 milióna slov, čo bol v tom čase štandardný podklad na prípravu frekvenčného slovníka. A hoci išlo o náročné manuálne spracovanie, v materiáli vymedzeného rozsahu aj obsahu sa mnohé (problematické) javy nevyskytovali vôbec alebo sa dali vyriešiť pomerne jednoducho.

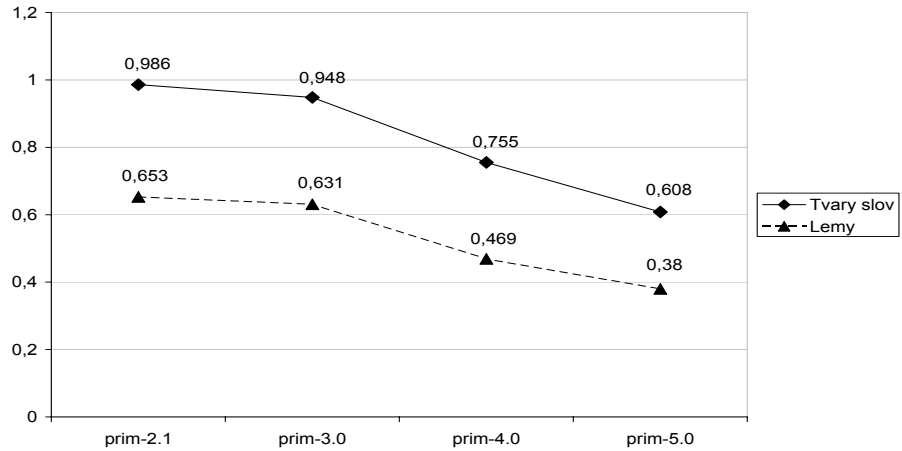
Jazykový materiál takého rozsahu, aký je možné zhromaždiť a analyzovať v korpusových databázach, je na jednej strane vynikajúcim prostriedkom na objektívne skúmanie jazyka a reálneho fungovania jazykových prostriedkov, no na druhej strane otvára nové otázky v lingvistike i v oblasti matematiky a informatiky (teória pravdepodobnosti, zákon veľkých čísel a pod., porov. napr. Králík, 2006). Pri budovaní a využívaní korpusov sa jazykový systém a jeho jednotky jednak akoby úplne obnažujú (každú entitu treba osobitne posúdiť a „naučiť" počítačové nástroje, ako ju identifikovať a správne zaradiť), jednak sa ukazujú v množstve čiastkových i komplexných vzťahov, ktoré sa bežnou ľudskou analýzou nedajú v takom rozsahu odhaliť. Svoje limity však majú aj korpusové možnosti. Predstavy zo začiatkov korpusovej lingvistiky, že stámiliónové databázy obsiahnu všetky jazykové prostriedky a ich tvary i rôzne použitia v dostatočne veľkom počte, že ak sa v menšom korpuse nachádza konkrétne slovo x-krát, tak v desaťnásobne väčšom korpuse sa dané slovo bude nachádzať takisto približne desaťkrát častejšie alebo že sa z korpusu dá na počkanie vybrať zoznam neologizmov či špecifických sémantických charakteristík slov, sa nedajú naplniť, no občas sa vyskytujú ešte aj v súčasnosti. Ich prejavom býva napr. častá otázka, ktoré slovo v danom jazyku je najmenej frekventované. Z údajov získaných z korpusov SNK môžeme o najnižšej frekvencii uviesť iba to, že v takmer 720-miliónovom korpuse verzie 5.0 je vyše 2,7 milióna jednotlivých lem, pričom frekvenciu 1 má približne 50 % z nich.

Samotným korpusovým lingvistom sa na základe porovnávania jednotlivých databáz postupne spresňuje spôsob a rozsah pribúdania unikátnych jazykových jednotiek do korpusu. Z porovnania štyroch posledných verzií SNK sa ukazuje, že pri viac ako dvojnásobnom zväčšení korpusu od verzie 2.1 po verziu 5.0 sa takmer o polovicu znížil podiel unikátnych (jednotlivých) výskytov slov a lem.

