Literatúra

CVRČEK, Václav a kol.: Mluvnice současné češtiny. I. Praha: Nakladatelství Karolinum 2010. 354 s.
Český národní korpus – úvod a příručka uživatele. Ed. J. Kocek – M. Kopřivová – K. Kučera. Praha: Ústav Českého národního korpusu FF UK 2000. 159 s.
Frekvenční slovník češtiny. Ed.: F. Čermák – M. Křen. Praha: Nakladatelství Lidové noviny 2004. 596 s.
GAJDOŠOVÁ, Katarína: Lingvistická analýza automatizovanej anotácie datívu adjektíva v heterogénnych menných skupinách v Slovenskom národnom korpuse. In: Slovo – Tvorba – Dynamickosť. Ed.: M. Šimková. Bratislava: Veda 2010, s. 309 – 330.
Gramatika a korpus / Grammar & Corpora 2005. Ed. F. Štícha – J. Šimandl. Praha: Ústav pro jazyk český AV ČR 2007. 304 s.
Korpus jako zdroj dat o češtině. Ed. P. Karlík. Brno: Masarykova univerzita 2004. 208 s.
Možnosti a meze české gramatiky. Ed. F. Štícha. Praha: Academia 2006. 304 s.
Studie z korpusové lingvistiky. Ed. F. Čermák – J. Klímová – V. Petkevič. Acta Universitatis Carolinae. Philologica 1997, č. 3 – 4. Praha: Karolinum 2000. 531 s.
ŠIMKOVÁ, Mária: Výberový slovník termínov z počítačovej a korpusovej lingvistiky. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity 2006, s. 187 – 193. Dostupný aj na www: http://korpus.juls.savba.sk/publications/block1/2006-simkova-vyberovy%20slovnik%20terminov/2006-simkova-vyberovy%20slovnik%20terminov.pdf
ŠULC, Michal: Korpusová lingvistika. První vstup. Praha: Karolinum 1999. 94 s.

KORPUSOVÝ POHĽAD NA SPÁJATEĽNOSŤ SLOV

(Kolokace. Studie z korpusové lingvistiky. Svazek 2. Ed. F. Čermák, M. Šulc. Praha: Nakladatelství Lidové noviny – Ústav Českého národního korpusu 2006. 452 s. ISBN 80-7106-863-2)

Daniela Majchráková
Slovenský národný korpus Jazykovedného ústavu Ľudovíta Štúra SAV
Panská 26, 813 64 Bratislava
e-mail: danam@korpus.sk

V roku 2006 vyšiel druhý zväzok Studií z korpusové lingvistiky s názvom Kolokace. Táto zbierka 16 štúdií je prvým českým zborníkom venovaným danej tematike, hoci čiastkové výstupy sa objavili už skôr.[1]

V úvodnom príspevku *Kolokace v lingvistice* (s. 9 – 16) František Čermák čitateľa oboznamuje so základnými poznatkami týkajúcimi sa kolokácií, čím mu poskytuje potrebnú teoretickú bázu pre jednotlivé štúdie zamerané na čiastkové otázky. Kolokácie autor považuje za kľúč k poznávaniu spôsobov, pravidiel a hraníc správa-

[1] Spomeňme napríklad štúdiu F. Čermáka Syntagmatika slovníku: typy lexikálních kombinací. In: Čeština – univerzália a specifika 3. Eds. Z. Hladká, P. Karlík. Brno: Vydavatelství Masarykovy univerzity 2001, s. 223 – 232.

