# FREKVENCIA LEXIKÁLNYCH JEDNOTIEK CUDZIEHO PÔVODU V SLOVENČINE

***Radovan Garabík – Agáta Karčová***

*Jazykovedný ústav Ľudovíta Štúra SAV*
*Bratislava*

GARABÍK, Radovan – KARČOVÁ, Agáta: Frequency of Lexical Units of Foreign Origin in Slovak. Slovak Language, 2019, Vol. 84, No 1, pp. 26 – 46.

**Abstract:** The article describes a method to analyse contemporary Slovak vocabulary with regard to the origin of the words. By using statistical data from a representative corpus of modern written language and etymological information we arrive at reasonably confident estimation of the ratio of loanwords in common Slovak vocabulary and the provenance of lexical borrowings. We demonstrate some of the findings in tables and charts, providing information that is interesting to non-linguistically oriented members of Slovak population (who are sometimes vocal in expressing their attitudes to the perceived amount of loanwords in the Slovak language), but can be also inspiring for further research in philology or linguistics.

**Key words:** corpus, etymology, loanwords, Slovak language, statistics, vocabulary

## ÚVOD

Slovná zásoba slovenčiny, rovnako ako každého živého funkčného jazyka, podlieha rôznym typom zmien. Jej neustála aktualizácia je nevyhnutná pre zachovanie jazyka schopného spĺňať všetky svoje funkcie, predovšetkým komunikatívnu a kognitívnu, na ktoré sú v súčasnosti najmä v súvislosti so zrýchleným tempom vedecko-technického pokroku kladené stále vyššie nároky. Dynamika jazyka sa najviac vníma a je najľahšie overiteľná na lexikálnej úrovni.

K prirodzeným spôsobom obohacovania slovnej zásoby patrí aj preberanie slov z fondov iných jazykov. Podľa J. Dolníka (2003, s. 165 – 166) sa medzery v nominačnej sústave odstraňujú preberaním, keď sa nedá utvoriť domáci výraz, alebo keď je domáci výraz menej vhodný, za bežný dôvod môžeme považovať aj preberanie v prospech diferenciácie a variabilnosti jazyka. Spôsoby preberania slov, slovných spojení, frazém, ich adaptácia v prijímajúcom jazyku, formovanie obsahu i rozsahu každého nového pojmu a sémantické posuny lexém sú predmetom skúmania nielen lingvistiky, ale aj viacerých ďalších disciplín, výskum lexikálneho fondu organicky prepája prístup synchrónie a diachrónie.

Existujú rôzne názory na preberanie slov z cudzích jazykov. Ako upozorňuje M. Ološtiak (2009, s. 87 – 95), mnohé z nich, stavajúce sa odmietavo k tomuto spôsobu obohacovania slovnej zásoby, sú argumentačne slabo podložené, dokonca „sa skĺza... do roviny voluntaristickej kritiky. ... Navyše, opierajú sa o jednotlivosti, ktoré sa prezentujú ako všeobecné trendy v rámci spisovného vyjadrovania.“ Ďalší lin-

gvisti, naopak, poukazujú na to, že preberanie je prirodzeným javom, napr.: „...náš jazyk nie je ohrozovaný nijakým iným jazykom a žije svoju tzv. normálnu sociolingvistickú situáciu“ (Ondrejovič, 2010, s. 5), vyzdvihujú ďalšie tvorivé možnosti tohto postupu: „Globalizácia jazyka však prináša nielen anglicizmy, hybridy, jazykové glokalizmy, ale umožňuje aj hravé narábanie s jazykom, tvorivosť pri zakomponovaní cudzích jazykových prvkov do systému domáceho jazyka, obohacovanie slovenčiny“ (Škvareninová, 2015, s. 44). Pomer cudzích slov v slovenčine, resp. množstvo novších výpožičiek, ostáva v diskusných aj odborných príspevkoch väčšinou nekvantifikovaný. V tejto súvislosti je namieste otázka, ako by sme mohli objektívne charakterizovať slovnú zásobu slovenčiny z hľadiska pôvodu slov. Samozrejme, každý odhad množstva prevzatých slov závisí od zvolených kritérií a od spôsobu, ako definujeme slová domáceho a cudzieho pôvodu, pričom podstatnú úlohu zohráva aj stanovenie miery a hĺbky využitia výsledkov etymologického výskumu. V tomto príspevku sa pokúsime o štatistické spracovanie analýzy relevantnej vzorky slovnej zásoby slovenčiny a o názorné zachytenie pomeru prevzatých slov a slov domáceho pôvodu v slovenskom jazyku na základe dostupných prameňov.

## 1. SELEKCIA A PRIMÁRNE SPRACOVANIE TEXTOVÉHO MATERIÁLU

### 1.1. Zdrojové východiská výskumu

Na začiatku našej práce bolo potrebné určiť relevantnú vzorku slovenského jazyka, ktorá by z frekvenčného hľadiska dostatočne spoľahlivo reprezentovala slovnú zásobu súčasnej slovenčiny. Využili sme textové dáta Slovenského národného korpusu, konkrétne korpus *prim-7.0-public-vyv* s rozsahom 340 708 046 pozícií, sprístupnený v r. 2015, v ktorom sú vo vyváženej miere zastúpené publicistické, odborné a beletristické texty súčasnej slovenčiny. Vo všeobecnosti platí, že výskyt lexikálnych jednotiek v texte prirodzeného jazyka sa zhruba riadi Zipfovým zákonom[1], (Zipf, 1936; Piantadosi, 2014; Garabík a kol., 2018; Corral a kol., 2015), ktorý hovorí, že počet výskytov lexikálnej jednotky v korpuse je nepriamo úmerný rangu tejto jednotky v inverznom usporiadaní podľa počtu výskytov:

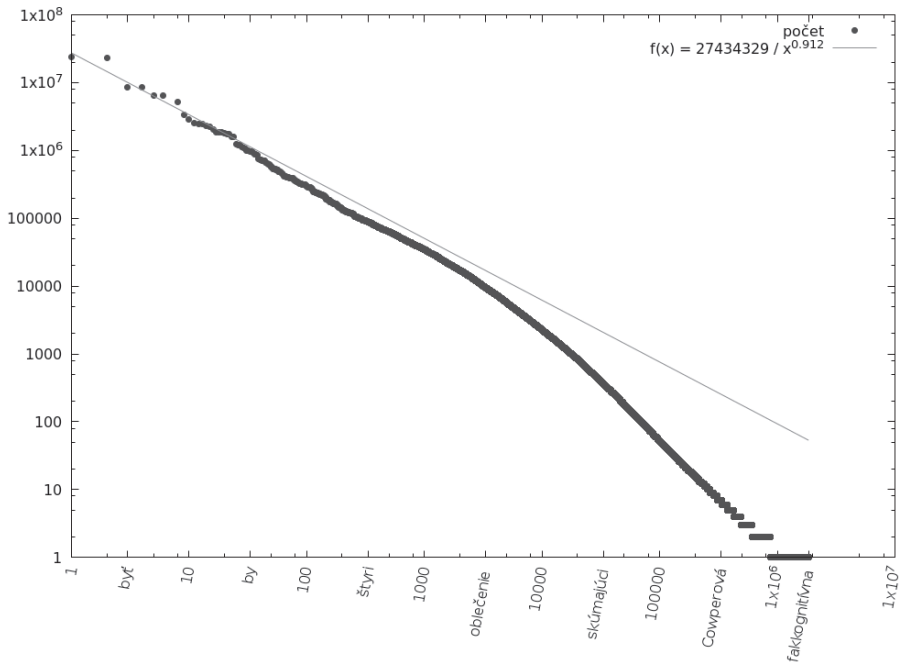
$$f = \frac{K}{r^b} \tag{1}$$

kde $f$ je počet výskytov slova v korpuse (frekvencia), $r$ jeho rang, $b$ parameter distribúcie udávajúci rýchlosť poklesu, $K$ multiplikatívna konštanta normalizujúca frekvenciu na

[1] V skutočnosti je Zipfov zákon (pôvodne na úrovni slovných tvarov, ale platný aj na úrovni lem) len hrubou aproximáciou skutočnej distribúcie počtov slov; existujú rôzne spresnenia a teoreticky zdôvodniteľné odhady. Podrobnejšia analýza by ale bola nad rámec tohto článku, keďže Zipfov zákon uvádzame len na ilustráciu pomerného zastúpenia najčastejších lexikálnych jednotiek v jazyku.

veľkosť daného korpusu. Aplikácia Zipfovho zákona na náš korpus *prim-7.0-public-vyv* je graficky znázornená na obr. 1; funkcia vystihuje výskyty slov v korpuse len v oblasti najfrekventovanejších slov (dá sa povedať, že do rangu 2 000). Mocninový koeficient b=0,912 vhodne využijeme pri odhadoch váhovaného Kendallovho $\tau$ koeficientu.

