FILOLOGIJA 30—31, Zagreb 1998

UDK 801.3
Izvorni znanstveni članak
Primljen 5. V. 1998.
Prihvaćen za tisak 15. VI. 1998.

Boris Pritchard
Pomorski fakultet
Studentska 2, HR-51000 Rijeka

O KOLOKACIJSKOM POTENCIJALU RJEČNIČKOG KORPUSA

Kolokacije, njihovo utvrđivanje u korpusu te način prezentiranja na razini rječničke makrostrukture i leksikografske natuknice, jedno su od najaktualnijih pitanja suvremene leksikografije. Tom se pitanju pridaje nejednako značenje u općim rječnicima, ali je posebice važno u stručnim rječnicima. Rad se bavi mjerilima utvrđivanja i odabira kolokacija kao jednog od formalnih oblika višečlanih leksičkih jedinica u rječničkom korpusu za jednojezične i dvojezične (englesko-hrvatske) opće i stručne rječnike. Na temelju korpusa *Bank of English* (*COBUILD-Collins*) dani su primjeri odabira kolokacija s istim glavnim leksičkim jedinicama za opće i stručne rječnike.

Uvod

Kolokacije su jedno od otvorenih teoretskih i praktičkih pitanja moderne leksikografije. To se odnosi prvenstveno na problem njihove definicije, formalna sredstva za njihovo identificiranje i razvrstavanje, kao i na pitanje koje vrste kolokacija valja uvrstiti u rječnik odnosno u koje vrste rječnika.

U leksikografskom smislu kolokacije su posebna vrsta višečlanih leksičkih jedinica. Termin *višečlana leksička jedinica* (multi-word lexical unit, multiwords) jest opći leksikografski naziv za sve sintagmatskim odnosima nastale leksičke kombinacije[1], koje se u rječniku obrađuju na razini natuknica (headword), postajući tako sastavnim dijelom rječničke nomenklature, ili na razini podnatuknica unutar članka za pojedine natuknice. Nerijetko se javljaju i na obje razine. Kolokacije su, nadalje, specifične za pojedine jezike pa stoga u različitim jezicima imaju i različit leksikografski status. Važnost kolokacija, također, nije jednaka za opće u odnosu na stručno obilježene rječnike.

[1] O nazivu i značenju termina *višečlana leksička jedinica* (*multi-word lexical units*) vidi Zgusta 1970, Ivir 1974, Landau 1987, Cowie 1992, Štambuk 1990 itd).

Uz klasične izvore i metode utvrđivanja i uvrštavanja kolokacija i drugih višečlanih leksičkih jedinica u rječnike, koje su uglavnom bile proizvoljne i ovisile o jezičnoj intuiciji i lingvističkom pristupu sastavljača rječnika, danas se za njihovu identifikaciju i uvrštavanje koriste (a) konkordancije odnosno isječci stvarnoga teksta s odabranom riječju (node) i (b) strojno čitljive leksičke baze podataka iz jednojezičnih, a danas sve češće i višejezičnih leksikografskih korpusa. Na temelju leksikografskog vrednovanja zasnovanoga na lingvističkom znanju i jezičnoj intuiciji sastavljača iz prvih, kao i posebnim matematičkim postupcima (algoritmima) iz drugih, dobivaju se kolokacije, tj. kombinacije riječi koje su povezane izvjesnim stupanjem očekivana međusobna pojavljivanja (cf. *mutual expectancy*, Stubbs 1996).

Uz prikaz teoretskih razmatranja o kolokacijama u ovom radu bit će riječi o kolokacijskom potencijalu, tj. o zastupljenosti kolokacija u rječnicima i njihovim leksičkim bazama podataka te o nekim načinima računalom podržanoga generiranja kolokacija, s ciljem njihova leksikografskoga vrednovanja i uvrštavanja u rječnik.

2. Kolokacije i višečlane leksičke jedinice

Kao što je već napomenuto, usprkos mnogim istraživanjima, lingvističkih i leksikografskih (posebice korpusnih), teško je ili gotovo nemoguće doći do obuhvatnije i opće prihvatljive definicije kolokacije. To je posljedica različitih pristupa istraživanju kolokacija, ali i različite tipološke prirode pojedinih jezika. Osim toga u leksikografiji se danas, posebice u korpusnoj, kolokacije povezuju ili čak izjednačuju s pojmom višečlanih leksičkih jedinica. Istraživanje višečlanih leksičkih jedinica, pa samim time i kolokacija, primjerice, lakše je i stoga češće u jezicima s jednostavnijom morfološkom strukturom (npr. u engleskom jeziku). Definiciji kolokacija najčešće se pristupa kao leksičkoj svezi nastaloj iz tri vrste ograničenja (a) iz specifičnog semantičkog odnosa dviju riječi, tj. strukture značenja članova kolokacije (collocational meaning), (b) iz distribucijskih odnosa članova kolokacija (collocational range) i (c) tzv. *pravih* kolokacijskih odnosa koji se ne mogu svrstati pod nijedan od gornjih odnosa (cf. Palmer 1981:77–79). Upravo taj, treći odnos, prema kojemu su kolokacije proizvoljne i neovisne o značenju, upućuje na nezaobilaznu potrebu proučavanja leksičkih sveza i značenja na tzv. *kolokacijskoj razini* budući da dio značenja leksema ovisi »ne o njihovoj funkciji u određenom kontekstu situacije, već o njihovoj sklonosti za supojavljivanjem u tekstu« (Lyons 1977:612). Istraživanja utemeljena na firthijanskim i hallidayanskim teoretskim postavkama provode se u posljednjih petnaestak godina u korpusnoj lingvistici pa su njihovi rezultati značajni za suvremenu leksikografsku teoriju i praksu.

Prema Benson–Benson–Ilsonu (1986) kolokacije idu u red višečlanih le-

ksičkih jedinica, koje se razvrstavaju od najslobodnijih otvorenih sveza ili slobodnih kombinacija (*free combinations*), preko kolokacija (*collocations*), prijelaznih sveza (*transitional combinations*), do idioma i zatvorenih složenica (*compounds*).

Otvorene sveze, odnosno slobodne ili otvorene kombinacije riječi (cf. unrestricted collocates, Carter 1987, Benson–Benson–Ilson 1986) najpogodnije su za leksikografsko oprimjerivanje uporabe leksema. Zbog svoga idiosinkrazijskog karaktera one omogućuju širok i najslobodniji izbor drugih leksičkih jedinica što se uz njih supojavljuju, pa su stoga važan čimbenik leksičkoga stvaranja (Lyons 1995:535 takve leksičke kombinacije naziva sintaktičkim složenicama, *syntactic compounds*). Njihovi su članovi najmanje semantički povezani i međusobno ovisni, što se očituje na prozodijskoj razini, npr. u jakom, primarnom naglasku na svakom članu sveze (u općem jeziku: *big house, small house, country house,* ili u stručno obilježenu vokabularu: *dry cargo, liquid cargo, dangerous cargo*). Čestota uporabe slobodnih kombinacija, međutim, dovodi do povećanog stupnja povezanosti značenja njihovih članova, a time i do manje paradigmatske zamjenjivosti (a to je obilježje kolokacija). Zbog beskonačnih mogućnosti njihova stvaranja u pravilu se slobodne leksičke kombinacije ne uvrštavaju u nomenklaturu općih. Bez obzira na još otvorena teoretska pitanja njihova definiranja i utvrđivanja, a zbog jačeg stupnja leksikalizacije, kolokacije imaju opravdana mjesta u rječnicima, posebice u općim dvojezičnim rječnicima, a nezaobilazne su u stručnim dvojezičnim rječnicima.

