Milica Gačić

KORIŠTENJE KONKORDANCIJA U IZRADI PEDAGOŠKIH RJEČNIKA STRUKE

Stručni članak. Primljen 1. 2. 1985.

UDK 37:800.866(038):801.82:681.3

Za efikasniju nastavu jezika struke nužna je izrada materijala s kojima će se postići optimalno učenje jezika. Korištenje kompjutora može nam znatno olakšati izradu pedagoških rječnika struke — bilo u klasičnoj formi standardnog rječnika — s tim da su obuhvaćene najrelevantnije i najčešće natuknice koje se koriste u određenoj struci, ili u obliku kontekstualnog rječnika struke koji može pružiti izuzetno vrijedne primjere upotrebe leksičke jedinice, kolokacije ili idioma.

Korištenje konkordancija je jedan od metodoloških postupaka koji može pridonijeti rješavanju bitnih pitanja u vezi s izradom rječnika struke jer pružaju kontekstualne podatke o riječima.

Prva dekada automatske izrade konkordancija teksta su pedesete godine ovoga stoljeća. Među pionirima u korištenju te metodologije svakako je Roberto Busa, teolog koji je proučavao radove Tome Aquinskoga. Njegovim radom postavljeni su temelji pa je 1956. godine u mjestu Gallarate ustanovljen Centar za automatizaciju literarnih analiza (Centro per L'Automazione dell' Analisi Letteraria).[1] U tom centru 1964. godine radilo je šezdeset ljudi.

U radu *Nieuve wegen in der lexicologie* (1963), holandski autor F. de Tollenaere naglašava da su leksikografija, izrada indeksa riječi i izrada konkordancije tri aktivnosti što pripadaju u *leksikologiju* koju definira kao »teoriju rječnika koja uključuje semantiku i morfologiju, tvorbu riječi, etimologiju, sinonimiju itd.«

Pod pojmom pedagoškog rječnika razumijevamo funkcionalni profil rječnika određene struke (u pravilu dvojezični) koji obuhvaća leksičku građu što optimalno zadovoljava potrebe nastave i svladavanja jezika struke. To znači da takav rječnik treba da sadrži stupnju nastave stranog jezika primjeran broj leksičkih jedinica optimalne korisnosti s obzirom na mogućnost razumijevanja stručne informacije na stranom jeziku. Time je pedagoški rječnik prije svega pomoć nastavniku zbog ograničenog dometa empirije i intuicije, te studentu zbog optimalizacije uloženog truda i sredstava u učenju stranog jezika.

Više je načina prikupljanja rječničke građe. Kad je riječ o problematici rječničke građe za potrebe nastave jezika struke, izuzetnu nam pomoć može pružiti kompjutorska obrada reprezentativnog korpusa jezika struke. U uvje-

[1] Burton, D. M., str. 12.

tima kad za određenu struku postoji jezični korpus unesen na magnetsku traku (ili drugi medij), relevantne podatke za izradu pedagoškog rječnika struke možemo dobiti ako kompjutorskom obradom dobijemo ove ispise:

(1) abecednu listu riječi,
(2) listu riječi po frekvenciji,
(3) konkordanciju teksta.

U sklopu raspoloživog vremena i u skladu s psiho-fizičkim mogućnostima studenata, kroz 120 sati nastave možemo očekivati usvajanje dosta ograničenog broja riječi. Da bismo utvrdili leksički okvir unutar kojega se treba kretati da bi se pažnja posvetila komunikacijski važnijim dijelovima leksičke građe, kao jedan od osnovnih kriterija može nam poslužiti frekvencijska lista leksičkih jedinica.

Korisnost utvrđivanja frekvencija jezičnih sredstava uočena je i korištena još u starom vijeku, a prva brojanja riječi rađena su za potrebe slijepih i stenografije. Za pedagoške potrebe prva brojanja riječi obavljena su u prvoj polovici 20. stoljeća.

Frekvencijski podaci o riječima u općem jeziku koristili su se u nizu tečajeva i projekata za učenje jezika. No vrijednost takvih podataka mnogo je upotrebljivija za pedagoške potrebe u jeziku struke jer jezik struke ima ograničen leksički repertoar koji je moguće dosta pouzdano utvrditi korištenjem dobrog korpusa.

