FILOZOFICKÁ FAKULTA UNIVERZITY KOMENSKÉHO
Katedra slovenského jazyka

Daniela Majchráková

Vyčleňovanie lexikalizovaných spojení pomocou štatistických nástrojov

Diplomová práca

Bratislava 2005

Čestne vyhlasujem, že som diplomovú prácu vypracovala samostatne s použitím uvedenej literatúry.

Bratislava 25. 4. 2005 **.....**

Daniela Majchráková

Ďakujem svojej diplomovej vedúcej PhDr. Márii Šimkovej za usmerňovanie, poskytnuté rady a pomoc, ako aj za trpezlivosť a venovaný čas, Mgr. Alexandre Jarošovej, CSc., za podnetné myšlienky a rady a oddeleniu Slovenského národného korpusu, predovšetkým PhDr. Márii Šimkovej a RNDr. Radovanovi Garabíkovi, že mi umožnili uskutočniť môj výskum.

Obsah

Úvod

Založením oddelenia Slovenského národného korpusu v Jazykovednom ústave Ľ. Štúra SAV v roku 2002 sa otvorili nové možnosti v lingvistickom skúmaní súčasného slovenského jazyka. Využívanie korpusu a jeho nástrojov môže

napomôcť k odhaľovaniu či potvrdzovaniu rôznych jazykových javov a umožniť preskúmanie problematických oblastí, akou je aj oblasť lexikalizovaných spojení.

Mgr. Alexandra Jarošová, CSc., vo svojej štúdii *Problém vyčleňovania ustálených lexikalizovaných spojení pomocou štatistických nástrojov* (1999) poukázala na možnosť počítačového spracovania lexikalizovaných spojení, ich automatizovaného extrahovania zo súboru textov a identifikácie pomocou korpusových nástrojov. Keďže táto myšlienka je v súčasnosti vďaka Slovenskému národnému korpusu uskutočniteľná, Mgr. Alexandra Jarošová, CSc., navrhla preskúmať takýto prístup k problematike lexikalizovaných spojení.

Pod vedením vedúcej oddelenia Slovenského národného korpusu PhDr. Márie Šimkovej a za pomoci programátora tohto oddelenia RNDr. Radovana Garabíka sme sa rozhodli v našej diplomovej práci túto myšlienku zrealizovať.

Budeme sledovať dve línie. Prvou je problematika automatizovaného vyhľadávania dvojíc a trojíc slov z korpusu, ktorým získame vhodný materiál na jazykovú analýzu; budeme skúmať výhody a nevýhody tohto prístupu. Druhou je následná analýza materiálu, v ktorej sa zameriame predovšetkým na rozbor lexikalizovaných spojení.

V úvodnej kapitole predstavíme možnosti korpusovej lingvistiky a jej rôzne štatistické nástroje slúžiace na ľahšiu identifikáciu ustálených spojení.

V ďalšej kapitole sa zameriame na problematiku lexikalizovaných spojení, predstavíme rôzne lingvistické názory, pričom budeme vychádzať najmä z prác A. Jarošovej. Osvetlíme pojem kolokácia a dáme ho do vzťahu k pojmu lexikalizované spojenie.

Tretia kapitola bude metodologickou časťou, v ktorej vymedzíme cieľ a postupnosť krokov nášho výskumu, ktoré podrobnejšie rozpracujeme v ďalších kapitolách.

V štvrtej kapitole sa zameriame na otázku výberu vhodného korpusového materiálu, do úvahy zoberieme predovšetkým otázku reprezentatívnosti, vyváženosti korpusu.

V ďalšej kapitole opíšeme postup automatizovaného spracovania korpusového materiálu, teda vyčlenenie dvojíc a trojíc tokenov a ich následnú úpravu tzv. negatívnym filtrom na vyhovujúci materiál možných ustálených spojení.

Lingvistická analýza získaného korpusového materiálu je náplňou šiestej kapitoly. V nej podáme opis „ručného“ spracovania zoznamu dvojíc a trojíc slov, teda výberu relevantných, ustálených spojení a vyčleníme termíny, frazémy a voľné spojenia, ktoré sa v zozname kolokácií vyskytli, aby sme získali zoznam možných lexikalizovaných spojení.

V ďalšej kapitole sa pokúsime urobiť analýzu vyčlenených lexikalizovaných spojení, vymedziť jednotlivé typy. Budeme sa pritom opierať o relevantné lingvistické práce analyzované v 2. kapitole.

V poslednej kapitole zhodnotíme výsledky našej práce a zamyslíme sa nad prínosmi a nedostatkami počítačového prístupu k lexikalizovaným spojeniam.

1. Možnosti korpusovej lingvistiky

„Korpusová lingvistika je zvláštna jazykovedná disciplína, ktorá systematicky pracuje s korpusom a jeho nástrojmi, študuje zásady práce s ním, aby lepšie poznala funkcie a štruktúru jazyka“ (Čermák, 2000, s. 17).

Predmetom korpusovej lingvistiky je teda korpus ako „cielene zhromaždený, vnútorne štruktúrovaný rozsiahly súbor textov, ktoré sú elektronicky uložené a spracovateľné“ (Šimková, <http://korpus.juls.savba.sk/korpus/biblioteka/pubikacie/presov1.html>).

Korpus je všeobecne považovaný za náhradu kartotečných lístkov. Rozdiel medzi kartotékou a korpusom je ten, že excerpčná kartotéka sa buduje zväčša výberovým spôsobom, obsahuje údaje len o „vybratých lexikálnych jednotkách z vybratých viet excerpovaných textov“ (Benko, 1997, s. 298). Na druhej strane jazykový korpus „sprístupňuje všetky výskyty všetkých slovných tvarov spracúvaného textu“ (tamže). Teda poskytuje „absolútny kontextový výskyt každého slova a tvaru tak, ako sa nachádza v zhromaždených textoch“ (Šimková, c. d.).

Najjednoduchší spôsob ako z korpusu získať požadované lingvistické informácie je vyhľadávanie slova alebo tvaru v kontextových použitiach pomocou niektorého zo špeciálnych konkordančných (vyhľadávacích) programov (Šimková, c. d.).

Na ilustráciu si uvedieme príklad, ako sa dajú z korpusu, v našom prípade zo SNK (verzia prim1), získať podnetné jazykové informácie pomocou vyhľadávaní konkordancií príslušných výrazov. Sledovali sme, či sa slovné spojenia *nákupná horúčka* a *vlna nákupov* nachádzajú v rovnakých kontextoch, a teda či sa používajú ako navzájom synonymné výrazy.

Korpusový manažér Manatee s klientom Bonito: ukážka konkordancie s hľadaným heslom *nákupná horúčka*:

zimnými sviatkami prepadá obyvateľstvo <nákupnej horúčke> , v obchodoch sa vytvárajú tlačenice
kvalitnom bývaní . Predávajúci využívajú <nákupnú horúčku> na trhu s nehnuteľnosťami a snažia sa
britské supermarkety sa pripravujú na <nákupnú horúčku> , pretože zákazníci sa budú chcieť
nižšími cenami v čase predvianočnej <nákupnej horúčky> . Výsledok je ten , že majú menej zákazníkov
spoločností . Dúfajme , že vianočná <nákupná horúčka> postihne aj investorov , ktorí budú chcieť
Život ma naučil nepodliehať pred Vianocami <nákupnej horúčke> . Obchody sú otvorené aj priamo na Štedrý

Konkordancie so slovným spojením *vlna nákupov*:

poklesom kurzu rubľa , sa strhla ďalšia <vlna nákupov> dolárov zo strany ruských obyvateľov devízové
pozície . Ako inak, tieto narástli po silnej <vlne nákupov> zahraničnej meny do značných rozmerov
odrazilo už počas rána , keď prišla <vlna nákupov> eura voči britskej libre . Ďalšou správou
forint . „ Dopoludnie sa nieslo v dvoch <vlnách nákupov> voľnej meny uskutočňovaných domácimi
Preto v pondelok v ranných hodinách sa <vlnou nákupov> dolár vyšplhal zo 125 , 60 na 126 , 50
posledných 10 rokov . Tento fakt vyvolal <vlnu nákupov> americkej meny oproti nemeckej marke .

Na základe uvedených výskytov spojení vidíme, že sa spojenia *nákupná horúčka* a *vlna nákupov* nachádzajú v odlišnom kontextovom okolí. Spojenie *vlna nákupov* sa používa výlučne v spojitosti s obchodovaním s menou (dokonca vo všetkých 22 výskytoch v korpuse). Môžeme sa zamyslieť, či táto viazanosť len na určitý kontext by mohla byť východiskom lexikalizácie spojenia. Pri zvažovaní, či dané spojenie je alebo nie je ustáleným, si môžeme pomôcť sledovaním rôznych štatistických hodnôt.

1.1. Štatistické nástroje korpusovej lingvistiky

Jazykový korpus predstavuje pre lingvistu veľké množstvo jazykového materiálu. Aby sa v ňom dokázal orientovať, potrebuje vhodný vyhľadávací program, ktorý mu v krátkom čase nájde všetky výskyty požadovaného slova alebo slovného spojenia.

Slovenský národný korpus umožňuje vyhľadávať v niekoľkých verziách korpusu (prim0.1, prim0.1-public, prim0.2, prim1) pomocou korpusového manažéra Manatee s klientom Bonito. Konkordančný program jednotlivé výrazy nielen zobrazí tak, ako sa v konkrétnych kontextoch nachádzajú, ale poskytne o nich aj rôzne frekvenčné a štatistické údaje.

Základná informácia, ktorú používateľ korpusu získava po zadaní hľadaného slova alebo spojenia slov, je absolútny počet výskytov požadovaného výrazu v danom korpuse. Ďalšie štatistické údaje sa týkajú frekvencie spoločného výskytu dvoch výrazov, ktoré sa nachádzajú buď v bezprostrednej blízkosti, alebo v rámci istého intervalu, kontextu. Kontext je zadaný počet slov pred alebo za hľadaným výrazom.

Hodnota **absolútnej frekvencie** v kontexte, označená $f\ (x,\ y)$, udáva počet výskytov ľubovoľného slova *y* v kontexte slova *x* (http://ucnk.ff.cuni.cz).

Relatívna frekvencia v kontexte $f_R\ (x,\ y)$ vyjadruje, koľko percent zo všetkých výskytov slova *y* sa nachádza v kontexte slova *x*, teda $f_R\ (x,\ y) = f\ (x,\ y)\ /\ f\ (x)\ .\ 100\%$ (tamže).

Ďalšie veličiny, ktoré popisujú spoločný výskyt dvoch slov, sú MI-score a T-score.

MI-score (Mutual information score), tzv. hodnota vzájomnej usúvzťažnenosti, je štatistický pojem v teórii informácie, kde sa do vzájomného pomeru dostáva pravdepodobnosť spoločného výskytu výrazov *x* a *y* s pravdepodobnosťou, ktorú dostaneme sčítaním pravdepodobnosti výskytu slova *x* s pravdepodobnosťou výskytu slova y (tamže):

$$\mathbf{I\ (x,\ y) = \log_2} \frac{P\ (x,\ y)}{P\ (x)\ P\ (y)}$$

MI-score nás upozorňuje na vzájomnú príťažlivosť dvoch výrazov, pričom vyzdvihuje slová málo frekventované (nízke hodnoty MI majú preto interpunkčné znamienka).

T-score, miera kontrastu, meria rozdiely v spájateľnosti dvoch slov. Vychádza zo štatistickej metódy testovania hypotéz pomocou tzv. t-testu. „V prípade kolokácií sa testuje, či jednotlivé počty výskytov jednotlivých slov a ich dvojíc zodpovedajú náhodnému rozloženiu slov v korpuse“ (http://ucnk.ff.cuni.cz).

Čím vyššia je hodnota T-score, tým je pravdepodobnejšie, že nejde o náhodné rozloženie slov, ale o pevnejšie, ustálenejšie spojenie.

Ilustrujme si jednotlivé štatistické hodnoty na príklade vybraných kolokácií so slovom *vlna* (vo všetkých tvaroch). Do úvahy sme brali len tie výrazy, ktoré sa nachádzali bezprostredne vedľa kľúčového slova, naľavo alebo napravo od neho.

Tab. č. 1: Kolokácie so slovom *vlna* (výsledok hľadania zápisu vln.{1,3}|vĺn) zoradené podľa absolútnej frekvencie, kontext = <-1, 1>:

	MI-score	**T-score**	**Relatívna frekvencia v %**	**Absolútna frekvencia**
.	1.577	30.5	0.0173	2104
,	1.449	27.17	0.01583	1838
privatizácie	10.89	34.67	11.03	1203
prvej	9.325	33.52	3.72	1127

novej	7.827	16.27	1.317	267
rozširovania	10.98	14.42	11.71	208
kritiky	10.12	13.63	6.452	186
násilia	9.105	12.31	3.192	152
tanečná	12.82	10.68	42.07	114
protestov	11.08	10.58	12.58	112
tlaková	10.63	9.214	9.169	85
odporu	9.416	9.042	3.961	82
zdvihla	10.46	8.994	8.141	81
záujmu	7.801	8.848	1.293	79
horúčav	11.93	8.364	22.65	70
povodňová	12.4	8.061	31.4	65

Všimnime si, ako sa jednotlivé štatistické hodnoty odrážajú na spájateľnosti dvoch lexikálnych jednotiek.

Už na prvý pohľad je zrejmé, že z absolútnej frekvencie sa ustálenosť spojenia vyvodiť nedá. Slovo *vlna* sa v korpuse najviac vyskytuje vedľa interpunkčného znamienka „bodka“. Vieme, že ustálené spojenie tvoriť nemôžu, naznačuje to aj nízka relatívna frekvencia. Vysoké hodnoty dosahuje kombinácia vlna + bodka pri koeficiente T-score, čo je však spôsobené tým, že interpunkčné znamienka sú v korpuse zvyčajne rozložené rovnomerne. Túto stránku rieši MI-score, v tomto prípade pochopiteľne veľmi nízke vzhľadom na ostatné kolokácie.

V ďalších uvedených výrazoch sú jednotlivé koeficienty pomerne vysoké. Za povšimnutie by stálo porovnanie relatívnej frekvencie pri spojeniach *vlna záujmu* a *vlna horúčav*, kde je zaujímavá veľmi nízka hodnota v prípade *záujmu* a pomerne vysoká v prípade *horúčav*. Na vyvodenie záveru by bola potrebná hlbšia analýza. Domnievame sa však, že sledovanie vysokej relatívnej frekvencie by mohlo pomôcť identifikovať slová so zníženou kolokabilitou (spájateľnosťou).

Pri výskume tohto typu si treba uvedomiť, že vysoké hodnoty v štatistických tabuľkách nezaručujú, že dané spojenie je ustáleným, môžu byť však pomocným testovacím nástrojom pri rozlíšení voľných a ustálených spojení.

2. Lexikalizované spojenie

Našou úlohou v tejto kapitole bude definovať pojem lexikalizované spojenie tak, ako nám to umožňujú výsledky doterajšieho lingvistického bádania o tejto problematike. Vymedzíme vlastnosti a jednotlivé typy lexikalizovaných spojení. Ďalej sa zameriame na pojem kolokácia a zamyslíme sa, či a ako korešponduje s pojmom lexikalizované spojenie.