Tab. č. 1: Pribúdanie unikátnych tvarov slov a lem v korpusoch SNK

Prim	**Počet jednotiek**	**Počet jednotl. tvarov slov**	Podiel	**Počet jednotl. lem**	Podiel
2.1	**294 087 581**	**2 898 794**	0,986 %	**1 920 646**	0,653 %
3.0	**339 063 215**	**3 214 612**	0,948 %	**2 139 956**	0,631 %
4.0	**526 082 640**	**3 969 719**	0,755 %	**2 467 595**	0,469 %
5.0	**719 499 235**	**4 376 090**	0,608 %	**2 711 202**	0,38%

Graf č. 1: Pribúdanie unikátnych tvarov slov a lem v korpusoch SNK



V uvedených výpočtoch vyzerá na prvý pohľad paradoxne vzájomný pomer jednotlivých výskytov lem a tvarov slov. Napr. paradigmy slovenských substantív by mali obsahovať 12 tvarov (ak vylúčime vokatív a abstrahujeme od rovnakého zakončenia v niektorých tvaroch), paradigmy adjektív a adjektiválií 36 tvarov, paradigmy slovies 25 rôznych tvarov, čiže počet tvarov slov v korpuse by sme mohli očakávať v priemere aspoň 20 x vyšší ako počet lem. Je však vyšší iba približne 1,56 x.[2] Tento nepomer je spôsobený častou tvarovou homonymiou, nepoužívaním všetkých potenciálnych

[2] V korpuse prim-2.1 je 1,51 x vyšší, v 3.0 presne 1,5 x, v korpusoch 4.0 a 5.0 je počet tvarov 1,61 x vyšší ako počet lem.

tvarov a predovšetkým veľkým počtom textových jednotiek, ktoré majú iba jeden, základný tvar (slová z neohybných slovných druhov, neohybné cudzie slová, skratky, číslice, značky, symboly).

Rôzne rozsahy korpusov umožňujú porovnávať aj (kumulatívne) rozloženie jazykových jednotiek. F. Čermák (1997, s. 21) uvádza, že z 54 tisíc slov zachytených vo Frekvenčnom slovníku češtiny kolektívu M. Těšitelovej prvých desaťtisíc najfrekventovanejších slov predstavuje viac ako 90 % spracovaných textov. M. Šulc (1999, s. 15) cituje G. Kennedyho, ktorý odhadoval, že v 100-miliónovom korpuse bude takmer 95 % slovných výskytov reprezentovať asi 8 tisíc lexikálnych jednotiek, zvyšných 5 % by však mohlo predstavovať okolo pol milióna lexikálnych jednotiek. Z našich výpočtov vyplýva, že v takmer 720-miliónovom korpuse sa nachádza 4,376 milióna jednotlivých výskytov slovných tvarov a z nich je vyše polovica (2,339 mil.) s frekvenciou 1 (53,44 %), čo predstavuje 0,325 % všetkých jednotiek v korpuse.[3] 5 % korpusu bude teda vo väčších rozsahoch textov tvoriť zrejme oveľa väčšie množstvo jednotiek. Podiel najfrekventovanejších jednotiek na rozsahu celého korpusu takisto závisí od objemu textov a od toho, či sa započítavajú všetky textové jednotky (tokeny) alebo iba slová. V nasledujúcej tabuľke sa nachádza porovnanie kumulatívnych frekvencií prvých desiatich, dvadsiatich, tridsiatich a prvých sto najfrekventovanejších lem (zahŕňajúcich iba slová) v materiáli spracovanom v Mistríkovom Frekvenčnom slovníku slovenčiny a v posledných dvoch základných elektronických korpusoch. Z verzie 5.0 uvádzame na porovnanie aj kumulatívne frekvencie pri zahrnutí všetkých textových jednotiek vyskytujúcich sa v prvej stovke najfrekventovanejších lem.