Složenice, čiji su semantička priroda i leksikografski status, posebice u engleskom jeziku, također otvoreni raznim tumačenjima, kako leksikografska praksa pokazuje, obavezno se uvrštavaju u rječničku nomenklaturu, tj. među (glavne) natuknice (main entry, headword), dok ustaljeni izrazi (fixed phrases) i idiomi imaju mjesto u frazeološkom ili nekom drugom, sekundarnom bloku članka natuknice.

Kod **kolokacija**, leksičkih sveza nastalih više ili manje predvidljivim sintagmatskim supojavljivanjem riječi, čija su osnovna obilježja opetovanost (rekurentnost) i ustaljenost, zamjenjivost jednog od elemenata kolokacije, za razliku od otvorenih kombinacija, mnogo je niža, zbog ograničenja na izbor mogućih riječi što s drugim članom kolokacije stoje u sintagmatskim odnosima. Ta ograničenja na neki način određuju značenje drugoga člana kolokacije (Ivir–Tanay 1976), s time što značenja svakog člana kolokacije i dalje ostaju prozirna. Kolokacijska ograničenja (collocational restrictions), kao jedno od uvjeta istoznačnosti (Lyons 1995:62) te kao važan diskursni čimbenik (cf. opetovanost i kolokacije kao elemente leksičke kohezije, Halliday–Hasan 1976), određuju i kolokacijski raspon pojedinih riječi, odnosno skup konteksta u kojemu se mogu pojaviti (collocational range). Riječi općenita značenja imaju veći kolokacijski raspon od riječi specifična, užega značenja (npr. u engleskom glagol *bury*, 'zakopati', uzima za svoj objekt više imenica, e.g. *man*,

treasure; head, face, feelings, memories, negoli npr. glagol *entomb* ili *inter*). Isto je i s višeznačnim riječima (*pay: money, freight, cheque, bill*, ali i *pay: attention, last respects, a visit*). I vrlo kratka usporedba s hrvatskim jezikom upućuje na distribucijske razlike odgovarajućih imenica s glagolima 'zakopati' (*čovjeka, *glavu, lice?**, *osjećaje, sjećanja*) i 'platiti' (*?novac, vozarinu, ček, račun; *pozornost*).

Kolokacije se razlikuju od slobodnih leksičkih sveza po tome što su semantička ograničenja kod kolokacija proizvoljna i specifična za pojedini jezik i što ne proizlaze iz propozicijskoga značenja riječi (Baker 1992:14), te su stoga zanimljive za dvojezičnu leksikografiju. Tako se npr. pojam za isti sadržaj/referent 'pokvariti se' (*go bad, stale, rotten*) u različitim uvjetima/ograničenjima leksikalizira različito. Primjerice u engleskom jeziku prisutnost različitih riječi uvjetuje različit izbor kolokata, kao i obrnuto: značenje riječi uvjetovano je izborom kolokata: cf. *eggs — addled; butter — rancid*). Za razliku od engleskog u hrvatskom je leksem *pokvaren* prihvatljiv kolokat s imeniciom *jaje* kao i s imenicom *maslac*. U engleskom su **rancid egg* odnosno **addled butter* neprihvatljive ili barem neprikladne, malo vjerojatne kolokacije. Slično je i s tzv. otvorenim kolokacijama (flexible collocations, Smadja et al.1996), npr. *pay attention* prema *posvetiti pozornost/pažnju* u hrvatskom, ali npr. kolokacije *draw attention, attract attention* formalno korespondiraju u oba jezika (*privući pažnju/pozornost)* dok se *catch attention* može svrstati u red *forsiranih* kolokacija (*marked* — Baker 1992:51) ili otvorenih leksičkih sveza.

Kolokacije, kao što je spomenuto, ovise i o prirodi pojedinih jezika (language-specific). Hrvatski jezik npr. uz ekvivalente spomenutih riječi (*maslac* i *jaja*) bira samo jedan kolokat (*pokvaren, pokvariti se*), iako su u nekim registrima (razgovorni hrvatski jezik) mogući i drugi leksemi (*gramzljiv*). Nadalje, *zakoni* se u hrvatskom *donose*, a u engleskom *prosljeđuju* (*pass*), *teretnica* se *predočuje* (na provjeru), roba ili teret se *predaje/isporučuje* ili *preuzima*. U engleskom se za isti pojam uz termin *bill of lading* supojavljuje drugačiji glagol (*produce*, ev. *present*) itd. Opis na kolokacijskoj razini ukazuje na značajne razlike između hrvatskog i engleskog jezika, te se proučavanje kolokacija i njihovih prijevodnih ekvivalenata nameće kao jedna od najvažnijih zadaća u obosmjernim rječnicima ta dva jezika. To je naročito važno u dvojezičnim općim i stručnim rječnicima, posebice rječnicima ili glosarima pravnog ili poslovnog nazivlja. Osim razlika uvjetovanih nepostojanjem formalne korespondencije na kolokacijskoj razini, što je čest uzrok stvaranju lažnih parova ili neprikladnosti leksičkog odabira u prevođenju, proučavanje kolokacija u ta dva jezika dokazuje da se pojedini pojmovi (koncepti) različito leksikaliziraju i gramatikaliziraju u engleskom i hrvatskom. Tako se npr. mnogi pojmovi iz poslovnog, pomorskog ili pravnog jezika u engleskom leksikaliziraju nominalizacijom u kolokacije, najčešće s "polugramatičkim" glagolima *make i take* i leksičkom riječju (imenicom) kao objektom: *make arrangements, make an assessment, make a claim, make delivery of, make use of; take delivery of* (koji se,

premda rječe, leksikaliziraju i odgovarajućim glagolima), dok u hrvatskoj normi prevladava glagolski oblik (*dogovoriti/organizirati, prosuditi, žaliti se / reklamirati, predati/isporučiti, koristiti/rabiti, preuzeti / primiti*).

