SLOVENČINA V ČÍSLACH

Mária ŠIMKOVÁ

O jazyku a jeho vlastnostiach sa dozvedáme z písmen/hlások, slov, viet, z celej komunikácie, jej kontextu vrátane účastníkov komunikácie, a to zvyčajne prostredníctvom deskripcie – jazykového opisu konkrétnych jednotiek. Rozmach interdisciplinarity, ktorý priniesli nové technológie a metódy objavujúce sa od polovice minulého storočia, zasiahol aj jazykovedu, a tak poznávanie jazyka napomáhajú disciplíny ako psycholingvistika, neurolingvistika, sociolingvistika, etnolingvistika, matematická lingvistika, kognitívna lingvistika, filozofia jazyka. Niektoré z nich sa v podobe čiastkových disciplín začínali rozvíjať už koncom 19. storočia. V tom čase sa stabilizovala aj kvantitatívna lingvistika, súčasť matematickej lingvistiky, v ktorej sa na analýzu zákonitostí fungujúcich v jazyku využívajú štatistické metódy, teória pravdepodobnosti, teória množín, náhodné procesy a pod. V súčasnosti sa tieto metódy aplikujú na výskum prirodzeného jazyka pomocou počítačových nástrojov pracujúcich na rozsiahlych materiálových databázach – jazykových korpusoch.

Slovenský jazyk môžeme takto analyzovať v rámci databázy ***Slovenského národného korpusu*** (http://korpus.juls.savba.sk/), kde je k dispozícii jednak základný všeobecný korpus písaných textov ***prim*** (od r. 2009 vo verzii 4.0), jednak viacero špecializovaných podkorpusov, napr. tvorených textami výlučne jedného štýlu (***inf*** – publicistika, ***prf*** – odborné texty, ***img*** – umelecké texty), ale aj osobitný korpus prepisov zvukových záznamov štandardnej spontánnej a polospontánnej komunikácie ***hovor***. Vzhľadom na to, že jednotlivé databázy už dosahujú dostatočné rozmery (uvádzame ich ďalej v príslušných tabuľkách) a ich frekvenčné analýzy poskytujú relevantné výsledky, pripravujeme v Jazykovednom ústave Ľ. Štúra SAV v Bratislave frekvenčný slovník súčasnej slovenčiny, v ktorom budú zhrnuté kvantitatívne charakteristiky slovenského jazyka posledných desaťročí. Podobný slovník už pre slovenčinu existuje: na materiáli 1 milión slov ho ručne pripravil v r. 1969 J. Mistrík (Frekvencia slov v slovenčine; ďalej FSS) a poskytol tak možnosť na porovnávanie jeho a súčasných zistení a istým spôsobom aj dynamiky slovenčiny v rozpätí pol storočia (prevažná väčšina ním spracúvaných textov je z konca 50. a začiatku 60. rokov 20. storočia).

Za materiálové zdroje pre tento príspevok sme vybrali verejne prístupné korpusy a podkorpusy, pričom pracujeme s východiskovými frekvenčnými zoznamami lem – ide o absolútnu frekvenciu, teda celkový počet výskytov všetkých reálnych tvarov slov zhrnutých v základnom tvare, napr. *ktorý* zahŕňa tvary *ktorého, ktorému, ktorom, ktorým, ktorí, ktorých, ktorými*. Termínu lema zodpovedá

v lexikografickom opise lexéma – heslové slovo, ktoré sa v slovníkoch takisto uvádza v základnom tvare.

Tab. č. 1: Prvých 10 najfrekventovanejších lem v písaných korpusoch a v hovorenom korpuse

	FSS	prim-4.0-all	prim-4.0-inf	prim-4.0-prf	prim-4.0-img	s-hovor-2.0
	1969; 1 mil.	2009; 526 mil.	2009; 330 mil.	2009; 85 mil.	2009; 89 mil.	2010; 680 tis.
1.	a	byť	v	a	byť	byť
2.	byť	a	byť	byť	a	to
3.	sa	v	a	v	sa	a
4.	v	sa	na	sa	na	že
5.	na	na	sa	na	v	sa
6.	on	to	ktorý	s	to	tak
7.	ten	ktorý	to	ktorý	on	v
8.	že	s	s	z	že	ja
9.	z	že	z	to	si	na
10.	ako	z	že	že	ja	no