Naša analýza sa bude zameriavať na slová zo začiatku takéhoto usporiadania, t. j. z počiatočnej časti Zipfovej distribúcie. Predpokladáme, že najfrekventovanejšie slová vnímajú používatelia jazyka ako najviac „normálne“, bežné, sú zároveň jadrom[2] aktívne využívanej slovnej zásoby. Hoci výsledky analýzy z tejto oblasti nemôžeme paušálne vztiahnuť na celý jazyk, predsa odrážajú vplyv cudzích jazykov práve na najdôležitejšiu časť slovnej zásoby. Nami prezentovaný postup možno, prirodzene, použiť aj na analýzy vzoriek v strednej časti Zipfovej distribúcie, prípadne na jej chvoste, a tak ohodnotiť vplyv cudzích jazykov aj na „nie takú dôležitú“, resp. menej frekventovanú časť slovnej zásoby slovenčiny.




Obr. 1: Platnosť Zipfovho zákona v korpuse *prim-7.0-public-vyv.* Na vodorovnej osi je rang lemy v korpuse a niekoľko príkladov slov s daným rangom; na zvislej osi počet výskytov danej lemy. Výskyty sú preložené funkciou f(x) (v logaritmicko-logaritmickej mierke znázornená priamkou).

[2] Jadro slovnej zásoby definujeme predovšetkým na základe frekvenčných parametrov, čo však v zásade neodporuje jeho zaužívanému vymedzeniu aj na základe ďalších charakteristík (porov. napr. Ondrus – Horecký – Furdík, 1980, s. 28), ako to dokladajú aj príklady uvedené v celom príspevku.

Z korpusu *prim-7.0-public-vyv* sme vyselektovali vzorku 2 000 najfrekventovanejších textových jednotiek (lem)[3], ktoré majú celkový počet výskytov 256 974 945 (výskyt týchto slov vo všetkých tvaroch paradigmy), čo pokrýva cca 75,4 % rozsahu celého korpusu[4]. Táto vzorka je teda založená na relatívne vyváženom[5], aktuálnom a dostatočne rozsiahlom zdrojovom materiáli a je použiteľná na ďalšie analýzy.

Pre zaujímavosť uvádzame porovnanie váhovaného Kendallovho $\tau$ (Shieh, 1998) pre prvých 2 000 najfrekventovanejších lem korpusu *prim-7.0-public-vyv* a iných korpusov slovenského jazyka[6]. Pri porovnávaní sa ignorovali rozdiely medzi interpunkciou a veľkosťou písmen. Jednotlivým lemám sme priradili váhu[7] zodpovedajúcu ich rangu podľa funkcie (1), s koeficientom $b$ rovným 0,912 (porovnaj obr. 1), pričom váhy boli multiplikatívne. V tabuľke 1 sú (v tomto poradí): korpus použitý pri tvorbe Frekvenčného slovníka slovenčiny na báze Slovenského národného korpusu, korpus beletrie (vrátane prekladov), korpus originálnej slovenskej beletrie, korpus odborných textov, korpus publicistiky, hovorený korpus, a nakoniec vyvážený korpus *prim-8.0-public-vyv*, ktorý vznikol až po ukončení hlavnej etapy anotovania našich dát.

| *korpus* | $\tau$ |
|---|---|
| prim-7.0-frk | 0,436 |
| prim-7.0-public-img | 0,292 |
| prim-7.0-public-img-sk | 0,263 |
| prim-7.0-public-prf | 0,261 |
| prim-7.0-public-inf | 0,193 |
| s-hovor-6.0 | -0,002 |
| prim-8.0-public-vyv | 0,485 |

Tab. 1: Porovnanie váhovaného Kendallovho $\tau$ pre prvých 2 000 najfrekventovanejších lem v niekoľkých korpusoch.

### 1.2. Primárne triedenie textových jednotiek

Text ako jazykovo-tematická štruktúra, jazykový znak svojho druhu (porov. Encyklopédia jazykovedy, s. 452) je tvorený slovami, pre prehľadnosť a čitateľnosť

[3] Lemu môžeme zjednodušene chápať ako slovo v základnom tvare, ktoré reprezentuje všetky tvary paradigmy. Do množiny lem zaraďujeme aj znaky interpunkcie a číslice.

[4] Do tejto množiny zaraďujeme aj znaky interpunkcie, pri ktorých neurčujeme ich etymologický pôvod (čo by sme, prísne vzaté, mohli, pretože interpunkcia je prevzatá, dokonca ani slovenská abeceda a číslice nie sú domáceho pôvodu) a v ďalších štatistikách s nimi nepočítame.

[5] V zmysle chápania pojmu *vyváženosť* aplikovanom v Slovenskom národnom korpuse.

[6] Pripomíname, že hodnota $\tau$ koeficientu rovná 1 zodpovedá absolútnej korelácii (identite) zoznamov, hodnota 0 situácii, keď sú zoznamy nezávislé.

[7] Uvedomujeme si, že táto váha je založená na Zipfovej distribúcii iba jedného korpusu, a teda jej použitie univerzálne nie je podložené. Dané porovnanie však slúži len ako vedľajšia ilustrácia; dôsledná analýza zhodnosti rangov rôznych korpusov by vysoko prekračovala rámec tohto článku. Tabuľka poukazuje skôr na vhodnosť či nevhodnosť použitia Kendallovho $\tau$ na porovnávanie frekvenčných zoznamov, než na podobnosť korpusov.

textu oddelenými interpunkčnými znakmi, ktoré majú v korpusovom spracovaní platnosť samostatných textových jednotiek (tokenov). Zo získanej vzorky 2 000 lem sme v prvej fáze vyčlenili netextové jednotky: interpunkciu, ďalšie neslovné elementy (napr. matematické znaky) a číslice (rímske a arabské), ako aj znak &, spolu 102 lem (zodpovedajúcich 27,7 % výskytom tokenov v texte).

Ďalej sme zo vzorky vytriedili všetky propriá (77 lem), ktoré v rámci lexikónu každého jazyka predstavujú špecifickú časť slovnej zásoby. Ide o frekventované toponymá, antroponymá a etnonymá, pričom sme do tejto skupiny z praktických dôvodov zaradili aj vzťahové adjektíva utvorené od vlastných mien (napr. *európsky, slovenský, rímsky, spišský*). V základných štatistických spracovaniach sme abstrahovali od týchto slov, v ďalšej fáze výskumu sme však na porovnanie zaradili do zoznamov aj spracované vlastné mená. Problematike proprií z etymologického hľadiska sa budeme dôkladnejšie venovať v nasledujúcom texte.

V primárnej fáze bolo potrebné sústrediť pozornosť aj na všetky textové jednotky, ktoré sú z istého hľadiska špecifické, resp. príznakové, a to skratky a skratkové slová (spolu 28 lem), ako aj 1-písmenové lemy, pri ktorých je vysoká miera homonymie, a preto si vyžadujú pri spracovaní osobitnú pozornosť.

## 2. METODOLÓGIA URČOVANIA PÔVODU APELATÍV

### 2.1. Pramenný materiál pre určovanie pôvodu slov

Po vytvorení relevantnej vzorky slov a následnom odfiltrovaní interpunkčných znamienok, neslovných elementov, špecifických značkových slov a proprií sa začala ďalšia podstatná a časovo náročná fáza výskumu. Sústredili sme sa na koncipovanie vhodnej metodológie na zachytenie pôvodu slov, ktorá mala zohľadniť spôsob spracovania lexiky slovenského jazyka v dostupných a relevantných prameňoch, možnosť štatistického spracovania, ako aj potrebu prehľadnosti a zrozumiteľnej demonštrácie výsledkov výskumu.

