Nelze tvrdit, že by titul přestával hrát důležitou úlohu v rámci kompozice literárního díla, stále funguje jako vývěsní štít díla i autora, ne nepodstatná je též jeho atraktivizační funkce, kdy tituly slouží jako „reklama“ na dílo (i v tom případě, je-li vztah mezi titulem a dílem jen volný nebo nejasný).

**Literatura**

Encyklopedický slovník češtiny. Red. P. Karlík, M. Nekula, J. Pleskalová. 1. vyd. Praha: Nakladatelství Lidové noviny 2002. 604 s.

HODROVÁ, Daniela: …na okraji chaosu… Praha: Torst 2001. 865 s.

PETERKA, Josef: Teorie literatury pro učitele. Praha: UK-PedF 2001. 287 s.

Slovník spisovné češtiny. Red. F. Daneš – J. Filipec – J. Machač – V. Mejstřík. 2. vyd. Praha: Academia 1998. 647 s.

VŠETIČKA, František: Podoby prózy. Olomouc: Votobia 1997. 248 s.

## Štatistické metódy vyhľadávania verbo-nominálnych kolokácií v korpuse a analýza kolokačných mier

Daniela Majchráková

Slovenský národný korpus,
Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied, Bratislava

### 1 Úvod

Kolokácie sú kombinácie slov, ktoré sú na základe frekvencie výskytu v istých textoch vyčleňované a zoraďované pomocou rozličných nástrojov korpusovej lingvistiky. Pojem kolokácia môžeme chápať v širšom a užšom zmysle podľa toho, či zohľadňujeme kvalitatívne alebo kvantitatívne kritérium. Buď je to frekventované spojenie, t. j. zdôrazňujeme len frekvenčné hľadisko spojenia, alebo berieme do úvahy aj jeho sémantický aspekt. Sémantická kompaktnosť, celistvosť komponentov slovného spojenia je pre posudzovateľa kolokácie, lingvistu, nevyhnutná požiadavka jej identifikácie. Pojem kolokácia má teda širší záber a pokrýva rôzne kombinácie slov, od idiomatických až po typické frekventované spojenia. *„Kolokace jsou prvním klíčem, jehož pomocí se dostáváme k poznání způsobů, pravidel a hranic chování lexémů“* (Čermák, 2006b, s. 10). Kolokácie nám napomáhajú odhaliť kombinatorickú schopnosť slova, jeho syntagmatický potenciál, vďaka čomu sa môžeme viac priblížiť k jeho aktuálnej platnosti v jazyku, k jeho významu (Čermák, 2006b). A. Jarošová definuje pojem kolokácia ako *„nocionálne ustálené spojenie alebo typické spojenie“*, čím pokrýva *„systémové jednotky (lexikalizované spojenia nominálne aj verbo-nominálne spojenia s významom adverbií a synsémantík) a útvary*

*na hranici textu a systému (typické spojenia)"* (Jarošová, 2007, s. 99 – 100). Pojem kolokácia vymedzuje na základe funkčného kritéria; kolokáciu tiež nazýva lexikalizovaným spojením (Jarošová, 2007).

V súčasnosti sa ako naliehavá vynára otázka spôsobu lexikografického zachytenia kolokácií a spôsobu využitia štatistických nástrojov pri ich identifikácii v rozsiahlych korpusových dátach. Na registráciu ustálených kolokácií extrahovaných z korpusu na základe frekvenčných kritérií je zameraný projekt *Slovníka slovenských kolokácií*, ktorý vypracoval a vedie P. Ďurčo (2007).

V nadväznosti na spomínaný projekt sa v príspevku zameriame na možnosti štatistickej identifikácie ustálených spojení a budeme sa snažiť odpovedať na otázku, do akej miery sa môže lingvista – lexikograf spoľahnúť na štatistické hodnoty kolokačných mier a čo hovoria jednotlivé štatistické hodnoty, odhliadnuc od zložitých matematických výpočtov, o spájateľnosti daného slova alebo o jeho postavení v korpuse. Rovnako sa budeme zaoberať otázkou, či je vysoká hodnota absolútnej frekvencie nevyhnutnou podmienkou určenia ustáleného slovného spojenia.