2.1. Pojem lexikalizované spojenie v kontexte rôznych lingvistických prác

Lexikalizované spojenie ako špecifický jazykový jav sa všeobecne pokladá za ustálené slovné spojenie, ktoré v procese lexikalizácie nadobudlo platnosť samostatnej lexikálnej jednotky a tým sa zaradilo do slovnej zásoby daného jazyka. V súvislosti s lexikalizovaným spojením sa hovorí o „zániku syntaxe“ medzi

komponentmi spojenia (Mlacek, 1984, s. 36) alebo o „deaktualizácii syntagmatickej štruktúry“ (Dolník, 2003, s. 152), preto sa tento jav týka nielen lexikológie, ale zasahuje aj do kompetencií syntaxe a morfológie.

Pojem lexikalizované spojenie na jednej strane zahŕňa všetky slovné spojenia, ktoré sú produktom lexikalizácie, na druhej strane ho môžeme chápať ako samostatný typ ustáleného spojenia, ktorý na osi medzi termínmi a frazeologizmami zastáva pevné miesto akéhosi stredného spojovacieho článku.

Pre ďalšiu prácu s termínom lexikalizované spojenie je nutné vymedziť si základné atribúty, ktoré nám pomôžu odlíšiť jednotlivé typy ustálených spojení a predovšetkým ich postavia do opozície k voľne utvoreným spojeniam.

„Termín ustálené slovné spojenie zastrešujúci frazeologické aj nefrazeologické jednotky naznačuje, že práve ustálenosť by mala byť základnou vlastnosťou odlišujúcou tento typ od spojení voľne vytvorených v procese rečovej komunikácie“ (Jarošová, 2000a, s. 141).

Vychádzajúc z analýzy konceptu ustálenosti A. Jarošová (2000a, s. 141 – 143; 2000b, s. 487) vyčlenila tieto základné vlastnosti lexikalizovaných spojení:

- dispozičnosť, resp. reprodukovanosť,
- nominačnosť,
- kolokačná anomálnosť.

Vymenované atribúty autorka považuje za spoločné aj pre frazémy a termíny, pričom sú v jednotlivých typoch ustálených spojení zastúpené v nerovnakej miere.

Reprodukovanosť (2000a, s. 141) alebo „hotovosť frazeologickej jednotky už pred rečovým procesom“ (Mlacek, 1984, s. 35) sa viaže na fakt, že dané spojenie sa radí k systémovým jednotkám, teda sa stáva súčasťou lexikálnej zásoby a ako takáto samostatná jednotka sa opakovane používa, reprodukuje v komunikácii. S reprodukovanosťou súvisí spomínaná absencia syntaktického vzťahu medzi komponentmi spojenia, keďže spolu vytvárajú samostatný lexikálny výraz.

Nominačnosť (Jarošová, 2000a, s. 141; podľa J. Horecký: Návrh na vymedzenie frazémy. In: Frazeologické štúdie II. Bratislava 1997), predstavuje pomenovaciu funkciu ustálených spojení. Každé slovné spojenie ako pomenovacia jednotka pomenúva istú entitu alebo istý postoj k realite. J. Mlacek v súvislosti s

„nefrazeologickými združenými pomenovaniami“, ktoré sa čiastočne kryjú s pojmom lexikalizované spojenia, hovorí o ich vzniku ako o „presnej a cieľavedomej pomenovacej činnosti ľudí, ktorí príslušný výraz umelo vytvárajú pre konkrétnu potrebu“ (Mlacek, 1984, s. 40).

„Anomálna kombinatorika (kolokabilita)“ (Jarošová, 2000a, s. 141; podľa F. Čermák: Česká přirovnání. In: Slovník české frazeologie a idiomatiky. Praha 1983), dôležitý aspekt ustálenosti, znamená, že v ustálených spojeniach dochádza k obmedzenej spájateľnosti (kolokabilite), prípadne k monokolokabilite jedného z komponentov.

Medzi lexikalizovanými spojeniami nachádzame také, ktorých komponenty disponujú oslabenou spájateľnosťou až monokolokabilitou, napr. *baliaci papier*, *pitná voda*, a tiež spojenia, ktorých komponenty majú otvorenú kolokačnú paradigmu, napr. lexikalizované spojenia s kategoriálnymi slovami typu *kruhy*, *vlna*, *sila* a pod. (tamže).

Popri týchto troch atribútoch ustálených spojení A. Jarošová vymedzuje ešte jednu vlastnosť – **nedoslovnosť**, resp. **nemotivovanosť** spojenia komponentmi (tamže). „Významová nerozložiteľnosť“ (Mlacek, 1984, s. 37) je viditeľná najmä pri frazeologických jednotkách a v ich prípade úzko súvisí s princípom obraznosti (c. d., s. 39), napr. *nosiť drevo do lesa* znamená „robiť niečo nadarmo, zbytočne“. Táto vlastnosť je však viditeľná aj v prípade niektorých lexikalizovaných spojení, ktoré autorka vyčleňuje ako „lexikalizované spojenia vyznačujúce sa motivačnou nezreteľnosťou“ typu *čierna skrinka*, *čierna mágia*, *modrá knižka*, v ktorých sledujeme sémantickú transponovanosť jedného alebo obidvoch komponentov (Jarošová, 2000b, s. 489). J. Dolník tvrdí, že príklady spojení typu *sviatočný šofér*, *suchá strava*, *studená misa*, či *túlavé nohy* sú dôkazom neostrej hranice medzi lexikalizáciou a frazeologizáciou. „Stupňovanie idiomatizácie znamená stupňovanie výraznosti frazeologizácie“ (Dolník, 2003, s. 153).

S rozlíšením frazeologizmov, termínov a lexikalizovaných spojení sú značné problémy, nejestvuje definícia, ktorá by medzi nimi vytýčila presnú hranicu.

A. Jarošová (2000a, s. 145 – 146) nadviazaním na vlastnosť nominačnosť a s použitím metodologického princípu J. Dolníka (1997, s. 36) – princípu funkčnej

separácie – sa pokúsila o vytvorenie deliacej čiary medzi jednotlivými ustálenými spojeniami.

Pomocou neho vyčlenila frazeologizmy ako ustálené spojenia, ktorých funkčným určením „je pomenúvať bežné situácie citovo hodnotiacim spôsobom“ (Jarošová, 2000a, s. 145).

Funkčným určením termínov je potom nocionálne pomenúvať pojmovo jednoznačne spracované triedy predmetov, javov a príznakov; pričom pomenovania sú podriadené definičnej vlastnosti termínu (tamže).

A napokon funkčným určením lexikalizovaných spojení je „primerane pomenovať špecifikovanú alebo zovšeobecnenú jednotlivinu“ (tamže). Zo spomínaných atribútov je preň teda príznačná najmä nominačnosť ako najvýraznejšia vlastnosť. „Lexikalizované spojenia pomenúvajú primerane komplexným spôsobom všeobecne známe, konvenčne vymedzené pojmy“ (Jarošová, 2000b, s. 490).

2.1.1. Delenie lexikalizovaných spojení

J. Mlacek (1984, s. 40) rozlišoval dva základné typy združených pomenovaní nefrazeologickej povahy, a to na základe opozície obraznosť – neobraznosť.

Medzi obrazné združené pomenovania zaradil spojenia *labutia pieseň*, *gordický uzol*, *trójsky kôň*, *tuhý tabak*, *kameň úrazu*, *vlčí mak*; medzi združené pomenovania s nepreneseným významom podľa neho patria napr. *podstatné meno*, *osobný vlak*, *kysličník uhličitý*. Treba dodať, že toto delenie zahŕňa tak lexikalizované spojenia, ako aj termíny.

Nefrazeologické združené pomenovania sa podľa J. Mlaceka môžu univerbizovať a meniť na jednoslovné pomenovania, napr. *obývacia izba – obývačka*, *panelový dom – panelák*. Autor tvrdí, že práve podliehanie univerbizácii vytvára deliacu čiaru medzi frazeologickými jednotkami a ich nefrazeologickými náprotivkami.

Základné delenie lexikalizovaných spojení ako „neterminologických viacslovných pomenovaní bez prvku obraznosti“ podľa J. Kačalu (1997b, s. 193 – 194) vyplýva z ich slovesnej alebo substantívnej povahy. Za podstatu lexikalizovaného spojenia považuje spojenie kategoriálneho slova

1. substantíva s prídavným menom v úlohe zhodného prívlastku alebo s podstatným menom v úlohe nezhodného prívlastku, napr. *kvalifikovaná sila*, *čitateľská verejnosť*, *proces likvidácie*,
2. slovesa s podstatným menom ako predmetu alebo s príslovkou v úlohe príslovkového určenia, napr. *dať príležitosť*, *prijať rozhodnutie*, *vzbudzovať podozrenie*, *dať do poriadku*.

A. Jarošová (2000b, s. 485) vyčlenila tieto základné typy lexikalizovaných spojení:

1. menné spojenia so spresňujúcim prívlastkom, napr. *pitná voda*, *hladká múka*, *zubná kefka*, *písací stôl*, *detská izba*,
2. menné spojenia s kategoriálnym slovom, napr. *diplomatické kruhy*, *vodné dielo*, *sekretárske sily*,
3. verbonominálne spojenia s kategoriálnym slovom, napr. *poskytnúť pomoc*, *získať dôveru*, *prejaviť záujem*, *spáchať zločin*,
4. predložkovo adjektívne spojenia, napr. *za mlada*, *od mala*, *za horúca*, *na bielo*, *za živa*, *za slobodna*,
5. predložkovo substantívne spojenia, napr. *na počesť*, *na úkor*, *v protiklade k*, *vzhľadom na*, *bez ohľadu na*.

Ďalšie delenie, ktoré ponúka A. Jarošová (2000b, s. 487 – 489), sa týka otázky motivovanosti, teda sémantickej transponovanosti komponentov spojenia. Na základe toho autorka rozlišuje lexikalizované spojenia

1. vyznačujúce sa motivačnou zreteľnosťou, napr. *slovenský jazyk*, *sprchovací kút*, *platobná karta*, *trhací kalendár*,
2. spojenia s čiastočnou motivačnou zreteľnosťou, napr. *ľadový čaj*, *burza práce*, *pasívny fajčiar*, *prvá pomoc*,
3. spojenia s motivačnou nezreteľnosťou, napr. *francúzska posteľ*, *čierna skrinka*, *čierna kronika*, *modrá knižka*.

2.2. Lexikalizované spojenie alebo kolokácia?

V korpusovej lingvistike sa v súvislosti s ustálenými spojeniami používa termín **kolokácia** (ang. collocation). Za kolokáciu sa považuje skupina významovo príbuzných slov, ktoré sa v texte nachádzajú veľmi často alebo až typicky spoločne, prevažne v rovnakom kontexte (Pecina – Holub, 2002, s. 3).

Kolokácie predstavujú kombinácie slov, ktoré sú na základe frekvencie ich výskytu v istých textoch vyčleňované a zoraďované pomocou rozličných nástrojov korpusovej lingvistiky.

Pojem kolokácia môžeme chápať v širšom alebo v užšom zmysle. Buď ako akékoľvek spojenie dvoch alebo viacerých slov s výrazne frekventovaným spoluvýskytom, pričom môže ísť o spojenia lexikálne nerelevantné, alebo výlučne ako spojenia, ktoré sa nielen najčastejšie opakujú, ale sú aj sémanticky významné.

Práve v súvislosti s druhým typom P. Pecina a M. Holub (2002, s. 6) vo svojej výskumnej správe hovoria o **„sémanticky signifikantnej kolokácii“**, teda

A o slovnom spojení vyjadrujúcom istý význam, ktorý typicky býva vyjadrovaný práve týmto spôsobom a iným je často vyjadriteľný len ťažko alebo vôbec, alebo

B o slovnom spojení, ktoré obsahuje sémanticky súvisiace slová.

V prípade variantu A autori hovoria o kolokáciách prvého druhu. Ako príklady uvádzajú idiomatické frázy, vlastné mená, technické pojmy typu *natáhnout bačkorky* (vo význame zomrieť), *přijít k sobě*, *chodit kolem horké kaše*, *trestní rejstřík*, *Sluneční soustava*, *Josef Novák*, *visutá lanovka*, *tlustá kniha*, *bílé víno*, *zbraně hromadného ničení*, *desetinná čárka*, *udělat rozhodnutí*, *nová verze*, *dětský lékař* (c. d., s. 3).

Kolokácia druhého druhu (variant B) je založená len na sémantickej príbuznosti, podobnosti či súvislosti jej komponentov.

Z lingvistického hľadiska sú pre nás relevantné len tie kolokácie, ktoré vystupujú ako samostatné lexikálne a sémantické jednotky. Pri definovaní slovných spojení si nevystačíme len so sémantickou príbuznosťou, preto je nutné, aby sme zobrali do úvahy aj ďalšie kritériá.

Z uvedených príkladov variantu A vidíme, že jednotlivé kolokácie by sme mohli zaradiť medzi ustálené slovné spojenia (okrem vlastných mien a niektorých voľných spojení, ktoré sa tam vyskytli), potrebujeme však poznať vlastnosti, ktoré autori P.

Pecina a M. Holub týmto kolokáciám priraďujú, a porovnať ich s charakteristikami ustálených spojení v časti 2.1.

Kolokácie majú podľa citovaných autorov nasledujúce charakteristické vlastnosti, pričom v jednotlivých kolokáciách nie sú zastúpené rovnako (c. d., s. 7 – 8):

1. **nekompozičnosť** – presný a jednoznačný význam kolokácie sa nedá odvodiť z jednotlivých významov jej komponentov,
2. **nesubstituovateľnosť** – komponenty kolokácií nemožno nahrádzať synonymnými výrazmi,
3. **nemodifikovateľnosť** – kolokácie nemožno modifikovať, nemožno uberať alebo pridávať nejaký ďalší lexikálny prvok do kolokácie,
4. **vnútorná štruktúra** – pri kolokáciách prvého druhu existuje medzi jej komponentmi syntaktická závislosť, v dvojslovných kolokáciách sa rozlišuje riadiace slovo a závislé slovo,
5. **preklad do iných jazykov** – častým znakom kolokácií je, že nie je možný ich doslovný preklad do cudzích jazykov,
6. **kolokačný kontext** – kolokačný kontext je kontext slova, v ktorom sa môžu nachádzať slová, ktoré s ním tvoria kolokáciu. Komponenty kolokácie nemusia byť len slová v tesnej blízkosti, ale môžu sa vyskytovať spolu v rámci istého intervalu.

Z hľadiska ustálených spojení sú pre nás dôležité vlastnosti 1. – 4., dokonca by sme ich pri posudzovaní jednotlivých spojení mohli zohľadniť. Vlastnosť „nekompozičnosť“ zjavne korešponduje s „nedoslovnosťou, resp. nemotivovanosťou“ (porov. 2.1.); podnetné sú aj atribúty „nesubstituovateľnosť“ a „nemodifikovateľnosť“. Vlastnosť „preklad do iných jazykov“ zdôrazňuje napr. aj J. Mlacek (1984, s. 37) v súvislosti s frazémami, my sa pozrieme aj na jej druhú stránku – pokúsime sa túto vlastnosť využiť pri hľadaní a porovnávaní inojazyčných ekvivalentov k ustáleným spojeniam na posudzovanie miery ich ustálenosti.