Jednou zo základných podmienok čo najspoľahlivejšieho určenia pôvodu slov vo vybranej vzorke bolo vybrať kľúčové lexikografické diela a odborné publikácie, ktoré spracúvajú etymológiu slov v slovenčine v súlade so stavom súčasného poznania v tejto disciplíne a rovnako v želanom rozsahu tak, aby sa v porovnateľnej miere a na jednotnom základe dali vyabstrahovať etymológie všetkých skúmaných slov. Aj keď etymológia ako najstaršia disciplína historickej lingvistiky má na Slovensku tradíciu, o syntetizujúcom lexikografickom diele zameranom osobitne na lexiku slovenského jazyka možno hovoriť až od r. 2015, keď vyšiel *Stručný etymologický slovník slovenčiny* v autorstve Ľ. Králika. Tento slovník, aj keď s prívlastkom *stručný*, obsahuje explikácie všetkých lexém apelatívnej slovnej zásoby, ktoré sa vyskytli v našej vzorke najfrekventovanejšie používaných slov v slovenčine. Keďže vzniku tohto diela predchádzal dlhoročný výskum erudovaného autora, mohli sme jeho die-

lo považovať za relevantný východiskový prameň, popri ktorom sme využili ďalšie doplňujúce diela (vybrané etymologické slovníky češtiny, slovenčiny a ďalších slovanských jazykov a Slovník cudzích slov, 2005).

### 2.2. Spôsob zachytenia etymológie vybraných apelatív

Cieľom našej štúdie je podať čo najpresnejší obraz o zložení slovnej zásoby slovenčiny z hľadiska pôvodu jej jednotiek. Nezahŕňa preto (resp. zahŕňa len v minimálnej miere) konfrontačný výskum vyplývajúci z odlišných interpretácií v mnohých lexikografických dielach i odborných štúdiách, nejednoznačných či viac-menej nerekonštruovateľných či málo pravdepodobných výkladov etymológie slov. Naopak, smeruje k prehľadnému podaniu vyabstrahovaných, zrozumiteľných a v aktuálnom stupni poznania všeobecne prijímaných faktov.

V priebehu koncipovania tohto prístupu k danej problematike bolo potrebné odpovedať na viaceré podstatné otázky: Ako prehľadne a spoľahlivo zachytiť na základe existujúcich relevantných prameňov pôvod jednotlivých slov tak, aby sme podali pravdivý obraz o zložení skúmaného materiálu? Akým spôsobom a nakoľko môžeme zjednodušiť komplexnú informáciu o pôvode slov tak, aby takéto zachytenie bolo relevantné? Akým spôsobom je potrebné zohľadniť diachrónne hľadisko vývoja jednotlivých slov tvoriacich slovnú zásoby slovenčiny? Ako pracovať s istou formou „neistoty", s ktorou sa pri explikáciách v etymologických dielach stretávame veľmi často?[8] Akým spôsobom spracovať informácie tak, aby sa dali ďalej štatisticky spracovať?

Nasledovala ďalšia fáza výskumu, ktorá bola najnáročnejšia z hľadiska overovania nosnosti skoncipovanej metodológie, ako aj z časového aspektu. Postupovali sme tak, že sme sa sústreďovali predovšetkým na východiskový jazyk, z ktorého dané slovo pochádza (pracovne ho nazývame *štartovací jazyk*), a jazyk, z ktorého bolo dané slovo bezprostredne prevzaté do slovenčiny, resp. praslovančiny (tzv. *posledný jazyk*). Zachytávali sme aj priebeh preberania lexémy medzi štartovacím a posledným jazykom, i keď v štatistických spracovaniach sme túto informáciu nevyužili; podobne sme nevyužili informáciu o prípadoch, keď je štartovací jazyk totožný s posledným (tieto údaje môžu poslúžiť v ďalších výskumoch). Na zachytenie údajov o pôvode slov sme vyvinuli jednoduchý popisný formálny minijazyk, ktorý sa skladá z nasledujúcich znakov a slov (ako jednotiek formálnych jazykov):

- alfabetické znaky, použité na:
  - zápis skratiek označujúcich jednotlivé jazyky (pridržiavali sme sa konvencie zavedenej v Stručnom etymologickom slovníku slovenčiny, 2015),

[8] Miera neistoty je vyjadrená relativizáciou vyjadrení pomocou slov *pravdepodobne, zrejme, azda, možno, prevzatie (s možným prostredníctvom z jazyka X)*. Ide o prirodzenú súčasť explikácií v etymologických lexikografických dielach, keďže viaceré formy, prevzatia a podobne sú nerekonštruovateľné, nepriezračné, resp. by ich jednoznačné určenie mohlo narušiť vnímanie zložitosti niektorých skúmaných javov či zastrieť otvorené možnosti na ich ďalšie preskúmanie.

    - zápis lexém vybranej vzorky,
- špeciálne slovo `*dezamb`,
- znak tabulátora: `U+0009 <control-0009>`,
- znak `U+003D EQUALS SIGN =`,
- číslice 0 až 9,
- znak šípky doľava[9]: `U+2190 LEFTWARDS ARROW ←`,
- čiarka: `U+002C COMMA ,`,
- znak `U+002B PLUS SIGN +`,
- znak `U+0023 NUMBER SIGN #`,
- znak `U+0025 PERCENT SIGN %`,
- znak konca riadku: `U+000A <control-000A>` (slúži ako oddeľovač záznamov o jednotlivých slovách).[10]

Tento minijazyk zrozumiteľne znázorňuje prvopočiatočný pôvod lexémy, ďalší (prípadný) priebeh jej preberania do ďalších jazykov až po následné bezprostredné prevzatie do slovenčiny. V schematizovaných zápisoch je štartovací jazyk na posledom mieste (úplne napravo), ostatné možné jazyky, ktoré slovo postupne preberali, sú oddelené šípkami doľava, reťazec ukončuje posledný jazyk, pred ktorým je umiestnený ako oddeľovač tabulátor, na prvom mieste (úplne naľavo) sa vždy nachádza skúmané slovo v základnom tvare a aktuálnej pravopisnej podobe.

Špeciálne slovo `*dezamb` označuje, že lema je homonymná, pričom dané homonymá s rôznymi významami pochádzajú z rôznych jazykov. Uvádza sa na začiatku reťazca jazykov, nasleduje za ním čiarka, po nej jazyky prefixované číselným údajom uvádzajúcim počet výskytov z daného jazyka pre každé homonymum a znak rovná sa `=`.

Ak ide o slovo zložené z niekoľkých zložiek pochádzajúcich z rôznych jazykov, tieto uvádzame konkatenáciou s oddeľovačom, a to znakom plus `+` v poradí, v akom tieto zložky nasledujú v zápise slova (t. j. zľava doprava).

Znak `#` označuje neurčiteľný pôvod, znak `%` citátový výraz.

Príklad zápisu v nami definovanom minijazyku:

```
ďakovať       hnem
charakter     l←g
káva          tur←perz←arab
von           *dezamb,11=slk,9=n
```

[9] Napr. v široko rozšírenom grafickom prostredí X11 možno tento znak štandardne zadať z klávesnice využitím *Compose* mechanizmu.

[10] Súbory sú v znakovej sade Unicode, resp. ISO/IEC 10646, v kódovaní UTF-8. Uvádzame aj názvy znakov podľa ISO/IEC 10646.

V prípade, že išlo o slovo domáceho, resp. praslovanského pôvodu, priamo za slovom nasleduje znak konca riadku. Pri nejasnom pôvode, resp. diskutabilnej časti etymologickej charakteristiky, sme konkrétne nejasnosti overili vo viacerých prameňoch alebo v prípade menšej pravdepodobnosti danej čiastkovej informácie sme túto skutočnosť zámerne vynechali. Pri explikáciách, v ktorých sa uvádza možný pôvod z dvoch, príp. viacerých rôznych jazykov, sme sa priklonili k pravdepodobnejšej verzii.