Tab. 2: Kumulatívna frekvencia lem slov a lem všetkých jednotiek nachádzajúcich sa v prvej stovke najfrekventovanejších slov/jednotiek podľa Frekvenčného slovníka slovenčiny (ďalej FSS) a na báze posledných dvoch verzií základných korpusov SNK

Poradie	**FSS** 1969, 1 mil.	**prim-4.0** 2009, 550 mil.	**prim-5.0** 2011, 700 mil.	
celkovo prvých	lemy slov	lemy slov	lemy slov	lemy all (+ interpunkcia, číslice...)
1. – 10.	26,67 %	17,98 %	14,15 %	25,53 %
11. – 20.	8,47 %	5,74 %	4,57 %	6,22 %
1. – 20.	35,14 %	23,72 %	18,72 %	31,74 %
21. – 30.	5,52 %	3,51 %	2,71 %	3,91 %
1 – 30.	40,66 %	27,23 %	21,42 %	35,65 %
1. – 100.	56,13 %	37,81 %	29,76 %	46,16 %

[3] V tejto skupine sa nachádzajú prevažne chybné tvary slov, značky, symboly, slová z iných abecedných sústav a iba približne 14 %, teda necelých 330 tisíc jednotiek, tvoria naozaj reálne slovenské slová.

V malom súbore textov (1 milión slov spracovaných vo FSS) už prvých 100 slov predstavuje vyše 56 % celého rozsahu materiálu. Vo veľkých korpusoch je to necelých 38 %, resp. necelých 30 % – medzi korpusmi 4.0 a 5.0 je rozdiel 150 miliónov jednotiek, čo v súčte prvých 100 najfrekventovanejších slov predstavuje rozdiel vyše 8 %. Keď sme však započítali aj neslovné tokeny (interpunkcia, číslice, značky), podiel prvých 100 jednotiek na celom rozsahu textov aktuálnej verzie SNK prim-5.0 sa priblížil k 50 %.

Jedným z najčastejšie sledovaných výsledkov frekvenčnej analýzy textov konkrétneho jazyka sú práve najfrekventovanejšie slová. Porovnanie prvých desiatich slov v rôzne veľkých zdrojoch písaných textov pochádzajúcich z rôznych období súčasnej slovenčiny ukazuje, že v tejto skupine má stabilné postavenie až 7 lem: *byť, a, v, sa, na, z, že*. Frekvenčný slovník založený na starších textoch prevažne umeleckej literatúry obsahuje na prvých miestach ešte zámená *on, ten, ako*, novšie korpusy zase zámeno *ktorý* (často vo funkcii spojky), zámeno *to* a predložku *s*. V korpuse hovorenej slovenčiny (posledný stĺpec) sa v prvej desiatke vyskytuje menej predložiek (nie sú tu *s, z*), zato viac zámen (*tak, ja*; zámeno *to* je dokonca na 2. mieste) a preniklo sem aj ďalšie sloveso – *mať*. Zloženie tejto prvej desiatky a poradie slov v nej predstavuje základný obraz o jazyku, no poukazuje aj na špecifické vlastnosti jednotlivých druhov korpusov/materiálových báz a spôsobov komunikácie.