nia sa lexém a definuje ich ako zmysluplné spojenia slov, ktorých vznik je podmienený vzájomnou kolokabilitou (spájateľnosťou) a kompatibilitou (sémantickou viazanosťou). Každé slovo slovnej zásoby je jedinečné tým, že sa spája so špecifickou skupinou slov (kolokátov), s nimi vytvára kolokácie, ktoré spolu tvoria jeho kolokačnú paradigmu. F. Čermák predostiera svoje členenie kolokácií ako lexikálnych kombinácií vyskytujúcich sa v textoch a vyčleňuje niekoľko typov spojení od ustálených, sémanticky zviazaných (*cestovní kancelář, ležet ladem, černá díra*) až po voľné, náhodné spojenia (*virové hrátky, třeskutě vtipný*). Pojem kolokácie sa podľa autora spája s korpusmi textov, z ktorých sú identifikované na základe rôznych metód korpusovej lingvistiky. Vzájomný spoluvýskyt dvoch slov sa vyhodnocuje pomocou štatistických mier ako MI-score, t-score, Z-score, cost-criterion a ďalších. Z nich najčastejšie používané asociatívne miery na skúmanie spájateľnosti dvoch slov sú MI-score a t-score. Kým MI-score, miera vzájomnej informácie, nachádza najmä zriedkavejšie spojenia, t-score, miera kontrastu, je metóda vhodná pre frekventovanejšie slová.[2]

Nasledujúce príspevky zborníka sme v našom prehľade rozdelili podľa tematických okruhov. Korpusovým metódam identifikácie kolokácií v textoch sa v zborníku podrobne venujú autori troch štúdií. Analýzu kolokačných mier predstavil v príspevku *Kolokační míry a čeština: srovnání na datech Českého národního korpusu* (s. 223 – 248) Michal Křen. Autor sa zameral na šesť štatistických mier, a to MI-score, t-score, log-likelihood, dice, odds a X^2. Jeho výskumná metóda spočívala v aplikácii jednotlivých mier na bigramy obsahujúce vybrané kľúčové slová (päť slovných tvarov: *hlava, jazyk, křížem, když, jako* a päť lem: *nebezpečí, odpověď, pozornost, ohled, doprava*). Na základe analýz rozdelil kolokačné miery do troch skupín: prvú tvoria miery MI-score a odds, ktoré nachádzajú kolokácie výnimočné, málo frekventované, často však aj nesystémové (*zbystřit pozornost, křížem nařízneme*), v druhej skupine sú t-score a log-likelihood, ktoré pomáhajú identifikovať veľmi frekventované a často systémové kolokácie (*hlava nehlava, křížem krážem*). Tretia skupina štatistických mier, dice a X^2, tvorí podľa autora umiernený prechod medzi dvoma predchádzajúcimi krajnými pólmi a môže byť vhodná na hľadanie kolokácií. Jedným z možných prístupov je aj kombinácia jednotlivých mier, napríklad MI-score a t-score, typická pre lexikografiu.

Cieľom štúdie Václava Cvrčka *Metoda zjišťování kolokační platnosti frekventovaných bigramů pomocí ranku* (s. 36 – 55) bolo ukázať spôsob ako odlíšiť kolokácie, ktoré majú pevnejšiu väzbu danú systémovými jazykovými vzťahmi, od frekventovaných kombinácií slov, spojení bez týchto väzieb. Autor zisťoval kolokačnú

[2] Bližšie v publikácii ČERMÁK, František – BLATNÁ, Renata a kol.: Jak využívat Český národní korpus – studijní příručka. Praha: Lidové noviny – Ústav Českého národního korpusu 2005 alebo na http://ucnk.ff.cuni.cz/.

platnosť vzorky sto najfrekventovanejších bigramov pomocou tzv. substitučnej metódy. Vychádzal z predpokladu, že pokiaľ je daný bigram kolokáciou a má v korpuse určitú frekvenciu, potom pozmenený bigram, vytvorený napríklad substitúciou, by mal mať frekvenciu výrazne odlišnú. Skúmal a porovnával štatistické miery kolokačných párov *cestovní kancelář – cestovní ruch, mít čas – trávit čas, milion korun – milion dolarů, valná hromada – valné shromáždění*. Podľa autora je možné túto metódu aplikovať len na veľmi frekventované bigramy.