Z prehľadu desiatich najčastejších slov v slovenčine vidíme, že prvých päť má v našom jazyku stabilné postavenie bez ohľadu na čas, veľkosť materiálu či jeho štýlovú príslušnosť. Sú to jednohláskové alebo najviac jednoslabičné predložky (*v, na*), spojka (*a*), pomocné sloveso (*byť*) a viacfunkčné slovo *sa*. V hovorenom korpuse sa však do prvej pätice na úkor predložiek dostali zámeno *to* a spojka i častica *že*. V druhej polovici sa takisto nachádzajú spojky, predložky a zámená, ale v staršom materiáli obsahujúcom vyše 50 % textov z vtedy prestížnej umeleckej literatúry (FSS), v podkorpuse súčasnej umeleckej literatúry a v hovorenom korpuse už registrujeme viaceré rozdiely oproti ostatným trom korpusom. V umeleckej literatúre a v hovorenej reči sa miesto predložiek *s, z* a zámena *ktorý* uprednostňujú osobné zámená *ja* (img, hovor), *on* (FSS, img), ukazovacie zámeno, spojka a častica *tak* (hovor), spojka a častica, resp. pomocné, výplnkové slovo *no* (hovor). Prvých desať slov spravidla pokrýva až okolo 18 % textu (FSS uvádza 18,6 %, v prim-4.0-all to je 17,98 %), t. j. takmer každé piate slovo v texte je jedným z uvedených desiatich najfrekventovanejších slov.

Ak si však porovnáme desať lem, ktoré sú na 991. – 1000. mieste, zistíme, že v tejto desiatke sa nachádzajú v každom type korpusu iné lemy, iba jedna (*odborný*) sa nachádza na týchto miestach v dvoch korpusoch (all a inf). Poradie slov – lem v strednom pásme frekvencie už poskytuje viaceré informácie o vlastnostiach konkrétneho textu alebo súboru textov.

Tab. č. 2: Zoznam lem vo frekvenčnom pásme 991 – 1000 v písaných korpusoch

	FSS	prim-4.0-all	prim-4.0-inf	prim-4.0-prf	prim-4.0-img
	1969; 1 mil.	2009; 526 mil.	2009; 330 mil.	2009; 85 mil.	2009; 89 mil.
991.	použiť	pozornosť	štvrtý	rýchly	pieseň
992.	požiadavka	papier	stanica	južný	ján
993.	predstaviteľ	použiť	podpísať	chyba	zbaviť
994.	slávny	čistý	odborný	nech	úzky
995.	slobodný	odborný	hotel	kedy	nízky
996.	umenie	investor	myšlienka	verzia	tisíc
997.	usmiať sa	termín	vhodný	výrobný	vedomie
998.	vysvetliť (si)	využívať	porovnanie	maďarský	rásť
999.	deväť	tlak	britský	29	ťahať
1000.	drevený	bežný	výrazne	new	osobný

Značné rozdiely sú tu aj v zastúpení slovných druhov: v podkorpuse odborných textov nie je v tejto desiatke ani jedno sloveso, kým FSS a img ich majú po tri, all dve, inf jedno; v posledných dvoch menovaných korpusoch je zase vyšší podiel podstatných mien (all päť, inf štyri), v img je medzi podstatnými menami v tejto desiatke aj vlastné meno *Ján*. V prf sú miesto slovies zámeno *kedy* a spojka i častica *nech*. Korpus odborných textov sa od ostatných odlišuje aj prítomnosťou číslice (*29*) a cudzieho slova v tejto desiatke (*new* – zrejme ide o jeho výskyt v cudzojazyčných bibliografických záznamoch).

Hovorený korpus je zatiaľ príliš malý (na rozdiel od desiatok a stoviek miliónov jednotiek v súčasných písaných korpusoch s-hovor-2.0 disponuje iba vyše polmiliónom jednotiek), preto sme z neho na porovnanie vybrali desať lem z rozpätia 91. – 100. miesta: *proste, dobrý, vidieť, dva, možno, vtedy, tiež, dostať, aký, iný*. Ich zloženie z hľadiska slovných druhov sa značne odlišuje od vybranej desiatky v písaných korpusoch: žiadne podstatné meno, dve slovesá, jedno prídavné meno, tri zámená, jedna číslovka, tri častice.

Špecifické vlastnosti korpusov písaných textov oproti korpusu tvorenému z prepisov hovorených prejavov sa potvrdzujú aj pri pohľade na prvých sto slov. V celom tomto pásme sa tak ako v prvej desiatke nachádzajú prevažne predložky a spojky, ale výraznejšie sú zastúpené už aj zámená, častice, ba aj základné slovné druhy, z ktorých sme sa osobitne zamerali na podstatné a prídavné mená a slovesá.