Ovdje ćemo navesti samo neke podatke koje ističu razni autori u vezi s korištenjem frekvencija. Tako P. Guiraud[2] kaže:

100 prvih riječi pokriva 60% teksta,
1 000 prvih riječi pokriva 85% teksta,
4 000 prvih riječi pokriva 97,5% teksta.

V. Pravda[3] navodi da minimalni vokabular od približno 2 200 riječi pokriva, prema provedenim ispitivanjima, oko 95% teksta, a u priručniku *Languages for Special Purposes*[4] autori navode da lista od 2 000 općih riječi i 500 specijaliziranih termina pokriva oko 90% stručnog teksta.

U vezi s dosad provedenim frekvencijskim ispitivanjima jezika struke treba naglasiti da je minimalne vokabulare u jeziku struke lakše utvrditi na osnovi podataka o frekvenciji jer, s obzirom na uži izbor leksičkih sredstava, postoje manje razlike u distribuciji.

Osvrnut ćemo se na primjer jedne takve analize.[5] Analizirajući frekvencijsku listu korpusa kriviČnih disciplina utvrdili smo da bi pedagoški bilo opravdano obraditi leksičke jedinice do minimalne vrijednosti frekvencije f = 4 u korpusu od 120 000 pojavnica. Za tu smo se vrijednost odlučili zato što smatramo da taj broj javljanja neke riječi u korpusu nije slučajnost. Tako dobivamo određeni broj leksičkih jedinica koje se mogu svrstati u tri kategorije:

(a) opće riječi,
(b) opće riječi u stručnoj upotrebi,
(c) stručne riječi.

[2] Guiraud, P., str. 10.
[3] Pravda, V., str. 270.
[4] *Languages for Special Purposes*, str. 34.
[5] Gačić, M., *Frekvencijsko-leksička analiza engleskog kao jezika struke kriviČnih disciplina*, neobjavljena doktorska disertacija, Zagreb, 1982.

Prvu kategoriju čine uglavnom riječi obuhvaćene prijašnjim stupnjevima učenja, pa u okviru nastave jezika struke nije predviđen proces njihovog aktivnog usvajanja, već samo ponavljanja, no one treba da budu uvrštene u pedagoški riječnik. Druge dvije kategorije mogu se obuhvatiti nastavom jezika struke, odnosno nužno im je posvetiti pažnju da bi bile aktivno usvojene.

Osnovni lingvistički postupci kojima treba podvrći tako relativno mehanički određen frekvencijski popis jest *lematizacija* (odnosno postupak svođenja kosih oblika na kanonski rječnički oblik — jedinica za imenice, infinitiv za glagole, pozitiv za pridjeve).

Tako dobivene rječničke natuknice treba svrstati abecednim redom. Broj leksičkih jedinica u lematiziranoj varijanti smanjuju je otprilike 26% u engleskom, a prema nekim podacima 20% u njemačkom jeziku.[6] U našem slučaju (u krivičnim disciplinama) broj leksičkih jedinica sveden je sa 3 379 na 2 494 jedinica.

Abecedni popis može sada, osim kao osnova za rječničke natuknice, poslužiti za uspješnije pronalaženje odgovarajuće leksičke jedinice u konkordanciji.

Ž. Bujas[7] ovako definira konkordanciju: »Konkordanca je tako priređen tekst da je svaka pojedina riječ tog teksta (bez obzira na to koliko se puta ponavljala, gdje se javljala i u kojem obliku dolazila) prikazana abecednim redom sa svojim kontekstom. Veličina konteksta u konkordancama je proizvoljna, ali — premda kontekst može obuhvaćati cijelu strofu, rečenicu ili odlomak — obično je u opsegu jednog stiha ili jednog retka kompjuterske liste (8 do 10 riječi ispred i iza konkordirane riječi)«.

Za potrebe analize jezika struke krivičnih disciplina izrađena je potpuna tzv. normalna kontekstualna konkordancija. Ispis sadrži ove podatke za svaku riječ: oznaku za knjigu, broj stranice u knjizi, broj sloga u ispisu teksta (red), tekst koji prethodi, razmakom istaknutu riječ koju konkordiramo i tekst koji slijedi.