Tab. 3: Prvých desať najčastejších lem slov vo FSS a v rôznych základných korpusoch

	FSS	**prim1**	**prim-4.0**	**prim-5.0**	**s-hovor-3.0**
	1969, 1 mil.	2004, 200 mil.	2009, 550 mil.	2011, 720 mil.	2011; 1 643 tis.
1.	a	byť	byť	byť	byť
2.	byť	v	a	a	to
3.	sa	a	v	v	a
4.	v	sa	sa	sa	že
5.	na	na	na	na	sa
6.	on	ten	to	to	v
7.	ten	ktorý	ktorý	ktorý	tak
8.	že	s	s	s	ja
9.	z	z	že	z	na
10.	ako	že	z	že	mať

Hoci aktuálny korpus prim-5.0 a jeho podkorpusy majú už značný rozsah, ktorým sa prirodzene eliminuje výskyt náhodných javov, a hoci je hlavný korpus zložený prevažne z publicistických textov (73 %), rozloženie prvých desiatich najfrekventovanejších lem v podkorpusoch utvorených podľa troch hlavných štýlov nie je totožné s hlavným korpusom ani v jednom prípade. Najviac sa odlišuje podkorpus umeleckých textov, v ktorom sa v prvej desiatke nenachádzajú (rovnako ako v kor-

puse hovorenej reči) predložky *s, z* a zámeno *ktorý*, miesto nich sú tu zámeno *ja* (rovnako ako v korpuse hovorenej reči), zámeno *on* a polyfunkčné slovo *si*. V podkorpuse slovenských umeleckých textov sú v porovnaní s podkorpusom všetkých umeleckých textov v prvej desiatke zastúpené tie isté lemy, iba so zmeneným poradím slov *to – on* a *ja – si*. V tejto tabuľke uvádzame v poslednom stĺpci na porovnanie výsledky frekvenčnej analýzy menšieho podkorpusu zloženého prevažne z umeleckých textov r. 1955 – 1989. Zastúpenie prvých ôsmich lem slov v tomto podkorpuse je rovnaké ako v ostatných podkorpusoch umeleckých textov, odlišné je len poradie dvoch slov: *byť – a* a *na – v*; *on – to* sú v takom poradí ako v slovenských umeleckých textoch. Na 9. a 10. mieste sa nachádzajú predložky *s, z*, ktoré sa v podkorpusoch umeleckých textov v prvej desiatke neumiestnili.

Tab. 4: Desať najčastejších lem slov v korpusoch z rôznych štýlov a období

	prim-5.0-prf	**prim-5.0-inf**	**prim-5.0-img**	**prim-5.0-skimg**	**r55az89-2.0**
	2011; 82 mil.	2011; 515 mil.	2011; 100 mil.	2011; 32 mil.	2011; 44 mil.
1.	a	v	byť	byť	a
2.	byť	byť	a	a	byť
3.	v	a	sa	sa	sa
4.	sa	na	na	na	v
5.	na	sa	v	v	na
6.	ktorý	ktorý	to	on	on
7.	s	to	on	to	to
8.	z	s	že	že	že
9.	to	z	si	ja	s
10.	že	že	ja	si	z

Väčšina slov s najvyššou frekvenciou sa podieľa aj na zložení najfrekventovanejších dvojíc a trojíc slov.[4] Nachádzajú sa v nich jednak vo vzájomných kombináciách (*sa v, sa na, je to, že sa, a v; že je to*), jednak v kombinácii s ďalším vysokofrekventovaným slovom (*nie, by*). V dvojiciach a najmä v trojiciach slov sa v prvej desiatke nachádza už aj viacero podstatných mien, ktoré v spojení s predložkou predstavujú základnú časovú (*v roku, v tom čase*) a priestorovú (*na Slovensku*) orientáciu a v spojení s dvoma predložkami poukazujú na vysoké využitie sekundárnych predložiek (*v súvislosti s, v porovnaní s, na rozdiel od*). Prítomnosť substantíva *strana* v spojení *na druhej strane*, ktoré je na 3. mieste medzi trigramami v korpuse prim-5.0, nám objasňuje, že vysoké umiestnenie tohto polysémického slova v prvej stovke najfrekventovanejších slov (porov. tab. č. 6) nebude súvisieť iba s významom „politická organizácia“ či „priestor, plocha“.