Tab. č. 3: Zastúpenie podstatných mien, slovies a prídavných mien v prvých sto najfrekventovanejších lemách (stĺpce obsahujú slovo, poradie podľa absolútnej frekvencie okrem FSS, kde je poradie podľa relatívnej frekvencie, a hodnotu absolútnej frekvencie)

FSS			prim-4.0-all			prim-4.0-skimg			s-hovor-2.0		
1 mil.; poradie je uvedené podľa relatívnej frekvencie			526 mil.			26 mil.			680 tis.		
človek	48.	2142	*rok*	26.	1655462	*človek*	47.	61153	*človek*	35.	2553
rok	62.	1724	*človek*	59.	740659	*ruka*	66.	38425	*rok*	59.	1638
čas	69.	1422	*Slovensko*	86.	517865	*rok*	83.	33268	*vec*	88.	926
deň	71.	1416	*čas*	92.	481088	*deň*	85.	32366	*mať*	13.	6160
svet	84.	1276	*strana*	96.	453741	*oko*	86.	32244	*vedieť*	38.	2360
život	89.	1138	*mať*	23.	2414658	*život*	87.	31701	*ísť*	40.	2279
vec	90.	1173	*môcť*	40.	1089966	*žena*	96.	29065	*povedať*	45.	2131
ruka	91.	1357	*povedať*	66.	657189	*mať*	25.	134244	hovoriť	48.	1893
voda	96.	1201	*musieť*	78.	584054	*povedať*	44.	63427	*nevedieť*	53.	1765
mať	13.	7119	*chcieť*	80.	545389	*môcť*	51.	48688	*môcť*	57.	1687
môcť	31.	3285	*nebyť*	82.	526940	*chcieť*	53.	47192	*dať*	62.	1555
vedieť	42.	2920	*ísť*	83.	521257	*ísť*	54.	46620	*robiť*	65.	1462
ísť	46.	2392	*veľký*	56.	781375	*vedieť*	55.	46375	*prísť*	67.	1363
musieť	50.	2022	*nový*	63.	713173	*nebyť*	58.	44044	*musieť*	69.	1297
dať	51.	1891	*slovenský*	71.	631342	*musieť*	68.	37777	*nebyť*	71.	1225
chcieť	52.	2283	*dobrý*	90.	485079	*prísť*	74.	34229	*chcieť*	73.	1197
povedať	54.	2348	*ďalší*	91.	481761	*dať*	76.	33892	*myslieť*	79.	1068
vidieť	60.	1773				*vidieť*	84.	33031	*nemať*	82.	993
celý	49.	1740				*nemať*	97.	28685	*vidieť*	93.	868
veľký	53.	1845				*nevedieť*	98.	28272	*dostať*	98.	787
nový	76.	1352				*začať*	100.	28046	*veľký*	89.	924
starý	97.	1012				*veľký*	80.	33647	*dobrý*	92.	875
dobrý	100.	1017				*celý*	92.	30315			

V celej prvej stovke najfrekventovanejších lem ide často o krátke, jednoslabičné slová (*rok, deň, oko, čas, vec; mať, ísť, dať*) z jadra slovnej zásoby. Z podstatných mien sa vo všetkých korpusoch na prvých miestach umiestnili lemy *človek* a *rok*, zo slovies sú to modálne slovesá (+ pomocné sloveso *byť*, ktoré bolo v prvej desiatke), z prídavných mien je to lema *veľký*. Príznačná je vysoká frekvencia podstatného mena *Slovensko* a prídavného mena *slovenský* vo veľkom všeobecnom korpuse, v ktorom prevládajú publicistické texty (65 % všetkých textov).

Porovnanie zastúpenia vybraných troch hlavných slovných druhov v prvej stovke v jednotlivých korpusoch ukazuje ich najvyšší podiel v uvedenom pásme vo FSS, skimg a hovor, v posledných dvoch korpusoch sú aj najvyššie výskyty slovies na tejto ploche:

FSS: 9 podstatných mien, 9 slovies, 5 prídavných mien; spolu 23
prim-all: 5 podstatných mien, 7 slovies, 5 prídavných mien; spolu 17
prim-skimg (pôvodné slovenské texty): 7 podstatných mien, 14 slovies, 2 prídavné mená; spolu 23
hovor: 3 podstatné mená, 17 slovies, 2 prídavné mená; spolu 22.