V ďalšom príspevku *Statistické metody hledání frazémů a idiomů v korpusech* (s. 94 – 106) sa František Čermák zaoberal možnosťami identifikácie frazeologických jednotiek v korpuse. Vytvoril zoznam automaticky vygenerovaných bigramov typu substantívum – adjektívum z korpusu SYN2000. Daný výstup následne filtroval podľa hodnoty MI-score. Ručným triedením lingvistického materiálu zistil, že frazémy je možné nájsť na celej škále MI-score. Dôvodom je fakt, že frazémy obsahujúce veľmi frekventované komponenty, dosahujú nízke hodnoty MI-score, ako napríklad frazémy s častým komponentom *hlava* (frazéma *čistá hlava* preto dosahuje hodnotu MI-score 2,617). Na identifikáciu frazém danej adjektívno-substantívnej štruktúry autor považuje za vhodné pásmo MI-score 18 až 7, pričom na úrovni MI-score 7 je možné pozorovať výrazný prechod od frazém k bežným kolokáciám.

Anna Čermáková v štúdii *Kolokace a valence některých případů substantiv* (s. 107 – 141) sa pokúsila opísať valenciu českých substantív z lexikologického hľadiska. Autorka neopisuje valenciu iba z formálneho hľadiska, ale všíma si aj javy s ňou spojené, napríklad kolokačné profily slov, ktoré vychádzajú z analýzy korpusových dát. V príspevku sa venuje trom okrajovým javom valencie substantív: koordinovaným substantívnym výrazom (*civilizace a demokracie, tolerance a respektování*), nominatívnej substantívnej valencii, ktorú tvoria dve po sebe nasledujúce substantíva s rovnakým denotátom (*chudák holka*), konkurencii genitívu a atributívneho adjektíva pri substantívach (*porost trávy – trávní porost*) a niekoľkonásobnej genitívnej valencii. V prípade tejto valencie zistila, že spoločnou charakteristikou analyzovaných slov je ich sémantická nesamostatnosť a silná tendencia ku genitívnej valencii (*vyhlášení stávkové pohotovosti od příštího týdne*). Štúdia predstavuje len čiastkový výskum, v neskoršej syntetickej práci (Valence českých substantiv, 2009), ktorej recenziu pripravujeme, podala autorka rozšírený výklad tejto problematiky.

Ako uvádza Milena Hnátková, cieľom jej príspevku *Typy a povaha komponentů neslovesných frazémů z hlediska lexikálního obsazení* (s. 142 – 167) bolo podrobiť analýze kombinácie slovných druhov v ustálených spojeniach, prevažne frazémach, z hľadiska slovnodruhového a lexikálneho obsadenia. Vychádzala zo zoznamu slovných spojení uvedených v Slovníku českej frazeológie a idiomatiky (SČFI). Najprv sa autorka venovala základným pojmom ako frazéma, kolokabilita, monokolokabi-

lita, kompatibilita a polysémia, ktoré vysvetľovala na príkladoch adjektív. V osobitnej časti príspevku sa venovala adjektívam ako komponentom frazém a rozdelila ich do niekoľkých sémantických skupín. Nakoniec vytvorila niekoľko slovnodruhových štruktúr na základe lexikálneho obsadenia frazém, konkrétne substantívne konštrukcie (*staré zlaté časy, agent provokatér, plnou parou vpřed*) a koordinácie (*nebe a dudy, oáza klidu a míru, tělem i duší*). Výsledkom tejto sondy bolo konštatovanie, že automatické spracovanie textu nám dáva možnosť automatického vyhľadávania frazém, ale ich automatická identifikácia je často nespoľahlivá.

Problematike lexikografického zachytenia kolokácií sa venovali autori Aleš Klégr a Pavlína Šaldová v štúdii *Kolokační faux amis* (s. 168 – 177). Nastolením otázky kolokačných faux amis (tzv. zradných kolokácií ako translatologického problému) podčiarkujú potrebu skúmať prekladovú ekvivalenciu predovšetkým na rovine syntagmatických jednotiek. Rozlišujú tri typy kolokačných faux amis: transpozičné, na základe typu transpozície, napr. *třít bídu*/*be poor*, ktoré vyžadujú slovnodruhový posun; modulačné, ktoré vyžadujú lexikálnu obmenu, napr. *vykročit pravou nohou*/*put best foot forward*; a adaptačné, ktoré je potrebné zásadne preformulovať, napr. *být solí v očích někomu*/*be a thorn in his flesh*. Samotní autori považujú toto delenie za hrubé a neúplné, možno ho použiť ako základ na podrobnejšiu analýzu zradných kolokácií a hľadanie ich ekvivalentov. Rovnako poukazujú na to, že na rozdiel od „zradných slov" tvoria kolokačné faux amis dynamické celky, ktoré je v texte zväčša ťažké identifikovať.