Vlastnosti, ktoré sú z hľadiska ustálených slovných spojení alebo konkrétne lexikalizovaných spojení neadekvátne, sa týkajú vnútornej štruktúry a kolokačného kontextu kolokácií. Autori tvrdia, že medzi komponentmi spojenia existujú syntaktické vzťahy, čo by znamenalo, že kolokácie sú „rozložiteľné“, a teda voľné.

Rovnako sledovať kolokačný kontext by bolo relevantné len pri voľných spojeniach. Táto vlastnosť je navyše v rozpore s vlastnosťou „nemodifikovateľnosť“, čo však súvisí s tým, že pojem kolokácia je evidentne širší a zahŕňa aj spomínané kolokácie variantu B.

Priblížme si ešte jedno pomerne podrobné delenie kolokácií, ako nám ho ponúka F. Čermák (http://ucnk.ff.cuni.cz).

Kolokácie v širšom zmysle nazýva lexikálnymi kombináciami, ktoré klasifikuje na základe dištinkcie systém – text (ustálenosť – neustálenosť) a pravidelný – nepravidelný (závisí od „rozdielu medzi gramatickými a sémantickými pravidlami“).

Delenie kolokácií podľa F. Čermáka (tamže):

A Systémové

1. pravidelné:

a) termínové kolokácie (viacslovné termíny): *cestovná kancelária*, *lodná doprava*, *kyselina sírová*,

b) propriálne kolokácie (viacslovné propriá): *Kanárske ostrovy*, *Stredozemné more*, *Veľká Británia*,

2. nepravidelné:

idiomatické kolokácie (idiómy a frazémy): *ležať ladom*, *len aby*, *bez okolkov*, *staré dobré Anglicko*.

B Textové

3. pravidelné:

a) bežné kolokácie (gramaticko-sémantické kombinácie): *letná dovolenka*, *ľahká odpoveď*, *teplé podnebie*, *staroveké pamiatky*, *koloniálna nadvláda*,

b) analytické kombinácie tvarov (analytické formy): *šiel by*, *bol zapísaný*, *spomenul si*,

4. nepravidelné:

a) individuálne metaforické kolokácie (autorské metafory): *treskúco vtipný*,

b) náhodné kombinácie susedné: *že v*, *ktorý určite*,

c) iné kombinácie (ktoré sa nedajú zaradiť do žiadnej skupiny).

C Textovo-systémové

bežné uzuálne kolokácie: *prať bielizeň*, *žehliť bielizeň*, *sušiť bielizeň*.

Na základe tohto delenia vidíme, že kolokácie sú naozaj širokým pojmom a zahŕňajú tak spojenia ustálené, ako aj náhodné kombinácie slov. Terminologicky je delenie ustálených spojení (systémových kolokácií) niečím nové, chýba v ňom zmienka o lexikalizovaných spojeniach, ktoré autor zaraďuje medzi periférium frazém a do istej miery ich stotožňuje s pojmom kvázifrazéma.

V našej práci za kolokáciu budeme považovať spojenia dvoch alebo viacerých slov nachádzajúcich sa v texte bezprostredne vedľa seba, pre ktoré je charakteristická istá miera ustálenosti, či už v súvislosti so spomínanými vlastnosťami 1. – 4. podľa autorov P. Pecina – M. Holub alebo s ich častým výskytom. Pojem kolokácie nebudeme používať len v spojitosti s lexikalizovanými spojeniami, ale aj s termínmi, frazeologizmami a tiež s frekventovanými voľnými spojeniami.

3. Vyčleňovanie lexikalizovaných spojení pomocou štatistických nástrojov

3.1. Stanovenie cieľa

Našou úlohou bude využiť možnosti, ktoré poskytuje korpusová lingvistika, na získanie čo najvhodnejšieho lingvistického materiálu na skúmanie lexikalizovaných spojení.

Pokúsime sa na vzorke vybraných textov a s využitím automatizovaného vyhľadávania najfrekventovanejších kolokujúcich bigramov a trigramov urobiť analýzu lexikalizovaných spojení. Využijeme rozličné nástroje korpusovej lingvistiky, napr. vyhľadávanie v Bonite, zamyslíme sa nad využitím štatistických nástrojov vyhodnocujúcich frekvenciu spoločného výskytu dvoch lexikálnych jednotiek. Opierať sa však budeme predovšetkým o doterajšie lingvistické poznatky o tejto problematike.

3.2. Výskumné metódy

Postupnosť našich krokov môžeme rozdeliť do niekoľkých fáz:

Výber vhodného jazykového materiálu
Výber vstupných textov potrebných na jazykovú analýzu.

Automatizovaná extrakcia bigramov a trigramov
Získanie reťazcov najfrekventovanejších dvojíc a trojíc slov nachádzajúcich sa v textoch pomocou štatistických nástrojov.

Negatívny filter

Predspracovanie získaného „surového“ materiálu, odstránenie nežiaducich tokenov, prípadne zmena tokenov v pôvodnom, kontextovom tvare na lemmy.

„Ručná“ selekcia kolokácií

Triedenie kolokujúcich bigramov a trigramov ľudskou silou, subjektívne a intuitívne.

Hodnotenie získaných kolokácií

Zaradenie kolokácií pomocou metódy komparácie a s využitím teoretických poznatkov a metód korpusovej lingvistiky do jednotlivých typov ustálených spojení.

Vyčlenenie lexikalizovaných spojení a ich analýza

Pokus o vytvorenie formálnych kritérií na odlíšenie lexikalizovaných spojení od voľne utvorených spojení, od termínov a frazeologizmov.

Zhodnotenie výsledkov

4. Výber jazykového materiálu

Slovenský národný korpus (ďalej SNK) vo verzii prim1, v čase prípravy našej práce v rozsahu takmer 190 miliónov tokenov, poskytuje základný materiál na výskum súčasného slovenského jazyka. Tak ako všetky národné korpusy aj SNK sleduje jeden cieľ: dosiahnuť vyváženosť korpusu, teda rovnomerné zastúpenie textov rozličných štýlov a žánrov, aby výsledky lingvistických analýz neboli skreslené. Korpus by mal byť reprezentatívnou vzorkou prirodzeného jazyka.

Kritérium reprezentatívnosti sme zohľadňovali aj my pri výbere vhodného jazykového (korpusového) materiálu.

Otázka reprezentatívnosti sa týka predovšetkým dvoch stránok korpusu:

– veľkosti a

– štýlovo-žánrového rozvrstvenia.

Zdá sa, že čím je korpus väčší, tým viac jazykových informácií nám môže poskytnúť. Na druhej strane pracovať s veľkým množstvom materiálu je prácne a zdĺhavé.

Pri rozhodovaní sa, aký veľký korpus vybrať pre naše ciele, sme zobrali do úvahy poznatok, že pri sledovaní frekventovaných javov postačí aj „menší rozsah dát, obsahujúci primerané množstvo bežných jazykových prostriedkov“ (Šimková, 2004, s. 208).

Keďže ustálené spojenia sú „frekventované“, reprodukované jednotky, viazané na časté používanie, z hľadiska reprezentatívnosti je pre nás vyhovujúci aj menší korpus (menší napríklad ako súčasný prim1).

Aby bol korpus reprezentatívny, musí byť tiež vyvážený, rovnomerne „namiešaný“, to znamená, že texty v ňom obsiahnuté by mali pochádzať z rôznych zdrojov a mali by mapovať rozličné sféry verejnej komunikácie.

Ak hovoríme o vyváženosti, tak predovšetkým o štylistickej, pretože „viac rôznorodosti v typoch textov sa premietne do širšej reprezentatívnosti typov jazykových javov, pričom vzorky textov musia byť dostatočne dlhé, aby dokázali distribúciu jazykových javov spoľahlivo reprezentovať“ (Šimková, 2004, s. 210;

podľa D. Bibera: Reprezentativnost v projektu korpusu. In: Studie z korpusové lingvistiky. Praha 2000).

Z tohto hľadiska bude pre nás ideálnym zdrojom materiálu vyvážený korpus Slovenského národného korpusu (prim-vyv).

Vyvážený korpus môžeme definovať ako podkorpus (subcorpus), ktorý je podmnožinou, statickou časťou väčšieho korpusového celku (Benko, 1997, s. 297).

Prvá pracovná verzia vyváženého korpusu SNK obsahuje 12 miliónov tokenov. Token je arbitrárna jednotka textu, je to akýkoľvek reťazec znakov medzi dvoma medzerami (slová, písmená, číslice) a zahŕňa aj jednotlivé znaky interpunkcie (http://korpus.juls.savba.sk/publikacie/Tagset-aktualny.pdf).

Vyvážený korpus SNK vznikol náhodným výberom textov publicistického a umeleckého štýlu, ktorý vykonal počítač, pričom odborná literatúra sa použila v plnom rozsahu.

Percentuálne zastúpenie textov vo vyváženom korpuse SNK:

60 % publicistická literatúra,

20 % umelecká literatúra,

20 % odborná literatúra.

5. Predspracovanie korpusového materiálu

V týchto dvoch fázach pracovného postupu uplatníme výlučne automatizované predspracovanie textového materiálu. Aby sme z 12-miliónového korpusu dokázali vyčleniť ustálené, resp. lexikalizované spojenia, potrebujeme počítačový program, ktorý nám v priebehu niekoľkých minút dokáže kombinácie slov zobraziť. Túto stránku našej práce zabezpečil programátor SNK RNDr. Radovan Garabík, ktorý všetky naše požiadavky pre počítač realizoval.

V predchádzajúcej kapitole sme hovorili, že SNK ako súbor jazykových dát je podrobený segmentácii, čiže tokenizácii, na základe ktorej za jednotku textu považujeme každý znak alebo skupinu znakov medzi dvoma medzerami a interpunkciu. Preto ak použijeme formuláciu „extrakcia bigramov, resp. trigramov“, máme na mysli vyčlenenie dvojíc (trojíc) tokenov stojacich tesne vedľa seba, teda nielen slov, ale aj značiek, bodiek, čiarok, číslic a podobne.

Z toho vyplýva, že program vyznačí veľké množstvo spojení, ktoré sú z nášho pohľadu nerelevantné. Veľa dvojíc a trojíc tokenov obsahuje interpunkciu a „spojenia“ s predložkami alebo spojkami vzhľadom na to, že spomedzi desiatich najfrekventovanejších tokenov v 12-miliónovom vyváženom korpuse SNK majú dominantné postavenie bodka a čiarka, za nimi nasledujú výrazy *a*, *sa*, *v*, *na* a ďalšie interpunkčné znamienka. Potrebujeme preto upraviť naše požiadavky pre počítač tak, aby sme získali čo najprehľadnejší jazykový materiál zbavený týchto nerelevantných dvojíc.

Pri predspracovaní korpusového materiálu sme postupovali nasledovnými krokmi.

5.1. Automatizovaná extrakcia bigramov a trigramov

Naša požiadavka, ktorú sme zadali počítaču, bola vytvoriť reťazec dvojíc a trojíc tokenov nachádzajúcich sa v korpuse bezprostredne vedľa seba, zoradených podľa frekvencie od kombinácií s najväčším výskytom až po kombinácie s jedným výskytom.

V korpuse, ktorý sme použili na vyčlenenie bigramov, neboli upravované nijaké z nastavení, to znamená, že tokeny ostali v pôvodných tvaroch s rozlíšením malých a veľkých písmen. Trigramy sa skladali len z malých písmen.

Pri zvažovaní ako upraviť získaný materiál ešte pred selekciou sme sa zamýšľali nad možnosťou zmeniť tokeny v kontextových, pôvodných tvaroch na lemmy, „slovníkové“ tvary tokenov (http://korpus.juls.savba.sk/publikacie/Tagset-aktualny.pdf). Takáto voľba by predstavovala niekoľko výhod aj nevýhod. Na jednej strane by bol získaný reťazec podstatne menší a najmä pri menných spojeniach by sa ukázala ich skutočná frekvenčná hodnota (nebola by rozdelená medzi jednotlivé tvary daného spojenia). Na druhej strane niektoré kolokácie zložené z tokenov v základnom tvare by boli ťažšie identifikovateľné, najmä v prípade trigramov by sa materiál značne zneprehľadnil.

Dôvod, pre ktorý sme pôvodné tokeny neupravovali na lemmy, bol aj ten, že nám to umožňuje každú kolokáciu posudzovať a hodnotiť v tvare, v akom sa v korpuse skutočne nachádza, teda v „prirodzenej kontextovej podobe“ (Čermák, 2000, s. 15).

Vytvorený zoznam bigramov a trigramov bol podľa našich očakávaní veľmi neprehľadný a obsahoval veľa pre nás nepotrebných spojení.

Ukážka reťazca bigramov (vľavo sa nachádza frekvenčná hodnota absolútneho výskytu v prim-vyv) :

že som	1908
tomu ,	1899
viac ako	1898
USA -	1898
0)	1891
. Nie	1882
. Do	1868
a jeho	1867
v tejto	1864
s tým	1863
. 25	1856
to je	1831
, potom	1802
, 18	1795
, keby	1793
. Pre	1791
, 20	1780

: 4 1778
. Tento 1774
mil . 1768
Sk . 1767
, 17 1754
1) 1750
. Čo 1748
, 7 1745
by si 1740
že si 1731
, prečo 1729

Rozsah zoznamu bigramov predstavoval zhruba štyri milióny dvojíc, trigramov bolo asi tri milióny. V prvých šesťsto prípadoch sa nachádzali prevažne spojenia „interpunkčné znamienko + slovo“, najmä „interpunkčné znamienko + spojka“ (najvyššiu frekvenciu v bigramoch dosiahlo spojenie „čiarka + že“). Pri prehliadaní bigramov a trigramov sme narazili na ďalšie tokeny, ktoré bolo vhodné na lepšiu orientáciu v reťazci odstrániť (v hojnom počte sa tam vyskytovali číslovky a rôzne skratky). Preto sme pristúpili k druhému kroku.

5.2. Negatívny filter

Negatívny filter znamená, že zo zoznamu tokenov nachádzajúcich sa v texte vylúčime tie, ktoré sú z nášho hľadiska nerelevantné. Pritom sme postupovali tak, že sme si ho zvolili zvlášť pre bigramy a zvlášť pre trigramy.

Negatívny filter pre bigramy:

1. všetky znaky okrem písmen: číslice, interpunkcia, špeciálne značky,
2. tvary *by*, *sa*, *si*, *nie*,
3. všetky prvotné predložky a aj vokalizované tvary: *s*, *so*, *z*, *zo*, *na*, *v*, *vo*, *po*, *do*, *u*, *o*, *k*, *ku*, *bez*, *cez*, *pri*, *pre*, *pred*, *predo*, *nad*, *nado*, *od*, *odo*, *pod*, *podo*, *za*, *proti*, *medzi*, *okrem*,
4. rôzne skratky: *USA*, *Sk*, *FC*, *PC*, *USD*, *ČR*, *NATO*, *SME*, *TASR*, *ČTK*, *SR*, *EÚ*.