Pre hovorenú komunikáciu i pre umeleckú literatúru je všeobecne typické vyššie využitie slovies, v bežnej reči aj zámen, ktorých je v hovorenom korpuse v prvej stovke 21, teda takmer toľko ako slov z troch hlavných slovných druhov spolu (napr. *ja, on, ten, ako, taký, čo, tam, tá, my*).

Celková frekvencia slovných druhov v jednotlivých korpusoch potvrdzuje doterajšie čiastkové zistenia (do prehľadu sme zahrnuli aj ručne morfologicky anotovaný korpus ***r-mak*** vzhľadom na takmer stopercentne správnu ručnú anotáciu oproti cca 96-percentnej správnosti automatizovanej anotácie v korpusoch prim a hovor):

Tab. č. 4: Frekvencia slovných druhov

	FSS	prim-4.0-all	r-mak-3.0	s-hovor-2.0
	1969; 1 mil.	2009; 526 mil.	2008; 1,2 mil.	2010; 680 tis.
1.	substantíva	substantíva	substantíva	slovesá
2.	slovesá	slovesá	slovesá	zámená
3.	zámená	predložky	predložky	substantíva
4.	adjektíva	adjektíva	zámená	spojky
5.	predložky	zámená	adjektíva	častice
6.	spojky	spojky	spojky	predložky
7.	príslovky	častice	príslovky	adjektíva
8.	častice	príslovky	častice	príslovky
9.	číslovky	číslovky	číslovky	číslovky
10.	citoslovcia	citoslovcia	citoslovcia	citoslovcia

V korpusoch založených na písaných textoch sú na prvých miestach z klasických slovných druhov substantíva a slovesá, v súčasných textoch (prim-4.0, r-mak-3.0) sú na treťom mieste predložky (na rozdiel od FSS, kde sú tretie zámená), čo môže svedčiť o intelektualizácii, zabstraktňovaní prehovorov, resp. môže byť ovplyvnené väčším zastúpením publicistických a odborných textov v súčasných korpusoch. Stabilné je postavenie spojok (zhodne 6. miesto), číslovieok a citosloviec (9. a 10. miesto zhodne aj s hovoreným korpusom). Celkovo nízky výskyt citosloviec má pritom v hovorenom korpuse percentuálne podstatne väčší podiel (0,23 %) ako v prim-4.0 (0,058 %). V hovorenom korpuse sa dalo predpokladať, že na prvých miestach budú slovesá a zámená, vyššie umiestnenie v porovnaní s ostatnými korpusmi majú aj spojky a častice, nižšie naopak predložky.

Pri príslovkách sa ukazuje rovnako v písaných korpusoch, ako aj v hovorenom korpuse, že ich reálne využitie je podstatne nižšie ako potencialita v systéme (derivácia od adjektív či substantív).

Frekvenčné údaje o jednotlivých slovách, ich spájateľnosti (kolokabilite), o gramatických kategóriách a pod. prinášajú najmä z veľkých materiálových zdrojov, akými sú korpusy, cenné objektívne informácie o jazyku. Často potvrdia doterajšie pozorovania (napr. špecifické vlastnosti konkrétnych štýlov), no poskytujú aj nové poznatky a dotvárajú tak celkový obraz o jazyku. S číslami a poradiami je však potrebné pracovať v kontexte poznania celého systému jazyka a jeho štruktúry, vyhodnocovať ich vo vzťahoch a súvislostiach.

Vzhľadom na existenciu podobných zdrojov a analýz v iných jazykoch je možné realizovať na tejto rovine aj rôzne porovnávacie výskumy. Keď sa pozrieme na prvých desať najfrekventovanejších slov v češtine (Frekvenčný slovník češtiny, 2004), zistíme, že sa v podstate zhodujú s tými, ktoré sú najčastejšie v slovenčine: *a, v/ve, se, být, na, ten, s/se, že, z/ze, který*. Porovnajme v oboch jazykoch napríklad najčastejšie slová s náhodne vybraným rovnakým zakončením.

