Janko Pečnik
Dunaj

UDK 81'374:004.5/.6

RAČUNALNIK PRI JEZIKOVNI OBDELAVI – TERMINOLOŠKE BAZE IN SLOVARJI

Osebni računalnik se je razvil v nujno potrebno sredstvo tako v jezikoslovnem raziskovanju kot v pisanju in oblikovanju besedil. S kvantitativno metodo – program za analiziranje besedil, napisan v Paradoxu – je možno dokazati jezikovne značilnosti določene zvrsti: fonološke, morfološke, leksikalne posebnosti idr. Na podlagi teh podatkov lahko naredimo selekcijo tistega gradiva, ki je pomembno za nadaljnjo polavtomatizirano terminološko obdelavo. Ta metoda omogoča črpanje podatkov, ki so dokumentirani v izhodiščnem gradivu, obdelavo podatkov in pravočasno reakcijo na nove jezikovne izzive. Terminološka baza je temelj za izdelavo splošnih slovarjev, tako za prevajalce kot za učitelje in učence.

računalniško podprta obdelava besedila, polavtomatična analiza, slovar, terminološka baza, slovarski podatek, morfološki podatek, fonološki podatek, konkordanca

Personal computers have become an essential tool for both linguistic studies and the practical application of languages. With a quanitative approach, i.e. a computer programme for the analysis of texts, reliable data, such as phonological, morphological, lexical information, and other data, can be obtained on the characteristics of a specific type of text. Based on this information, the text elements relevant for the development of terminological data bases can be selected and also processed in a semi-automatic way. This method offers the advantage that clear and documented results can be obtained and processed, and that new linguistic challenges can be identified at an early stage. According to the criteria employed and the structure of the data base, dictionaries can be generated from such a data base.

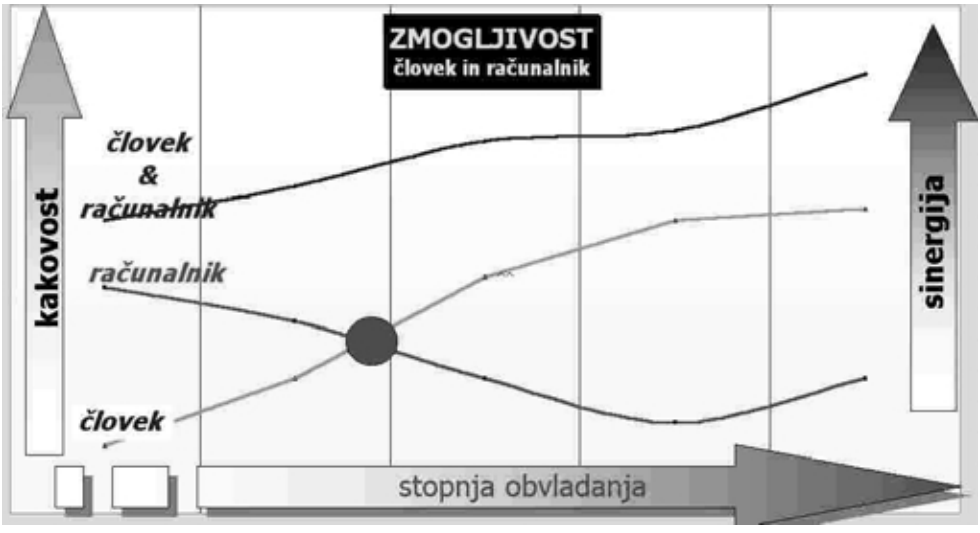
electronic word processing, semi-automatic analysis, dictionary, terminological data base, dictionary information, morphological information, phonological information, concordance

Uvod

Majhnim jezikovnim skupinam nikdar ne bo uspelo izdelati ustreznih slovarjev za vsa strokovna področja. Večje stroke, ki razpolagajo z večjimi finančnimi sredstvi, so in bodo v boljšem položaju. Prevajalci, učitelji in vsi ostali, ki skrbijo za nemoteno komuniciranje znotraj teh skupin in preko njih, bodo še naprej prepuščeni sami sebi.