[4] Pri analýze bigramov a trigramov sa vychádza z konkrétnych tvarov slov, z ktorých sa tvoria náhodné, voľné, ale aj zmysluplné zoskupenia rôznej pevnosti.

Tab. č. 5: Prvých desať najfrekventovanejších dvojíc a trojíc slov

	prim-4.0	**prim-5.0**
	2009; 550 mil.	2011; 720 mil.
1.	nie je	nie je
2.	sa v	je to
3.	sa na	sa v
4.	je to	v roku
5.	že sa	sa na
6.	som sa	a v
7.	by sa	by sa
8.	a v	že sa
9.	v roku	o tom
10.	na to	na Slovensku

prim-4.0	**prim-5.0**
2009; 550 mil.	2011; 720 mil.
že je to	v tomto roku
v súvislosti s	v súvislosti s
v tomto roku	na druhej strane
v porovnaní s	v porovnaní s
to nie je	nie je to
Na druhej strane	na rozdiel od
na druhej strane	že je to
by som sa	to nie je
Nie je to	v tom čase
že by sa	v tomto prípade

Okrem spojok, predložiek, zámen a pomocného slovesa *byť* sa na popredných miestach vo frekvencii – v prvej stovke – nachádzajú aj slová z ďalších slovných druhov. Z podstatných mien sú to najmä všeobecné pomenovania (*človek, spoločnosť, svet, život*) a slová pomenúvajúce časové jednotky (*rok, čas, deň*), ale aj vlastné meno *Slovensko* a spomínané substantívum *strana*. Zo slovies sú najfrekventovanejšie modálne slovesá a základné slovesá pohybu a hovorenia. Z adjektív sa vo všetkých uvedených korpusoch v prvej stovke nachádzajú adjektíva *veľký* a *dobrý*, okrem nich sú ešte časté prívlastky *nový, ďalší* a *slovenský*. Zastúpenie prídavných mien medzi prvými sto najfrekventovanejšími slovami je oproti podstatným menám a slovesám nižšie, v hovorenom korpuse možno až prekvapujúco nízke,[5] no zodpovedá v ňom celkovo nižšiemu umiestneniu prídavných mien medzi slovnými druhmi (7. miesto) v hovorenej komunikácii (porov. ďalej v tab. č. 7). Väčšina vysokofrekventovaných plnovýznamových slov, okrem *Slovensko, slovenský, spoločnosť* a niektorých negatívnych foriem slovies, je jedno- a dvojslabičná, čo korešponduje so všeobecnou tendenciou k ekonomizácii jazyka a jazykového prejavu.

Napriek tomu, že na prvých miestach vo frekvencii slov a lem sú najčastejšie predložky a spojky, rozloženie frekvencie klasických slovných druhov v slovenských písaných textoch ukazuje, že na 1. mieste sú substantíva a na 2. slovesá. V hovorenej reči sú vzhľadom na dejovosť, osobnosť a kontextovosť tohto druhu komunikácie na 1. mieste slovesá, na 2. zámená a podstatné mená sú až na 3. mieste. Častejšie sú tu aj častice (na 6. mieste). V súčasných písaných textoch sú na 3. mieste predložky, v staršom materiáli, spracovanom vo Frekvenčnom slovníku, sú na 3. mieste zámená a predložky až na 5. mieste. Poukazuje to na značnú dynamiku predložkového systému a funkcií jednotlivých predložiek, čo súvisí s tendenciou k inte-

[5] Väčší podiel tu majú v porovnaní s korpusmi písaných textov častice a osobné zámená.

lektualizácii a abstraktnosti jazykového vyjadrovania. Do nasledujúcej tabuľky sme zahrnuli aj ručne morfologicky anotovaný korpus, v ktorom by malo byť určenie slovných druhov na hranici bezchybnosti. Aj keď je tento korpus v porovnaní s ostatnými súčasnými korpusmi veľmi malý (1,2 mil. jednotiek), vidíme oproti veľkému korpusu prim-5.0 (resp. aj prim-4.0, v ktorom je poradie slovných druhov také isté ako v prim-5.0) iba minimálne rozdiely v umiestnení prídavných mien a zámen (v 4.0 a 5.0 sú prídavné mená na 4. mieste, zámená na 5., v ručne anotovanom korpuse je to naopak) a v umiestnení častíc a príslovíek (v 5.0 sú častice na 7. mieste, príslovky na 8., v ručne anotovanom korpuse je to naopak).