**2 Extrakcia kolokácií**

Centrom nášho záujmu budú verbo-nominálne spojenia, konkrétne spojenia slovesa a podstatného mena a slovesa, prídavného mena a podstatného mena. Môžeme ich prirovnať k lexikalizovaným verbo-nominálnym spojeniam (systémovým jednotkám) alebo typickým spojeniam *„na hranici textu a systému"* podľa A. Jarošovej (2007, s. 100), alebo k lexikalizovaným spojeniam typu sloveso a podstatné meno, resp. príslovka, teda pôjde o „*neterminologické viacslovné pomenovania bez prvku obraznosti"* podľa J. Kačalu (1997, s. 193 – 194), alebo tiež k analytickým verbo-nominálnym spojeniam podľa Ľ. Kralčáka (2005). Samozrejme očakávame, že medzi získanými kolokáciami sa môžu vyskytovať aj frazeologické spojenia – idiomatické kolokácie (Čermák, 2006, s. 12). V našom kontexte budeme teda za kolokáciu považovať idiomatické spojenia, ustálené typické a takisto veľmi frekventované spojenia, pričom sa vyhneme analýze termínov a proprií.

Ako jazykový materiál sme si zvolili dáta Slovenského národného korpusu, konkrétne verziu prim-3.0-snk-all z roku 2007 v rozsahu 350 miliónov textových jednotiek. Z korpusu boli vybrané n-gramy, t. j. postupnosti tokenov (textových jednotiek), na základe morfologických kritérií. Automatizovane boli vygenerované tieto kombinácie lem: verbum + substantívum (V – S), verbum + prepozícia + substantívum (V – P – S), verbum + adjektívum + substantívum (V – A – S), verbum + prepozícia + adjektívum + substantívum (V – P – A – S). Jednotlivé bigramy, trigramy a tetragramy obsahovali slová v základných tvaroch (lemy) a boli zoradené podľa absolútnej frekvencie v korpuse. Napriek tomu, že sa kolokáty nie vždy dajú zovšeobecniť na lemy, keďže sú primárne tvorené slovnými tvarmi, jazykový materiál extrahujeme na základe jazykových vzorcov, ktoré nevyhnutne počítajú so základnými tvarmi slov. V prvej fáze budeme

kolokácie získavať s ohľadom na ich formálnu stránku, v ďalšej sa zameriame na sémantické a funkčné kritériá, podľa ktorých budeme dané kolokácie triediť. Budeme sa snažiť zodpovedať tieto kľúčové otázky: Ako súvisí frekvencia kolokácie s jej ustálenosťou? Ako sa ustálenosť kolokácie odráža v jednotlivých štatistických hodnotách kolokačných mier?

**3 Analýza kolokácií**

Zoznam najfrekventovanejších kolokácií typu verbum – substantívum ukazuje, že najčastejšie sa v slovenčine vyskytujú kombinácie so slovesom *byť* (porov. tabuľka č. 1). Tieto spojenia predstavujú len analytické tvary so slovesom *byť*, napr. *byť čas, byť dôvod, byť problém, byť hráč.* Na druhej strane je však zoznam bohatý na tzv. lexikalizované spojenia *„kategoriálneho slovesa s podstatným menom"* (Kačala, 1997, s. 193 – 194) alebo *„analytické verbo-nominálne spojenia"* (Kralčák, 2005), napr. *dať pozor, klásť otázka, položiť otázka, venovať pozornosť, mať obava, utrpieť zranenie*, ktoré by sme rovnako mohli pomenovať typickými spojeniami, ku ktorým môžeme priradiť ich rovnako typické cudzojazyčné ekvivalenty (*venovať pozornosť – pay attention, dať pozor – take care* a pod.). Takto zvolený spôsob získavania jazykového materiálu sa môže osvedčiť pri skúmaní a získavaní verbo-nominálnych spojení.