Z iznajdbo in izpopolnitvijo osebnih računalnikov je tudi za jezik napočil novi vek, ker računalniki niso samo sredstvo za pisanje in oblikovanje besedil, temveč

pokorni hlapci pri raziskovanju jezikovnih vprašanj. Tehnično so že toliko dozoreli, da delajo dokaj stabilno, zmorejo kar precejšno količino podatkov in so se uveljavili v vseh podjetjih, šolah in gospodinjstvih. Njihova razširjenost, sorazmerno nizka cena in velika zmogljivost odpirajo nove možnosti tako na raziskovalni kot tudi na uporabniški ravni. Treba bo najti samo ravnovesje med tistim, kar lahko prepustimo računalniku, in tistim, kar mora napraviti človek sam. Rezultati matrične preveritve argumentov za računalniško oziroma ročno analiziranje in obdelavo jezikovnega gradiva v grafiki jasno ponazarjajo, da je skupna pot najboljša rešitev.



Če seštejemo kriterije (hitrost, točnost, količina idr.) vidimo, da je na vseh stopnjah jezikovnega obvladovanja smiselna uporaba računalnika in da so vsa prerekanja in ugibanja, ali bo računalnik nadomestil človeka ali ne, jalova.

Analiziranje

Zgornja preglednica, vsakodnevna praksa, pomanjkanje ustreznih slovarjev, pomanjkanje časa me je privedlo k temu, da sem uporabnost osebnega in prenosnega računalnika praktično preveril. Relacionalna podatkovna baza[1] se mi je ponudila kot primerno računalniško orodje za kvantitativne analize izbranih besedil Ive Andrića, Zije Dizdarevića, hrvaških uradnih listov in slovenskih vojaških uradnih listov.

Namen raziskav je bil:

a) preveriti zmogljivost osebnega in prenosnega računalnika pri jezikovnih raziskavah;

[1] © Paradox – Borland

b) razviti tako za raziskovanje kot tudi za praktično uporabo primeren računalniški program;
c) vse to z namenom, da bi čim več tega lahko uporabili pri vsakdanjem delu, pri raziskovanju sinhronih in diahronih jezikovnih pojavov, pri prevajanju, pouku, pri izdelavi učnih pripomočkov in slovarjev.

Izid tega podviga je bila ugotovitev, da je računalnik tisto sredstvo, s katerim lahko kljubujemo jezikovnim izzivom sodobnega časa. S pomočjo relacionalne podatkovne baze je možno manipulirati z velikimi količinami jezikovnih podatkov. Čim bolj vešča sta uporabnik ali raziskovalec, tem podrobnejši in točnejši so jezikovni izidi.

Ker so besedila pisana v različnih oblikah, pisavah in formatih, so pri računalniški obdelavi potrebne določene konvencije, ki obdelavo šele omogočijo. To so:

a) Sistemsko pogojeni format: bibliografski podatki, označevanje strani, delitev besed itd. Brez tega sistem ne more prepoznati, kaj je del besedila in kaj ne. Na primer: ali je številka, ki jo program najde, oštevilčenje strani ali je del besedila. Bibliografski znak in oštevilčenje strani potrebuje sistem za dokumentiranje podatkov. S tema dvema podatkoma so povezana vsa gesla in kolokacije v terminoloških bazah, kar omogoča in olajšuje evalvacijo terminološke baze. Vsak podatek ima interno oznako, v katerem besedilu, na kateri strani in vrstici se je nahaja.
b) Sprememba grafemov v foneme. To je potrebno za fonemske analize: razporeditev, redukcija in pogostnost fonemov, dolžina besed idr.
c) Določanje stavčne meje. Za stavčne analize je potrebno, da se ročno preverijo stavčne meje, ker stavčni znaki sami za računalnik niso zanesljiv kriterij za stavčno mejo.

Po oblikovanju besedila sistem sprejme besedilo za kvantitativne analize. Prvi korak je avtomatična izdelava raznih konkordanc, ki so potrebne za nadaljnja raziskovanja. Nato sledi lematizacija, ki poteka le na pol avtomatično. To pomeni, da sistem na podlagi primerjav že obdelanih besedil samostojno vnese že obstoječe podatke iz podatkovne baze oziroma v primeru nejasnosti zahteva odločitev od uporabnika. Nove besede, ki še niso lematizirane, mora uporabnik ročno dodati. Z vsakim novim postopkom se dopolnjuje podatkovna baza. Obenem se pri postopku lematizacije lemam dodajo slovnični, etimološki, stilistični kriteriji in za stroko pomembne informacije.