Vylúčením kondicionálovej morfémy *by*, zvratného *sa* a *si* sme zo zoznamu odstránili analytické tvary slovies, ktoré sú z hľadiska nášho pozorovania kolokácií

nepodstatné. Za vyradenie tokenov *sa* a *si* hovoril aj fakt, že v zozname najfrekventovanejším tokenov vyváženého korpusu sa lexéma *sa* nachádzala na poprednom 4. mieste a *si* na 19. mieste. Treba však pripomenúť homonymiu tvarov *sa* a *si*; keďže *sa* je aj skráteným tvarom od zámena *seba* a *si* tvarom sponového slovesa *byť*, v reťazci bigramov nám tak budú chýbať spojenia typu *vidím sa*, *si doma*, *bol si* a pod. S časticou *nie* je situácia podobná. Spojenia, v ktorých sa môže nachádzať, sú napr. *nie je*, *povedať nie*, *nie práve*, *teraz nie* a pod. Nie je však vylúčené, že jej odstránením z nášho zoznamu prídeme o možné ustálené spojenia. Za nepotrebné sme tiež považovali zaradenie spojok do negatívneho filtra vzhľadom na to, že odstránením interpunkčných znamienok, resp. čiarok sa ich podstatná časť v analyzovanom materiáli nevyskytne.

Vylúčenie základných predložiek zoznam bigramov takisto podstatne zúži (predložka *v* je na 5. mieste a *na* obsadilo 6. miesto medzi najfrekventovanejšími tokenmi), prídeme však o predložkovo substantívne a predložkovo adjektívne spojenia typu *na úkor*, *na počesť*, *bez problémov*, *na bielo*, *za mlada*, *za horúca* (porovnaj 2.1.).

Okrem toho, rozlišovanie veľkých a malých písmen spôsobí, že predložky, ktoré sa v korpuse nachádzajú na začiatku vety, sa napriek negatívnemu filtru v reťazci bigramov objavia.

Negatívny filter pre trigramy:

1. všetky znaky okrem písmen,
2. skratky: usa, sk, fc, pc, usd, čr, nato...

Výber skratiek v negatívnom filtri sa ukázal ako správny, keďže sa už v analyzovanom materiáli nevyskytovali. Zachovanie veľkých a malých písmen v prípade bigramov sa spočiatku javilo ako chybný krok, neskôr sme objavili aj klady takéhoto rozhodnutia.

Ukážka upraveného reťazca trigramov:

v súvislosti s 978
že je to 819
v tomto roku 788
v porovnaní s 722
a ja som 667
v tomto prípade 535

bez ohľadu na 495
v tom čase 476
v tejto súvislosti 469
a tak sa 466
pokiaľ ide o 436
po prvý raz 434
v banskej bystrici 406
že ide o 405
v týchto dňoch 394
v ktorom sa 389
v súlade s 385
keď som sa 382
a potom sa 371
a to je 349
v minulom roku 346
podľa jeho slov 341
a keď sa 340
v new yorku 329
v tej chvíli 324
v prvom rade 316
v poslednom čase 305

Takto získaný materiál nám môže ďalej slúžiť na jazykovú analýzu. Pre porovnanie, z nevytriedeného reťazca bigramov v rozsahu asi štyroch miliónov dvojíc sme použitím negatívneho filtra dostali zhruba 571 tisíc bigramov a z vyše troch miliónov trigramov asi 760 tisíc trojíc.

6. Analýza korpusového materiálu

V tejto kapitole opíšeme ďalšie dve fázy pracovného postupu. Od automatizovaného spracovania korpusového materiálu prejdeme k lingvistickému – pokúsime sa z počítačom vytriedených bigramov a trigramov vyselektovať

kolokácie, ktoré sú z hľadiska nášho skúmania relevantné, a následne ich zatriediť do jednotlivých typov ustálených spojení.

6.1. „Ručná“ selekcia kolokácií

Predspracovaním korpusového materiálu sme získali pomerne rozsiahly zoznam bigramov a trigramov a našou úlohou bolo urobiť jeho opätovnú selekciu, ktorá už nebude vykonávaná počítačom, ale výlučne manuálne. Z toho dôvodu bolo triedenie kolokácií prácnejšie a časovo náročnejšie a na rozdiel od automatizovanej selekcie založené na čisto subjektívnom výbere.

Do osobitného textového súboru sme preniesli spojenia, ktoré sme považovali za ustálené, lexikalizované. Kritérium, ktoré sme si zadali pre vyčlenenie kolokácií bolo: Vybrať spojenia, ktoré sa vyznačujú istou mierou ustálenosti. To znamená, že do úvahy prichádzajú aj spojenia, ktoré by sme pri ďalšej analýze mohli zaradiť medzi voľné. Ide napríklad o spojenia, ktoré F. Čermák nazýva „bežnými uzuálnymi kolokáciami“ (2.2.), teda na hranici medzi ustálenými a voľnými spojeniami. Cieľom bolo získať čo najširšie spektrum kombinácií slov, ktoré spadajú pod označenie ustálené či lexikalizované (v širšom zmysle) spojenie.

Nami vykonaný výber kolokácií sa neopieral o žiadnu teóriu, do značnej miery bol subjektívny, a teda aj nedokonalý. Nie je vylúčené, že pri takom dlhom reťazci sme uprednostnili len isté typy spojení a iné prehliadli.

Samotná selekcia kolokácií bola napriek predchádzajúcej filtrácii jazykového materiálu pomerne náročná, orientácia v zozname bigramov a trigramov vyžadovala značnú mieru sústredenosti, materiál stále obsahoval veľa pre nás nepodstatných kombinácií slov.

V bigramoch sa vedľa seba nachádzali náhodné spojenia typu *že sú*, *divákmi rozhodoval*, *naraz je*, *ale aj*, *zamyslieť prečo*, voľné spojenia *niekoľko rokov*, *budúci týždeň*, *mladých ľudí*, *moju hlavu*, i kolokácie typu *dobrú noc*, *kupónová privatizácia*, *majstrovstvá sveta*, *hlava štátu*, pri ktorých sme zvažovali, či ich zaradiť medzi ustálené spojenia. Z bigramov sme vyčlenili aj také dvojice slov, ktoré sú (pravdepodobne) súčasťou troj- a viacslovného ustáleného spojenia, napr.

kolokácia *špinavých peňazí* sa bude posudzovať v trojslovnom spojení „*pranie špinavých peňazí*“, alebo kolokácia *spoločnú reč* v spojení *nájsť spoločnú reč*. Pri trigramoch sme postupovali rovnako.

Z vyše 571 tisíc bigramov, čo predstavovalo ešte stále veľmi rozsiahly zoznam kombinácií, bolo zhruba 200 tisíc kombinácií s jedným výskytom v korpuse, 200 tisíc s dvoma výskytmi a aj pri frekvencii šesť výskytov to pre nás predstavovalo neprehľadný materiál. Pri analýze bigramov sme prišli po pozíciu 22 tisíc (po frekvenciu 17) a vyčlenili 650 kolokácií, na ktorých sme ešte vykonali úpravu (pozn. D. M.: Konečný zoznam vytriedených bigramov pozri v časti Príloha.). Pritom platilo, že so znižujúcou sa frekvenciou počet ustálených spojení narastal.

Naopak, v trigramoch sme vyčlenili z 6 tisíc trojíc tokenov len 64 spojení, prišli sme po frekvenciu 14 (pozri Príloha). Znamenalo to, že s nižšou frekvenciou klesal aj počet nájdených ustálených spojení (pri frekvencii 13 už boli zriedkavé).

Z týchto poznatkov vyplýva, že veľkosť korpusu je dostačujúca na hľadanie ustálených dvojíc slov, ale pre trojslovné spojenia je potrebný väčší korpus.

Práca s reťazcom bigramov poukázala na nedostatky predchádzajúcej fázy spracovania materiálu. Potvrdilo sa, že rozlišovanie veľkých a malých písmen spôsobuje, že niektoré spojenia sa v rovnakých tvaroch nachádzali v reťazci dvakrát, podľa toho, či v korpuse stáli na začiatku (s veľkým písmenom na začiatku) alebo v strede či na konci vety, napríklad v tabuľke najfrekventovanejších bigramov sa na prvom mieste nachádzalo spojenie „ *je to*“ a na druhej pozícii bigram „*Je to*“. Zachovávanie veľkých písmen malo aj svoju výhodu najmä pri identifikácii proprií alebo spojení, ktoré figurovali na začiatku vety vo funkcii uvádzacích častíc.

6.2. Hodnotenie kolokácií

V tejto fáze prejdeme k lingvistickému rozboru získaného materiálu. Vytriedený zoznam kolokácií sa na prvý pohľad môže javiť ako značne rovnorodý materiál, či už po stránke gramatickej (ide najmä o adjektívno substantívne spojenia), alebo z hľadiska príslušnosti k istému štýlu (najpočetnejší je výskyt termínov pochádzajúcich z publicistických textov, čo je spôsobené dominantným zastúpením

publicistickej literatúry vo vyváženom korpuse). Do istej miery sa pod tento stav mohla podpísať aj subjektívnosť výberu kolokácií.

Ukážeme však, že aj zdanlivo nevyvážený materiál môže poskytnúť podnetné lingvistické informácie a potvrdiť alebo objaviť niektoré zaujímavé jazykové javy.

V nasledujúcich podkapitolách rozoberieme vzťah jednotlivých vybraných kolokácií, konkrétne vzťah lexikalizovaných spojení k termínom, frazeologizmom a voľným spojeniam.

6.2.1. Viacslovné termíny a lexikalizované spojenia

Vzhľadom na to, že texty vo vyváženom korpuse boli v prevažnej miere publicistického charakteru, termíny, ktoré sa v zozname kolokácií vyskytli, pochádzajú najčastejšie z tohto zdroja. Ide predovšetkým o termíny z oblasti práva a ekonomiky typu *právna norma*, *kapitálový trh*, *cenné papiere* početne zastúpené práve v publicistike napriek tomu, že ich vlastná štýlová sféra je odborná literatúra.

Tento jav korešponduje s procesom nazývaným terminologizácia, pod ktorým môžeme rozumieť „prenikanie odborných termínov do textov, najmä popularizačných“ (Horecký – Buzássyová – Bosák, 1989, s. 260).

J. Dolník (2003, s. 173 – 174) v tejto súvislosti hovorí o štýlovej transpozícii. „Štýlová transpozícia je prenášanie lexikálnej jednotky z vlastnej štýlovej domény do nevlastnej štýlovej domény, pričom toto prenášanie je zámerné a dosahuje sa ním osobitný komunikačný efekt“ (Dolník, 2003, s. 173).

V našom prípade ide o výskyt viacslovných termínov v publicistických textoch, čo môžeme označiť ako príznakové použitie daných lexikálnych jednotiek (porov. c. d., s. 173), keďže ich vlastnou štýlovou doménou sú odborné texty.

Vysoký výskyt termínov typu *jadrové zbrane*, *kupónová privatizácia*, *národnostná menšina*, *zdravotné stredisko* mimo odborných textov sa na druhej strane dotýka procesu determinologizácie, „ktorým sa pôvodne úzko odborný termín, s presne vymedzeným významom a miestom v istom systéme pojmov, vyberá zo systému pojmov, dostáva sa do širokého používania a tým stráca definitorickú a systémovú jednoznačnosť“ (Horecký – Buzássyová – Bosák, 1989, s.

260). Determinologizácia ako „organické zaradenie termínu do sústavy a komunikačného prostredia slov-netermínov“ predstavuje „sémantickú obmenu slova vnútri neodborného jazyka“, čiže prechod od definície k lexikálnemu významu alebo od odborného pojmu k lexikálnemu (Dolník, 2003, s. 77).

Vychádzajúc z týchto poznatkov sa zamyslíme, či prechod viacslovného termínu z odborného komunikačného prostredia do publicistického (resp. hovorového) môže znamenať jeho posun bližšie k lexikalizovanému spojeniu.

Najprv si vymedzíme jednotlivé atribúty termínov, aby sme ich v zozname kolokácií dokázali identifikovať.

J. Dolník vymenúva tieto vlastnosti termínov (2003, s. 175 – 180):

1. definovanosť – definičné určenie termínu, ktoré vyplýva z potreby odbornej komunikácie a zo samotnej povahy pojmu;
2. jednoznačnosť – znamená jednoznačnosť termínu v rámci daného terminologického systému;
3. ustálenosť – v zmysle definičného určenia termínu, pričom sa vyžaduje aj ustálenosť jeho formy;
4. systémovosť (ústrojnosť) – požiadavka formálnej (štruktúrnej) pravidelnosti vo vzťahu k zaradeniu daného pojmu do príslušnej sústavy pojmov;
5. motivovanosť – požiadavka motivovanosti pri tvorení termínov;
6. derivatívnosť (nosnosť) – ak je možnosť voľby z viacerých podôb pojmu, preferuje sa podoba, ktorá sa vyznačuje vyšším stupňom derivačnej flexibility;
7. akonotatívnosť – znamená, že konotatívnosť termínu sa považuje za nežiaducu vlastnosť.

A. Jarošová (2000b, s. 486; podľa Horecký, J.: Základné problémy terminológie. Kultúra slova, 1974, s. 129 – 132) zdôrazňuje dva aspekty termínov – definičné vymedzenie a miesto v systéme pojmov.

Na základe uvedených vlastností sme v zozname kolokácií hľadali spojenia, ktoré spĺňajú dané kritériá. Vyčlenili sme tri druhy termínov:

– termíny vlastné vedným odvetviam nachádzajúce sa prevažne v odborných textoch (T_1),

– termíny jednotlivých odvetví charakteristické frekventovanosťou výskytu v neodborných, publicistických textoch (T_2),

– termíny udomácnené v bežnej slovnej zásobe (T/L).

Tab. č. 2: Jednotlivé kolokácie v základných (nominatívnych) tvaroch

T_1	T_2	T/L
rečový orgán	hrubý domáci produkt	životné prostredie
lyrický subjekt	informačné technológie	politická strana
	platobná bilancia	červená karta
	právna norma	verejná mienka
	kapitálový trh	zemný plyn
	cenné papiere	elektrická energia
	Ústavný zákon	životná úroveň
		slovná zásoba
		detský domov
		základná škola

Tabuľku tvoria termíny, ktorých pojmy síce majú presne vymedzený význam a svoje pevné miesto v sústave pojmov, či už v ekonómii, práve, pedagogike, napriek tomu nie sú úplne rovnocenné. Jednotlivé termíny sa v rôznych komunikačných prostrediach vnímajú odlišne.

Spojenia *rečový orgán* a *lyrický subjekt* ako jediné termíny v zozname kolokácií pochádzajú z čisto odbornej literatúry. Sú to termíny, ktoré sa v neodborných textoch zvyčajne nevyskytujú a ak sa vyskytujú, vnímajú sa ako príznakové. Frekvenčná distribúcia kolokácie *lyrický subjekt* ukázala, že spojenie sa v prim-vyv vyskytuje spolu 32-krát, z toho sa iba štyrikrát nachádza v denníku SME (v literárnej prílohe), v ostatných prípadoch výlučne v odborných literárnych časopisoch a publikáciách.

Na druhej strane sú termíny *kapitálový trh*, *hrubý domáci produkt*, *právna norma*, s ktorými sa používateľ síce stretáva prostredníctvom médií, no napriek

tomu je preňho definičný význam jednotlivých pojmov pomerne vzdialený. Sú to predovšetkým spojenia charakteristické veľkou frekventovanosťou v publicistike.