Na sljedećoj stranici dajemo primjer opisane konkordancije.

[6] Teubert, W., str. 296.
[7] Bujas, Ž., *Kompjuterska konkordanca . . .*, str. 35.

S A T - SISTEM ZA ANALIZU TEKSTA

ZAGREB, ENGELSOVA 6b

DAN: 26 SVIBANJ 1981 STR: 434

K O N K O R D A N C A T E K S T A

KNJG	STR	RED	TEKST KOJI PRETHODI	RIJEČ	TEKST KOJI SLIJEDI
K9	17	14	O THIS END, OFFICERS USING CYCLES, MOTOR-CYCLES AND VANS	CARRY	OUT SPOT CHECKS AROUND THE BANKS. IF THE HEADS
K9	45	15	T ONLY IN CONNECTION WITH ACTIVE POLICE WORK BUT ALSO TO	CARRY	OUT TASKS OF AN ADMINISTRATIVE AND FINANCIAL N
K9	53	04	ENQUIRIES ALONG THE RIGHT LINES. IT IS ALSO POSSIBLE TO	CARRY	OUT SEARCHES FOR ACCOMPLICES, I.E. TO COMPILE
K3	150	03	ILITARY TRAINING AND DISPLAY; THE WIDESPREAD PRACTICE OF	CARRYING	GUNS OR KNIVES; THE CONCERN WITH PERSONAL H
K3	588	06	IMPROVING LOCKING DEVICES OR DOING AWAY WITH THE NEED OF	CARRYING	CASH. THESE IDEAS DO NOT REPRESENT AN ABRUP
K6	188	20	THIS NATURE - IN ORDER TO REDUCE THE VISIBILITY OF THOSE	CARRYING	OUT THE CHECKS, AND THE DETAILS OF CARS FOU
K7	189	02	ALLERGY, E.G. SNIFFLES, RUNNING EYES; OR A PERSON MAY BE	CARRYING	A HYPODERMIC NEEDLE BECAUSE HE OR SHE IS DI
K8	38	18	SCRIBED IN CHAPTER 15. SUSPICIOUS STAINS, OR THE OBJECTS	CARRYING	THEM, MUST BE REMOVED TO THE LABORATORY FOR
K8	38	21	FF AND PLACED IN A GLASS VIAL. CRACKS IN THE FLOOR, SOIL	CARRYING	BLOOD, AND OTHER SPORTS THAT MAY COLLECT AN
K8	122	23	ION CAN BE AVOIDED BY CAREFUL EXAMINATION OF THE SURFACE	CARRYING	THE IMPRESSION, NOTING THE LOCATION ON THE
K3	120	03	THE STREET. (ACCORDING TO THE FBI, IN 42 PERCENT OF ALL	CARS	STOLEN THE KEY HAS BEEN LEFT IN THE IGNITION.)
K3	372	03	AT A LARGE SHARE OF CITIZEN COMPLAINTS RADIOED TO PATROL	CARS	DID NOT INVOLVE VIOLATIONS OF THE CRIMINAL LAW
K3	384	08	AVOID DIRECT EYE CONTACT PERSON WEARING COAT ON HOT DAYS	CARS	WITH MISMATCHED HUB CAPS, OR DIRTY CAR WITH CLE
K3	600	12	TO ELIMINATE OBVIOUS OPPORTUNITIES FOR CRIMINALS. LOCKED	CARS	WELL -LIGHTED STREETS, ALARM SYSTEMS, AND PROP
K3	600	19	ON SCREENS; PROVIDING MORE EFFECTIVE LOCKING DEVICES FOR	CARS	HOUSES, AND OFFICES; SURROUNDING RESIDENTIAL A
K4	14	29	S AND EXPLOSIVES. 12. RECOGNITION AND RECOVERY OF STOLEN	CARS	13. TRAFFIC REGULATION AND CONTROL. 14. PURSUI
K4	44	04	CATION BETWEEN POLICE HEADQUARTERS AND THE ROVING PATROL	CARS	IS PROBABLY THE MOST IMPORTANT OF THEM ALL. THR
K6	188	20	ITY OF THOSE CARRYING OUT THE CHECKS, AND THE DETAILS OF	CARS	FOUND INSECURE WERE RECORDED; FROM THIS INFORMA
K6	188	22	ION IT WAS POSSIBLE TO IDENTIFY WHAT PROPORTION OF THOSE	CARS	LEFT INSECURE BELONGED TO VISITORS TO THE CITY,
K8	410	10	MS OF PASSANGER MILES, OR EVEN IN TERMS OF THE NUMBER OF	CARS	ON THE ROAD. THUS, IT MUST BE ACKNOWLEDGED THAT
K8	466	11	STRIAL ACTIVITIES. LARGE STORAGE TANKS AND RAILROAD TANK	CARS	OCCASIONALLY EXPLODE WHILE THEY ARE BEING PAINT
K6	3	09	OWN AREA, SOMETHING WHICH I ARRANGED TO DO WITH SHERIFF	CARSON'S	SPONSORSHIP. FOR FOUR MONTHS, FOR FOUR HOUR
K8	26	19	BE IN THE KIT. PARTICULARLY IN OUTDOOR CRIMES, CARDBOARD	CARTONS	ARE VERY USEFUL. THESE ALSO MAY USUALY BE OB
K7	61	10	OSSIBLE TO DETERMINE, FROM THE LOCATION OF THE RECOVERED	CARTRIDGE	CASES AND A KNOLEDGE OF THE WEAPON THAT FI