V záujme prehľadnosti a potrebnej abstrakcie sme nerozlišovali časové hľadisko: informácia, či slovom disponuje slovenčina napr. od 17. storočia alebo až od 19. storočia, príp. bolo prevzaté či písomne doložené v omnoho staršom období, tu nie je zachytená. Rovnako sme z praktických dôvodov nerozlišovali mieru adaptácie slov a ich vnímanie a rozlišovanie používateľmi súčasnej slovenčiny na osi *zdomácnené – cudzie*. V súvislosti s tým sme tiež zjednodušili zložité explikácie niektorých slov, pri vývine a prevzatiach ktorých figurovala latinčina v rôznych podobách, napr. ako starolatinčina, stredoveká latinčina či ľudová latinčina, a analogicky aj ďalšie jazyky v ich historickej podobe. Naopak, zachovali sme rozlišovanie variantov jazykov z hľadiska ich územného rozčlenenia, ak sú vo všetkých relevantných dielach zachytené (napr. sme samostatne určovali pôvod z dolnonemeckého a hornonemeckého jazyka, zaznamenanú informáciu sme nezjednodušovali na: „*slovo nemeckého pôvodu*“).

Po spracovaní celej vzorky sa ustálil zoznam jazykov, z ktorých slovenčina prevzala značnú časť slovnej zásoby (3 z uvedených jazykov – *bavorský*, *románsky* a *perzský* – sa vyskytujú len v strede postupnosti jazykov, ktorými slovo „putovalo“, preto sa vo výsledkoch nenachádzajú). Zoznam použitých skratiek jazykov:

| | |
|---|---|
| a | anglický |
| akkad | akkadský |
| arab | arabský |
| bavor | bavorský |
| čes | český |
| dnem | dolnonemecký |
| egypt | egyptský |
| f | francúzsky |
| g | grécky |
| germ | germánsky |
| hebr | hebrejský |
| hnem | hornonemecký |
| l | latinský |
| maď | maďarský |
| n | nemecký |
| perz | perzský |
| poľ | poľský |

| | |
|---|---|
| r | ruský |
| román | románsky |
| slk | slovenský (zapisujeme iba pri kompozitách, resp. viacslovných skratkách) |
| šp | španielsky |
| t | taliansky |
| turk | turkický |
| tur | turecký |

### 2.3. Prípady homonymie a ich riešenie

Samozrejme, v priebehu práce sa vyskytli mnohé ďalšie prípady nejednoznačností vyplývajúce z prirodzenej zložitosti jazyka ako funkčného dynamického systému. Osobitnú pozornosť sme venovali homonymii, ktorá sa objavuje ako sprievodný jav pri rozširovaní slovnej zásoby (porov. Horecký a kol., s. 338). Išlo hlavne o určovanie pôvodu homonymných lexém a pomeru ich výskytu, ktorý sme určili ručnou dezambiguáciou (zjednoznačňovaním) homoným na náhodne vybranej vzorke konkordancií daných lem v korpuse, typicky s veľkosťou 20 výskytov. Napr. homonymá *trieda*[1] a *trieda*[2] majú tieto odlišné významy vyjadrené parafrázami: *1. skupina s istými spoločnými znakmi*; *2. široká ulica*. Z nich lexéma s 1. významom bola vytvorená podľa českého *třída*, zatiaľ čo homonymná lexéma s 2. významom má pôvod v taliančine. Pri tejto leme sme preto určili na menšej vzorke dvadsiatich náhodne vybratých výskytov v *prim-7.0-vyv*, koľkokrát sa slovo v textoch vyskytuje v každom z významov. V tomto konkrétnom prípade sme zistili, že všetky kontexty svedčia o použití slova trieda v 1. význame, čo sme považovali za dostatočný dôvod na odfiltrovanie 2. významu a zároveň s ním aj nezachytenie druhého možného pôvodu homonyma[11]. Podobne sme postupovali aj pri lexémach *banka*, *čelo* a ďalších homonymách, ktorých etymológia je odlišná.

Vysoká variabilita sa ukázala pri 1-písmenových slovách. Po preskúmaní konkordancií z ďalšieho spracovania sme vylúčili 18 z nich (napr. *b, j, t, u*), vzhľadom na ich vysokú homonymiu a nevyhranenosť v ich jazykovom používaní (nevstupujú tak do štatistiky a percentuálne zastúpenie v tab. 3 až 10 je vzhľadom na korpus bez týchto tokenov). Do základnej vzorky slov sa dostali len neplnovýznamové slová *a*, *v*, *s*, *z*, *k*, *o* na základe presnejšej analýzy – po ručnom preskúmaní vzorky konkordancií (s veľkosťou 500 výskytov) danej jednopísmenovej lemy. Ostatné jednopísmenové tokeny sa vyskytovali prevažne v netextových kontextoch ako súčasti vzorcov, preklepy atď. Problematické boli aj viaceré neplnovýznamové jednoslabičné slová, skratky a skratkové slová, ktoré sa paralelne vyskytujú v slovenských textoch

[11] Pre ilustráciu, v prípade, ak dezambiguáciou pôvodu slov určíme, že všetkých 20 slov má rovnaký pôvod, 90%-ný interval spoľahlivosti za predpokladu aproximácie hypergeometrického rozdelenia výberu slov vo vzorke rozdelením binomickým je (17.21, 20].

ako viacvýznamové skratky (napr. *st, sk, ms, de*), príp. ako cudzie slová a skratky tvarovo zhodné so slovenskými adverbiami, pronominami a pod. (napr. *von, nato*).[12] Pri všetkých sme určili pôvod homoným a ich pravdepodobný pomer výskytu, v prípade potreby aj na väčšej vzorke, napr.

```
ms      *dezamb,15=l+slk,1=l+a,2=a,2=slk
nato    *dezamb,11=slk,9=a
```

K nehomonymným skratkám, ktorým zodpovedá vo väčšine prípadov 1 slovo, resp. slovné spojenie, sa zaradilo 20 z nich (*resp, dr, ing, tasr* a i.), ich pôvod sme teda určili podobne ako pri ostatných apelatívach.

### 2.4. (Ne)zachytenie citátových výrazov

Vo vzorke 2 000 najčastejších slov sa našli také, ktoré nepatria medzi domáce ani prevzaté slová, ale fungujú ako súčasti cudzojazyčných názvov a dlhších importovaných kontextov. Táto skutočnosť prirodzene vyplýva z charakteru písaných korpusov, v korešpondencii s bežne vydávanými textami, v ktorých sa kratšie časti textu (najčastejšie cudzojazyčné názvy v rámci viet) nefiltrujú. Všetky takéto slová (*the, of, new, and* a skratka *USD*) pochádzajú z angličtiny[13]. Vzhľadom na to, že tieto slová nemôžeme považovať za súčasť slovnej zásoby slovenčiny, bolo potrebné vyčleniť ich zo skupiny slovenských slov, preto sme na ich vymedzenie použili značku %.[14] O istej miere adaptácie, resp. používaní v pôvodnej pravopisnej podobe, sme uvažovali len pri skratkách *USA* a *www.*

### 2.5. Zachytenie kalkov, polokalkov, kompozít a kvázikompozít

Lexikón slovenského jazyka tvoria aj rôzne slová, ktoré vznikli kalkovaním, t. j. napodobením štruktúry slova v cudzom jazyku. Z pragmatického hľadiska môžeme v prípade kalkov hovoriť o slovách, pri tvorbe ktorých sa využil z lexikálneho hľadiska „fond domáceho jazyka“, aj keď inšpirácia cudzím výrazom je pri ich vzniku nesporná. V našom zostručnenom opise ich považujeme za slová rovnocenné slovám domáceho pôvodu. Výnimku tvoria polokalky, pri ktorých sa každá príslušná časť – preložená i prevzatá – určuje osobitne, aj keď nemusí ísť o slovo zložené z koreňových morfém 2 plnovýznamových slov (napr. `nádherný slk+hnem`).

[12] Pri lematizácii sme všetky lemy spracovávali v tvare s malými písmenami (v ďalšom texte a tabuľkách tohto článku sú z estetických príčin upravené propriá do ortografickej podoby); bodka za skratkou sa hodnotí ako samostatná lema. Pri určovaní skratiek sme využili o. i. Slovník skratiek a značiek (2004).

[13] Hoci skratka USD je diskutabilná – v konečnom (či primárnom?) dôsledku *United* pochádza z latinčiny, *States* je takisto latinizmus a *Dollar* je germanizmus.

[14] V morfologickom značkovaní vypracovanom v oddelení Slovenského národného korpusu sa značkou % označuje citátový výraz, ktorého definícia nie je totožná s jeho chápaním v tradičnej slovenskej lingvistike (porov. Garabík – Šimková, 2012).