Najfrekventovanejšie kolokácie typu verbum – prepozícia – substantívum tvoria zhodne s predchádzajúcou skupinou kolokácií prevažne spojenia slovesa *byť*, v zozname tiež nájdeme typické spojenia, *„lexikalizované spojenia kategoriálneho slovesa s podstatným menom ako predmetu alebo príslovkového určenia"* (Kačala, 1997, s. 194), napr. *zísť na um, mať na svedomí, brať na vedomie, prichádzať do úvahy, dať za pravdu, utkvieť v pamäti*. Na rozdiel od skupiny kolokácií verbum – substantívum sa môžeme stretnúť aj s výskytom niekoľkých idiomatických spojení s istou mierou obraznosti *vŕtať v hlave, visieť vo vzduchu, praskať vo švíkoch, voditi (niekoho) za nos*.

**Tabuľka č. 1:** Príklady najfrekventovanejších kolokácií typu verbum – substantívum a verbum – prepozícia – substantívum v poradí, v akom sa vyskytovali v prim-3.0-snk-all, uvedené podľa absolútnej frekvencie.

| **kolokácie typu verbum-substantívum** | | | **kolokácie typu verbum-prepozícia-substantívum** | | | |
|---|---|---|---|---|---|---|
| 15841 | byť | pravda | 7000 | byť | v | poriadok |
| 13058 | mať | možnosť | 6281 | mať | k | dispozícia |
| 12691 | mať | pocit | 5735 | byť | k | dispozícia |
| 11496 | mať | pravda | 5676 | mať | na | myseľ |
| 10983 | byť | súčasť | 5672 | byť | v | súčasnosť |
| 10959 | mať | záujem | 5054 | byť | na | Slovensko |
| 10470 | mať | právo | 5001 | Vranov | nad | Topľa |
| 8456 | Vladimír | mečiar | 4477 | byť | v | rok |
| 8272 | mať | problém | 3715 | mať | na | starosť |
| 7392 | byť | problém | 3706 | byť | na | program |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7304 | byť | človek | 3623 | byť | v | súlad |
| 7236 | vysoký | tatra | 3522 | brať | do | úvaha |
| 7091 | byť | člen | 3407 | byť | v | podstata |
| 6360 | byť | výsledok | 3347 | byť | v | rozpor |
| 6020 | byť | čas | 3046 | mať | za | následok |
| 5730 | byť | dôvod | 2935 | zísť | na | um |
| 5675 | byť | ľudia | 2821 | odpovedať | na | otázka |
| 5267 | mať | šanca | 2631 | mať | v | ruka |
| 5202 | pokrútiť | hlava | 2585 | prísť | o | život |
| 4944 | mať | zmysel | 2544 | mať | v | úmysel |
| 4927 | otvoriť | dvere | 2496 | držať | v | ruka |
| 4631 | byť | situácia | 2444 | byť | na | svet |
| 4574 | dávať | pozor | 2395 | prísť | na | rad |
| 4307 | byť | otázka | 2226 | byť | v | skutočnosť |
| 4235 | mať | čas | 2120 | byť | v | porovnanie |
| 4107 | byť | svedok | 2089 | dostať | do | ruka |
| 4062 | byť | vec | 1948 | byť | na | miesto |
| 3985 | byť | predmet | 1917 | mať | na | svedomie |
| 3904 | byť | fakt | 1900 | vzniknúť | v | rok |
| 3606 | byť | pán | 1722 | byť | v | prípad |