Slika na zaslonu pri lematiziranju

Znaki na zaslonu, ki imajo svojo številčno vrednost in kriteriji iz postopka lematizacije omogočajo računanje z jezikovnimi podatki.[2]

Po postopku lematizacije so na razpolago najvažnejši statistični podatki raziskovalnega besedila kot na primer:

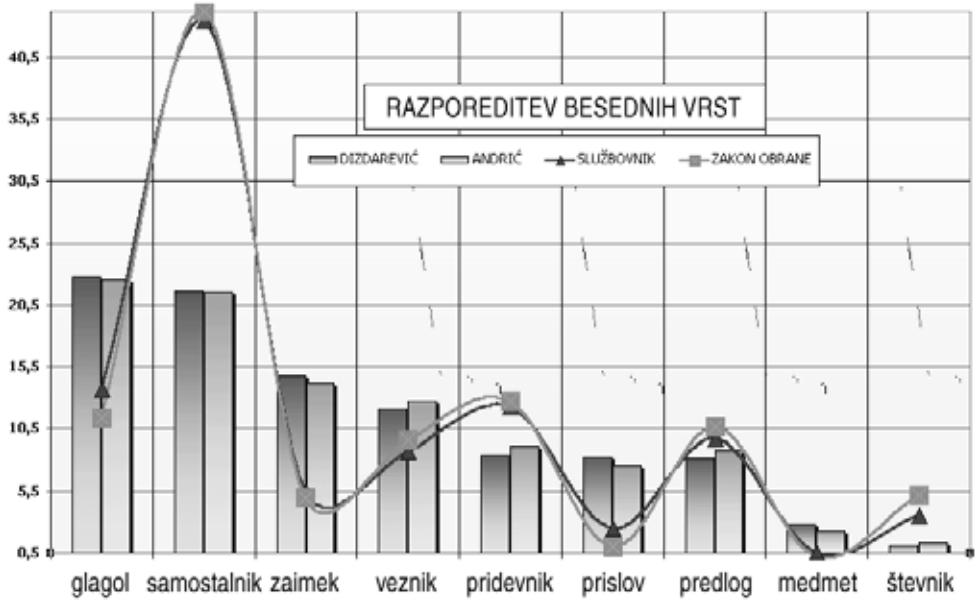
– avtosemantični in sinsemantični delež besedišča v besedilu;
– pogostnost fonemov ali črk;
– etimološke informacije – razporeditev prevzetih besed;
– razporeditev besednih vrst;
– razporeditev besednih oblik;
– razporeditev lem;
– dolžine besed in stavkov;
– konkordance (navadna, odzadnja);
– konkordanca stavčnih znakov;
– lematizirana konkordanca.

Te informacije so standardizirane in so na razpolago za vsako v program vneseno besedilo. Že te preglednice same so dobro merilo za primerjalno ocenjevanje lastnosti besedila. Podatki dajejo vpogled v splošne jezikovne zakonitosti ter zajemajo avtorjeve jezikovne lastnosti. Možne so še druge kvantitativne informacije, ki pa so odvisne od raziskovalnega cilja. Cilj narekuje kriterije, ki se morajo pri postopku lematizacije dodati.[3] Različni kriteriji omogočajo v najkrajšem času prikrojiti konkordance, jih omejiti na določene besedne vrste, prevzete besede, nestandardne jezikovne prvine, stilsko zaznamovano besedišče itd.

[2] Nekaj teh statističnih podatkov vsebuje članek D. Pečnik *Računalnik pri jezikovni obdelavi – možnosti in dimenzije na osnovi literarnih besedil* v tem zborniku (str. 487–494).

[3] Npr. literarna analiza: prostor, čas, figure ...

Temeljna informacija za terminološko bazo je informacija o besedni vrsti. Besedne vrste in semantični kriteriji služijo za selekcijo lematizirane konkordance, ki je pri nadaljnjem postopku temelj za slovarsko obdelavo besedila. Zvrstnost besedila določa razporeditev in pogostnost besednih vrst in s tem vpliva na razporeditev gesel v terminološki datoteki. V datoteko pride samo to, kar je dokumentirano v besedilu. Računalnik ne dovoljuje samovoljnega vnašanja ali izpuščanja podatkov. To se pravi, da sta besedilo in terminološka baza neločljivi enoti. Besedilo se odslikava v terminološki datoteki. Izbira neprimernega besedila spremeni zaporedje in vsebino namenske terminološke baze. Ne glede na zgornje kriterije grafika kaže, da je za izdelavo terminološke baze oziroma standardnih slovarjev potrebna predhodna analiza besedila.