K7	61	12	RESTS ON THE FACT THAT AUTOMATIC WEAPONS DISCHARGE THEIR	CARTRIDGE	CASINGS AT CHARACTERISTIC ANGLES AND FOR C
K7	69	02	E. BOTH SPECIMENS FACE THE SAME DIRECTION. ONE BULLET OR	CARTRIDGE	CASE WILL BE TURNED UNTIL AN AREA CONTAINI
K7	69	04	NING PROMINENT STRIAE IS FOUND. THEN THE OTHER BULLET OR	CARTRIDGE	CASE WILL BE ROTATED IN SEARCH OF THE SAME
K8	382	02	A REPEATING WEAPON THAT, FOLLOWING THE DISCHARGE OF ONE	CARTRIDGE,	WILL AUTOMATICALLY EJECT THE EXPENDED CAR
K8	382	03	OF ONE CARTRIDGE, WILL AUTOMATICALLY EJECT THE EXPENDED	CARTRIDGE	CASE, LOAD A NEW CARTRIDGE INTO CHAMBER, A
K8	382	03	OMATICALLY EJECT THE EXPENDED CARTRIDGE CASE, LOAD A NEW	CARTRIDGE	INTO CHAMBER, AND REMAIN COCKED FOR THE NE
K8	382	12	AND MARKS THE PRIMER; THE EXTRACTOR, WHICH WITHDRAWS THE	CARTRIDGE	FROM THE CHAMBER; THE EJECTOR, WHICH MAY I
K8	382	13	THE CHAMBER; THE EJECTOR, WHICH MAY INDENT AND MARK THE	CARTRIDGE	CASE; AND POSSIBLY THE CHAMBER, BECAUSE IN
K8	382	14	Y THE CHAMBER, BECAUSE IN RARE INSTANCES IT MAY MARK THE	CARTRIDGE	CASE AT THE TIME OF THE DISCHARGE OR DURIN
K8	382	16	OVAL OF THE CASE. A PISTOL MAY LEAVE ADDITIONAL MARKS ON	CARTRIDGE	CASES, SOME ASSOCIATED WITH THE CLIP OR MA
K8	394	12	NGE CANNELURE, WHICH SERVES AS A RETAINER INTO WHICH THE	CARTRIDGE	CASE IS INDENTED TO HOLD THE BULLET IN PLA
K8	398	19	NSWER TO THE QUESTION 'FROM WHERE WAS THE BULLET FIRED?'	CARTRIDGE	AND SHELL CASES ALTHOUGH FIRED CASES ARE L
K8	402	02	DIFFERENT FROM THOSE OF MOST OTHER MARKS ON THE SHELL OR	CARTRIDGE	CASE, AND THEIR LENGTH IS OFTEN CONSIDERAB
K8	402	07	REPEATING AND AUTOMATIC WEAPONS ARE DESIGNED TO REMOVE A	CARTRIDGE	OR SHELL FROM THE MAGAZINE, FORCE IT (USUA
K8	402	09	O THE LATTER. WHILE THESE OPERATIONS ARE PROCEEDING, THE	CARTRIDGE	OR SHELL IS FORCED OVER A NUMBER OF METALL
K8	402	16	T BASIS FOR COMPARISON IDENTIFICATION. TO SUMMARIZE: THE	CARTRIDGE	OR SHELL CASE USUALLY CARRIES MARKINGS THA
K7	61	17	HAPED ARE CHARACTERISTIC FORMS OF FIRING PINS ON RIMFIRE	CARTRIDGES.	MOST FIRING PINS MAY BE SLIGHTLY OFF CEN
K7	56	09	S USED FOR RIMFIRE ONLY FIGURE 4-6: COMPONENT PARTS OF A	CARTRIGE	(COURTESY AM. INST. OF APPLIED SCIENCE). PR
K7	69	01	ACTLY AS IT WAS AT THE TIME OF THE CRIME. EACH BULLET OR	CARTRIGE	IS MOUNTED ON A MECHANICAL STAGE. BOTH SPEC
K1	3	22	R PROCEEDINGS, CIVIL OR CRIMINAL, IN RESPECT TO THE SAME	CASE	(OFFENCES AGAINST THE PERSON ACT 1861, SECTIONS
K1	3	23	S AGAINST THE PERSON ACT 1861, SECTIONS 44, 45). IN THIS	CASE,	THEREFORE, IF THE PERSON AGGRIEVED WISHES TO T
K1	9	12	POLICY OF THE CRIMINAL LAW TO PREVENT. FOR EXAMPLE, IN A	CASE	OF MURDER THE ACTUS REUS IS CONDUCT WHICH CAUSE
K1	15	24	VEHICLE THE ACCUSED WAS UNAWARE OF THIS OCCURRENCE. THIS	CASE	WAS A STRONG ONE BECAUSE SECTION 22 RE-ENACTED
K1	18	05	FECT THE COMPANY WITH LIABILITY. A STILL MORE REMARKABLE	CASE	WAS THAT OF ALLEN V. WHITEHEAD,D.C., IN WHICH T
K1	18	16	S VICARIOUS LIABILITY. TWO OF THE FIVE LAW LORDS IN THIS	CASE	WERE, HOWEVER, PREPARED TO ACCEPT THAT AN EXCEP
K1	24	12	TS; ONLY PARLIAMENT CAN AMEND THEM. ALTHOUGH M'NAGHTEN'S	CASE	WAS ONE OF DELUSIONAL INSANITY AND THE JUDGES'