Zaujímavé spojenia nachádzame v zozname kolokácií verbum – adjektívum – substantívum a verbum – prepozícia – adjektívum – substantívum, v ktorých sa vyskytuje podstatne viac idiomatických spojení. Frekvenčné hodnoty kolokácií poukazujú na stereotypné alebo časté používanie typických spojení, ako aj obrazných pomenovaní, napríklad *nájsť spoločnú reč, podať pomocnú ruku, byť v plnom prúde, vidieť na vlastné oči, ťahať za kratší koniec*, *byť na dobrej ceste, brať na ľahkú váhu.*

Niektoré spojenia však môžu mať skutočné frekvenčné hodnoty vyššie, ako ich máme zaznamenané v zozname. Je to spôsobené buď tým, že jednotlivé komponenty konkrétneho ustáleného spojenia môžu variovať, napr. *zobrať/brať/vziať na ľahkú váhu, ťahať/potiahnuť za kratší koniec*, alebo pozičným umiestnením jednotlivých komponentov spojenia v rámci vety, napr. *Tam už* ***bola*** *zábava* ***v plnom prúde***. Zoznam kolokácií preto nemusí vždy zobrazovať skutočnú absolútnu frekvenciu ustáleného spojenia a aj poradie jednotlivých kolokácií je len orientačné.

**Tabuľka č. 2:** Príklady najfrekventovanejších kolokácií typu verbum – adjektívum – substantívum a verbum – prepozícia – adjektívum – substantívum v poradí, v akom sa vyskytovali v prim-3.0-snk-all, uvedené podľa absolútnej frekvencie.

| **kolokácie typu verbum-adjektívum-substantívum** | | | | **kolokácie typu verbum-prepozícia-adjektívum-substantívum** | | | | |
|---|---|---|---|---|---|---|---|---|
| 1228 | byť | vysoký | čas | 691 | byť | na | dobrý | cesta |
| 942 | mať | veľký | šanca | 648 | obviniť | z | trestný | čin |
| 931 | mať | veľký | význam | 615 | vidieť | na | vlastný | oko |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 893 | podať | trestný | oznámenie | 591 | ťahať | za | krátky | koniec |
| 763 | studiť[1] | Academica | Slovaca | 505 | byť | v | plný | prúd |
| 715 | začať | trestný | stíhanie | 469 | byť | s | veľký | pravde-podobnosť |
| 645 | trvať | udržateľný | rozvoj | 466 | byť | v | posledný | čas |
| 641 | byť | veľký | problém | 406 | byť | v | konečný | dôsledok |
| 639 | venovať | veľký | pozornosť | 400 | byť | v | minulý | rok |
| 585 | mať | veľký | vplyv | 382 | byť | na | druhý | strana |
| 557 | byť | veľký | záujem | 378 | byť | v | každý | prípad |
| 543 | nájsť | spoločný | reč | 370 | byť | do | istý | miera |
| 535 | byť | každý | deň | 370 | prejsť | na | druhý | strana |
| 535 | mať | veľký | problém | 367 | byť | do | značný | miera |
| 530 | dávať | veľký | pozor | 348 | študovať | na | vysoký | škola |
| 521 | mať | iný | názor | 333 | dosiahnuť | v | minulý | rok |
| 479 | byť | veľký | rozdiel | 320 | byť | na | vysoký | úroveň |
| 470 | byť | celý | deň | 319 | byť | v | istý | zmysel |
| 463 | byť | hlavný | dôvod | 313 | vyskočiť | na | rovný | noha |
| 459 | Dukla | banský | Bystrica | 304 | byť | v | veľký | miera |
| 459 | klásť | veľký | dôraz | 295 | byť | z | veľký | časť |
| 450 | podať | dobrý | výkon | 293 | byť | v | posledný | rok |
| 449 | mať | dobrý | nálada | 283 | skončiť | na | druhý | miesto |
| 443 | mať | plný | ruka | 282 | obvinený | z | trestný | čin |
| 434 | podať | pomocný | ruka | 270 | obžalovať | z | trestný | čin |
| 425 | byť | dlhý | rok | 260 | byť | po | celý | čas |
| 419 | dopustiť | trestný | čin | 245 | chcieť | za | každý | cena |
| 418 | mať | veľký | radosť | 241 | vstúpiť | do | európsky | únia |
| 412 | byť | hlavný | téma | 236 | byť | na | druhý | deň |
| 402 | utrpieť | ťažký | zranenie | 228 | brať | na | ľahký | váha |