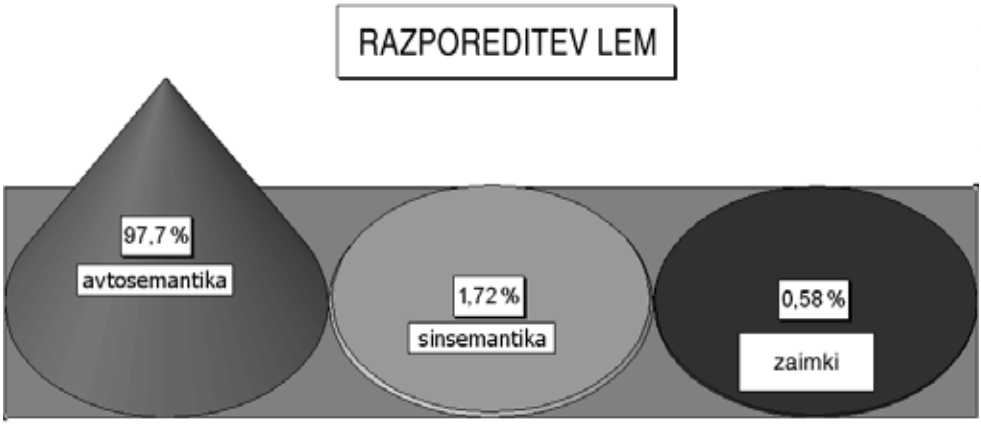


Terminološka podatkovna baza

Uporabnik elektronskih podatkovnih baz in slovarjev narekuje strukturo baze. Komercialni sistemi imajo v glavnem prikrojeno strukturo za prevajalce in komaj upoštevajo potrebe učiteljev in učencev. Prevajalec išče v bazah ustreznice in razlage v ciljnem jeziku. Učitelj, ki poučuje jezik po komunikacijskem načelu, išče poleg tematsko vezanega besedišča tudi informacije o pogostnosti, učenec med drugim tudi slovnične in pravopisne informacije. Če podatkovna baza te točke v svoji strukturi in vsebini upošteva, je samo še tehnično vprašanje spojiti zahtevane elemente baze, jih ustrezno oblikovati in iztisniti na zaslonu ali jih prirediti za knjižno obliko. To pomeni, da je podatkovna baza načelno obsežnejša od knjižnih

verzij. Zaradi tega velja tudi načelo: česar ni v datoteki, ni mogoče ne najti niti iztisniti.

Postopki:

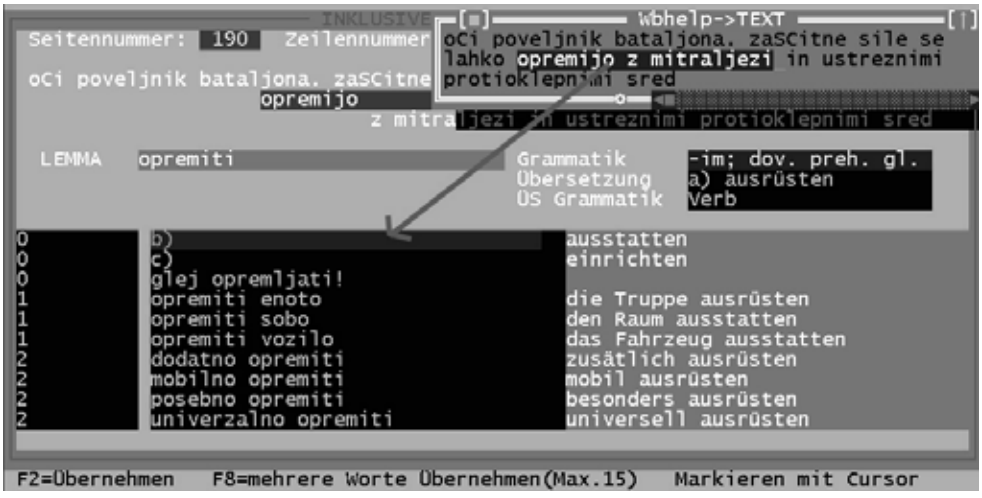


Po računalniški analizi besedila in končani lematizaciji sledi selekcija. To pomeni, da se iz lematizirane konkordance izločijo vsi deli, ki za terminološko bazo niso potrebni. Pri strokovnem slovarju so to sinsemantični del in določeni deli besedila kot topografska imena, zaimki in števniki. Na ta način se skrči konkordanca, pomembna za slovarsko obdelavo, na manj kot 60 %.