Tab. 6: Najfrekventovanejšie podstatné mená, slovesá a prídavné mená nachádzajúce sa v prvej stovke lem slov vo vybraných korpusoch (číslica vyjadruje poradie lemy podľa absolútnej frekvencie)

FSS		**prim-4.0-all**		**prim-5.0-all**		**s-hovor-3.0**	
1969; 1 mil.		2009; 526 mil.		2011; 720 mil.		2011; 1 643 tis.	
človek	48.	*rok*	26.	*rok*	26.	*človek*	34.
rok	62.	*človek*	59.	*človek*	56.	*rok*	54.
čas	69.	*Slovensko*	86.	*Slovensko*	79.	*vec*	91.
deň	71.	*čas*	92.	*deň*	82.	*dieťa*	92.
svet	84.	*strana*	96.	*čas*	91.	*mať*	12.
život	89.	*mať*	23.	*strana*	96.	*vedieť*	37.
vec	90.	*môcť*	40.	*spoločnosť*	100.	*ísť*	40.
ruka	91.	*povedať*	66.	*mať*	21.	*povedať*	42.
voda	96.	*musieť*	78.	*môcť*	40.	*hovoriť*	47.
mať	13.	*chcieť*	80.	*povedať*	66.	*môcť*	49.
môcť	31.	*nebyť*	82.	*musieť*	76.	*dať*	60.
vedieť	42.	*ísť*	83.	*chcieť*	77.	*nevedieť*	64.
ísť	46.	*veľký*	56.	*ísť*	80.	*prísť*	66.
musieť	50.	*nový*	63.	*nebyť*	83.	*robiť*	67.
dať	51.	*slovenský*	71.	*veľký*	57.	*musieť*	69.
chcieť	52.	*dobrý*	90.	*nový*	59.	*chcieť*	70.
povedať	54.	*ďalší*	91.	*slovenský*	72.	*nebyť*	71.
vidieť	60.			*dobrý*	85.	*nemať*	82.
celý	49.			*ďalší*	88.	*myslieť*	84.
veľký	53.			*celý*	97.	*dostať*	93.
nový	76.					*vidieť*	97.
starý	97.					*veľký*	86.
dobrý	100.					*dobrý*	98.