Vzhľadom na daný materiál môžeme konštatovať, že so vzrastajúcim počtom komponentov kolokácie v kombinácii s vyššími hodnotami frekvenčnej distribúcie sa zmenšuje pravdepodobnosť, že daná kolokácia predstavuje náhodné spojenie. Na druhej strane frekvenčné hodnoty spoluvýskytu v bigramoch sú vyššie ako v trigramoch a tetragramoch, keďže pravdepodobnosť spoluvýskytu jednotlivých komponentov v trigramoch a tetragramoch je menšia a vzácnejšia.

**Vzťah medzi frekventovanosťou a ustálenosťou: skupina verbum – adjektívum – substantívum**

Pri porovnávaní dvoch vzoriek kolokácií sme sa zamerali na kolokácie na krajných pozíciách frekvenčnej škály: v zozname trigramov so štruktúrou V – A – S sme sledovali prvých 1000 kolokácií a konfrontovali sme ich s kolokáciami,

[1] V zozname n-gramov sa môžu v dôsledku nedokonalej automatizovanej morfologickej anotácie korpusu vyskytovať spojenia typu studiť Academica Slovaca.

ktoré sa v zozname nachádzali na pozícii 10000 až 11000. Analýzou materiálu sme sa snažili zistiť, nakoľko súvisí frekventovanosť kolokácie s jej ustálenosťou a naopak.

V prvých 1000 trigramoch typu V – A – S sme identifikovali frazémy, lexikalizované a typické spojenia, napr. *nájsť spoločnú reč, podať pomocnú ruku, mať plné zuby* (niečoho)*, neveriť vlastným očiam, mať ťažké srdce* (na niekoho)*, mať zviazané ruky, zvíťaziť zdravý rozum, naliať čistého vína, šiť horúcou ihlou, naskakovala husia koža, zachovať chladnú hlavu, položiť základný kameň.*

Zo zoznamu 1000 kolokácií sme vybrali spolu 29 ustálených spojení. Najviac spojení (spolu 18) sa nachádzalo v prvých 400 kolokáciách, v ďalších, s frekvenčnou distribúciou 110 výskytov, ich bolo podstatne menej. Môžeme konštatovať, že so znižujúcou sa frekvenciou sa znižoval aj počet nájdených ustálených spojení.

Z analýzy kolokácií nachádzajúcich sa v zozname na mieste 10000 až 11000 vyplynulo, že išlo prevažne o typické spojenia s frekvenciou 14 výskytov v korpuse: *strácať zdravý rozum, nemať šťastnú ruku, mať silné reči, ponúkať pomocnú ruku, zaujať tvrdý postoj, urobiť malý zázrak, rozpútať hotové peklo, dostať sladkú odmenu.* Našli sme však aj spojenie s preneseným frazeologickým významom *je hodená rukavica.*

Je teda zrejmé, že existuje relácia medzi frekvenciou a ustálenosťou slovného spojenia. Analýza materiálu potvrdila, že ustálené spojenia inklinujú k vysokým štatistickým hodnotám. Napriek tomu nie každé frekventované spojenie musí byť automaticky ustálené a aj medzi najmenej frekventovanými spojeniami možno rovnako nájsť ustálené slovné spojenia. Prikláňame sa k názoru, že ustálené spojenia sa môžu nachádzať na celej škále frekvenčnej distribúcie, v centre aj na frekvenčnej periférii (porov. Ďurčo, 2007).