U primjeru vidimo da se riječ *cartridge(s)* javlja 20 puta. Da se kolokacija *cartridge case* javlja osam puta, da se izraz *cartrige or shell* (odnosno obrnuto) javlja četiri puta. Naravno, da bismo kolokacije unijeli u pedagoški rječnik, potrebno je najprije definirati što je to kolokacija, na čemu se ovdje nećemo zadržavati.

Konkordancija će nam pružiti nezaobilazne pokazatelje za utvrđivanje značenja riječi (što je riječ »stručnija«, to je jednoznačnija), a posebno kad je u pitanju razdvajanje homografa koji su izrazito česti u općem engleskom jeziku, a nešto manje česti u stručnom engleskom jeziku. Zatim, konkordancija je nezaobilazan izvor za utvrđivanje kolokacije i idioma u kojima se javljaju riječi koje smo izabrali.

Danas je već moguća automatska lematizacija korpusa, npr. u Institutu für deutsche Sprache u Mannheimu razrađen je paket programa Lemma koji je 90% algoritam, što znači da ne pokriva svaku nepravilnost njemačkog jezika.

Spomenuti kompjutorski ispisi omogućuju također stvaranje ekscerpta za rječničke kartice, i tako dolazimo do pitanja što treba da sadrži rječnička natuknica u pedagoškom rječniku. Smatramo da treba da sadrži ove informacije:

— o ortografiji (upozoriti na varijante spelovanja, na primjer)
— o transkripciji (izgovoru)
— o funkciji (morfološkoj i sintaktičkoj)
— o frekvenciji
— o značenju (semantičku)
— o najčešćim kolokacijama, složenicama i idiomima.