Tab. č. 5: Najfrekventovanejšie slová zakončené na *-oba, -ota, -ava* v slovenčine a češtine

	sl. ***-oba***	čes. ***-oba***	sl. ***-ota***	čes. ***-ota***	sl. ***-ava***	čes. ***-ava***
1.	oba	doba	hodnota	hodnota	bratislava	hlava
2.	osoba	oba/voba	sobota	sobota	hlava	výstava
3.	výroba	osoba	teplota	teplota	príprava	doprava
4.	doba	výroba	istota	jistota	predstava	příprava
5.	podoba	podoba	robota	hmota	postava	představa
6.	choroba	zásoba	lehota	bota	výstava	úprava
7.	zásoba	choroba	jednota	jednota	doprava	postava
8.	chudoba	žaloba	hmota	porota	trnava	obava
9.	žaloba	nádoba	porota	nejistota	ústava	oprava
	nádoba	obdoba	neistota	čistota	obava	ústava
	obdoba 15.	chudoba 14.	čistota 13.	robota 25.	úprava 11.	
					oprava 17.	

Vo všetkých troch stĺpcoch sa v obidvoch jazykoch medzi prvými desiatimi najčastejšími slovami s daným zakončením nachádzajú v ôsmich až deviatich prípadoch rovnaké výrazy, čo predstavuje 80 – 90-percentnú zhodu. Väčšie rozdiely sú len v prípade vlastných mien (*Bratislava, Trnava*), ktoré v češtine, prirodzene, nie sú také frekventované (v českom frekvenčnom slovníku sa navyše neuvádzajú medzi apelatívami, ale sú zoradené v osobitnom slovníčku), resp. pri slove typickom pre jeden jazyk (*bota*). Slová, ktoré sa v jednom z jazykov umiestnili

v prvej desiatke, no v druhom nie, uvádzame pri danom jazyku na konci stĺpca aj s uvedením poradia. Zoznam najfrekventovanejších slov s daným zakončením v slovenčine a češtine predstavuje názornú ukážku príbuznosti a blízkosti dvoch jazykov, ktorých nositelia si navzájom dobre rozumejú aj vďaka podobnostiam v slovnej zásobe.

Kvantitatívna lingvistika zaznamenala v posledných desaťročiach značný rozvoj aj v súvislosti s tvorbou rozsiahlych jazykových korpusov (rádovo stovky miliónov až miliardy jednotiek) a s vývojom počítačových nástrojov na spracúvanie a vyhodnocovanie jazykových dát. Slovenský národný korpus a jeho podkorpusy už v súčasnosti takisto predstavujú vhodný zdroj na aplikáciu kvantitatívnych metód, ktoré môžu dobre poslúžiť pri skúmaní reálneho stavu súčasnej písanej i hovorenej podoby slovenského jazyka, pri riešení viacerých aktuálnych otázok i v porovnávacích výskumoch.

Literatúra a zdroje

Frekvenční slovník češtiny. Praha: Nakladatelství Lidové noviny 2004. 595 s. + CD

MISTRÍK, Jozef: Frekvencia slov v slovenčine. Bratislava: Vydavateľstvo SAV 1969. 728 s.

Slovenský hovorený korpus – s-hovor-2.0. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2010. Dostupný z WWW: https://data.juls.savba.sk/oral/.

Slovenský národný korpus – prim-4.0-public. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2009. Dostupný z WWW: http://korpus.juls.savba.sk.

Resumé

SLOVAK IN NUMBERS

Statistical (quantitative) analysis in linguistics has been a reality since the end of the 19th century. Their results have not just been performed on linguistic (especially lexicographic) usage, but also on didactics, psychology, logopedia, neurology and so on, which at present are particularly used for computer language processing. With the expansion of information technology and modern methods, the frequency of attributes, words, forms, constructions and such are being examined: they are discovering more and more amounts of linguistic data (from hundreds of millions to billions of units) with ever increasingly detailed and precise results. Within the framework of projects in the Slovak National Corpus (http://korpus.juls.savba.sk), those who are interested in the Slovak language have several types of corpora available to them: written – spoken, monolingual – multilingual, most

current – older, broad – specific. All are lemmatised and morphologically annotated which provides effective processing of data. The frequency of linguistic resources has been forming the objective picture of the Slovak language and its dynamics since the middle of the 20th century.
In the paper the most frequented words and word classes are compared in selected subcorpora of written and spoken Slovak. Analysis of the material will confirm a stable bearing on the most frequented units, detect differences between older and newer texts and signal common features of publicistic and specialised styles as well as artistic literature and spoken language.
Further information on frequency parametres of modern Slovak will be supplied by an upcoming frequency dictionary.

PhDr. Mária Šimková
Jazykovedný ústav Ľ. Štúra SAV Bratislava, Slovenský národný korpus
e-mail: simkova.maria@korpus.juls.savba.sk