Međusobna semantička privlačivost i leksička povezanost elemenata u kolokacijama je stupnjevite naravi, pa tako razlikujemo slobodnije kolokacije (unrestricted), uobičajene (familiar) i čvrste (restricted, fixed) kolokacije. Bakerova (1992:51) po tom kriteriju dijeli kolokacije na obične i naglašene, forsirane (marked collocations), te za potonje daje primjer: *Could peace break out after all?* nasuprot kolokatu *war* kao uobičajenom argumentu za glagole *break out* ili *prevail*. Funkcija je takvih kolokacija proizvodnja stilskih efekata i naglasaka u prozi, oglasima itd.

Semantička prozirnost kolokacija također je stupnjevite prirode (cf. Carter 1987), od najtransparentnijih kolokacija (*discharge cargo, pay freight*), preko poluprozirnih (*make the berthing arrangements, lift embargo, take a chance, do business*) do vrlo neprozirnih *make good the distance/course (prijeći udaljenost/ /ploviti kursom), call at a port (ticati luku, pristajati), distance run (prijeđeni/prevaljeni put)*, ili metaforičkih izraza: *shoot the sun (određivanje pozicije smjeranjem Sunca), mean business (biti ozbiljan, ne šaliti se)*.

Ireverzibilnost je također jedna od osobina kolokacija, posebno čvrstih, fiksnih kolokacija *cash and carry, hit and miss, trial and error method, No Cure — No Pay, terms and conditions*).

Biskup (1992), kao i većina autora, razlikuje gramatičke i leksičke kolokacije. Gramatičke kolokacije sastoje se od kombinacije: leksička/tematska riječi + gramatička riječ, npr. prijedložne skupine te konstrukcije »to-inf/gerund«). Leksičke kolokacije, koje dijeli u pet vrsta (vidi također Benson—Benson—Ilson 1986b), obilježene su relativnom fiksnošću oblika i neidiomatičnošću značenja. Za razliku od idioma značenje pojedinih članova kolokacija može se izlučiti (dekodirati), dok su idiomi semantički nedjeljive cjeline (*non-compositional phrases*, Verstraten 1992).

Može se stoga zaključiti da kolokacije zatvaraju širok raspon leksičko-semantičke povezanosti dvaju članova leksičkih sveza između dvije krajnosti: slobodnih leksičkih sveza (kombinacija) i idioma nastalih postupnim "isušivanjem" (*draining away*, Cowie 1992) značenja sastavnica kolokacije, pa je ono (značenje) preneseno na izraz (kolokaciju) kao cjelinu. U kolokacija s relativno nižom ograničenošću izbora kolokata veći je stupanj zamjenjivosti riječi što s odabranom riječju (node) stoje u sintagmatskom odnosu, dok je ta ograničenost to veća što je privlačivost, odnosno kolokabilnost veća. Također, u semantičkom smislu opće (generičke) riječi imaju veći kolokacijski raspon (broj kolokata s kojima se pojavljuju) nego njihovi hiponimi (cf.: *burry* vs. *entomb, inter; deliver a bill of lading* vs. *despatch, present, endorse a bill of lading*). Višeznačne riječi također imaju veći kolokacijski raspon: *deliver: a bill of lading, cargo, ship, document, letter, lecture* itd.

Kolokacijska analiza i kolokacijski potencijal rječnika i leksičkih baza

Na firthijanskoj tradiciji razvio se kontekstualni pristup proučavanju kolokacija — sintagmatskih leksičkih sveza nastalih iz tzv. sociologije riječi (Barnbrook 1996:87–106). Računalna lingvistika, posebice razvitak korpusne lingvistike u posljednjih desetak godina, omogućila je na temelju kvantitativnih analiza brz, izravan pristup kvalitativnoj, lingvističkoj i leksikografskoj analizi leksičkih sveza.

Dobri rječnici (bilo jednojezični ili višejezični, opći ili stručno obilježeni) ovise, između ostaloga, o njihovu kolokacijskom potencijalu, tj. ponuđenu izboru kolokata rječničke natuknice, kao i o leksikografskom upućivanju na kolokacijska ograničenja, što korisniku rječnika omogućuju proizvodnju "naslućenih" ali ipak ovjerenih kolokacija. U tradicionalnim rječnicima izražavanje kolokacijskog potencijala bilo je uglavnom rezultat lingvističke analize odabranih riječi ili pak jezične intuicije sastavljača. Pojavom računalnih korpusa ta se analiza počela oslanjati na konkordancijskim ispisima odnosno velikim brojem ovjerenih primjera uporabe odabrane riječi (key-word). Računalna kolokacijska analiza rječničkih korpusa zasniva se na »formalnom prepoznavanju iste riječi (node) koja se neposredno zamjećuje ako se konkordancije očitavaju okomito« (Tognini-Bonelli 1996:201). Prema istom autoru jedini apstraktni postupci što su u tom procesu potrebni, jesu računalno prepoznavanje kolokata, tj. leksema koji su sintagmatski raspoređeni u odnosu na odabranu riječ (node) i koji su određeni čestotom pojavljivanja. Kako bi se dobili kolokati, valja utvrditi kontekstualni raspon odabrane riječi (context span). Uobičajeni raspon iznosi po dvije leksičke jedinice (odvojene razmakom) s obje strane odabrane riječi. Međutim, i takva računalom podržana, prvenstveno kvantitativna analiza iziskivala je nakon toga sporu i napornu sastavljačevu analizu konkordancijskih redova i nije omogućavala brzo i jednostavno donošenje generalizacija. Konkordancije, naime, impliciraju kvantitativnu procjenu (analizu) pojedinih pojava ključnih riječi u kontekstu unutar nekog korpusa i predstavljaju preliminarni kvantitativni postupak.

Tek je naknadnom primjenom kolokacijske analize, s pomoću niza računskih metoda i postupaka (algoritama), omogućen izravniji i kraći pristup kvalitativnoj analizi višečlanih leksičkih sveza kolokacijskog tipa, bez sastavljačeva vlastita prosuđivanja.

Suština računalne leksikografske analize kolokacija (»the occurrence of two or more words within a short space of each other in a text«, Sinclair 1991:170) svodi se na (računalom izračunatu) mjeru po kojoj se stvarni sintagmatski oblik pojavljivanja takvih leksičkih sveza razlikuje od očekivana oblika. Pritom su ključne riječi: *oblik* (*pattern*) vs. *pojavljivanje* (*occurrence*) i *mjera* (*measure*) vs. *očekivanje* (*expectancy*), a to su i neki od pojmova teorije vjerojatnosti.

U kolokacijskoj analizi računalom se formalno određuju i pronalaze posebni leksički odnosi zasnovani na očekivanom međusobnom pojavljivanju leksičkih jedinica. Drugim riječima, prisutnost određene leksičke jedinice pretpostavlja vjerojatnost neslučajnog pojavljivanja druge/drugih leksičkih jedinica. Ono što se izlučuje kolokacijskom analizom jesu leksičke jedinice i s njima povezana leksikografska značenja (lexicographic sense), kao i distribucija lingvističkih oblika i njihovih značenja, pa je tako dobiveno značenje zanimljivo i lingvistima.