Možemo sažeti da je osnovni proces dobivanja pedagoškog rječnika korištenjem automatski izrađenih konkordancija ovakav:

I — unos teksta
— izolacija riječi pojavnica u tekstu
— njihovo rangiranje po frekvenciji
— lematizacija
— abecedno sređivanje
— provjera u rječniku
— provjera u konkordanciji.

II Ispis natuknice:
— obrada (izgovor, vrsta riječi)
— izdvajanje homografa
— traženje kolokacija
— obrada na drugi jezik
— provjera
— konačna verzija.

Vrijedno je još spomenuti da je jedna od mogućih varijanti pedagoškog rječnika za izradu kojeg nam je vrlo korisna konkordancija tzv. kontekstualni rječnik.

Autori kontekstualnog rječnika geogologije[8] oslonili su se na korpus od 40 000 riječi te na osnovi njega sastavili rječnik za uvođenje u geološke studije od 1 500 leksičkih jedinica uz sintaktičku i semantičku analizu.

Shodno iznesenim stavovima zalažemo se za kompjutorsku obradu lingvističkih podataka samo kao za jedan od načina prikupljanja objektivnih podataka (npr. o frekvenciji, distribuciji i rangu riječi) koji može bitno pomoći lingvistima da, koristeći se podacima do kojih nije moguće doći, intuicijom dadu doprinos procesu učenja jezika.

U ovom slučaju to se odnosi na unapređivanje leksikografskog rada u domeni svladavanja stranog jezika struke.

LITERATURA

Bujas, Ž., »Concordancing as a Method in Contrastive Analysis.« *SRAZ*, 23, str. 49—62, 1967.

Bujas, Ž., »Kompjuterska konkordanca Gundulićeva »Osmana«, *Filologija*, 7, str. 35—59, 1973.

Burton, D. M., »Automated Concordances and Word Indexes: the Early Sixties and the Early Centres.« *Computer and the Humanities*, 15, str. 83—100, 1981.

Burton, D. M., »Automated Concordances and Word Indexes: the Fifties«. *Computer and the Humanities*, 15, str. 1—14, 1981.

Burton, D. M., »Automated Concordances and Word Indexes: the Process, the Programs, and the Products«. *Computer and the Humanities*, 15, str. 139—154, 1981.

Centre for Information on Language Teaching and Research, *Languages for Special Purposes*, London, CILT for British Association for Applied Linguistics, 1971.

Descamps, J. L. i dr., *Dictionnaire contextuel de français pour la géologie*. Paris, Didier, 1976.

Gačić, M., »Primjena kompjutora u pripremi materijala za analizu jezika struke«. *Zbornik II znanstvenega srečanja računalniška obdelava lingvističnih podatkov*, str. 335—344, Bled, 1982.

Guiraud, P., *Les caractères statistique du vocabulaire*, Paris, PUF, 1954.

Parunak, H. V. D., »Prolegomena to Pictorial Concordances«, *Computers and the Humanities*, 15, str. 15—16, 1981.

Pravda, V., »Expériences d'enseignement des langues de spécialité à l'Université Charles à Prague«, u *Les langues de spécialité analyse linguistique et recherche pédagogique*, Strasbourg, 1970.

Teubert, W., »Corpus and Lexicography«, *Zbornik II znanstvenega srečanja računalniška obdelava lingvističnih podatkov*, str. 275—301, Bled, 1982.

[8] Descamps, J. L. i dr.

COMPUTER CONCORDANCES IN COMPILING PEDAGOGICAL LSP DICTIONARIES

Summary

Concordances are standard tool in modern lexicology. Pedagogical LSP dictionaries are first of all functional types of dictionaries which enable the user to decode text efficiently using minimal lexical resources.
To compile the pedagogical LSP dictionary we need first of all frequency lists of the words used in a certain field (with the distribution data if possible) and concordance of the KWIC type. Using those data we have to: lemmatize the word lists, separate homographs and establish the most frequent and important collocations and idioms.
All the data should be listed by the alphabetical order, explained in the other language, chequed and presented in the final dictionary form.
Another form of the pedagogical LSP dictionary could be so called contextual dictionary of the given field, which is today almost impossible to compile without using concordances.