**Analýza kolokačných mier: skupina verbum – substantívum**

Medzi štatistické metódy identifikácie kolokácií, tzv. kolokačné miery, patrí napríklad MI-score, t-score, Dice koeficient, Log-likelihood, Z-score, cost-criterion. Tieto metódy vyhodnocujú vzájomný spoluvýskyt dvoch slov. Prostredníctvom korpusového manažéra Manatee a klienta Bonito môžeme štatisticky vyhodnocovať vzájomný spoluvýskyt slov na základe MI-score a t-score.

Ako uvádza F. Čermák (2006b, s. 13), MI-score je miera vzájomnej informácie, meria silu asociácie medzi dvoma slovami, pri ktorých ide o podiel pravdepodobnosti výskytu týchto dvoch slov spolu a výskytu každého z týchto slov nezávisle. Podľa autora frekventovanejšie slová sú v tabuľke postavené nižšie ako slová menej frekventované, ktoré majú hodnotu MI-score vyššiu.

Na druhej strane t-score je test signifikantnosti založený na rozdiele medzi pozorovaným a predpokladaným výsledkom, tzv. miera kontrastu (Čermák, 2006b, s. 13). Na základe t-score sa testuje, či zistené počty výskytov jednotlivých slov a ich dvojíc zodpovedajú náhodnému rozloženiu slov v korpuse. Čím

väčšia je hodnota t-score, tým je menej pravdepodobné, že ide o náhodné rozloženie slov, a naopak, tým pravdepodobnejšie je, že ide o pevnejšie, ustálenejšie spojenie (Český národní korpus – úvod a příručka uživatele, 2000, s. 60). Tieto tvrdenia sa budeme snažiť overiť zodpovedaním otázky, čo nám kolokačné miery hovoria o danej kolokácii.

Z prvých 500 bigramov typu V – S sme vybrali 12 verbo-nominálnych lexikalizovaných spojení a 12 voľných spojení. Tieto spojenia mali absolútnu frekvenciu v korpuse do 700 výskytov. Rovnako sme porovnali vzorku deviatich voľných spojení a piatich typických spojení vybraných z 500 kolokácií nachádzajúcich sa v zozname na miestach 10000 až 10500. Vo všetkých prípadoch sme zisťovali hodnoty kolokačných mier a navzájom ich porovnávali.

Z tabuľky č. 3 je zrejmé, že ustálené spojenia majú hodnotu MI-score podstatne vyššiu ako voľné spojenia, pričom hranica ustálených spojení sa pohybovala okolo hodnoty 9. V prípade kolokácie *položiť otázku* to však tak nie je napriek tomu, že ide o veľmi typické verbo-nominálne lexikalizované spojenie.

Hodnota t-score je relatívne vysoká u oboch typov slovných spojení, okrem prípadov *mať priateľa* a *byť pacientom*, čo je pravdepodobne spôsobené širokou spájateľnosťou verb *mať* a *byť*. Viacvýznamové slová spravidla „produkujú" veľké množstvo kolokátov a platí, že čím majú viac významov, tým viac kolokácií s daným slovom tvoria, keďže každý jednotlivý význam slova viaže na seba špecifické kolokáty (porov. Čermák, 2006a, s. 930 – 931). Monosémantické a monokolokabilné slová sú z tohto hľadiska uprednostnené a majú zväčša vyššie hodnoty štatistických mier.