Proučavanje kolokacija dobivenih računalnim putem iz korpusa obuhvaća postupak koji pretpostavlja: (a) identificiranje formalnih oblika leksičkih jedinica (word-forms), (b) utvrđivanje čestote njihova apsolutnoga pojavljivanja, (c) čestote njihova supojavljivanja s drugim leksičkim jedinicama i (d) statističko mjerenje važnosti odnosno značenja (significance) takvoga supojavljivanja (cf. Clear 1993, 1996). Isti autor kolociranje drži za jednostavno supojavljivanje leksičkih oblika (jedinica) u tekstu, dok sklonost kolokacija da se razvijaju u »prepoznatljive dijelove govornikove leksičke riznice« (Clear 1993:273) naziva procesom stereotipizacije kolokacije. Upravo mjerenje statističke značajnosti leksičkoga supojavljivanja doprinosi kvalitativnom pristupu analizi kolokacija.

Za svako izvlačenje neke riječi iz korpusa potrebno je prethodno odrediti (a) raspon teksta s ključnom riječju u središtu, (b) najnižu frekvenciju ključne riječi i njenih kolokata u tekstu (korpusu) i donijeti odluku o provođenju lematizacije. Neki autori drže da raspon predstavlja sintaktičku kategoriju (atribut + imenica, glagol + imenica/objekt), za što je prethodno potrebno u tekstu provesti precizno obilježavanje vrsta riječi (tagging) i sintaktičkih funkcija (parsing), ali je za korpusna istraživanja (npr. u projektu COBUILD) usvojeno načelo kvantitativnoga određivanja raspona: 2+riječ+2. Općenito se drži da, u velikim korpusima, čestota pojavljivanja ključne riječi i nekog kolokata niža od 3 nije značajna za kolokacijsku analizu. Pitanje lematizacije ovisi, između ostaloga, i o tipološkoj naravi jezika. U jezicima s morfološki razvijenim sustavima, kao npr. hrvatski, lematizacija je nužna, dok u slučaju engleskog jezika leksikografska praksa pokazuje da se lematizacija može provesti i selektivno.

Za brzo i selektivno nalaženje kolokacija (retrieval) u korpusu relevantne su tri vrste statističkog mjerenja, od kojih drugo i treće pripadaju teoriji informacija:

1) obična čestota supojavljivanja (raw frequency)

2) *t*-score: vrijednost vjerojatnosti, odnosno našeg vjerovanja u međusobnu povezanost dvije riječi, tj. ključne riječi i njenih kolokata (vidi Church—Hanks 1990 i Barnbrook 1996), i

3) *MI-value*: mjera za jačinu međusobne obavijesnosti između dviju riječi (MI-value).

Rezultati tih mjera relevantni su prvenstveno na vrlo velikim jezičnim korpusima. U ovom radu korišten je danas najveći računalni korpus pisanoga i razgovornoga engleskog jezika, *The Bank of English*, koji predstavlja zbir od preko 320 milijuna riječi (prosinac 1997) u 13 potkorpusa. Na korpusu se mogu izvoditi brojne leksikografske operacije, od pozivanja željene riječi ili njihovih najraznovrsnijih kombinacija, pretraživanja po vrstama riječi i afiksima, traženja kolokata u redovima od 120 i više znakova pa do dobivanja ispisa cijelih tekstova (članaka, knjiga i podacima o potkorpusima) u kojima se odabrana riječ pojavljuje.

Primjenom prve mjere (*raw frequency*) u korpusu Bank of English dobiva se 50 ili više kolokata odabrane (ključne) riječi poredanih prema aritmetičkoj čestoti.

Druga mjera (*t-score*) mjeri stupanj vjerojatnosti da će se dvije riječi supojaviti u odnosu na njihovo slučajno pojavljivanje. Njome se mjeri kolika je vjerojatnost da će mijenjanje čestote kolokata neke ključne riječi biti statistički značajno. Temelji se na izračunu čestote, ali ispravlja nepotrebno obilje leksikografskih informacija (kolokata) što se dobivaju čistim frekvencijskim postupcima. Za usporedbu prvih dviju mjera navodi se djelomični ispis frekvencija (list of frequency) i ispis dobiven na temelju primjene postupka *t-score*, također za riječ *proper*:

Čestotni popis		t-score	
the	6936	a	35,585285
a	5098	without	19,960988
to		for	
and		to	
of		not	
in		and	
for		job	15,009968
<p>		care	13,548637
is, etc.		place, etc.	
itd.		itd.	

Prva leksička riječ u čestotnom popisu *place*, s frekvencijom 281, tek je 23. riječ po redu, dok se u popisu prema mjeri *t-score* leksičke riječi nalaze pri vrhu ljestvice (*job, care, place, use, training, way, procedures* itd.).

Treća mjera (MI-value) pokazuje količinu informacije/obavijesti koju jedna riječ pruža o vjerojatnosti supojavljivanja druge riječi i tu vrijednost uspoređuje s utvrđenom čestotom pojavljivanja svake od tih riječi. Npr. ako se riječ *pit* pojavljuje 10 puta u korpusu od 10 milijuna riječi, vjerojatnost njena pojavljivanja iznosit će 0,000001 (tj. jedanput na milijun riječi). Utvrdimo li za-

tim da se u istom korpusu riječ *pith* pojavljuje 5 puta i da u svakom tom pojavljivanju iza nje slijedi riječ *pit*, otkrit ćemo da će vjerojatnost pronalaženja/ /pojavljivanja riječi *pit* iznositi 0,5 (tj. pet puta od ukupno 10). Slijedi dakle da nam riječ *pith* pruža mnogo obavijesti o vjerojatnosti da ćemo uz nju naći i riječ *pit*. Vjerojatnost je dakle porasla s 0,000001 na 0,5. Isto vrijedi i za obrnuti odnos tih dviju riječi. *MI*-vrijednost dakle pokazuje količinu informacije koju nam nalaženje jedne riječi pruža o prisustvu druge u danom korpusu. Na primjer, za odabranu riječ (node) — pridjev *proper*, naredba »C« (*MI-score*) u 320-miljunskom COBUILD-Collinsovu korpusu Bank of English ispisat će sljedeće MI-vrijednosti:

unverified	3	10,470693
gradely	19	10,432215
toggling	4	10,110255
bortnowska	3	9,209701
charlies	6	8,624680
macnaught	14	8,497910
spelling	25	7,976301
...		...
nouns	56	7,711822
holder	70	5,562775
procedures	100	5,220294