Spojenia *školský rok*, *základná škola*, *detský domov* sa vnímajú v bežnej komunikácii ako nepríznakové, sú súčasťou každodenného života, a preto si používateľ neuvedomuje aj ich terminologickú platnosť. Vymenované viacslovné termíny sú výsledkom procesu terminologizácie v jej druhom význame: procesu, ktorým sa slovo bežnej reči, spravidla štylisticky neutrálne, stáva termínom v niektorom vednom odbore (Horecký, J. – Buzássyová, K. – Bosák, J., 1989, s. 260). Preto majú tieto kolokácie bližšie k lexikalizovaným spojeniam ako k termínom.

V spojeniach typu *zemný plyn*, *elektrická energia*, *akciová spoločnosť* sa stretávame s tzv. medzisystémovou polysémiou, čiže pretváraním odborného pojmu na jazykový význam. Prechodom termínu do neodbornej komunikácie sa definične vymedzený pojem zjednodušuje (Dolník, 2003, s. 78). Napríklad spojenie *červená karta* sa primárne zaraďuje do športovej terminológie, kde je význam tohto pojmu presne definovaný. Bežný používateľ jazyka má o pojme *červená karta* základné informácie a priraďuje mu význam „vylúčenie hráča z hry“, pričom mu presné a jasné podmienky, za akých toto „vylúčenie“ môže nastať, nemusia byť známe. Tie sú náplňou definície tohto pojmu. Dôkazom udomácnenia spojenia *červená karta* v bežnej slovnej zásobe je aj fakt, že toto spojenie sa používa tiež v prenesenom význame a funguje ako frazeologická jednotka („dostať červenú kartu“).

A. Jarošová tvrdí, že sémantická vágnosť je charakteristická práve pre lexikalizované spojenia a je spôsobená „absenciou pevného systému súradných a nadradených pojmov“ (Jarošová, 2000b, s. 487). Vzťah lexikalizovaných spojení a termínov teda v súlade s A. Jarošovou môžeme chápať ako opozíciu sémantickej vágnosti a definičnosti.

6.2.2. Frazémy a lexikalizované spojenia

V analyzovanom zozname kolokácií sme vyčlenili frazeologizmy na základe týchto doterajších poznatkov o tejto problematike.

J. Mlacek (1984, s. 46) definuje frazeologickú jednotku ako „ustálené slovné spojenie, ktoré sa vyznačuje obraznosťou a nerozložiteľnosťou svojho významu, ako aj expresívnosťou“.

J. Horecký (1997, s. 78 – 80) považuje frazému predovšetkým za pomenovaciu jednotku pomenúvajúcu istú situáciu, pričom ju charakterizoval týmito vlastnosťami:

1. reprodukovanosť – frazéma vystupuje ako lexikálny výraz;
2. štruktúrovanosť – frazéma je súbor jazykových výrazov, ktoré sú usporiadané, štruktúrované;
3. variantnosť – ňou je sprevádzaná vlastnosť štruktúrovanosť;
4. celostnosť – je základnou obsahovou vlastnosťou frazémy, ktorá funguje ako celok;
5. transformovanosť – vo frazéme ide o transformáciu pôvodnej, primárnej výpovede;
6. modálnosť – prejavuje sa pri výstavbe frazémy;
7. expresívnosť – frazéma vyjadruje postoj hovoriaceho.

Pripomeňme si, že na základe princípu funkčnej separácie A. Jarošová (2000a, s. 145) vyčlenila frazeologické jednotky ako pomenovania „bežných situácií citovo hodnotiacim spôsobom“ (porov. 2.1.).

Zo zoznamu kolokácií sme vybrali spojenia, ktoré disponovali uvedenými vlastnosťami, a tiež spojenia, ktoré sa istými vlastnosťami približovali k frazémam. Vytvorili sme dve skupiny kolokácií:

– vlastné frazeologické jednotky (F),

– ustálené spojenia (F/L), ktoré podliehajú frazeologizácii, čiže sa u nich prejavuje istý stupeň idiomatickosti (Dolník, 2003, s. 153; porov. 2.1.).

Tab. č. 3.

F	F/L
čierna diera	(tovar) z druhej ruky
(mať vo) vlastných rukách	Stratená ríša
(ako blesk) z jasného neba	horúca linka
(robiť si) ťažkú hlavu	pranie špinavých peňazí
(trafiť) do čierneho	deň otvorených dverí
(mať) plné ruky práce	prírodné kino
(nemať) ani potuchy	studená vojna
(zažiť na) vlastnej koži	babie leto
(priviesť niekoho) do varu	čierna skrinka

V prípade kolokácie *čierna diera* (nie astronomického pojmu) ide o nesúčtovosť významov obidvoch komponentov spojenia. Znamená to, že dané spojenie je frazéma napriek tomu, že po formálnej stránke sa javí ako typické lexikalizované spojenie s „farebným adjektívom“ (Jarošová, 2000b, s. 498) .

V prípade ustálených spojení v skupine F/L môžeme sledovať sémantickú transponovanosť aspoň jedného z komponentov, napr. v spojení *horúca linka* je prenesený význam adjektíva *horúci*. Keďže lexikalizované spojenie pomenúva všeobecne známe pojmy, spojenie *pranie špinavých peňazí*, pôvodom ekonomický pojem, by sme mohli nazvať lexikalizovaným – pomenúva „istú nelegálnu činnosť“. Súčasne v ňom však zreteľne cítiť istý hodnotiaci postoj k pomenovanému. Takisto spojenie *deň otvorených dverí* sa viaže k určitej udalosti, vnímame ho ako ustálené, teda lexikalizované napriek tomu, že je preň charakteristická sémantická nedoslovnosť vlastná predovšetkým frazémam. V spojení *studená vojna* sa istý hodnotiaci aspekt v ňom pôvodne prítomný dostal do úzadia a spojenie funguje už len ako pomenovacia jednotka.

Hranica medzi lexikalizovanými spojeniami a frazeologickými jednotkami nie je podľa A. Jarošovej vedená na osi obraznosť – neobraznosť, ale ich odlíšenie sa odvíja od prítomnosti alebo neprítomnosti citovo hodnotiaceho postoja subjektu k pomenovanému (Jarošová, 2000a, s. 147).

J. Kačala vytyčuje značne ostrú hranicu medzi frazémami a lexikalizovanými spojeniami. Pripomeňme si, že J. Kačala definuje lexikalizované spojenie ako dvojčlennú sústavu majúcu v dominantnom postavení kategoriálne sloveso alebo podstatné meno (Kačala, 1997b, s. 193 – 194; 1997c, s. 99; porov. 2.1.1.). Práve kategoriálne slová považuje za konštitutívny prvok a za najvýraznejší znak, ktorý odlišuje lexikalizované spojenia od „pestrej výstavbovej stránky“ frazeologických jednotiek (Kačala, 1997c, s. 97). Na základe nami analyzovaných typov lexikalizovaných spojení však musíme konštatovať, že takto vymedzené kritérium

nie je dostačujúce a prakticky nezahŕňa trojčlenné spojenia, ktoré sa v našej tabuľke vyskytli.

Keďže „hranica medzi lexikalizáciou a frazeologizáciou je priestupná“ (Dolník, 2003, s. 153; porov. 2.1.), za dostačujúce kritérium na rozlíšenie lexikalizovaných spojení od frazeologizmov považujeme v súlade s A. Jarošovou funkčné určenie lexikalizovaných spojení ako pomenovacích jednotiek „špecifikovaných a zovšeobecnených jednotlivín“ a frazém ako pomenovaní bežných situácií „s (emocionálno-)hodnotiacim prvkom vo význame“ (Jarošová, 2000a, s. 145; porov. 2.1.).

6.2.3. Voľné a lexikalizované spojenia

Voľné spojenia ako „aktuálne utvárané syntagmatické reťazce“ (Dolník, 2003, s. 152) stoja v opozícii k lexikalizovaným spojeniam, ktoré sa považujú za reprodukované, opakovane používané jednotky v komunikácii. V súlade s A. Jarošovou (2000a, s. 141; porov. 2.1.) môžeme za rozhodujúcu vlastnosť, ktorá rozlíši lexikalizované spojenia od voľne vytvorených, považovať ustálenosť. Autorka vyčlenila tri aspekty ustálenosti: reprodukovanosť (dispozičnosť), oslabenú spájateľnosť a nominačnosť (porov. 2.1.).

Podľa toho, či sa tieto vlastnosti v jednotlivých analyzovaných kolokáciách nachádzali, alebo v nich absentovali, sme vyčlenili nasledujúce voľné spojenia a k nim sme na porovnanie priradili kolokácie, ktoré sme považovali za lexikalizované (ustálené):

– skupina kolokácií, ktorú sme pracovne vyčlenili pri triedení trigramov a bigramov ako vzorku voľných spojení (V_1),

– voľné spojenia, ktoré sme zaradili do zoznamu kolokácií pôvodne ako ustálené (V_2).

Tab. č. 4

V_1	V_2

mladí ľudia	v bezprostrednej blízkosti
nové manželstvo	v konečnom dôsledku
budúci týždeň	horizont udalostí
druhý polčas	praktické zázraky
moja hlava	zelený jazyk
istý čas	do značnej miery
základná časť	konflikt záujmov
trojčlenná skupina	rozvoj bývania
	balík opatrení
	poslanecký klub

Skupinu V_1 tvoria kolokácie, ktoré môžeme charakterizovať ako voľné spojenia v pravom zmysle slova, ako vytvárané pre aktuálnu potrebu, nereprodukované. Sú to spojenia, ktoré nepomenúvajú žiadnu špecifickú jednotlivinu, sú náhodne utvorené v procese komunikácie.

Do skupiny V_2 sme začlenili spojenia, ktoré sme pri selekcii bigramov a trigramov považovali za ustálené a intuitívne sme ich zaradili do zoznamu kolokácií. Pozrieme sa preto, čo majú tieto spojenia spoločné s ustálenými, resp. s lexikalizovanými spojeniami.

Jednotlivé voľné spojenia môžeme rozdeliť do troch skupín:

1. Spojenia, ktoré sú charakteristické veľkou frekventovanosťou, a to nielen v publicistických textoch. Ide o spojenia typu *do značnej miery*, *v konečnom dôsledku*, *z dlhodobého hľadiska*, *v bezprostrednej blízkosti*, *diametrálne odlišné*, *v poslednom čase* a pod.. Sledovali sme frekvenčnú distribúciu spojenia *do značnej miery* v korpuse (vo verzii prim-vyv s absolútnou frekvenciou 87 výskytov) a zistili sme, že sa vyskytuje nielen v publicistike, ale aj v odbornej a umeleckej literatúre.
2. Pomenovacie jednotky známe len istému okruhu používateľov alebo individuálne, originálne spojenia. Našli sme dve kolokácie takéhoto typu: *zelený jazyk* a *praktické zázraky*. Obidve kolokácie boli nerovnomerne rozložené v korpuse, nachádzali sa len v istom titule. Spojenie *zelený jazyk* sme v prvej fáze vyčlenili ako ustálené spojenie aj

z toho dôvodu, že svojou formou pripomína adjektívno substantívne lexikalizované spojenie (Jarošová, 2000b, s. 485; porov.2.1.1.). Pri prehliadaní konkordancií s týmto spojením sme zistili, že výraz *zelený jazyk* je špecifické pomenovanie pre istý okultný „jazyk“, ktorého pojem však nie je všeobecne známy, a preto toto spojenie nespĺňa podmienku ustálenosti.

3. Publicizmy, ktoré sú viazané prevažne na vlastnú štýlovú vrstvu, čím nespĺňajú podmienku všeobecnej reprodukovanosti. Sú to spojenia typu *balík opatrení*, *horizont udalostí*, *poslanecký klub*, *vládna garnitúra*, *bankový sektor*. Mnohé publicizmy vznikli ako synonymné výrazy k niektorým spojeniam, napr. spojenie *poslanecký klub* sa v publicistike obmieňa so spojením *politická strana*. A. Jarošová v tejto súvislosti hovorí o „publicistickom synonyme“ (2000a, s. 148), ktoré vzniklo z potreby variovať frekventovane používané pomenovanie.

7. Vyčlenenie lexikalizovaných spojení a ich analýza

Aj po vylúčení termínov, frazeologizmov a voľných spojení zo zoznamu kolokácií, zostalo na analýzu široké spektrum rôznych spojení, ktoré bolo potrebné roztriediť a klasifikovať. V tejto fáze sa teda zameriame na zvyšné kolokácie a zamyslíme sa, ktoré možno považovať za lexikalizované spojenia.

Budeme vychádzať z týchto už spomínaných vlastností lexikalizovaných spojení podľa A. Jarošovej (2000a, s. 141 – 143; 2000b, s. 487; porov. 2.1.):

1. reprodukovanosť – bude sa skúmať, či dané spojenie je jedinečným spojením spĺňajúcim vlastnosť samostatnej lexikálnej jednotky slovnej zásoby,
2. oslabená spájateľnosť jedného z komponentov, ktorá sa úzko viaže na predchádzajúcu vlastnosť v zmysle jedinečnosti vytvoreného spojenia,
3. nominačnosť – pomenúvacia funkcia, ktorej špecifiku autorka vidí v „pomenúvaní predmetov, javov a konvenčných pojmov špecifikujúcim alebo zovšeobecňujúcim spôsobom“ (Jarošová, 2000a, s. 149),
4. sémantická transponovanosť, čiže významový posun jedného z komponentov spojenia, či nerozložiteľnosť významu, teda istý stupeň idiomatickosti prítomný v niektorých lexikalizovaných spojeniach.

Vlastnosť, ktorú budeme v spojeniach skúmať, je kolokabilita (spájateľnosť). Budeme pozorovať, do akej miery sa v komponente spojenia prejavuje anomálna

spájateľnosť. Zo zoznamu kolokácií sme si vybrali tieto príklady: *informačná služba*, *olivový olej*, *pitná voda*, *bytová jednotka*, *jedálny lístok*, *písací stôl*, *vysoká škola*, *koncentračný tábor*, *citrónová šťava*, *cestovný ruch*, *očitý svedok*, *reprezentačný tréner*, *tiesňové volanie*, *obchodná sieť*, *predvolebná kampaň*, *čerpacia stanica*, *batožinový priestor*, *pešia zóna*, *odborový zväz*, *letecká doprava*, *voľný čas*.

Každá z uvedených kolokácií už na prvý pohľad spĺňa obe zo základných požiadaviek pre lexikalizované spojenia, konkrétne vlastnosť nominačnosť, t. j. všetky uvedené spojenia pomenúvajú „istú špecifikovanú alebo zovšeobecnenú jednotlivinu“ (Jarošová, 2000a, s. 145), a reprodukovanosť, čiže dané spojenia sa opakovane používajú v komunikácii. Vlastnosť anomálna kolokabilita však nie je vlastná všetkým uvedeným kolokáciám.

Spojenia *batožinový priestor*, *informačná služba* či *obchodná sieť* môžeme zaradiť k lexikalizovaným spojeniam, ako ich definuje J. Kačala, s kategoriálnym podstatným menom, ktoré má všeobecný až abstraktný lexikálny význam. Tieto spojenia sú charakteristické tým, že kategoriálny komponent zostáva v spojeniach fixný, obmieňajú sa len „špecifikačné, pohyblivé zložky“ (1997b, s. 195). To znamená, že sa u kategoriálneho podstatného mena otvorená kolokačná paradigma priam predpokladá.