Tab. č. 7: Zastúpenie slovných druhov vo FSS a rôznych korpusoch SNK

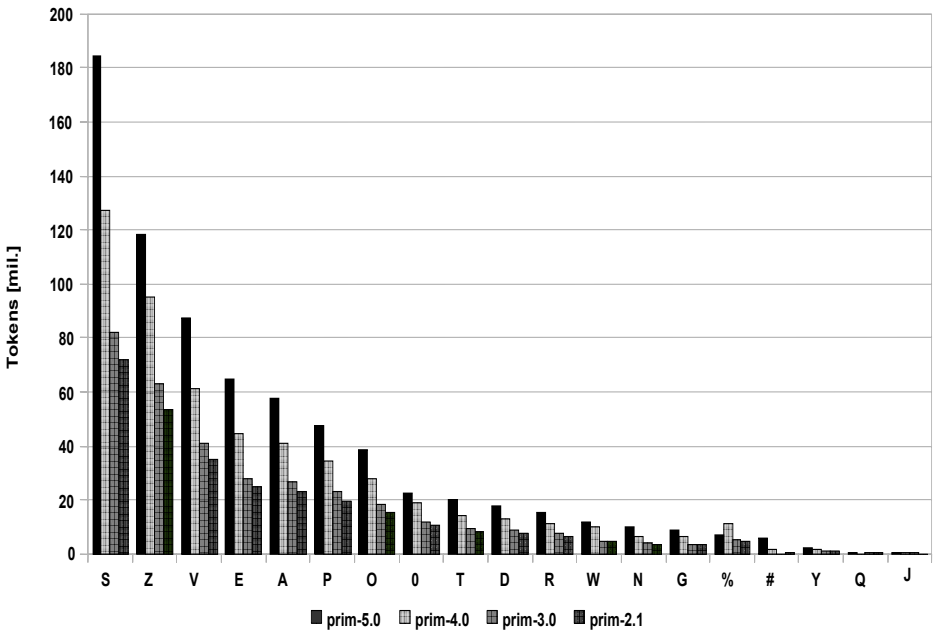
	FSS	**r-mak-3.0**	**prim-4.0-all**	**prim-5.0-all**	**s-hovor-3.0**
	1969; 1 mil.	2008; 1,2 mil.	2009; 526 mil.	2011; 720 mil.	2011; 1 643 tis.
1.	substantíva	substantíva	substantíva	substantíva	slovesá
2.	slovesá	slovesá	slovesá	slovesá	zámená
3.	zámená	predložky	predložky	predložky	substantíva
4.	adjektíva	zámená	adjektíva	adjektíva	spojky
5.	predložky	adjektíva	zámená	zámená	častice
6.	spojky	spojky	spojky	spojky	predložky
7.	príslovky	príslovky	častice	častice	adjektíva
8.	častice	častice	príslovky	príslovky	príslovky
9.	číslovky	číslovky	číslovky	číslovky	číslovky
10.	citoslovcia	citoslovcia	citoslovcia	citoslovcia	citoslovcia

Okrem klasických slovných druhov sa v korpusoch osobitne vyčleňujú aj ďalšie textové jednotky. V Slovenskom národnom korpuse je to podľa pravidiel morfologickej anotácie (http://korpus.sk/morpho.html) týchto 9 slovných tried: príčastia (G), zvratné *sa/si* (R), *by* (Y), interpunkcia (Z), skratky a značky (W), neurčiteľné slová alebo časti slov (Q), neslovné elementy (#), citáty (%), číslice (0). Ich podiel v jednotlivých verziách písaných korpusov ukazuje graf č. 2.

Podstatných mien, ktoré sú najsilnejšou pomenúvacou jednotkou a ku ktorým patrí značné množstvo vlastných mien, je dvakrát viac ako slovies. Z textových jednotiek, ktoré nepatria medzi štandardné slovné druhy, sú najfrekventovanejšie interpunkčné znamienka, a to dokonca hneď na 2. mieste (ich inventár je maličký, ide o 13 prvkov, ale frekvencia väčšiny z nich je veľmi vysoká), a číslice na 8. mieste. Stabilné je umiestnenie citosloviec na poslednom mieste – v písaných textoch je ich zastúpenie na hranici štatistickej odchýlky, predstavujú iba 0,04 % všetkých jednotiek, v korpuse hovorenej reči ich je registrovaných o niečo viac (0,24 %). Ukazuje sa, že citoslovcové výrazy sa v písaných textoch používajú skôr výnimočne a aj to zväčša v prekladovej literatúre, ale ani v hovorenej reči ich nie je jednoduché zaznamenať, keďže nahrávky v Slovenskom hovorenom korpuse v podstate neregistrujú nejaké emocionálne vypäté situácie. Výskyt onomatopojí sa z ich priameho použitia výrazne posúva do použitia vo funkcii vetných členov, ktorými bývajú podstatné a prídavné mená alebo slovesá. Na správnu anotáciu citosloviec v jednotlivých ver-

ziách korpusu písaných textov navyše výrazne vplýva kvalita morfologického analyzátora. Hoci sa veľkosť korpusu od verzie 3.0 po verziu 5.0 zdvojnásobila, počet slov označených ako citoslovce sa zmenšil o niekoľko tisíc jednotiek – ubudli spomedzi nich najmä dovtedy nesprávne zanotované cudzie slová a vlastné mená, ktoré sa medzitým doplnili do morfologického analyzátora, čím sa skvalitnila automatizovaná morfologická anotácia novej verzie korpusu.