Imenice koje bismo intuitivno očekivali (npr.: *nouns, motions, utilization, holder, functioning burial, precautions, authorization, procedures*), kao što se vidi iz drugog stupca, imaju mnogo nižu MI-vrijednost od očekivane, tj. one zbog više apsolutne čestote u korpusu pružaju manju količinu međusobne obavijesti. S druge strane kolokati riječi *proper* na čelu popisa pokazuju nedostatke takva mjerenja jer izbacuju na površinu čudne, neočekivane kombinacije riječi, koje zbog svoje opetovane ali često slučajne i sasvim idiosinkratske pojavnosti u jednom te istom tekstu imaju visok stupanj međusobne obavijesnosti. U korpusnoj računalnoj lingvistici ta je pojava poznata kao tzv. *whelks problem* (Kilgarriff 1997), a nastaje visokom čestotom neke rijetke riječi, npr. *whelk* (vrsta morskog puža) u korpusu čiji je sastavni dio npr. knjiga o takvim puževima. Ipak, MI-vrijednosti koristan su računalni postupak za dobivanje kolokacija jer ukazuju na stupanj međusobne povezanosti kolokata i ključne riječi, te su posebno dragocjene u stručno obilježenim rječnicima. Prema Sinclairu (1991) u slučaju kolokata koji u pravilu imaju nižu čestotu od odabrane riječi (*downward collocates*) primjena MI-mjerenja proizvodi i na vrh ljestvice stavlja kolokacije, primjerice iz područja (*field, subject-matter*) struke (npr. tehnike, prava, navigacije), arhaičke termine, te fiksne i gramatičke kolokacije. Nasuprot tome, kolokati čija je čestota veća od oda-

brane riječi (node), tzv. *upward collocates*), daju puno više obavijesti o semantičkoj prirodi neke riječi.

Ipak, s leksikografskoga su stajališta najkorisniji kolokacijski podaci što ih pružaju tzv. kolokacijske slike dobivene bilo čestotnim istraživanjima ili primjenom spomenutih metoda obavijesne teorije (MI-value i *t*-score). Za usporedbu u prilogu 1 dana je kolokacijska slika prema mjeri *t-score* za odabranu riječ (node) — pridjev *heavy*, a u prilogu 2 kolokacijska slika za istu riječ ali prema vrijednosti *MI* (mutual information value). U oba slučaja dana su tri stupca kolokata s obje strane (moguće je izlučiti do 6 kolokata s obje strane!). Kolokacijske kombinacije s *heavy* valja čitati / stvarati ne u linearnom slijedu već traženjem odgovarajućeg kolokata iz bilo kojeg reda u bilo kojem stupcu. Već sama površna analiza kolokacijskih slika daje nam prije svega sintaktičku i semantičku distribuciju pridjeva *heavy* i njegovih kolokata (priloga i imenica). To su imenice iz vojnog nazivlja (očito se radi o dominaciji novinarskih tekstova iz posljednjih nekoliko godina) odnosno semantičkih polja vojnih podstruka: vrste oružja (*weapons, artillery, guns*), vojne aktivnosti (*fighting, fire, shelling*), posljedica vojnih djelovanja (*losses, casualties, defeat, toll*). Tu su zatim imenice kojima se imenuju opće meteorološke pojave: *rain, seas, snow, pressure weather*) te svakodnevne pojave i radnje (*duty, price, burden, traffic, pressure*). Izuzetno je visoko rangirana kolokacija *heavy drinker(s)*, što ukazuje na vrlo visok stupanj leksikalizacije te kolokacije. S lijeve strane analizi se nameću tipični priložni intenzifikatori (*too, very* te nešto niže na ljestvici *really, quite, extremely* itd.).

Kolokacijska slika dobivena mjerom međusobne obavijesnosti (*MI-value*) pruža popis ne toliko čestih kolokata s pridjevom *heavy* već onih kolokata koji o riječi *heavy* pružaju najviše međusobnih informacija i obrnuto. To su leksikografski vrlo vrijedne kolokacije. Ovdje su na prvom mjestu specifične, rjeđe ali leksički čvršće kolokacije s područja meteorologije (*snowfalls, downpours, rains, rainstorms*), a slijede imenice o materijalima i teretima (*payloads, caseloads, workloads, crudes, metals*). Kolokacija *heavy drinker(s)* u objema je kolokacijskim slikama jednako visoko rangirana. Navedene kolokacijske slike pružaju i obilje relevantnnih gramatičkih informacije, npr. o čestoti i semantičkoj važnosti oblika množine imenica, itd.

Konačno, postavlja se pitanje kako leksikografski vrednovati strojno dobivene podatke o kolokacijskom ponašanju riječi. Korist je dvostruka: (a) kolokacijska informacija o odabranoj riječi daje nam semantički profil ključne riječi i pomaže izlučivanju pojedinih, bilo dominantnih ili sporednih značenja pojedine riječi, i (b) kolokacije govore o sintaktičkoj i semantičkoj distribuciji riječi, tj. o tipičnoj okolini i uporabi neke riječi. Svakako, ljudska analiza nije u stanju obraditi toliku količinu informacija u izuzetno kratkom vremenu, te je za tu svrhu stroj nezaobilazan. Pitanje je međutim koliko se strojnoj analizi može vjerovati.

Jedan od načina provjere pouzdanosti računalne analize jest usporedba njenih rezultata (dobivenih kolokacija) s kolokacijskim informacijama što nam ga pružaju rječnici koji su sastavljeni bez pomoći računala ili spomenutih statističkih mjerenja. Tu valja uzeti u obzir kolokate s lijeve i desne strane ključne riječi te raspon kolokacije, tj. broj znakova odnosno riječi odvojene razmakom (span) lijevo i desno od ključne riječi. Za usporedbu je uzeta riječ *berth* i njena obrada u CCELD, rječniku koji je sastavljen na temelju u ovom radu razmatranih računalnih istraživanja korpusa i CEDT, rječnika sastavljena na tradicionalnim načelima leksikografije.

U korpusu Bank of English *berth* javlja se ukupno 2174 puta, od čega 2010 puta kao imenica i 164 puta kao glagol (past participle, past tense, ing-form). Kolokacijska slika imenice *berth* prema mjeri *t-score* (vidi prilog 3) daje po tri kolokata s lijeve i desne strane. Zanemarujući funkcionalne riječi, s lijeve strane dobivaju se sljedeće najtipičnije kolokacije, tj. zbog čestote pojavljivanja najvjerojatnije očekivane statistički relevantne leksičke kombinacije: *wide, final, double, four, playoff, marina, fuel, test, vee, single, midfield, upper, team, wider, pilot, sick, twin, permanent BERTH*, dok s desne strane bilježimo mnogo više gramatičkih riječi (prijedloga, zamjenica, priloga, pomoćnih glagola) i glagola: *in, at, with, on, for, near; alongside, inside;* ali manje imenica *(marina, cabin, holder)* koje modificiraju imenicu *berth*. Analiza potkorpusa prema sadržaju i tematici ukazuje da prevladavaju tekstovi iz svakodnevna života (ephemera), npr. iz sporta (automobilskih trka, tenisa) te nautičkog turizma (čamci za razonodu, marine itd.).