V korpuse sme pozorovali „správanie sa“ spojenia *predvolebná kampaň*. Prekvapili nás pomerne vysoké štatistické hodnoty: MI = 12.132, T = 11.830. Dokonca spájateľnosť výrazu *predvolebná* s výrazom *kampaň* bola vyčíslená relatívnou frekvenciou 74 %, čo znamená, že výskyt výrazu *predvolebná* s výrazom *kampaň* pokrýva 74 % jeho celkového výskytu v korpuse. Na prvý pohľad by sa mohlo zdať, že sme narazili na oslabenú spájateľnosť. Avšak pri zisťovaní rozloženosti zvyšných 26 % výskytu, sme objavili zoznam ďalších slov, tvoriacich so slovom kampaň kolokáciu – *reklamná, sľuby, mítingy, prieskumy, preferencie, plagáty...*

Podobnými príkladmi sú spojenia *citrónová šťava, olivový olej, reprezentačný tréner*, ktoré síce neobsahujú všeobecné či abstraktné kategoriálne podstatné meno, môžeme ich zaradiť do skupiny spojení so značne otvorenou kolokačnou paradigmou. Napríklad lexéma *olej* sa spája s adjektívami *olivový*, *slnečnicový*,

rastlinný...; lexéma *šťava* sa môže spájať s celou škálou „zeleninových“ a „ovocných“ adjektív.

V spojeniach *písací stôl, pešia zóna, tiesňové volanie* môžeme hovoriť o „úzkej významovej špecializácii“ alebo o obmedzenej (viazanej) spájateľnosti (Jarošová, 2000a, s. 143) – *písací stôl/stroj/zošit*; *pešia zóna/turistika*; *tiesňové volanie*, *tiesňová linka, tiesňový hovor*.

Napokon sme sa pomocou štatistických nástrojov pokúsili vyčleniť monokolokabilné spojenia: relatívnu frekvenciu 100 % mali spojenia *očitý svedok*, *koncentračný tábor* a *čerpacia stanica*.

Kolokácie sme na základe našich analýz rozdelili do troch skupín:

- spojenia s otvorenou kolokačnou paradigmou (L/V),
- spojenia, v ktorých sa prejavuje anomálna spájateľnosť (L_1),
- monokolokabilné spojenia (L_2).

Tab. č. 5.

L/V	L_1	L_2
informačná služba	jedálny lístok	čerpacia stanica
letecká doprava	voľný čas	očitý svedok
predvolebná kampaň	písací stôl	koncentračný tábor
citrónová šťava	pešia zóna	
reprezentačný tréner	cestovný ruch	
olivový olej	vysoká škola	
obchodná sieť	tiesňové volanie	
bytová jednotka	pitná voda	
batožinový priestor		
odborový zväz		

A. Jarošová spojenia s kategoriálnymi podstatnými menami síce zaraďuje do jedného zo svojich delení lexikalizovaných spojení, napriek tomu ich považuje za periférium ustálených spojení práve pre relatívne otvorenú kolokačnú paradigmu (Jarošová, 2000a, s. 146).

Kolokácie, ktoré sme začlenili do skupiny L/V, na základe absencie prvku anomálnosti v spájateľnosti môžeme v súlade s A. Jarošovou zaradiť medzi spojenia, ktoré tvoria prechod medzi voľnými a lexikalizovanými spojeniami.

7.1. Typy lexikalizovaných spojení

Typy spojení, ktoré sa v analyzovanom materiáli vyskytli a nespadali do „kolónky“ termínov, frazém ani voľných spojení, sme podľa formálnych kritérií rozdelili do niekoľkých skupín a sledovali, ktoré možno považovať za lexikalizované spojenia.

A Menné spojenia

Aa V rámci tejto skupiny sme vyčlenili spojenia typu **adjektívum – substantívum** (**ADJ – S),** napr. *studená vojna, voľný čas, životná úroveň, kupónová privatizácia, umelý sneh, televízne noviny, čierna skrinka, jadrové zbrane, stratená ríša, prírodné kino, vysoká škola, čierne korenie, červená karta, starý kontinent, očitý svedok, židovský štát, pušný prach, dobré ráno*.

Na základe predchádzajúcich analýz sme do tejto skupiny nezaradili spojenia s otvorenou kolokačnou paradigmou, teda lexikalizované spojenia s kategoriálnym podstatným menom, ako ich vymedzuje J. Kačala (1997b).

Menné spojenia typu ADJ – S môžeme rozdeliť do troch skupín:

1. Lexikalizované spojenia, ktoré A. Jarošová (2000b, s. 487; 2000a, s. 142; porov. 2.1.1.) delí podľa stupňa sémantickej transponovanosti komponentov spojenia, čiže podľa toho, či komponenty vstupujú do spojenia vo svojich priamych „slovníkových“ významoch. Na základe tohto kritéria sme vyčlenili z nášho zoznamu
   - lexikalizované spojenia vyznačujúce sa motivačnou zreteľnosťou, napr. *manželský súhlas, čajová lyžička, televízne noviny;*
   - lexikalizované spojenia vyznačujúce sa čiastočnou motivačnou zreteľnosťou, napr. *pušný prach, vysoká škola, stará mama, prírodné kino;*
   - lexikalizované spojenia vyznačujúce sa motivačnou nezreteľnosťou, napr. *babie leto, čierna skrinka, horúca linka.*

2. Lexikalizované spojenia s platnosťou proprií a apelativizovaných proprií. Spojenia typu *stratená ríša, starý kontinent, židovský štát* J. Kačala (1997a, s. 41) nazýva obraznými viacslovnými názvami utvorenými ako náprotivky oficiálnych názvov a majúcimi platnosť vlastných mien.
   Spojenie *Televízne noviny*, názov televíznej spravodajskej relácie, sa tiež v korpuse nachádza ako proprium, domnievame sa však, že v prípade tohto spojenia môžeme hovoriť o apelativizácii, čiže procese, v ktorom sa z propria stáva apelatívum (porov. Horecký – Buzássyová – Bosák, 1989, s. 92). V prípade kolokácie *Prírodné kino* je situácia opačná; v korpuse sa dané spojenie vyskytovalo výlučne s veľkým začiatočným písmenom, napriek tomu funguje ako apelatívum a používa sa ako synonymný výraz k slovu *amfiteráter*
3. Z formálneho hľadiska zaraďujeme k spojeniam typu ADJ – S kontaktové formuly *dobré ráno, dobrú noc*, ktoré môžeme považovať za lexikalizované v širšom zmysle. Otázku začlenenia kontaktových formúl (aj podobného typu: *všetko dobré, ako sa máš, všetko najlepšie*) do skupiny vlastných lexikalizovaných spojení považuje A. Jarošová (2000b, s. 485) za otvorenú.

Ab V skupine **S – S gen.** sa nachádzajú menné spojenia, v ktorých druhý komponent je zo syntaktického hľadiska v úlohe nezhodného prívlastku: *majster sveta, hlava štátu, otras mozgu, výčitky svedomia, koniec koncov*.

V rámci tejto skupiny sme vyčlenili spojenia typu **S – ADJ – S gen.**, kde adjektívum so subtantívom v genitíve predstavujú rozvitý nezhodný prívlastok pre nadradené substantívum. Našli sme dve lexikalizované spojenia tohto typu: *pranie špinavých peňazí, deň otvorených dverí*.

B slovesno menné spojenia

Ba Do skupiny **V – S^{abstr}** sme zaradili spojenia kategoriálneho slovesa s abstraktným podstatným menom. Sú to lexikalizované spojenia *podať demisiu,*

venovať pozornosť, klásť otázky/odpor, spytovať svedomie, dať prednosť, mať pravdu, dávať pozor, spáchať zločin a pod.

Spojenia tohto typu nazýva napr. F. Čermák verbonominálnymi kvázifrazémami stojacimi na periférii frazém (Čermák, <http://ucnk.ff.cuni.cz>).

A. Jarošová v tejto súvislosti hovorí o „verbalizácii abstrákt pomocou slovies so všeobecnou sémantikou“ (Jarošová, 2000a, s. 143), či o regelmentovanej spájateľnosti (Jarošová, 1992. s. 121). „Ak chce používateľ opísať situáciu, v ktorej vystupuje ako aktant abstraktum, musí poznať jeho reglementovanú spájateľnosť, t. j. štylisticko-normatívnu spájateľnosť“ (Jarošová, 1992, s. 121).

Bb Zo spojení skupiny **V – S^{konkr}** sa v analyzovanom korpuse nachádzali len dve kolokácie – *pokrčiť plecami* a *pokrútiť hlavou*. Túto skupinu podľa A. Jarošovej tvoria spojenia, ktorých slovesá vyjadrujú typický pohyb, resp. vlastnosť určitých častí tela, napr. klipkať očami (Jarošová, 1995, s. 86 – 87).

C Iné spojenia

V tejto skupine sa nachádzajú spojenia, ktoré nespadajú do predchádzajúcich typov a zároveň sa vyznačujú istými spoločnými vlastnosťami.

Ca Spojenia s časticovou platnosťou – *inými slovami, jednoducho povedané, tak či onak* – vystupujú ako isté „gramatické spojenia“ majúce vo vete gramatický význam, resp. gramatický typ významu (porov. Kačala, 1993, s. 19). Spojenie *inými slovami* sa v zozname kolokácií vyskytovalo s veľkým začiatočným písmenom, čo indikuje, že ide o uvádzaciu časticu (podľa Oravec – Bajzíková – Furdík, 1988, s.203 – 204).

Pozrime sa na výskyt spojenia *tak či onak* v korpuse:

1. Je vecou historikov a ľudí , ktorí ho chcú vidieť <tak či onak> , aby o tom diskutovali . " Od akcie sa však
2. je však isté . Nech už parlament rozhodne <tak či onak> , každé , aj zamietavé rozhodnutie bude
3. minulosť . Komunistická nomenklatúra totiž <tak či onak> rešpektovala rodovú štruktúru spoločnosti a jej
4. protivníka . Vývoj udalostí v Čiernej Hore <tak či onak> opäť potvrdil , že najväčším nešťastím Srbov
5. a vzdaním sa komunistom , ktorí , <tak či onak> , hovoria o mieri a ich ideály vlastne nemajú
6. čo Stolpe spravil s kamennou tvárou . <Tak či onak> , zákon musí ešte podpísať spolkový

Z predchádzajúcich príkladov vidíme, že kolokácia *tak či onak* má ako častica niekoľko pozícií vo vete, napr. ako vytyčovacia či uvádzacia častica (Oravec – Bajzíková – Furdík, 1988, s. 206 – 207) a tiež, že vystupuje aj ako príslovka (v príklade č. 2 vo význame „akokoľvek“). Takže figuruje nielen ako gramatická, ale aj sémantická zložka vety (porov. Jarošová, 1999, s. 95; podľa D. Cruse: Lexical Semantics. Cambridge University Press 1986).

Dokonca môže mať spojenie *tak či onak* status minimálnej frazeologickej jednotky. V príklade č. 1 je naznačený istý hodnotiaci prvok vo význame spojenia *tak či onak* (*tak či onak* = „za každú cenu“).

Časticovú platnosť má aj spojenie *koniec koncov*, ktoré po formálnej stránke zaraďujeme k menným spojeniam (S – S gen).

Cb Spojenia s príslovkovou platnosťou: *skôr či neskôr, deň čo deň, hore nohami, znova a znova, tu a tam.*

Všetky vymenované spojenia sa vo vete vyskytujú ako príslovkové výrazy majúce špecifické významy (*tu a tam* = „občas“, *znova a znova* = „opakovane“).

Príslovkovú platnosť spojenia *hore nohami* si ukážeme na príkladoch:

1. Varte 8 minút a potom knedľu obráťte <hore nohami> . <hi> Knedľa , samozrejme , v skutočnosti
2. následkov po povodni , ktorá obrátila obec <hore nohami> 7 . augusta minulého roku . " Vybreženie
3. abstraktné , že divák nevie , či obraz nevisí <hore nohami> , ak poézia stráca jasnosť a krásu a mení
4. vierou v zázrak , že sa to pred voľbami obráti <hore nohami> , je odmietaním ochoty pozrieť sa

Z uvedených príkladov vidíme, že spojenie *hore nohami* sa vo vete používa vo význame „naopak, naruby“, konkrétne v príkladoch č. 1. a č. 3. Napriek tomu, že v danom spojení cítiť prvok obraznosti, považujeme ho za príslovkový výraz. V ostatných prípadoch môžeme hovoriť o frazeologickej platnosti daného spojenia, v príkladoch č. 2. a č. 3. je spojenie *hore nohami* súčasťou frazémy *byť hore nohami*.

Spojenia s časticovou a príslovkovou platnosťou považujeme za lexikalizované v tom zmysle, že vystupujú ako reprodukované jednotky opakovane používané v procese komunikácie, ako samostatné lexikálne formy a v prípade prísloviek aj ako sémantické jednotky. A. Jarošová (2000b, s. 485; porov. 2.1.1.) do svojho delenia lexikalizovaných spojení zahŕňa predložkovo adjektívne spojenia typu *za*

mlada, od mala, za horúca, ktorým pripisuje príslovkovú platnosť, a predložkovo substantívne spojenia typu *na počesť, na úkor, v protiklade k*, v ktorých prípade hovorí o posune k predložkovej platnosti. Vzhľadom na to, že zo zoznamu trigramov sme spojenia tohto typu nevyčlenili, nezaradili sme ich ani do nášho delenia lexikalizovaných spojení.

8. Zhodnotenie výsledkov

Analyzovaný zoznam kolokácií predstavoval rôznorodý jazykový materiál so širokým spektrom slovných spojení s prevahou výrazov pochádzajúcich z publicistických textov. Použitím metódy komparácie sme podľa miery ustálenosti a príslušnosti k istej štýlovej vrstve vyčlenili tieto typy spojení:

- frekventované spojenia typu *do značnej miery, v poslednom čase, rozvoj bývania, diametrálne odlišné;*
- spojenia tvoriace prechod medzi voľnými a ustálenými spojeniami typu *informačná služba, predvolebná kampaň, citrónová šťava;*
- vlastné lexikalizované spojenia typu *očitý svedok, stará mama, polievková lyžica, ročné obdobie;*
- „zlexikalizované“ viacslovné termíny typu *politická strana, zemný plyn, slovná zásoba;*
- viacslovné termíny charakteristické frekventovanosťou výskytu v publicistických textoch typu *hrubý domáci produkt, cenné papiere, kapitálový trh;*
- viacslovné termíny odborných textov typu *rečový orgán, lyrický subjekt*
- lexikalizované spojenia podliehajúce idiomatizácii typu *horúca linka, babie leto, čierna skrinka, studená vojna;*
- frazeologické jednotky typu *mať plné ruky práce, dostať niekoho do varu, zažiť na vlastnej koži.*

Po vylúčení voľných spojení, frazém a termínov, sme sa pokúsili o typologizáciu zvyšného materiálu, pričom sme skúmali, ktoré zo spojení sú lexikalizované.

Z formálneho hľadiska sme vyčlenili tieto typy lexikalizovaných spojení:

A Menné spojenia

- **ADJ – S** : *pracovný čas, ročné obdobie, manželský súhlas, priamy prenos, babie leto, tiesňový hovor, studená vojna, voľný čas, životná úroveň, kupónová privatizácia, umelý sneh, televízne noviny, čierna skrinka, jadrové zbrane, stratená ríša, prírodné kino, vysoká škola, čierne korenie, červená karta, starý kontinent, očitý svedok, židovský štát, pušný prach, stará mama, čajová lyžička* a pod.