Adverbiálnym kolokačným typom sa venovala v štúdii *Kolokace některých intenzifikačních adverbií* (s. 178 – 222) Marie Kopřivová[3]. Najprv skúmala postavenie adverbií medzi ostatnými slovnými druhmi, potom na základe frekvencie pripravila výber intenzifikačných adverbií, pri ktorých sledovala spájateľnosť s adjektívom a adverbiom. Vyhľadávané spojenia rozdelila do dvoch skupín. Prvú skupinu tvoria napríklad adverbiá *daleko, ohromně, šíleně, náramně*, ktoré sa objavujú s adjektívom alebo adverbiom menej než v polovici svojich výskytov. Druhú skupinu tvoria adverbiá ako *velmi, velice, mnohem, značně, podstatně*, ktoré sa prevažne spájajú s adjektívami a adverbiami. Zamerala sa tiež na typické kolokáty podľa hodnôt MI-score a t-score. Na základe analýz autorka napríklad zistila, že adverbiá *mnohem, daleko, podstatně* vyžadujú adverbium alebo adjektívum v komparatíve, adverbiá *náramně, ohromně, navýsost, neskonale* sa spájajú s kolokátmi s kladnými významami, adverbiá *značně, strašlivě, příšerne* sa spájajú s negatívnymi významami. Porovnaním kolokátov tvarov *hluboko – hluboce*, *vysoko – vysoce* a *široko – široce* zas zistila, že tvary *hluboko*, *vysoko* majú konkrétne významy, zatiaľ čo *hluboce*, *vysoce*, *široce* majú významy abstraktné.

[3] M. Kopřivová je aj autorkou publikácie Valence českých adjektiv (2006), ktorá je štvrtým zväzkom edície Studie z korpusové lingvistiky.

Problematikou lexikálnej sémantiky sa zaoberá František Čermák v štúdii *Polysémie a kolokace: případ adjektiva měkký* (s. 56 – 93). Ako autor uvádza, jej cieľom je kritický náhľad na hlavné problémy lexikografickej praxe vyplývajúce z potreby opísať významy polysémických lexém. Vychádza z poznatku, že všetky potrebné informácie o význame slova sú uložené v jeho syntagmatike, v kontextovom použití slova. Analyzuje konkordancie adjektíva *měkký* a v nich obsiahnuté kolokácie, pričom sa zameriava na kolokabilitu adjektíva s rôznymi typmi substantív. Výsledkom tohto rozboru je sémantický profil slova pozostávajúci zo štyroch skupín substantív, a teda štyroch významov adjektíva *měkký*. Autor zdôrazňuje, že k dokonalému spracovaniu sémantiky lexémy patrí i spoľahlivá kvantitatívna informácia, predovšetkým frekvencia.

Nasledujúce tri príspevky obsahovali rozbor rôznych adjektívnych kolokácií. Pavel Vondrička sa v štúdii *Významová spojitelnost adjektivních opozit* (s. 403 – 452) zaoberá štatistickým prehľadom najfrekventovanejších spojení opozitných stupňovateľných adjektív so substantívami. Vybral si niekoľko centrálnych párov adjektív s najvyššou frekvenciou výskytu vo Frekvenčnom slovníku češtiny (2004), napr. *veľký – malý, dobrý – zlý / špatný, dlhý – krátky, starý – mladý,* a zisťoval v korpuse SYN2000, s akými substantívami sa uvedené adjektíva najčastejšie spájajú. Na základe tejto analýzy chcel postihnúť skutočnú šírku významu jednotlivých adjektív a porovnať, nakoľko a v akom kontexte ide skutočne o opozitá a do akej miery sa použitie týchto adjektív líši na základe významu a kontextu. Spájateľnosť jednotlivých adjektívnych opozít analyzoval z hľadiska ich ustálenosti a sémantickej a pragmatickej motivácie.