Podobne sme postupovali pri určovaní pôvodu kompozít (vo vzorke sa nevyskytovali), kvázikompozít (televízia a←g+l) a skratiek (km *skr+f←g+f←g).

## 3. ETYMOLOGICKÁ CHARAKTERISTIKA PROPRIÍ

Medzi základné opozície lexikálnej zásoby patrí opozícia apelatívnosť – propriálnosť. Lexikológia sa zameriava na skúmanie predovšetkým apelatívnej lexiky. Onomastiku, skúmajúcu propriálnu lexiku, môžeme na základe rozličných kritérií chápať na jednej strane ako samostatnú vednú disciplínu využívajúcu vo výraznej miere poznatky z mnohých ďalších humanitných aj prírodných vied, na druhej strane ako špecifickú súčasť lexikológie (porov. Dolník, 2003, s. 5 – 6). Vlastné mená, ktoré sa vyskytli v nami skúmanej vzorke najčastejších 2 000 slov, tvorili malú skupinu (77 slov vrátane motivovaných vzťahových adjektív, t. j. 3,85 % unikátnych slov), ktorá ale nie je úplne zanedbateľná.

Skupinu najfrekventovanejších proprií tvoria bionymá (26 antroponým – ide o bežné krstné mená v nezdrobnenom a nezdomácnenom tvare, dve propriá označujúce ústrednú postavu kresťanstva – *Ježiš, Kristus*), toponymá, predovšetkým názvy miest (endonymá a exonymá, napr. *Londýn, Paríž, Rím*), choronymá (názvy krajín, svetadielu *Európa*, kraja *Spiš*) a oronymum (*Tatry*). Názvy slovenských miest sú vo viacerých prípadoch homonymné s hydronymami, čo z hľadiska určovania etymológie nepredstavovalo problém, keďže práve z názvov vodných tokov sa často odvodzovali názvy miest, etymológia oboch je preto rovnaká. Na okraji skupiny proprií stoja názvy príslušníkov národov, v našej vzorke sa nachádzajú len v malom počte (*Slovan, Slovák, Nemec, Žid* a odvodené adjektívum *židovský*).[15]

Môžeme konštatovať, že v porovnaní s určovaním pôvodu bežnej lexiky sa pôvod proprií stanovuje omnoho ťažšie, pretože v tejto oblasti sa často stretávame so slovami, ktoré slovenčina prevzala, resp. boli vytvorené ešte v predslovanskom období, vznikli ako slovanské názvy osôb aj geografických útvarov zväčša v dávnej minulosti. Z toho vyplýva omnoho väčšia náročnosť rekonštrukcie ich pôvodu, ako aj výraznejšia názorová rozdielnosť onomastikov, ktorá sa prejavuje v nezriedka odlišnej interpretácii výsledkov etymologických výskumov proprií, predovšetkým toponým.

Na túto malú vzorku slov sme pre spoľahlivejšie určenie pôvodu použili viacero slovenských a českých prameňov, ktoré sa o. i. líšia spôsobom aj hĺbkou spracovania problematiky. Niektoré pramene sa v zhode so svojím populárnovedným charakterom sústreďujú len na určenie východiskového jazyka (napr. meno *Ivan* je vysvetlené ako *stará slovan. podoba hebr. mena,* Majtán – Považaj, 1998, s. 131), rozsiahlejšie koncipovaná publikácia (Knappová, 2010) určuje aj jazyk, prostredníc-

[15] Druhy vlastných mien označujeme podľa monografie Sémiotika (Černý – Holeš, 2004) a Slovníka cudzích slov (2005).

tvom ktorého prijala slovenčina dané meno (uvedené antroponymum v tomto tvare z hebrejčiny prevzala gréčtina, z nej ruština, z ktorej bolo bezprostredne prevzaté do slovenčiny). Pri niektorých menách sa vo viacerých prameňoch zhodne uvádza dvojaký možný a pravdepodobný pôvod (napr. *Marián* – od mena *Mária*, resp. rodového mena *Márius*).

Najvyššou mierou názorovej nejednoznačnosti a rôznorodosti sa vyznačujú opisy etymológie názvov slovenských miest. Ako príklad pars pro toto môžeme uviesť pomenovanie rieky a mesta *Nitra*, jedného z najstaršie doložených slovenských názvov, a jeho rozličné explikácie v dielach R. Krajčoviča (2005, s. 18 – 20), J. Hladkého (2004, s. 147 – 148) a kolektívu autorov I. Lutterera, L. Kropáčka, V. Huňáčka (1976, s. 195 – 196), ako aj kolektívu zloženého z I. Lutterera, M. Majtána a R. Šrámka (1982, s. 212).

Výsledok porovnania etymologickej charakteristiky vo viacerých prameňoch bol v 11 prípadoch vrátane opísaných natoľko nejednoznačný, že sme pri týchto propriách zaznačili ich pôvod ako neurčiteľný (značkou #). Etymológiu zvyšných 65 proprií sme určili v schematizovanej podobe analogicky ako pri apelatívach.

Aby sme kvantifikovali aj podiel proprií v rámci často používaných prevzatých slov, ukážeme v tomto príspevku dve alternatívne množiny výsledkov: jednu, v rámci ktorej sme celú špecifickú skupinu propriálnych lexikálnych jednotiek nezaradili do výpočtov a štatistík; a druhú, do ktorej sme vlastné mená zaradili.

## VÝSLEDKY SPRACOVANIA

Všetky získané dáta (t. j. najfrekventovanejších 2 000 slov predstavujúcich v danom korpuse 256 974 945 tokenov) anotované na rovnakom metodologickom základe a jednotným spôsobom sme následne spracovali do podoby prehľadných tabuliek a grafov, z ktorých vyberáme najdôležitejšie časti. V nasledujúcich tabuľkách uvádzame percentuálne vyčíslený podiel slov domáceho pôvodu (t. j. slovenské slová), rovnako počet slov pochádzajúcich z konkrétnych cudzích jazykov. Pre prehľadnosť uvádzame len slová s podielom väčším ako 0,01 %, pričom rozlišujeme podiel štartovacích a posledných jazykov. Osobitne podávame výsledky analýz iba apelatívnej lexiky, podobne prezentujeme výsledky pre všetky slová, zahŕňajúc do štatistík apelatíva aj propriá. Všetky percentuálne zastúpenia boli počítané vzhľadom na počet tokenov korpusu pokrytých vzorkou 2 000 lem po odstránení interpunkcie, číslic, jednopísmenových tokenov vo funkcii neslovných elementov a vlastných mien (v prípade analýzy iba apelatív).

| | bez vlastných mien | s vlastnými menami |
|---|---|---|
| pokryté tokeny | 179 543 331 | 182 400 765 |

Tab. 2: Pokrytie korpusu analyzovanými slovami

| *Štartovací jazyk* | *Podiel slov [%]* | *Štartovací jazyk* | *Podiel slov [%]* |
|---|---|---|---|
| slovenský | 93,00 | francúzsky | 0,04 |
| latinský | 3,57 | turkický | 0,03 |
| grécky | 1,56 | dolnonemecký | 0,02 |
| hornonemecký | 0,62 | egyptský | 0,02 |
| český | 0,29 | keltský | 0,02 |
| anglický | 0,24 | ugrofínsky | 0,02 |
| germánsky | 0,18 | poľský | 0,01 |
| hebrejský | 0,14 | ruský | 0,01 |
| akkadský | 0,10 | čínsky | 0,01 |
| nemecký | 0,10 | arabský | 0,01 |
| | | taliansky | 0,01 |

Tab. 3: Percentuálny podiel lexikálnych jednotiek v sledovanej vzorke podľa štartovacieho jazyka bez vlastných mien.

| *Štartovací jazyk* | *Podiel slov [%]* | *Štartovací jazyk* | *Podiel slov [%]* |
|---|---|---|---|
| slovenský | 92,65 | francúzsky | 0,04 |
| latinský | 3,61 | keltský | 0,03 |
| grécky | 1,63 | ugrofínsky | 0,03 |
| hornonemecký | 0,62 | turkický | 0,03 |
| český | 0,29 | čínsky | 0,02 |
| hebrejský | 0,25 | dolnonemecký | 0,02 |
| anglický | 0,24 | egyptský | 0,02 |
| germánsky | 0,24 | poľský | 0,01 |
| akkadský | 0,14 | ruský | 0,01 |
| nemecký | 0,11 | arabský | 0,01 |
| | | taliansky | 0,01 |