Kolokacije dobivene mjerom MI-vrijednosti pružaju nešto specifičniju, ali obavijesno za leksikografa zanimljiviju kolokacijsku sliku (vidi Prilog 4), bez obzira na poneke idiosinkratske, sasvim slučajne kolokate visoke međusobne obavijesnosti s riječju *berth*: *vee, playoff, settee, ccs, finals, marina, wide, afl, midfield, final, twin, defensive, sleeping, empty, olympic, test, cup, single BERTH*. Kolokati s desne strane potvrđuju semantičku distribuciju (*BERTH cabins, attendant, holder, yacht, accommodation*) ali daju i više podataka o sintaktičkoj distribuciji: *single/double BERTH cabin, double-BERTH boat, BERTH alongside* itd.

Na temelju iz korpusa automatskim putem izvučenih kolokacija i njihova izbora te leksikografskom obradom konkordancija sastavljači su u CCED[2] (1. izdanje) ponudili sljedeći leksikografski opis leksema *berth*:

1. If you **give** something a **wide berth**, you avoid going near it because it is unpleasant or dangerous.
   EX: *The milkman has obviously given this place a wide berth.*
   GR: phrase: verb inflects
2. A **berth** in a harbour is a space by the quay where a ship can stay

[2] CCED = *Collins COBUILD English Dictionary*, Harper Collins Publ., 1995.

for a period of time
GR: count noun
SYN: mooring

3. When a ship **berths** or when someone **berths** a ship, the ship sails into a harbour and stops at the quay
EX: *The ship berths at noon.*
SYN: dock
4. A **berth** is a bed on a boat, train, or caravan
EX: *Miss Ryan was curled up in her berth.*
GR: countable noun
SYN: bunk

Natuknica *berth* u CEDT[3]:

berth n.
1. a bed or bunk in a vessel or train, usually narrow and fixed to a wall.
2. *Nautical.* a place assigned to a ship at a mooring.
3. *Nautical.* sufficient distance from the shore or from other ships or objects for a ship to manoeuvre
4. give a wide berth to — to keep clear of; avoid.
5. *Nautical* accommodation on a ship.
6. *Informal.* a job, esp. as a member of a ship's crew.

berth vb.
7. (tr.) *Nautical.*to assign a berth to (a vessel).
8. *Nautical.* to dock (a vessel).
9. (*tr.*) to provide with a sleeping place, as on a vessel or train.
10. (*intr.*) *Nautical.*to pick up a mooring in an anchorage

Najvjerojatnije iz statističkih razloga kolokacijska analiza u CCELD na prvo je mjesto među značenijima (senses) stavila kolokaciju odnosno izraz **give a wide berth**, što nije slučaj u tradicionalnom rječniku CEDT. Kolokacije su u oba rječnika sadržane u tekstu definicije, koja je u CCELD eksplicitna, a u CEDT slijedi načelo: od općenitog (*genus proximum*) prema posebnom (*differentia specifica*). Kolokacije su u CCELD glavni nositelji primjera (značenja) i naročito uporabe (sintaktička distribucija). U istom rječniku glagolska kolokacija (*berth* + imenica) leksikografski je vrlo istaknuta, što je zacijelo rezultat visoke vrijednosti vjerojatnosti njihova međusobnog pojavljivanja i međusobne obavijesnosti. CCELD jest rječnik sastavljen na osnovi čestotnih načela i opisanih statističkih mjera, ali se opis natuknice u CEDT-u ipak leksikografu čini sustavnijim. Koliko su jedan ili drugi opis prirodniji, valjalo bi

[3] CEDT = *Collins English Dictionary and Thesaurus*, Harper Collins Publ., 1993.

ispitati korisnike rječnika koji nisu ni jezični stručnjaci niti leksikografi. Konačno, zanimljivo je da nijedan rječnik ne nudi kolokacije ili višečlane leksičke jedinice na razini rječničke nomenklature.

S druge strane, dvojezični stručni rječnici uz imenicu *berth* daju kao natuknice brojne višečlane leksičke jedinice (od kolokacija preko složenica do frazeološkog bloka), npr. Dluhy (*Schiffstechnisches Wörterbuch*): *berth and space, berth cargo, berth charge, berth clause, berth crane, berth dues, berth party, itd.*, te u glavnom dijelu natuknice kolokacije: *appropriated berth, clear berth, fitting-out berth, lay-up berth, wide berth*). U pomorskom rječniku Kerchove 1961[4] nalazimo još: *berth note, berth owner, berth rates, berth terms, berth traffic; building berth, foul berth, loading berth, swinging berth*).

Zaključak

Proučavanje kolokacija izuzetno je važan leksikografski posao. S obzirom na to da je kolokacijski odabir specifičan za pojedine jezike, kolokacije moraju nužno naći mjesta u svakom rječniku, posebice u dvojezičnim općim rječnicima kao i u onima posebne namjene. Posebnu pozornost valja posvetiti kolokacijama s vrlo čestim, pa stoga i višeznačnim riječima i u tu svrhu koristiti sva sredstva leksikografskog opisa (uputnice, komentari itd.). Posebice je korisno dati informacije o kolokacijskom rasponu pojedine riječi općeg jezika.

Kolokacije se mogu identificirati i razvrstavati klasičnim putem, tj. na temelju poznavanja i vlastite intuicije sastavljača rječnika, te računalnim putem na velikim korpusima priređenima za raznovrsna leksikografska istraživanja, između ostalih i dobivanja kolokacijskih informacija i slika prikazanih u ovome radu, u što se je autor imao prilike i sam uvjeriti radeći s leksičkom bazom podataka *Bank of English* (tvrtke *Collins-COBUILD*) u Birminghamu. Računalno dobivanje kolokacija danas je nezaobilazan leksikografski postupak i pruža dobre rezultate na velikim korpusima. Stoga pri izradi velikog korpusa hrvatskog jezika (najmanje stomilijunski korpus) valja predvidjeti programsku mogućnosti utvrđivanja kolokacijskog potencijala hrvatskog jezika, ali i praktičkog pozivanja i ispisivanja kolokacija i kolokacijskih slika o čemu je bilo riječi u ovom tekstu. Na taj način korpus hrvatskog jezika bio bi usporediv sa sličnim korpusima drugih jezika u svijetu. S obzirom na postojanje potprograma za lematizaciju, to ne bi trebalo biti suviše teško provedivo.

[4] R. De Kerchove, *International Maritime Dictionary; An encyclopaedic dictionary of useful maritime terms and phrases, together with equivalents in French and German* (1961). New York – London – Toronto – Melbourne, Van Nostrand Reinhold Co.

Literatura

Baker, M. 1992. *In Other Words, a coursebook on translation*. London : Routledge.

Barnbrook, G. 1996. *Language and Computers*. Edinburgh : Edinburgh University Press.

Benson, M., E. Benson, R. Ilson (BBI) 1986. *Lexicographic Description of English*. Amsterdam : John Benjamins.