Cieľom príspevku Martina Stluku s názvom *Kolokace lexémů vel(i)ký a malý v nejstarší staročesky psané próze* (s. 362 – 373) bolo hľadanie kolokačných vzťahov adjektív *velký (veliký)* a *malý* vrátane ich derivátov v troch písomných pamiatkach staročeského písomníctva: v Pasionáli, v Živote Krista Pána a v Životoch svätých. Všetky tri rukopisy sú súčasťou diachrónnej časti Českého národného korpusu a sú čiastočne ručne lematizované. Autor určoval lexikálnu frekvenciu jednotlivých odvodenín daných adjektív (*maličký, malichný, větší*) a následne analyzoval kolokácie a ich distribúciu v textoch (*veliký mor, veliká pohroma, veliké utrpení*). Autorovi sa podarilo identifikovať aj niekoľko ustálených spojení v češtine z prelomu 14. storočia, ktoré mal podložené výraznejšou frekvenciou výskytu (*Veliká noc, malá hodinka*).

Věra Schmiedtová sa v štúdii *Volná a vázaná spojitelnost/kolokabilita názvů barev a jejich odstínů v češtine: analýza na základe ČNK* (s. 311 – 361) zamerala na kolokabilitu slov označujúcich farby, konkrétne spojení ich adjektívnych označení s adverbiami. Autorka zistila, že pri vyjadrovaní farebnej sýtosti sa používajú adverbiá ako *temně, bledě, sytě, měkce*; pri vyjadrovaní intenzity farby sa používajú ad-

verbiá *oslnivě, zářivě, matně, tlumeně* a pod. Pri vybraných ustálených spojeniach označujúcich farbu analyzovala, s akými objektmi sa spájajú. Vytvorila štyri kategórie: *člověk, příroda, výrobky* a *abstraktá*. V prípade kategórie človek zistila, že niektoré pomenovania smerujú k monokolokabilite, napríklad adjektíva *blond, blonďatý, snědý* a *prošedivělý* sa prevažne spájajú so substantívom *muž*. V poslednej časti príspevku sa autorka venovala prirovnaniam typu *mající barvu jako* a zistila, že k nim hovoriaci pristupuje tvorivo a okrem lexikalizovaných spojení nachádzala množstvo voľne utvorených spojení (*růžový jako dítě, hnědý jako čokoláda)*.

Spájateľnosti predložiek sa venovali dvaja autori. Vladimír Petkevič (s. 262 – 310) analyzoval *Složité předložkové skupiny (kolokace předložek a jmen)*. Ide o predložkové skupiny, ktoré obsahujú nominálne skupiny, kde po predložke nasleduje ako prvé meno v inom páde, než aký predložka valenčne vyžaduje, napr. *na deštěm smáčaném hřišti, ke státem řízenému hospodářství, k vámi citovanému bodu*. Cieľom autora bolo vytvoriť klasifikáciu týchto predložkových skupín na účely automatickej dezambiguácie textov v korpusoch češtiny, najmä na správnejšie určenie pádov mien. Jednotlivé typy podrobne klasifikoval, vytvoril hlavné štruktúrne typy zložitých predložkových skupín podľa jednotlivých pádov a každý typ doložil príkladmi z korpusu. Tieto poznatky boli použité na vytvorenie pravidiel, zapísaných v špeciálnom programovacom jazyku, na rozpoznávanie takýchto skupín slov v korpusoch.