Tab. 4: Percentuálny podiel lexikálnych jednotiek v sledovanej vzorke podľa štartovacieho jazyka vrátane vlastných mien.

| *Posledný jazyk* | *Podiel slov [%]* | *Posledný jazyk* | *Podiel slov [%]* |
|---|---|---|---|
| slovenský | 93,00 | český | 0,29 |
| latinský | 3,58 | germánsky | 0,21 |
| grécky | 0,94 | taliansky | 0,18 |
| hornonemecký | 0,63 | nemecký | 0,14 |
| francúzsky | 0,39 | hebrejský | 0,09 |
| anglický | 0,33 | akkadský | 0,05 |

| *Posledný jazyk* | *Podiel slov [%]* | *Posledný jazyk* | *Podiel slov [%]* |
|---|---|---|---|
| poľský | 0,02 | ugrofínsky | 0,02 |
| turkický | 0,02 | ruský | 0,02 |
| maďarský | 0,02 | keltský | 0,01 |
| španielsky | 0,02 | dolnonemecký | 0,01 |
| | | egyptský | 0,01 |

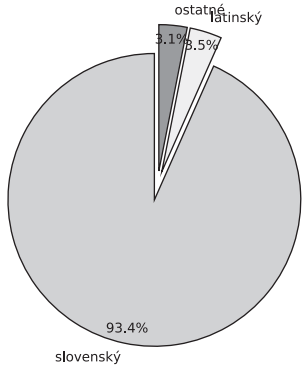
Tab. 5: Percentuálny podiel lexikálnych jednotiek v sledovanej vzorke podľa posledného jazyka bez vlastných mien.

| *Posledný jazyk* | *Podiel slov [%]* | *Posledný jazyk* | *Podiel slov [%]* |
|---|---|---|---|
| slovenský | 92,89 | turkický | 0,02 |
| latinský | 3,61 | ugrofínsky | 0,02 |
| grécky | 1,17 | poľský | 0,02 |
| hornonemecký | 0,63 | keltský | 0,02 |
| anglický | 0,30 | ruský | 0,02 |
| český | 0,29 | dolnonemecký | 0,01 |
| francúzsky | 0,28 | španielsky | 0,01 |
| germánsky | 0,21 | egyptský | 0,01 |
| taliansky | 0,14 | maďarský | 0,01 |
| nemecký | 0,13 | čínsky | 0,01 |
| hebrejský | 0,12 | perzský | 0,01 |
| akkadský | 0,07 | arabský | 0,01 |

Tab. 6: Percentuálny podiel lexikálnych jednotiek v sledovanej vzorke podľa posledného jazyka vrátane vlastných mien.

Informácie z predchádzajúcich tabuliek môžeme ukázať v zjednodušenej podobe v koláčovom grafe (obr. 2), ktorý vzhľadom na zaokrúhľovanie vyzerá približne rovnako pre rozdelenie slov z hľadiska štartovacieho aj posledného jazyka.

Keďže slová domáceho pôvodu tvoria prevažnú väčšinu najpoužívanejšej časti slovenskej lexiky (približne na úrovni 93 %), nižšie uvádzame tabuľky, v ktorých sú zachytené percentuálne podiely len cudzích slov vzhľadom na všetky cudzie slová, t. j. po odfiltrovaní slov domáceho pôvodu (ide teda o podrobnejšiu analýzu zvyšných cca 7 % lexiky).




Obr. 2: Percentuálne zastúpenie apelatív domáceho pôvodu, latinizmov a slov prevzatých z ostatných cudzích jazykov, bez vlastných mien. Tento graf je takmer zhodný pre štartovacie aj pre posledné jazyky, a zároveň je takmer zhodný s grafom zachytávajúcim vzorku slov vrátane vlastných mien.

| *Štartovací jazyk* | *Podiel prevzatých slov [%]* |
|---|---|
| latinský | 53,13 |
| grécky | 22,59 |
| hornonemecký | 9,44 |
| český | 4,44 |
| anglický | 3,71 |
| germánsky | 1,97 |
| nemecký | 1,30 |
| akkadský | 0,79 |
| francúzsky | 0,66 |
| turkický | 0,39 |
| hebrejský | 0,37 |
| dolnonemecký | 0,29 |
| egyptský | 0,28 |
| poľský | 0,16 |
| ruský | 0,16 |
| arabský | 0,15 |
| taliansky | 0,09 |
| španielsky | 0,07 |
| neurčiteľný | 0,02 |

Tab. 7: Percentuálny podiel prevzatých slov v sledovanej vzorke podľa štartovacieho jazyka bez vlastných mien.

| *Štartovací jazyk* | *Podiel prevzatých slov [%]* |
|---|---|
| latinský | 49,08 |
| grécky | 22,12 |
| hornonemecký | 8,39 |
| český | 3,95 |
| hebrejský | 3,36 |
| anglický | 3,30 |
| germánsky | 3,24 |
| akkadský | 1,94 |
| nemecký | 1,43 |
| francúzsky | 0,58 |
| keltský | 0,46 |
| ugrofínsky | 0,43 |
| turkický | 0,35 |
| čínsky | 0,27 |
| dolnonemecký | 0,26 |
| egyptský | 0,25 |
| poľský | 0,14 |
| ruský | 0,14 |
| arabský | 0,13 |
| taliansky | 0,08 |
| španielsky | 0,06 |
| neurčiteľný | 0,02 |

Tab. 8: Percentuálny podiel prevzatých slov v sledovanej vzorke podľa štartovacieho jazyka vrátane vlastných mien.

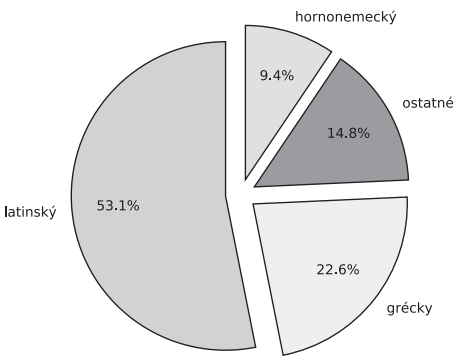
| *Posledný jazyk* | *Podiel prevzatých slov [%]* |
|---|---|
| latinský | 52,53 |
| francúzsky | 10,85 |
| hornonemecký | 9,73 |
| anglický | 6,40 |
| taliansky | 4,71 |
| český | 4,44 |
| grécky | 3,61 |
| germánsky | 3,03 |
| nemecký | 2,67 |
| poľský | 0,54 |
| maďarský | 0,53 |
| španielsky | 0,44 |
| turkický | 0,20 |
| ruský | 0,16 |
| turecký | 0,15 |
| neurčiteľný | 0,02 |

Tab. 9: Percentuálny podiel prevzatých slov v sledovanej vzorke podľa posledného jazyka bez vlastných mien.

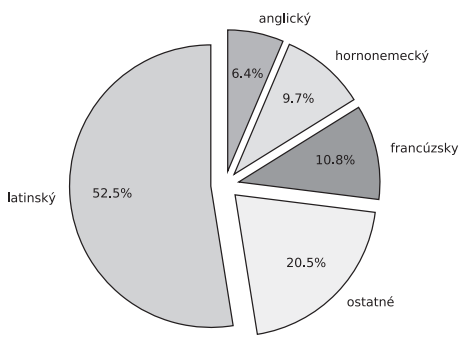
| *Posledný jazyk* | *Podiel prevzatých slov [%]* |
|---|---|
| latinský | 50,44 |
| francúzsky | 10,11 |
| hornonemecký | 8,66 |
| anglický | 5,69 |
| grécky | 5,50 |
| taliansky | 5,35 |
| český | 3,95 |
| germánsky | 3,53 |
| nemecký | 2,66 |
| hebrejský | 1,04 |
| poľský | 0,48 |
| maďarský | 0,47 |
| ugrofínsky | 0,43 |
| ruský | 0,40 |
| španielsky | 0,39 |
| keltský | 0,29 |
| perzský | 0,27 |
| turkický | 0,18 |
| turecký | 0,13 |
| neurčiteľný | 0,02 |

Tab. 10: Percentuálny podiel prevzatých slov v sledovanej vzorke podľa posledného jazyka vrátane vlastných mien.