**Tabuľka č. 3:** Výber z prvých 500 bigramov typu V – S (do 700 výskytov v korpuse).

| ustálené a typické spojenia | MI-score | t-score | voľné spojenia | MI-score | t-score |
|---|---|---|---|---|---|
| niesť zodpovednosť | 9,494 | 39,43 | mať priateľa | 1,039 | 13,57 |
| pokrútiť hlavou | 11,08 | 72,0 | byť pacientom | 5,228 | 8,687 |
| pokrčiť plecami | 13,38 | 57,62 | povedať riaditeľovi | 3,618 | 34,93 |
| položiť otázku | 7,402 | 46,13 | informovať hovorca | 7,681 | 26,31 |
| držať palce | 10,52 | 35,98 | stať predmetom | 5,181 | 29,62 |
| utrpieť zranenie | 11,07 | 41,63 | dostať meno | 3,751 | 26,02 |
| zmerať sily | 10,4 | 36,43 | potrebovať peniaze | 5,472 | 26,73 |
| zastávať názor | 9,747 | 39,16 | prosiť pán | 5,304 | 26,37 |
| lámať hlavu | 9,209 | 31,44 | znamenať koniec | 4,191 | 23,2 |
| vyraziť dych | 11,26 | 30,96 | hľadať cestu | 4,577 | 24,9 |
| mávnuť rukou | 10,51 | 28,69 | získať cenu | 4,017 | 22,29 |
| nadviazať kontakt | 10,33 | 29,37 | nájsť spôsob | 4,493 | 28,22 |

Ako vidieť z tabuľky č. 4, voľné spojenia majú nízke hodnoty MI-score, porovnateľné s frekventovanými voľnými spojeniami v tabuľke č. 3. Na druhej strane v kolokáciách *odslúžiť omšu*, *píliť drevo*, *strpčiť život*, *snímať hriech* sa

vyskytujú verbá s úzkou kolokabilitou, ktorá im „zaručuje“ vysoké hodnoty MI-score. V spojení *píliť drevo* má slovo *píliť* obmedzenú spájateľnosť, keďže viaže na seba kolokáty istej sémantickej skupiny. Rovnako je to v prípade kolokácie *strpčiť život*, ktorej relatívna frekvencia 55,56 nám hovorí, že v priemere v každom druhom prípade sa slovo *strpčiť* vyskytuje so slovom *život*.

**Tabuľka č. 4:** Výber z 500 bigramov z miest 10000 až 10500 s frekvenciou 66 až 64 výskytov v korpuse.

| voľné spojenia | MI-score | t-score | typické spojenia | MI-score | t-score |
|---|---|---|---|---|---|
| navrhnúť vytvorenie | 5,735 | 7,726 | odslúžiť omšu (r.f.11,9) | 11,47 | 8,121 |
| tvrdiť odborník | 3,421 | 7,693 | píliť drevo (r.f. 11,45) | 11,15 | 7,996 |
| komentovať tréner | 4,634 | 9,693 | snímať hriech | 9,955 | 7,992 |
| zlepšiť hra | 4,002 | 7,501 | strpčiť život (r. f. 55,56) | 9,351 | 7,734 |
| vyjadrovať myšlienka | 4,246 | 7,812 | utrácať peniaze | 9,049 | 8,047 |
| vysvetliť fakt | 4,33 | 7,778 | | | |
| vrátiť sila | 1,016 | 4,076 | | | |
| rozhodovať dátum | 5,842 | 7,922 | | | |
| predstaviť súbor | 3,443 | 7,208 | | | |

Keď porovnáme hodnoty t-score v oboch tabuľkách, zistíme, že rozdiely sú priepastné, a to bez ohľadu na to, či ide o spojenie ustálené alebo voľné. Z toho vyplýva, že t-score nám v podstate nehovorí o sémantickej zviazanosti spojenia, ale skôr o tom, ako je spojenie frekventované vzhľadom na ostatné spojenia v korpuse. Môžeme konštatovať, že so znižujúcou sa absolútnou frekvenciou spojení sa znižuje aj hodnota ich t-score. Frekventované kolokácie sú z tohto hľadiska zvýhodnené pred menej frekventovanými, keďže vykazujú zväčša vysoké hodnoty t-score, či už sú ustálené alebo voľné.