Benson, M., E. Benson, R. Ilson (BBI) E. 1986b. *Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam : John Benjamins.

Biskup, D. 1992. L_1 Influence on Learners' Renderings of English Collocations: A Polish/German Empirical Study. U knj. P. J. Arnaud and H. Bejoint, *Vocabulary and Applied Linguistics*. London : Mac Millan Academic & Professional Ltd., 85–93.

Carter, R. 1987. *Vocabulary, Applied Linguistics Perspectives*. London : Allen Unwin.

Church, K. W., and P. Hanks. 1990. Word Association Norms, Mutual Information & Lexicography, *Computational Linguistics*, 16, 1., 22–29.

Clear, J. 1993. From Firth Principles, Computational Tools for teh Study of Collocation. In: M. Baker, G. Francis, E. Tognini-Bonelli. *Text and Technology: In Honour of John Sinclair*. Amsterdam, Philadelphia : Benjamins Publ. Co.

Clear, J. 1996. »Grammar and nonsense«: or syntax and word senses. In: *KWHAA Konferenser* 36, Stockholm, 213–242.

Cowie, A.P. 1992. Multiword Lexical Units and Communicative Language Teaching. U knj. P. J. Arnaud and H. Bejoint. *Vocabulary and Applied Linguistics*. London : Mac Millan Academic & Professional Ltd., 1–12.

Crystal, D. 1987. *The Cambridge Encyclopedia of Language*. Cambridge : Cambridge University Press.

Halliday. M.A.K., R. Hasan. 1976. *Cohesion in English*. London and New York : Longman.

Ivir, Vladimir. 1974. *Teorija i tehnika prevođenja*. Sr. Karlovci : Centar »Karlovačka gimnazija«.

Ivir, V., V. Tanay. 1976. The Contrastive Analysis of Collocations: Collocational Ranges of Make and Take with Nouns and Their Serbo-Croatian Correspondents. *Reports 10, YSCECP*, Zagreb, 20–47.

Kilgarriff, A. 1997. Putting frequencies in the dictionary. *Journal of Lexicography*, Vol. 10, No. 2, 135–155.

Lyons, J. 1995. *Linguistic Semantics, An Introduction*. Cambridge: Cambridge University Press.

Palmer, F.R. 1981. *Semantics*. 2nd ed. Cambridge : Cambridge University Press.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Smadja, F., V. Hatzivassiloglou, K. McKeown. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, ACL.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford : Blackwell.

Štambuk, A. 1990. Korištenje terminološke baze podataka u leksikografskim i leksikološkim istraživanjima. U knj. *Informatička tehnologija u primijenjenoj lingvistici*. Zagreb : HDPLU, 99–104.

Tognini-Bonelli, Elena. 1996. Towards Translation Equivalence from a Corpus Perspective. *IJL*, Vol. 9, No. 3, 197–217.

Vertstraten, L. 1992. Fixed Phrases in Monolingual Learners' Dictionaries. U knj. P. J. Arnaud and H. Bejoint, *Vocabulary and Applied Linguistics*. MacMillan Academic & Professional Ltd., London, 28–40

Zgusta, Ladislav. 1970. *Manual of Lexicography*. Mouton.

Prilozi

1. *HEAVY* — Kolokacijska slika prema mjeri *t-score*

there	with	a	NODE	rain	and	the
has	was	the	NODE	fighting	on	volume
was	trading	too	NODE	metal	of	around
it	after	very	NODE	weapons	in	reported
some	is	under	NODE	artillery	from	on
because	came	of	NODE	than	<p>	sarajevo
were	been	with	NODE	losses	for	fire
they	tanks	and	NODE	casualties	presence	and
have	despite	their	NODE	duty	or	rain
had	paid	by	NODE	industry	around	snow
also	because	after	NODE	price	fire	heavy
points	come	is	NODE	rains	guns	their
edged	pay	some	NODE	burden	which	fighting
been	by	was	NODE	seas	such	lb
trading	of	so	NODE	handed	gun	artillery
oil	are	in	NODE	traffic	losses	million
forces	have	suffered	NODE	pressure	door	its
due	under	its	NODE	metals	but	un
tanks	carrying	inflicted	NODE	trading	burden	also
light	suffered	much	NODE	equipment	air	both
serbs	points	s	NODE	drinking	traffic	his
which	winds	really	NODE	guns	rain	continuing
heavy	days	quite	NODE	snow	pan	government
traffic	were	extremely	NODE	weather	load	flooding
troops	had	been	NODE	defeat	over	continued
strong	be	unusually	NODE	machine	casualties	rocket
are	air	particular	NODE	toll	pressure	light
spite	such	were	NODE	fire	vehicles	alcohol
shares	traffic	fairly	NODE	shelling	between	hydrogen
result	reports	his	NODE	drinkers	equipment	which
fire	carry	carrying	NODE	investment	against	furniture
to	fighting	rather	NODE	security	curtains	over
large	through	despite	NODE	periods	during	pan
already	caused	carry	NODE	weight	with	winds
fighting	took	slightly	NODE	ground	band	machine

2. *HEAVY* — Kolokacijska slika prema vrijednosti međusobne obavijesnosti (*MI-value*)

icesheets	exacts	30kg	NODE	snowfalls	breakable	kerrang
mudslide	deuterium	inflicting	NODE	snowfall	bombardmen	schmaltz
skittled	exacted	20kg	NODE	downpours	shellfire	howitzers
deuterium	lugged	gratifying	NODE	payloads	writedowns	castors
howitzers	capsized	unseasonab	NODE	rains	transporte	gudermes
tau	mortars	inflicted	NODE	impasto	flashlight	metallica
tritium	lugging	kilos	NODE	rainstorms	brocade	mortars
semtex	stretchers	incurring	NODE	drinker	putty	thundersto
cranes	tritium	unusually	NODE	rainstorm	bombardmen	sleet
protons	eyelids	moderately	NODE	lidded	cumbersome	sheeting
fission	elevators	inflict	NODE	caseloads	bracelet	volume
edged	burdened	exceptiona	NODE	petting	rimmed	flooding
caller	wodehouse	braved	NODE	artillery	soils	hydrogen
sludge	bulky	unbearably	NODE	hitters	shutters	depress
sofas	launchers	incur	NODE	croppers	brows	mortar
tugged	thundersto	unduly	NODE	workloads	heavier	artillery
vicary	underlines	mitsubishi	NODE	crudes	burdens	weaponry
eyelids	repulsed	impossibly	NODE	metals	sleet	vicary
fleets	heavier	5kg	NODE	machinegun	bulky	armour
grenade	tanks	polluting	NODE	breathers	brooch	sarajevo
helium	thicker	hauling	NODE	sedation	rains	leftist
hampered	hampered	hyundai	NODE	isotope	saucepan	scent
forecaster	winds	suffered	NODE	machinegun	riffs	discountin
marty	braved	mortars	NODE	drinkers	cp	rocket
dubrovnik	shouldered	withdraw	NODE	seas	pounded	gale
workload	declines	withstand	NODE	workload	exertion	cumbersome
deterred	overcast	thicker	NODE	shelling	polythene	slabs
monrovia	pla	bulky	NODE	bombardmen	riddick	ethiopian
tanks	taller	eyelids	NODE	weaponry	loosened	shelling
armored	leaden	excessivel	NODE	bombardmen	drinkers	mtv
floods	inflicting	taller	NODE	rainfall	casualties	burdens
seizing	trading	imposes	NODE	downpour	curtains	rains
tornado	nuclei	sustained	NODE	browed	downpour	bombardmen
armoured	outhbreaks	lifting	NODE	casualties	load	pounded
bihac	saddles	exceedingl	NODE	scratching	showers	thunder
soaking	reinforcem	unexpected	NODE	footfalls	burden	tamil