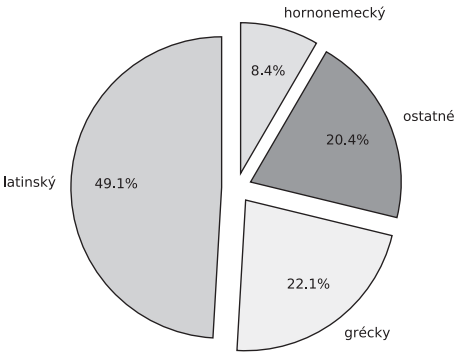
Vizuálne sa tieto údaje dajú znázorniť v grafoch, samostatne s prehľadom štartovacích a posledných jazykov:



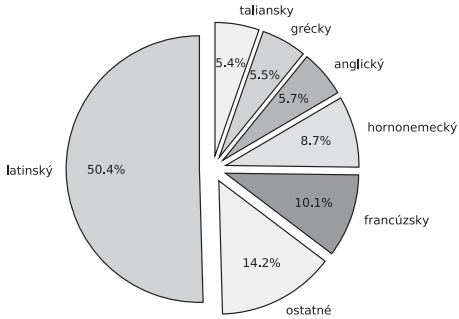

Obr. 3: Podiel prevzatých slov v sledovanej vzorke podľa štartovacieho jazyka bez vlastných mien.




Obr. 4: Podiel prevzatých slov v sledovanej vzorke podľa posledného jazyka bez vlastných mien.




Obr 5: Podiel prevzatých slov v sledovanej vzorke podľa štartovacieho jazyka vrátane vlastných mien.




Obr. 6: Podiel prevzatých slov v sledovanej vzorke podľa posledného jazyka vrátane vlastných mien.

V nasledujúcich tabuľkách uvádzame zoznamy niekoľkých najčastejších slov pochádzajúcich z cudzích jazykov, z ktorých má slovenčina najvyšší počet prevzatých slov: z latinčiny, gréčtiny, francúzštiny, angličtiny, hornej nemčiny a češtiny. Dôsledne opäť rozlišujeme medzi prevzatiami z tzv. štartovacieho, resp. posledného jazyka. Počet výskytov uvádzame v jednotkách na milión slov korpusu (partes per millionem; ppm).

| *prevzaté slová* | *ppm* |
|---|---|
| štát | 369 |
| situácia | 348 |
| firma | 320 |
| informácia | 309 |
| sociálny | 270 |
| milión | 254 |
| percento | 249 |
| autor | 248 |
| projekt | 248 |
| štátny | 240 |

Tab. 11: latinčina, štartovací jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| problém | 508 |
| škola | 441 |
| štát | 369 |
| systém | 353 |
| informácia | 309 |
| program | 304 |
| Ján | 277 |
| sociálny | 270 |
| Peter | 256 |
| percento | 249 |

Tab. 12: latinčina, posledný jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| problém | 508 |
| škola | 441 |
| systém | 353 |
| program | 304 |
| Peter | 256 |
| auto | 247 |
| politický | 246 |
| cirkev | 202 |
| typ | 183 |
| koruna | 179 |

Tab. 13: gréčtina, štartovací jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| európsky | 287 |
| euro | 278 |
| auto | 247 |
| Európa | 198 |
| fotografia | 106 |
| technický | 103 |
| telefón | 92 |
| technológia | 91 |
| Štefan | 84 |
| technika | 81 |

Tab. 14: gréčtina, posledný jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| tréner | 164 |
| de | 50 |
| st | 16 |

Tab. 15: francúzština, štartovací jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| situácia | 348 |
| politický | 246 |
| finančný | 192 |
| organizácia | 172 |
| ministerstvo | 167 |
| plán | 164 |
| minister | 161 |
| republika | 147 |
| politika | 146 |
| sezóna | 144 |

Tab. 16: francúzština, posledný jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| film | 277 |
| gól | 239 |
| klub | 186 |
| tím | 184 |
| futbalový | 83 |
| filmový | 76 |
| futbal | 71 |
| in | 57 |
| internet | 47 |
| NATO | 33 |

Tab. 17: angličtina, štartovací jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| film | 277 |
| gól | 239 |
| klub | 186 |
| tím | 184 |
| tréner | 164 |
| dolár | 103 |
| parlament | 93 |
| partner | 90 |
| futbalový | 83 |
| šport | 81 |

Tab. 18: angličtina, posledný jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| musieť | 1155 |
| chvíľa | 382 |
| cieľ | 272 |
| rámec | 209 |
| banka | 202 |
| nemusieť | 170 |
| považovať | 169 |
| budova | 136 |
| škoda | 112 |
| vážny | 90 |

Tab. 19: horná nemčina, štartovací jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| musieť | 1155 |
| chvíľa | 382 |
| cieľ | 272 |
| rámec | 209 |
| nemusieť | 170 |
| považovať | 169 |
| budova | 136 |
| škoda | 112 |
| biskup | 92 |
| vážny | 90 |

Tab. 20: horná nemčina, posledný jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| otázka | 429 |
| skupina | 382 |
| trieda | 113 |
| zvláštny | 96 |
| vzájomný | 79 |
| zabezpečiť | 79 |
| slečna | 69 |
| bezpečnosť | 68 |
| nebezpečný | 68 |
| vzdelávanie | 61 |

Tab. 21: čeština, štartovací jazyk

| *prevzaté slová* | *ppm* |
|---|---|
| otázka | 429 |
| skupina | 382 |
| trieda | 113 |
| zvláštny | 96 |
| vzájomný | 79 |
| zabezpečiť | 79 |
| slečna | 69 |
| bezpečnosť | 68 |
| nebezpečný | 68 |
| vzdelávanie | 61 |

Tab. 22: čeština, posledný jazyk

## ZÁVER

Z výsledkov nášho skúmania môžeme upriamiť pozornosť hlavne na fakt, že podiel prevzatých slov v našej vzorke, pričom máme na mysli slová prevzaté v ktoromkoľvek dejinnom období a na rôznej úrovni adaptácie, je menej ako 7 % (pričom táto vzorka lexikálne pokrýva cca ¾ korpusu, ak berieme do úvahy aj interpunkciu a číslice, alebo ½ korpusu bez uváženia interpunkcie a číslic). Pri analýze jazykového materiálu sme vychádzali z aktuálneho stupňa poznania v oblasti etymológie a onomastiky opísaného v slovníkových dielach a štúdiách. Skúmaná vzorka v rámci reprezentatívneho a rozsiahleho korpusu zahŕňa veľkú časť frekventovanej slovnej zásoby (2 000 najčastejších lexém a im zodpovedajúcich 256 974 945 slovných tvarov a neslovných elementov (tokenov) z korpusu *prim-7.0-public-vyv*, ktorého celková veľkosť je 340 708 046 tokenov), ktorú môžeme považovať za súčasť jadra slovenskej lexiky. Vzhľadom na mnohotvárnosť skúmaného jazykového materiálu sme zvolili primeraný spôsob spracovania tak, že sme na začiatku vyselektovali podstatnú časť slov a neslovných jednotiek (odfiltrovaním interpunkcie, vysoko homonymných jednopísmenových slov a skratiek). Následne sme osobitne skúmali jednotky apelatívnej lexiky, ďalej sme do výpočtov zahrnuli aj najčastejšie propriá a vzťahové adjektíva motivované vlastnými menami. Štatistickým spracovaním sme získali prehľadne usporiadané dáta, ktoré prinášajú relevantné informácie o zložení väčšej časti slovnej zásoby slovenčiny z hľadiska pôvodu, prehľad jazykov, z ktorých slovenčina prevzala najviac lexém, ako aj konkrétne zoznamy prevzatých slov s príslušnými frekvenciami.

Podľa očakávania je jazykom s najväčším lexikálnym vplyvom na slovenčinu latinčina, dokonca viac ako polovica všetkých prevzatých slov sú latinizmy, po nej nasleduje gréčtina. Vysoký podiel prevzatých slov zo (staro)hornonemeckého jazyka je spôsobený prevažne jednou, zato veľmi častou lexémou (*musieť*). Pri výskume sme zachytávali celý „priebeh“ preberania, od tzv. štartovacieho až po posledný jazyk, z ktorého slovenčina bezprostredne danú lexému prevzala. Z tohto hľadiska sa ako zaujímavý jazyk prejavuje francúzština, pretože z nej slovenčina priamo prevzala viacero slov, ktoré však nie sú pôvodom francúzske. Väčšina z frekventovaných galicizmov má pôvod v latinčine, resp. aj v gréčtine, ďalšie prevzal tento románsky jazyk z taliančiny či španielčiny a až cez francúzštinu boli importované do slovenčiny. Francúzština tak figuruje viac ako „sprostredkovateľský“ jazyk, nedá sa tiež obísť fakt, že časté prevzaté slová z tohto jazyka sú hlavne termíny z dôležitých oblastí finančníctva a politiky.