Graf č. 2: Zastúpenie slovných tried v rôznych verziách korpusov písaných textov slovenského jazyka



Porovnaním údajov o frekvencii slov a tvarov v súčasnej slovenčine podľa slovníka Frekvencia slov v slovenčine (Mistrík, 1969) a na základe analýz viacerých verzií a podkorpusov Slovenského národného korpusu dostávame obraz o stabilite a dynamike jazykového systému slovenčiny v rôznych sférach jej použitia za posledných 55 rokov. Štatistické informácie zhrnuté vo frekvenčných slovníkoch a zoznamoch napomáhajú lepšiemu, objektívnejšiemu poznaniu fungovania jazykových prostriedkov v komunikácii a vymedzeniu vlastností daného jazyka v porovnaní s typologicky odlišnými i blízkymi jazykmi. Príspevok je malou ukážkou možností štatistickej analýzy a prípravou na vydanie nového frekvenčného slovníka slovenčiny na báze materiálu Slovenského národného korpusu.

Literatúra

ČERMÁK, František: Jazyk a jazykověda. Přehled a slovníky. Praha: Pražská imaginace 1997. 460 s.
KRÁLÍK, Jan: Zamyšlení nad velkými výběry. In: Korpusová lingvistika: Stav a modelové přístupy. Ed.: F. Čermák – R. Blatná. Praha: Nakladatelství Lidové noviny / Ústav Českého národního korpusu 2006, s. 205 – 209.
MISTRÍK, Jozef: Frekvencia slov v slovenčine. Bratislava: Vydavateľstvo SAV 1969. 726 s.
MISTRÍK, Jozef: Retrográdny slovník slovenčiny. Bratislava: Univerzita Komenského 1976. 736 s.
MISTRÍK, Jozef: Frekvencia tvarov a konštrukcií v slovenčine. Bratislava: Veda 1985. 320 s.
ŠIMKOVÁ, Mária: Jazykové prostriedky vo vybraných dielach Martina Rázusa (analýza na báze textov Slovenského národného korpusu). In: Martin Rázus – politik, spisovateľ a cirkevný činiteľ. Ed. M. Pekník. Bratislava: Veda 2008, s. 238 – 249.
ŠIMKOVÁ, Mária: Slovenčina v číslach. In: Slovo o slove. 16. Zborník Katedry komunikačnej a literárnej výchovy Pedagogickej fakulty Prešovskej univerzity. Ed. Ľ. Liptáková – M. Andričíková – M. Klimovič. Prešov: Katedra komunikačnej a literárnej výchovy, Pedagogická fakulta Prešovskej univerzity v Prešove 2010, s. 182 – 189. Dostupné aj na: http://www.unipo.sk/public/media/11314/Slovo%20o%20slove%2016.pdf.
ŠIMKOVÁ, Mária – ĽOS, Miroslav: Frequency of Words and Forms in Contemporary Slovak (Based on the Slovak National Corpus). In: Slovko 2009. Počítačové spracovanie prirodzeného jazyka, korpusová lingvistika a gramatický výskum. Ed. J. Levická – R. Garabík. Brno: Tribun 2009, s. 340 – 348. Dostupné aj na: http://korpus.juls.savba.sk/~slovko/2009/Proceedings_Slovko_2009.pdf
ŠULC, Michal: Korpusová lingvistika. První vstup. Praha: Karolinum 1999. 94 s.