## 4 Záver

V príspevku sme zisťovali a overovali, akú úlohu môžu zohrávať štatistické informácie pri hodnotení ustálenosti spojenia. Analýzou materiálu sme dospeli k týmto zisteniam:

- Štatistické hodnoty môžu napomôcť lepšej identifikácii ustálených spojení; pri extrakcii na základe absolútnej frekvencie sa však uprednostňujú frekventovanejšie kombinácie slov, vďaka čomu môžu byť menej frekventované spojenia prehliadané.
- Vzhľadom na zvolený spôsob selekcie a identifikácie kolokácií sme prišli k záveru, že čím väčší počet komponentov má identifikovaná kolokácia, tým je viac pravdepodobné, že nejde o náhodné voľné spojenie.
- Množstvo dát je rozhodujúce: čím väčší korpus, tým uspokojivejšie výsledky prináša.
- Monosémantické a monokolokabilné slová inklinujú k vyšším hodnotám MI-score.

- Ustálené kolokácie, ktorých komponenty sú viacvýznamové slová, čiže slová so širokou spájateľnosťou, a slová veľmi frekventované môžu byť vďaka nízkej hodnote MI-score a t-score prehliadané.

Daný spôsob extrakcie kolokácií sa osvedčil ako vhodný na skúmanie verbonominálnych spojení. Štatistické nástroje umožňujú orientovať sa vo veľkom množstve dát, ktoré dnes korpusy obsahujú, pomáhajú selektovať a verifikovať lingvistický materiál. Pre lingvistu, lexikografa či kohokoľvek, kto sa pokúša odhaliť zákonitosti fungovania jazyka a kombinatorických schopností jeho jednotiek – slov, sú tieto korpusové nástroje, ako aj samotný korpus nevyhnutným predpokladom výskumu.

**Literatúra**


ČERMÁK, František: Collocations, Collocability and Dictionary. In: Proceedings XII EURALEX International Congress, Volume II., Torino, Italia, September 6th – 9th, 2006. Ed. E. Corino – C. Marello – C. Onesti. Alessandria: Edizioni dell‘ Orso 2006a, s. 929 – 937.

ČERMÁK, František: Kolokace v lingvistice. In: Kolokace. Studie z korpusové lingvistiky. Svazek 2. Ed. F. Čermák – M. Šulc. Praha: Nakladatelství Lidové noviny – Ústav českého národního korpusu 2006b, s. 9 – 16.

Český národní korpus – úvod a příručka uživatele. Ed. J. Kocek – M. Kopřivová – K. Kučera. Praha: Ústav českého národního korpusu 2000. 160 s.

ĎURČO, Peter: Collocations in Slovak (Based on the Slovak National Corpus). In: Computer Treatment of Slavic and East European Languages. Ed. J. Levická – R. Garabík. Bratislava: Tribun 2007, s. 43 – 50.

JAROŠOVÁ, Alexandra: Problém vymedzenia kolokácií. In: Jazykovedný časopis, 2007, roč. 58, č. 2, s. 81 – 102.

JAROŠOVÁ, Alexandra: Problém vyčleňovania ustálených lexikalizovaných spojení pomocou štatistických nástrojov. In: Jazykovedný časopis, 1999, roč. 50, č. 2, s. 94 – 100.

KAČALA, Ján: K statusu lexikalizovaných spojení. In: Slovenská reč, 1997, roč. 62, č. 4, s. 193 – 203.

Kolokace. Studie z korpusové lingvistiky. Svazek 2. Ed. F. Čermák – M. Šulc. Praha: Nakladatelství Lidové noviny – Ústav českého národního korpusu 2006. 453 s.

KRALČÁK, Ľubomír: Analytické verbo-nominálne spojenia v slovenčine. Synchrónno-diachrónny pohľad. Nitra: Filozofická fakulta Univerzity Konštantína Filozofa 2005. 204 s.

MAJCHRÁKOVÁ, Daniela: Vyčleňovanie lexikalizovaných spojení pomocou štatistických nástrojov. [Diplomová práca]. Bratislava: Filozofická fakulta 2005. 63 s.