Zahrnutie proprií do analyzovanej vzorky spôsobuje malý, ale viditeľný nárast hebrejčiny a tomu zodpovedajúci pokles podielu ostatných jazykov vzhľadom na častý hebrejský pôvod obľúbených krstných mien.

Slovenčina sa nám na základe zistených dát javí ako otvorený, živý a dynamický systém s prevažnou väčšinou slov domáceho pôvodu. Potvrdili sa tiež všeobecne

známe závery, že nepatrí k jazykom, do ktorých by výrazne prenikali prevzaté slová z ktoréhokoľvek jazyka, hoci ich podiel nie je celkom zanedbateľný.

Anotované dáta sú sprístupnené[16] a možno ich použiť v ďalších výskumoch, napríklad sa dá ľubovoľne rozšíriť oblasť pokrytia na slová z nižším výskytom, alebo sa môžu vyčleniť osobitné triedy slov (napr. pri skúmaní vybraných plnovýznamových slovných druhov, ale aj synsématík).

**Poďakovanie:**

Za cenné rady, pripomienky a zapožičanie odbornej literatúry vyjadrujeme vďaku pracovníkom JÚĽŠ SAV Ľ. Králikovi a I. Valentovej. Na prvotnom spracovaní apelatívnej lexiky na základe Stručného etymologického slovníka slovenčiny (2015) spolupracovali stážistky zo Záhrebskej univerzity Ana Grbavać a Marina Kolesarić, za čo im rovnako patrí poďakovanie.

**Literatúra**

CORRAL, Álvaro – BOLEDA, Gemma – FERRER-I-CANCHO, Ramon: Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts. In: PLoS ONE, 2015, 10(7). Dostupné na: https://doi.org/10.1371/journal.pone.0129031

ČERNÝ, Jiří – HOLEŠ, Jan: Sémiotika. Praha: Portál 2004. 368 s.

DERKSEN, Rick: Etymological Dictionary of the Slavic Inherited Lexicon. Leiden – Boston: Brill 2008. 726 s.

DOLNÍK, Juraj: Lexikológia. Bratislava: Univerzita Komenského 2003. 236 s.

Encyklopédia jazykovedy. Zost. J. Mistrík a kol. Bratislava: Obzor 1993. 513 s.

GARABÍK, Radovan – KMEŤOVÁ, Beáta – ŠIMKOVÁ, Mária – ZUMRÍK, Miroslav: Frekvenčný slovník slovenčiny na báze Slovenského národného korpusu. Bratislava: Veda 2017. 562 s.

GARABÍK, Radovan – KMEŤOVÁ, Beáta – KARČOVÁ, Agáta – BOBEKOVÁ, Kristína – MAJCHRÁKOVÁ, Daniela – CHLPÍKOVÁ, Katarína: Retrográdny slovník súčasnej slovenčiny – slovné tvary na báze Slovenského národného korpusu. Hl. red. R. Garabík. Veda 2018. 848 s.

GARABÍK, Radovan – ŠIMKOVÁ, Mária: Slovak Morphosyntactic Tagset. In: Journal of Language Modelling. Institute of Computer Science PAS, 2012, roč. 0, č. 1, s. 41 – 63.

Loanwords in the World's Languages: A Comparative Handbook. Ed. M. Haspelmath – U. Tadmor. Berlin: Mouton De Gruyter 2009. 1081 s.

HLADKÝ, Juraj: Hydronymia povodia Nitry. Trnava: Pedagogická fakulta Trnavskej univerzity 2004. 294 s.

HOLUB, Josef: Stručný etymologický slovník jazyka českého. 2. rozšír. vyd. pripravil I. Lutterer. Praha: SPN 1978. 528 s.

HORECKÝ, Ján – BUZÁSSYOVÁ, Klára – BOSÁK, Ján a kol.: Dynamika slovnej zásoby súčasnej slovenčiny. Bratislava: Veda 1989. 436 s.

ISO/IEC 10646:2017, Information technology – Universal Coded Character Set (UCS).

KNAPPOVÁ, Miloslava: Jak se bude vaše dítě jmenovat? 5. aktualiz. a rozšír. vyd. Praha: Academia 2010. 784 s.

[16] Dáta sú zverejnené pod licenciou *Creative Commons Attribution-ShareAlike 4.0 International* na https://www.juls.savba.sk/data.html

KRAJČOVIČ, Rudolf: Živé kroniky slovenských dejín skryté v názvoch obcí a miest. Bratislava: Literárne informačné centrum 2005. 230 s.
KRÁLIK, Ľubor: Stručný etymologický slovník slovenčiny. Bratislava: Veda 2015. 704 s.
LALÚCH, Róbert – KONCOVÁ, Monika: Slovník skratiek a značiek. Bratislava: Ikar 2004. 488 s.
LUTTERER, Ivan – KROPÁČEK, Luboš – HUŇÁČEK, Václav: Původ zeměpisných jmen: etymologický slovník 1000 vlastních jmen zemí, měst a přírodních objektů z celého světa. Praha: Mladá fronta 1976. 304 s.
LUTTERER, Ivan – MAJTÁN, Milan – ŠRÁMEK, Rudolf: Zeměpisná jména Československa. Slovník vybraných zeměpisných jmen s výkladem jejich původu a historického vývoje. Praha: Mladá fronta 1982. 373 s.
MACHEK, Václav: Etymologický slovník jazyka českého a slovenského. 1. vyd. Praha: Vydavateľstvo ČSAV 1957. 627 s.
MACHEK, Václav: Etymologický slovník jazyka českého. 3. vyd. Praha: Academia 1971. 866 s.
MAJTÁN, Milan – POVAŽAJ, Matej: Vyberte si meno pre svoje dieťa. 1. vyd. Bratislava: ART AREA 1998. 344 s.
OLOŠTIAK, Martin: O názoroch na preberanie cudzích lexikálnych jednotiek do slovenčiny. In: Jazyková kultúra na začiatku tretieho tisícročia. Ed. M. Považaj. Bratislava: Veda 2009, s. 87 – 95.
ONDREJOVIČ, Slavomír: K niektorým výzvam a petíciám na ochranu slovenského jazyka. In: Jazykovedný časopis, 2010, roč. 61, č. 1, s. 5 – 14.
ONDRUS, Pavel – HORECKÝ, Ján – FURDÍK, Juraj: Súčasný slovenský spisovný jazyk. Lexikológia. Bratislava: Slovenské pedagogické nakladateľstvo 1980. 232 s.
PIANTADOSI, Steven T.: Zipf's word frequency law in natural language: A critical review and future directions. In: Psychonomic Bulletin & Review, 2014, roč. 21, č. 5, s. 1112 – 1130.
REJZEK, Jiří: Český etymologický slovník. 1. vyd. Vozice: Leda 2001. 752 s.
SHIEH, Grace S.: A weighted Kendall's tau statistic. In: Statistics & Probability Letters, 1998, roč. 39, č. 1, s. 17 – 24.
Slovenský národný korpus – *prim-7.0-public-vyv*. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2015. Dostupný z www: http://korpus.juls.savba.sk.
Slovník cudzích slov. Akademický. 2., dopl. a uprav. slov. vyd. Bratislava: Slovenské pedagogické nakladateľstvo 2005. 991 s.
ŠKVARENINOVÁ, Oľga: Vplyv médií na globalizáciu slovenského jazyka. In: Jazyk v politických, ideologických a interkultúrnych vzťahoch. Sociolinguistica Slovaca 8. Ed. J. Wachtarczyková – L. Satinská – S. Ondrejovič. Bratislava: Veda 2015, s. 33 – 47.
ZIPF, George Kingsley: The Psycho-Biology of Language. An Introduction to Dynamic Philology. London: George Routledge & Sons Ltd. 1936. 336 s.