- **S – S gen.** : *majster sveta, hlava štátu, otras mozgu, výčitky svedomia, koniec koncov*

  S – ADJ – S gen : *pranie špinavých peňazí, deň otvorených dverí*

B Slovesno menné spojenia

- **V – S^{abstr}** : *podať demisiu, venovať pozornosť, klásť otázky/odpor, spytovať svedomie, dať prednosť, mať pravdu, dávať pozor, spáchať zločin*
- **V – S^{konkr}** : *pokrčiť plecami, pokrútiť hlavou*

C Iné spojenia

- **Spojenia s príslovkovou platnosťou**: *znova a znova, hore nohami, skôr či neskôr, deň čo deň, tu a tam, tak či onak*
- **Spojenia s časticovou platnosťou**: *inými slovami, jednoducho povedané, tak či onak*

Nesnažili sme sa o vytvorenie nového delenia lexikalizovaných spojení, vychádzali sme z materiálu a z doterajších poznatkov o tejto problematike. Neusilovali sme sa ani vyriešiť všetky otázky (otázku začlenenia kontaktových formúl medzi vlastné lexikalizované spojenia sme nechali otvorenú).

Naše delenie lexikalizovaných spojení nie je vyčerpávajúce, čo súvisí s povahou spracovávaného materiálu.

Záver

V našej práci sme sa pokúsili o preskúmanie jedného z prístupov k problematike lexikalizovaných spojení.

Úloha, ktorú sme si zadali bola s využitím korpusovej textovej databázy a štatistických metód vyextrahovať lexikalizované spojenia a následne tento jazykový materiál analyzovať. Na základe zvažovania otázky reprezentatívnosti sme si pre náš výskum vybrali vyvážený korpus SNK a pomocou automatizovanej extrakcie frekventovaných dvojíc a trojíc slov sme sa pokúsili o získanie čo najvhodnejšieho jazykového materiálu na skúmanie ustálených, resp. lexikalizovaných spojení. Použitím negatívneho filtra sme odstránili nežiaduce spojenia a následne podrobili zoznam bigramov a trigramov selekcii, pri ktorej sme vyčlenili ustálené spojenia. Ako sa ukázalo, značná prevaha publicistických textov vo vyváženom korpuse sa odrazila aj v „ručne“ selektovanom materiáli. Napriek tomu nám získaný korpusový materiál poskytol možnosť preskúmať rôzne typy ustálených spojení.

K analyzovanému materiálu sme pristupovali na základe doterajších poznatkov o problematike, využívali sme rôzne štatistické nástroje vyhodnocujúce spoločný výskyt dvoch slov, na základe vyhľadávania konkordancií príslušných výrazov sme pozorovali „správanie sa“ spojení v istých kontextových prostrediach. Analyzovaný materiál obsahoval niekoľko typov spojení, či už po stránke formálnej, či z hľadiska štýlovej príslušnosti, alebo miery ustálenosti spojení.

Výsledkom našej práce nebolo vytvorenie novej typológie lexikalizovaných spojení, ale analýza získaného materiálu a tým potvrdenie alebo vyvrátenie správnosti jeho výberu a spracovania. Môžeme skonštatovať, že daný spôsob získania materiálu ako jeden z možných je podnetný, no určite nie jediným.

Problematika lexikalizovaných spojení, ako sme ukázali, je otvorená, bádanie v tejto oblasti si vyžaduje ďalšie analýzy. My sme predstavili jednu z možností, ako pristupovať k lexikalizovaným spojeniam a jeden zo spôsobov, ako ju zrealizovať. Úlohou ďalších analýz by mohlo byť prepracovanie spôsobu získavania jazykového materiálu, či otázka vypracovania operačných kritérií, na základe ktorých by počítač dokázal identifikoval ustálené spojenia. Zaujímavé by bolo aj sledovanie vzťahu medzi frekventovanosťou výskytu spojenia a jeho ustálenosťou, čomu sme v našej práci nevenovali pozornosť. Možnosti využitia korpusu sú v tomto smere otvorené a nabádajú k pozorovaniu rôznych vzájomných súvislostí jazykových javov.

Zoznam použitej a citovanej literatúry

BENKO, V.: Korpus textov slovenského jazyka – súčasný stav a budúcnosť. In: Slovenčina na konci 20. storočia, jej normy a perspektívy. Sociolingvistica Slovaca 3. Ed. S. Ondrejovič. Bratislava: Veda 1997, s. 297 – 303.

ČERMÁK, F.: Jazykový korpus: Prostředek a zdroj poznání. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 15 – 37.

ČERMÁK, F.: Syntagmatika slovníku: typy lexikálních kombinací. <http://ucnk.ff.cuni.cz>

DOLNÍK, J.: Motivácia a hodnota termínu. In: Kultúra slova, 1983, č. 17, s. 133 – 140.

DOLNÍK, J.: Jazykové princípy vo výstavbe frazém. In: Frazeologické štúdie II. Ed. P. Ďurčo. Bratislava: Komisia pre výskum frazeológie pri Slovenskom komitéte slavistov 1997, s. 36 – 44.

DOLNÍK, J.: Lexikológia. Bratislava: Univerzita Komenského 2003.

GARABÍK, R. – GIANITSOVÁ, L. – HORÁK, A. – ŠIMKOVÁ, M.: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. <http://korpus.juls.savba.sk/publikacie/Tagset-aktualny.pdf>

HORECKÝ, J.: Návrh na vymedzenie frazémy. In: Frazeologické štúdie II. Ed. P. Ďurčo. Bratislava: Komisia pre výskum frazeológie pri Slovenskom komitéte slavistov 1997, s. 78 – 81.

HORECKÝ, J. – BUZÁSSYOVÁ, K. – BOSÁK, J. a kol.: Dynamika slovnej zásoby súčasnej slovenčiny. Bratislava: Veda 1989.

JAROŠOVÁ, A.: Spájateľnosť slov a jej odraz v slovníku. In: Jazykovedný časopis, 1992, roč. 43, č. 2, s. 116 – 125.

JAROŠOVÁ, A.: Monokolokabilné slová v slovenčine. In: Jazykovedný časopis, 1995, roč. 46, č. 2, s. 83 – 99.

JAROŠOVÁ, A.: Lexikografia a počítače – slovenský variant. In: Slovenčina na konci 20. storočia, jej normy a perspektívy. Sociolingvistica Slovaca 3. Ed. S. Ondrejovič. Bratislava: Veda 1997, s. 304 – 311.

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

JAROŠOVÁ, A.: Lexikalizované spojenie v kontexte ustálených spojení. In: Princípy jazyka a textu. Ed. J. Dolník. Bratislava: Univerzita Komenského 2000(a), s. 138 – 151.

JAROŠOVÁ, A.: Viacslovný termín a lexikalizované spojenie. In: Človek a jeho jazyk. 1. Jazyk ako fenomén kultúry. Ed. K. Buzássyová. Bratislava: Veda 2000(b), s. 481 – 193.

JAROŠOVÁ, A.: Národný korpus slovenského jazyka a jeho dimenzie. <http://korpus.juls.savba.sk/korpus/biblioteka/publikacie/>

KAČALA, J.: Kategoriálne slová v slovných spojeniach (príspevok k teórii jazykového významu). In: Jazykovedný časopis, 1993, roč. 44, č. 1, s. 14 – 24.

KAČALA, J.: Viacslovné pomenovania v slovnej zásobe. In: Zborník filozofickej fakulty Univerzity Komenského. Philologica XLV. Bratislava: Univerzita Komenského 1997(a), s. 33 – 42.

KAČALA, J.: K statusu lexikalizovaných spojení. Slovenská reč, 1997(b), roč. 62, č. 4, s. 193 – 203.

KAČALA, J.: Lexikalizované spojenia a frazeologické jednotky. In: Frazeologické štúdie II. Ed. P. Ďurčo. Bratislava: Komisia pre výskum frazeológie pri Slovenskom komitéte slavistov 1997(c), s. 95 – 102.

KOPŘIVOVÁ, M.: Využití korpusu při spracování frazeologie ve výkladovém slovníku. <http://ucnk.ff.cuni.cz>

Krátky slovník slovenského jazyka. Red. J. Kačala – M. Pisárčiková – M. Považaj. Bratislava: Veda 2003; <http://kssj.juls.savba.sk>

MLACEK, J.: Slovenská frazeológia. Bratislava: Slovenské pedagogické nakladateľstvo 1984.

ORAVEC, J. – BAJZÍKOVÁ, E. – FURDÍK, J.: Súčasný slovenský spisovný jazyk. Morfológia. Bratislava: Slovenské pedagogické nakladateľstvo 1988.

PECINA, P. – HOLUB, M.: Sémanticky signifikantní kolokace. Praha: Universitas Carolina Pragensis 2002.

ŠIMKOVÁ, M.: Možnosti využitia Slovenského národného korpusu na štúdium slovenského jazyka. In: Studia Academica Slovaca 33. Bratislava: Stimul 2004, s. 204 – 216.

ŠIMKOVÁ. M.: Slovenský národný korpus – východiská a plány. <http://korpus.juls.savba.sk/korpus/biblioteka/publikacie/presov1.html>

<http://korpus.juls.savba.sk>

<http://ucnk.ff.cuni.cz>

Príloha

A Zoznam vytriedených bigramov

1.	štátneho rozpočtu	455
2.	vládnej koalície	361
3.	životného prostredia	321
4.	poslednom čase	305
5.	prvý pohľad	300
6.	ľudských práv	277
7.	cenných papierov	273
8.	štátnej správy	268
9.	politických strán	263
10.	polievkové lyžice	257
11.	hlavného mesta	253
47.	kupónová privatizácia	119
48.	tlačovej besede	118
49.	zeleného jazyka	116
50.	Trestného zákona	114
51.	minister obrany	110
52.	zahraničných investorov	107
53.	Najvyššieho súdu	107
54.	svetový rekord	104
55.	štátny rozpočet	104
56.	elektrickej energie	104
57.	členských krajín	104
58.	citrónovej šťavy	103
59.	pridanej hodnoty	102

12. svetovej vojny 241
13. trestného činu 238
14. generálny riaditeľ 237
15. verejnej mienky 234
16. čiernej diery 223
17. konečnom dôsledku 222
18. národného majetku 219
19. verejnej správy 208
20. finančných prostriedkov 208
21. žlté karty 184
22. Ústavný súd 183
23. predsedu vlády 183
24. Televízne noviny 179
25. čierne korenie 167
26. parlamentné voľby 160
27. tlačovej konferencii 159
28. Policajného zboru 155
29. vysokých škôl 154
30. bez problémov 153
31. minister vnútra 148
32. cestovného ruchu 143
33. základného imania 138
34. Národnej rady 137
35. Svetového pohára 135
36. čajová lyžička 135
37. vládna koalícia 134
38. dozornej rady 134
39. Dobrú noc 134
40. životné prostredie 132
41. trestných činov 128
42. stará mama 128
43. sociálnych vecí 127
44. zahraničný obchod 126
45. majstrovstvách sveta 126
46. úrokových sadzieb 120
60. odňatia slobody 102
61. zdravotný stav 101
62. volebného zákona 101
63. kapitálového trhu 101
64. cennými papiermi 101
65. centrálnej banky 100
66. majster sveta 99
67. tajnej služby 98
68. akciovej spoločnosti 98
69. ministerstva financií 97
70. pracovných miest 96
71. trestnom konaní 95
72. bezprostrednej blízkosti 93
73. štátny tajomník 92
74. kapitálovom trhu 92
75. umelý sneh 91
76. právnických osôb 91
77. poslaneckého klubu 91
78. ZÁKLADNÉ ÚDAJE 90
79. Dobrodružný film 90
80. Dobré ráno 90
81. novely zákona 89
82. denná teplota 89
83. do varu 89
84. trestnej činnosti 88
85. Inými slovami 88
86. značnej miery 87
87. valnom zhromaždení 86
88. štátneho tajomníka 86
89. bežného účtu 86
90. zahraničné investície 85
91. ozbrojených síl 85
92. celé hodiny 85

93. domáceho produktu 84
94. valné zhromaždenie 83
95. daňových poplatníkov 83
96. čierne diery 83
97. ľudskej prirodzenosti 82
98. národnostných menšín 81
99. ústavných činiteľov 80
100. moderovaný blok 80
101. menovej únie 80
102. zdravotných poisťovní 79
103. vysoké školy 79
104. úrokové sadzby 79
105. členských štátov 79
106. Bieleho domu 79
107. verejného činiteľa 78
108. trestné stíhanie 78
109. Bezpečnostnej rady 78
110. olivového oleja 77
111. futbalového zväzu 77
112. ľudskú prirodzenosť 76
113. futbalovej ligy 76
114. opozičných strán 75
115. cenné papiere 75
116. generálneho tajomníka 75
117. dolu schodmi 72
118. obchodnej bilancie 72
119. volebného obdobia 71
120. vlastnej koži 71
121. olympijského výboru 71
122. fyzických osôb 71
123. Zábavný seriál 70
124. trestné oznámenie 70
125. platobnej bilancie 69
126. odborových zväzov 69
127. Čierna skrinka 69
128. stredných škôl 68
129. politickej scéne 68
130. rodinného domu 68
131. Horskej služby 68
132. hore schodmi 68
133. zahraničnej politiky 67
134. ústavného zákona 67
135. dávať pozor 67
136. celozrnnej múky 67
137. výtvarného umenia 66
138. úradu vyšetrovania 66
139. prezidentských volieb 66
140. Pobrežná hliadka 66
141. istým spôsobom 66
142. centre mesta 66
143. slovnej zásoby 65
144. športovej hale 64
145. Bábkové divadlo 64
146. verejných financií 63
147. sviatosti pokánia 63
148. má pravdu 63
149. Koniec koncov 63
150. životnej úrovne 62
151. slnečnicového oleja 62
152. Prírodné kino 62
153. jadrovej elektrárne 62
154. bankového sektora 61
155. valného zhromaždenia 60
156. starý otec 60
157. studenej vojny 59
158. motorových vozidiel 59
159. koniec koncov 59
160. zhodou okolností 58
161. Tehelnom poli 58
162. Priamy prenos 58
163. básnického prekladu 58
164. volebnej kampane 57
165. občianskej vojny 57
166. ľudského života 57
167. ľudská prirodzenosť 57
168. konkurznej podstaty 57
169. zdravotnej starostlivosti 56
170. reprezentačný tréner 56
171. mobilný telefón 56
172. demokratickej ľavice 56
173. pokutového kopu 55
174. voľného času 54
175. telefónne číslo 54
176. komunálnych voľbách 54
177. občianskej spoločnosti 53
178. koaličných partnerov 52
179. jadrových elektrární 52
180. horizontu udalostí 52
181. funkčné obdobie 52
182. zdravotné problémy 51
183. kapitálový trh 51
184. hromadného ničenia 51
185. fyzické osoby 51
186. funkčného obdobia 51
187. finančných zdrojov 51
188. cestovný ruch 51
189. domácom prostredí 51
190. územných celkov 50
191. svetovom šampionáte 50
192. zemného plynu 49
193. odpadových vôd 49
194. bezpečnostné opatrenia 49
195. pitnej vody 48
196. otvorených dverí 48
197. kandidátskych krajín 48
198. výberového konania 47
199. tretieho sektora 47
200. vplyvom alkoholu 47
201. sociálne zabezpečenie 47
202. riešenie problémov 47