Problematike viacslovných predložiek sa venovala Renata Blatná v štúdii *Víceslovné předložky ve funkci víceslovných spojek* (s. 17 – 25). Skúmala slovnodruhový prechod predložiek typu *na základě něčeho*, *vzhledem k něčemu*, *v souvislosti s něčím*, *nehledě na něco*. Podľa autorky v tomto prípade hovoríme o extenzii javu, ktorého úzus je príznačný najmä pre odbornú literatúru. Slovnodruhový prechod je typický tým, že valenčná pádová pozícia viacslovnej predložky sa obsadzuje najčastejšie zámenom *ten* v neutre singuláru v príslušnom páde a za ňou nasleduje klauza uvedená spojkou alebo iným spájacím výrazom, napr. *na základě toho, že …; vzhledem k tomu, že …; v souvislosti s tím, zda …; nehledě na to, že…* Autorka sa venovala podrobnej frekvenčnej analýze piatich vybraných štruktúrnych typov viacslovných predložiek. Jedným zo zistení je fakt, že najčastejším spájacím výrazom na začiatku vedľajšej klauzy je *že*, potom *který, co, kdo*, a v prípade adverbií *kdy* a *jak.*

Michal Šulc sa v príspevku *Slovní asociace versus korpusové kolokac*e (s. 374 – 402) snažil dokázať, že nejde o totožné pojmy. Slovné asociácie chápal z psycholingvistického hľadiska ako slovné spojenia získané na základe asociačného testu a textové kolokácie ako štatisticky signifikantné spojenia slov. V svojom výskume porovnával zoznam kolokácií vygenerovaných a zoradených podľa hodnoty miery MI-score so zoznamom slovných asociácií získaných experimentálne na vzorke 50 respondentov, ktorí mali zaznamenávať asociácie k 150 podnetovým slovám. Autor zistil, že 35 % slovných asociácií z prvých miest zoznamu sa nevyskytovalo ani

v strednom pásme kolokačného zoznamu (v niektorých prípadoch dosahovali hodnotu MI-score menšiu ako 6). Vzhľadom na výraznú diskrepanciu autor tvrdí, že slovné asociácie nemožno stotožňovať s textovými kolokáciami.

Poslednou štúdiou je práca Karla Kučeru *Neustálené kolokace tvarů s provedenými a neprovedenými hláskovými změnami ve starších českých textech* (s. 249 – 261). Jedným z rysov staročeských textov je, že vedľa seba stoja tvary s uskutočnenými aj s neuskutočnenými hláskovými zmenami typu *učenej vlaský král*, *takovou nepruvodnú věc*. Autor si kládol otázku, či usporiadanie vývojovo starších a vývojovo mladších foriem v rámci jednej kolokácie je náhodné alebo súvisí s ich vlastnosťami? Vychádzal pri tom zo štyroch alternatívnych hypotéz, ktoré sa týkali preferencií starších alternantov proti mladším, zadnejších proti prednejším, nižších proti vyšším a diftongických proti monodiftongickým. Na výskum použil vzorku kolokácií, ktoré boli vyextrahované z diachrónnej časti Českého národného korpusu. Kvantitatívnou analýzou 401 neustálených kolokácií obsahujúcich páry hláskových alternácií *ě – e, ó – uo, ý – ej, ú – ou, ie – í, uo – ú, é – í* zistil, že výber hláskových alternantov v kolokáciách nie je úplne náhodný a súvisí s tendenciou používať v prvej zložke kolokácie nižšie hláskové alternanty.

Tematický zborník Kolokace nám poskytuje prehľad výsledkov v oblasti výskumu spájateľnosti slov v češtine. Jednotlivé štúdie môžu byť inšpiratívne pre ďalšie bádanie v tejto oblasti aj v iných jazykoch, najmä poukázaním na možnosti práce s korpusovými databázami a štatistickými nástrojmi. Na Slovensku sa výskum kolokácií v súčasnosti orientuje skôr lexikograficky, jedným z jeho výsledkov je Slovník slovenských kolokácií (pripravovaný od roku 2008), ktorý bude čoskoro dostupný v tlačenej i elektronickej podobe. V slovníku sa spracúvajú kolokačné profily najfrekventovanejších substantív slovenčiny vytvorené pomocou štatistických nástrojov, z ktorých najefektívnejší je Sketch Engine. V čase zostavovania zborníka Kolokace tento nástroj ešte nebol k dispozícii.