3. *BERTH* — Kolokacijska slika prema mjeri *t-score*

Mon Mar 17 16:08:11 1997 1

a	a	wide	NODE	in	the	port
for	semi	a	NODE	<p>	a	final
given	for	final	NODE	at	pound	first
into	an	finals	NODE	with	air	conditione
give	grand	double	NODE	on	cabin	marina
to	quarter	four	NODE	marina	<h>	finals
an	into	2	NODE	cabin	<b>	world
clinch	play	eight	NODE	for	cruiser	national
dinner	his	its	NODE	and	sh	australian
him	final	six	NODE	holders	stake	open
pound	first	starting	NODE	against	graham	wc
persons	top	off	NODE	by	taking	galley
them	breakfast	two	NODE	when	but	berth
earned	double	playoff	NODE	cabins	cabins	harbour
gave	league	cup	NODE	cruiser	associatio	sheffield
berth	wide	marina	NODE	next	yacht	tournament
pp	left	fuel	NODE	</h>	meals	melbourne
fill	ccs	uarter	NODE	after	accommodat	6
ship	earn	test	NODE	alongside	with	grand
from	super	vee	NODE	boat	associatio	queensland
lay	her	s	NODE	forward	3	held
giving	queensland	single	NODE	sleeper	sunday	start
share	midshipmen	settee	NODE	yacht	there	gooch
australian	clinched	midfield	NODE	attended	o	serving
cup	clinch	my	NODE	inside	next	ship
back	sharing	bowl	NODE	is	season	by
take	at	upper	NODE	yesterday	he	boat
side	commodore	her	NODE	sleeping	sybille	super
clinched	uefa	team	NODE	or	aspirants	offers
cement	booked	wider	NODE	as	novotna	coach
sharing	runner	pilot	NODE	while	compartmen	beat
soccer	navy	6	NODE	compartmen	small	second
on	olympic	4	NODE	gunn	screw	high
gain	10	large	NODE	inthe	paramount	down
x	world	automatic	NODE	mobile	hastings	3
edge	towards	half	NODE	accomodat	marina	1
offers	season	per	NODE	last	monrovia	third
conference	book	mud	NODE	to	still	round
1	large	sick	NODE	unless	sailing	with
first	the	10	NODE	above	southampto	two
league	securing	back	NODE	but	crystal	former
near	scrum	our	NODE	outside	derby	7
cost	runners	afl	NODE	because	ocean	top
final	new	qualifying	NODE	hw	beting	six
her	approached	twin	NODE	compartmen	nervous	ccs
club	beneath	permanent	NODE	crumbled	occupied	westlake

4. *BERTH* — Kolokacijska slika prema vrijednosti međusobne obavijesnosti *(MI-value)*

Mon Mar 17 16:08:27 1997 1

clinch	midshipmen	vee	NODE	cabins	cruiser	wc
berth	ccs	playoff	NODE	marina	cabins	conditione
clinched	commodore	settee	NODE	cruiser	cabin	galley
cement	clinched	ccs	NODE	sleeper	sh	marina
persons	clinch	finals	NODE	cabin	associatio	berth
earned	semi	marina	NODE	attendant	yacht	gooch
sharing	uefa	wide	NODE	holders	meals	harbour
pp	securing	afl	NODE	compartmen	accomodat	port
dinner	scrum	midfield	NODE	gunn	stake	finals
fill	booked	final	NODE	yacht	graham	sheffield
soccer	runners	double	NODE	inthe	air	tournament
ship	runner	mud	NODE	alongside	associatio	melbourne
lay	grand	automatic	NODE	sleeping	sunday	serving
gain	earn	bowl	NODE	mobile	taking	ship
x	breakfast	fuel	NODE	accomodat	o	grand
given	sharing	ualifying	NODE	boat	season	boat
edge	quarter	metre	NODE	forward	pound	final
offers	approached	wider	NODE	inside		super
give	navy	pilot	NODE	unless	small	offers
giving	olympic	starting	NODE	against	3	coach
gave	beneath	upper	NODE	above	next	australian
share	earned	defensive	NODE	next	<h>	beat
australian	super	twin	NODE	outside	still	queensland
conference	double	bears	NODE	yesterday	the	held
cup	comfortabl	sleeping	NODE	at	another	open
league	wide	sick	NODE	in	long	start
near	ship	quarter	NODE	<p>	2	national
cost	wild	permanent	NODE	with	year	6
final	final	tournament	NODE	while	where	third
side	league	grade	NODE	when	many	round
club	play	eight	NODE	after	there	world
book	queensland	empty	NODE	until	then	first
looking	claimed	olympic	NODE	on	a	former
into	winning	test	NODE	</h>	when	7
car	forced	wing	NODE	left	any	top
pound	offered	four	NODE	again	could	second
for	claim	cup	NODE	since	but	six
him	top	single	NODE	today	an	high
a	allowed	300	NODE	for	my	30
take	towards	row	NODE	by	with	3
second	easy	v	NODE	last	two	away
back	season	200	NODE	because	new	john
	test	six	NODE	or	last	left
them	left	opening	NODE	where	he	state
an	book	card	NODE	and	if	down
four	large	lower	NODE	which	be	1

On the collocational power of the lexicographic corpus

Summary

Collocations, their retrieval from the corpus or lexical database, as well as their presentation within the dictionary word-list or at the level of a lexicographic entry, present a major issue in modern lexicography. This problem has been given different levels of importance in general dictionaries. It is, however, highly important in the compilation of technical, or subject-matter related dictionaries. The paper deals with the methods of determining and selecting collocations. Collocations are regarded as an important type of multi-word lexical units to be selected or retrieved from the corpora for single or bilingual (English-Croatian) dictionaries of both general and technical type. Examples of corpus-based collocations are given for the same head-words on the basis of the Bank of English run by COBUILD-Collins.

Ključne riječi: kolokacije, višečlane leksičke jedinice, korpus, statistička mjerenja, rječnik

Key words: collocations, multi-word lexical units, corpus, statistical measures, dictionary