203. Obrazové noviny 47
204. multikultúrnej výchovy 47
205. politických subjektov 46
206. hlavnú úlohu 46
207. herecký výkon 46
208. hlas znel 46
209. vládneho kabinetu 45
210. tesnej blízkosti 45
211. trest smrti 45
212. spoločenského života 45
213. plnom rozsahu 45
214. obchodného registra 45
215. vysokej úrovni 44
216. školského roka 44
217. osobných údajov 44
218. horskej služby 44
219. Hlavné námestie 44
220. tajných služieb 43
221. štátnom jazyku 43
222. štátne orgány 43
223. praktické zázraky 43
224. volebné obdobie 43
225. manželské puto 43
226. ani stopy 43
227. ani potuchy 43
228. akciová spoločnosť 43
229. vyslovenie nedôvery 42
230. rozvoja bývania 42
231. regionálneho rozvoja 42
232. programové vyhlásenie 42
233. lyrického subjektu 42
234. Lokálne vysielanie 42
235. životný štýl 41
236. zdvihol hlavu 41
237. prírodnej energie 41
238. podstatné meno 41
239. otvorenej spoločnosti 41
240. národnej kultúry 41
241. konflikte záujmov 41
242. dovoznej prirážky 41
243. spôsobil škodu 40
244. dopravných nehôd 40
245. pravidelné slovesá 39
246. podielových fondov 39
247. nájomných bytov 39
248. mobilných telefónov 39
249. jadrových zbraní 39
250. hlavných úlohách 39
251. domáceho prostredia 39
252. výtvarných umení 38
253. voľný čas 38
254. preferenčných hlasov 38
255. ľudské telo 38
256. Ľudové piesne 38
257. informačnej služby 38
258. dlhodobého hľadiska 38
259. devízovom trhu 38
260. zahraničného výboru 37
261. platobných kariet 37
262. olympijské hry 37
263. občianske práva 37
264. horizont udalostí 37
265. futbalový zväz 37
266. balík akcií 37
267. strategický investor 36
268. obchodného centra 36
269. kultúrneho dedičstva 36
270. dôchodkového systému 36
271. sójovej omáčky 35
272. rodinný príslušník 35
273. produktivity práce 35
274. organizovaný zločin 35
275. do lona 35
276. trhovej ekonomiky 34
277. strúčiky cesnaku 34
278. pracovné príležitosti 34
279. štvorcových metrov 34
280. kultúrnej identity 34
281. daňového zaťaženia 34
282. druhej ruky 34
283. komunistický režim 34
284. hlava štátu 34
285. emočných návykov 34
286. zdravého rozumu 33
287. zelenom jazyku 33
288. trestného stíhania 33
289. strednej školy 33
290. STOLNÝ TENIS 33
291. starého otca 33
292. stará matka 33
293. slovnej zásobe 33
294. reálnych miezd 33
295. jedným dychom 33
296. olympijských hrách 33
297. moderného umenia 33
298. množného čísla 33
299. informačný servis 33
300. jedným dychom 33
301. hraničnom priechode 33
302. finančné zdroje 33
303. finančné problémy 33
304. do čierneho 33
305. duchovného života 33
306. vodičský preukaz 32
307. zdravotných dôvodov 32
308. verejnú mienku 32
309. verejného života 32
310. verejného poriadku 32
311. Štedrý deň 32
312. spoločnú reč 32

313. právnych noriem 32
314. pracovných síl 32
315. postupom času 32
316. informačné technológie 32
317. finančná situácia 32
318. dokumentárny film 32
319. zlomených sŕdc 31
320. vojnových zločinov 31
321. vojenskej služby 31
322. tempo rastu 31
323. tabakových výrobkov 31
324. stavebného sporenia 31
325. štátnej pokladnice 31
326. starého kontinentu 31
327. špinavých peňazí 31
328. priameho kopu 31
329. predvolebnú kampaň 31
330. podnikateľské subjekty 31
331. organizovaný zločin 31
332. minimálnej mzdy 31
333. mierových síl 31
334. mierovej dohody 31
335. inými slovami 31
336. internetovej stránke 31
337. hypotekárnych úverov 31
338. hospodárskej súťaže 31
339. horských oblastiach 31
340. červená karta 31
341. základných práv 30
342. základnej zostavy 30
343. výkonnej moci 30
344. verejnom záujme 30
345. verejné otázky 30
346. umeleckej školy 30
347. svetovej literatúry 30
348. svetového rebríčka 30
349. svetelných rokov 30
350. súdnej sieni 30
351. strednej triedy 30
352. slnečné lúče 30
353. schválila návrh 30
354. národného parku 30
355. konflikt záujmov 30
356. hlasovacích lístkov 30
357. dennom poriadku 30
358. denné svetlo 30
359. červeného vína 30
360. ani korunu 30
361. životného minima 29
362. Zhodou okolností 29
363. pokrčila plecami 29
364. nadmorskej výške 29
365. menovej politiky 29
366. klásť otázky 29
367. dopravnej nehode 29
368. prvej triedy 28
369. petržlenovou vňaťou 28
370. písacom stole 28
371. osobných automobilov 28
372. manželského života 28
373. jazdecký kôň 28
374. Horúce linky 28
375. daňové priznanie 28
376. časovom horizonte 28
377. zlaté medaily 27
378. zazvonil telefón 27
379. vyšetrovacej väzbe 27
380. Veľká noc 27
381. umelecké dielo 27
382. stredných podnikateľov 27
383. sociálne dávky 27
384. pracovný deň 27
385. pracovné miesta 27
386. pohonných látok 27
387. pohonných hmôt 27
388. nočnom stolíku 27
389. očitých svedkov 27
390. petičnej akcie 27
391. otvorenými ústami 27
392. otvorené dvere 27
393. oficiálnu návštevu 27
394. obchodné centrum 27
395. jedálny lístok 27
396. Extrémne športy 27
397. domácej pôde 27
398. cestovnej kancelárie 27
399. celoživotné dielo 27
400. bez seba 27
401. živý tvor 26
402. výčitky svedomia 26
403. verejných prác 26
404. venovať pozornosť 26
405. televíznych staníc 26
406. televíznej obrazovke 26
407. telefonický rozhovor 26
408. spáchal samovraždu 26
409. rečových orgánov 26
410. národnostné menšiny 26
411. minerálnych látok 26
412. kyslej smotany 26
413. kukuričnej múky 26
414. koncentračný tábor 26
415. Digitálne správy 26
416. dávajú prednosť 26
417. bezpečnostnej služby 26
418. alkoholické nápoje 26
419. Vodné kasárne 25
420. vlastnú päsť 25
421. vlastnom záujme 25
422. všetko dobré 25

423.umelej hmoty 25
424.pozemského života 25
425.povodňovej aktivity 25
426.personálne zmeny 25
427.Občianska výchova 25
428.občianska vojna 25
429.otras mozgu 25
430.ľadovom hokeji 25
431.hore nohami 25
432.jadrové zbrane 25
433.batožinový priestor 25
434.zatvorenými dverami 24
435.základnej zostave 24
436.vlastného mena 24
437.vízovej povinnosti 24
438.uzavrieť manželstvo 24
439.tvorivé dielne 24
440.sójovú omáčku 24
441.prejavili záujem 24
442.Pokrútila hlavou 24
443.plnej sile 24
444.odňatím slobody 24
445.nákladné automobily 24
446.množnom čísle 24
447.leteckej spoločnosti 24
448.Kriminálny seriál 24
449.domácej palubovke 24
450.cukrovej repy 24
451.čistého zisku 24
452.bezpečnostných síl 24
453.živých tvorov 23
454.ženského pohlavia 23
455.železničnej stanice 23
456.výtvarné umenie 23
457.východného bloku 23
458.vnútorný hlas 23
459.vládneho programu 23
460.vatikánsky koncil 23
461.technické vybavenie 23
462.tak dosť 23
463.súdnou cestou 23
464.starom kontinente 23
465.sklonenou hlavou 23
466.sama sebou 23
467.rohovom kope 23
468.ročných období 23
469.prvej ruky 23
470.priemyselných parkov 23
471.pokladničná poukážka 23
472.múzických umení 23
473.motorové vozidlá 23
474.miesta činu 23
475.manželského súhlasu 23
476.kultúrne stredisko 23
477.knihy rekordov 23
480.kandidačnej listine 23
481.Katolícka cirkev 23
482.Dopravný podnik 23
483.cudzí jazyk 23
484.bývalého režimu 23
485.Blízky východ 23
486.bankový sektor 23
487.zorného uhla 22
488.Žitného ostrova 22
489.zlatého fondu 22
490.životnú úroveň 22
491.vypísanie referenda 22
492.Visegrádskej štvorky 22
493.Varšavskej zmluvy 22
494.ustálené spojenie 22
495.studený front 22
496.Štedrý večer 22
497.rajčiakového pretlaku 22
498.poznávacou značkou 22
499.pracovné príležitosti 22
500.podnikateľské aktivity 22
501.periodickej tlače 22
502.otvorenými očami 22
503.obchodnom dome 22
504.nákladných áut 22
505.ľudskej bytosti 22
506.končekmi prstov 22
507.jedálneho lístka 22
508.jazykové spoločenstvo 22
509.druhého čítania 22
510.Červený kríž 22
511.čerpacích staníc 22
512.Bezpečnostná rada 22
513.Živý plameň 21
514.založenými rukami 21
515.večného života 21
516.Západná konferencia 21
517.základných školách 21
518.výtvarných umelcov 21
519.výberové konanie 21
520.vchodových dverách 21
521.ročné obdobia 21
522.pomarančovej šťavy 21
523.PEŠIA ZÓNA 21
524.miešaných nápojov 21
525.hrozí nebezpečenstvo 21
526.hladkej múky 21
527.hárok papiera 21
528.duševný stav 21
529.detských domovov 21
530.čerstvom vzduchu 21
531.Babie leto 21

535.zimnom štadióne 20
536.zemepisných šírkach 20
537.zdravotných sestier 20
538.vzdušného priestoru 20
539.výberovej komisie 20
540.Úprimne povedané 20
541.tvorivú silu 20
542.technický stav 20
543.tanečnej hudby 20
544.svetovou vojnou 20
545.spodnú bielizeň 20
546.pušného prachu 20
547.právny zástupca 20
548.požiarnej ochrany 20
549.Pokrčil plecami 20
550.plnom prúde 20
551.plné ruky 20
552.osobného vlastníctva 20
553.odkladacej mriežke 20
554.občianskej výchovy 20
555.ľudskej rasy 20
556.ľudskej dôstojnosti 20
557.dýchacích ciest 20
558.dozorných radách 20
559.doslovný preklad 20
560.dobrej povesti 20
561.dobrej forme 20
562.do bodky 20
563.červený posun 20
564.bytovú výstavbu 20
565.božskej prirodzenosti 20
566.židovský štát 19
567.záujmové skupiny 19
568.voľnou technikou 19
569.voľného obchodu 19
570.vládna garnitúra 19
571.verejného osvetlenia 19
572.územného celku 19
573.telesne postihnutých 19
574.psích záprahov 19
575.mliečnych výrobkov 19
576.materinský jazyk 19
577.ľudovú nôtu 19
578.Kreslené filmy 19
579.káblovej televízie 19
580.hraničný priechod 19
581.gravitačné vlny 19
582.finančné toky 19
583.cukrového sirupu 19
584.súkromné školy 19
585.Starú mater 19
590.mäkké podnebie 19
591.krvný tlak 19
592.hromadnej dopravy 19
593.zatajeným dychom 18
594.tmavé okuliare 18
595.tlakovej výše 18
596.ťažkú hlavu 18
597.ťažké zranenia 18
598.sväté prijímanie 18
599.súdneho procesu 18
600.Stratená ríša 18
601.stolnom tenise 18
602.spytovanie svedomia 18
603.školských zariadení 18
604.slamený klobúk 18
605.rozhlasových staníc 18
606.petržlenová vňať 18
607.opačného pohlavia 18
608.nočnej košeli 18
609.nadmorskej výšky 18
610.jazdeckého koňa 18
611.inak povedané 18
612.hotelovej izbe 18
613.holým nebom 18
614.dobré znamenie 18
615.bytových jednotiek 18
616.bodový zisk 18
617.bezpečnostných zložiek 18
618.voľnú ruku 17
619.vnútorný pokoj 17
620.vážnej hudby 17
621.Tiesňové volanie 17
622.telefónnom zozname 17
623.svetelné lúče 17
624.podal demisiu 17
625.obchodných reťazcov 17
626.miesto činu 17
627.ľudského pokolenia 17
628.koncertné turné 17
629.Jednoducho povedané 17
630.dopravnej nehody 17
631.diametrálne odlišné 17
632.divadelnej hry 17
633.biologických zbraní 17

586.psích záprahov 19
587.promile alkoholu 19
588.petičných hárkov 19
589.občianskej vojne 19

B Zoznam vytriedených trigramov

1. v poslednom čase 305
2. spišská nová ves 271
3. v žiadnom prípade 203
4. druhej svetovej vojny 154
5. za každú cenu 125
6. ministerstva zahraničných vecí 120
7. a tak ďalej 117
8. fondu národného majetku 110
9. v krátkom čase 108
10. znova a znova 106
11. všetko v poriadku 104
12. v trestnom konaní 94
13. krok za krokom 92
14. v opačnom prípade 90
15. do značnej miery 87
16. hrubého domáceho produktu 82
17. skôr či neskôr 74
18. ministerstvo zahraničných vecí 71
19. v bezprostrednej blízkosti 71
20. viac či menej 67
21. tak či onak 61
22. trest odňatia slobody 60
23. druhej svetovej vojne 60
24. brať do úvahy 60
25. reformy verejnej správy 57
26. tu a tam 53
27. zo všetkých síl 51
28. za normálnych okolností 51
29. máme do činenia 50
30. deň čo deň 50
31. pod vplyvom alkoholu 46
32. z prvej ruky 43
33. neprichádza do úvahy 43
34. tatranskej horskej služby 42
35. v priamom prenose 40
36. pod jednou strechou 40
37. prieskumov verejnej mienky 37
38. v pravý čas 33
39. mleté čierne korenie 32
40. tvárou v tvár 31
41. so zatvorenými očami 31
42. vezmeme do úvahy 30
43. z druhej ruky 28
44. strechu nad hlavou 28
45. vo verejnom záujme 27
46. z celého srdca 26
47. pravom zmysle slova 24
50. burze cenných papierov 23
51. zbraní hromadného ničenia 23
52. za bieleho dňa 22
53. so založenými rukami 22
54. za zatvorenými dverami 21
55. základných ľudských práv 20
56. vo vlastných rukách 20
57. so sklonenou hlavou 20
58. neberie do úvahy 19
59. ani vo sne 19
60. so zatajeným dychom 19
61. pokrčila som plecami 18
62. plné ruky práce 17
63. za okrúhlym stolom 16
64. v drvivej väčšine 14
14

48. z jasného neba 23
49. spáchania trestného činu