Za tem sledi polavtomatično prevzemanje slovarskega gradiva iz preostale konkordance v slovarsko podatkovno bazo. Lekseme prevzema računalnik avtomatično. Vse ostale informacije (npr. oblikoslovne, slovnične) mora uporabnik ročno vstaviti. Pri ponovnem pogonu sledi samo še dopolnjevanje že obstoječih podatkov. V primeru dvojezične ali večjezične baze moramo dodati še ustreznice iz ciljnega jezika. Za zagotovitev standarda služijo normativni slovarji.

Naslednji korak je polavtomatično prevzemanje kolokacij iz konkordance v podatkovno bazo. Če kolokacija, ki je del stavka oziroma besedila v konkordanci, ni v osnovni obliki, jo moramo spremeniti v osnovno obliko. Dosedanje izkušnje so pokazale, da pride na eno geslo povprečno 10 kolokacij. Izid analize vojaško-pravnih besedil[4] je okrog 2000 gesel in 20.000 kolokacij.

[4] Na spletni strani MO RS objavljeni službeni listi.



Polavtomatično dopolnjevanje terminološke baze s kolokacijami

Predpogoj za ažuriranje in evalvacijo leksikografskih podatkov je standardizirano gradivo, kar pomeni izbiranje standardnih besedil in standardiziranje izidov v skupinskem delu jezikoslovcev in strokovnjakov dane stroke. Samo strokovnjaki določene stroke lahko potrdijo razlage kolokacij in omočijo najti ustreznice v ciljnem jeziku. Ustroj datoteke, razporeditev kolokacij po skupinah, slovnični in drugi kvalifikatorji omogočajo tudi zahtevnejšo iskanje in selektiranje za namensko uporabo, tako za prevajalca kot tudi za lektorja.

Priloga: Primer gesla *vojak*

voj<u>a</u>k -a m *der* Soldat; *čin – der* Rekrut; *der* Grundwehrdiener; *zastar. – der* Wehrmann; **~ obveznik** Wehrpflichtiger; **~ iz stalne sestave** Kadersoldat; **~ med služenjem vojaškega roka** Grundwehrdiener; **~ na izhodu** Soldat beim Ausgehen; **~ s činom** Charge; **~ v postroju** Soldat in der Formation; **~ v pripravljenosti** Soldat in Bereitschaft; **~ za posredovanje** Bereitschaftssoldat, Eingreifsoldat; **buditi ~a** den Soldaten wecken; **določiti ~a** den Soldaten bestimmen, einteilen; **dovoliti ~u** dem Soldaten genehmigen, erlauben; **izdati ~u dovoljenje** dem Soldaten die Erlaubnis ausstellen; **nadzorovati ~a** den Soldaten überwachen; **nadzorovati vstajanje ~a** die Tagwache des Soldaten überwachen; **nagraditi ~a** den Soldaten belohnen; **nasloviti ~a** den Soldaten ansprechen; **oborožiti ~a** den Soldaten bewaffnen; **odobriti ~u** dem Soldaten genehmigen; **odobriti ~u izhod** dem Soldaten den Ausgang genehmigen; **odpustiti ~a** den Soldaten entlassen; **odpustiti ~a iz rezervne sestave** den Soldaten entordern; den Soldaten aus der Miliz entlassen; **odrediti ~a** den Soldaten einteilen; **peljati ~a** den Soldaten führen; **poklicati ~a** den Soldaten aufrufen; **poslati ~a** den Soldaten entsenden; **postrojiti ~a** den Soldaten antreten lassen; **povišati ~a** den Soldaten befördern; **pozdraviti ~a** den Soldaten grüßen; **prebrati seznam ~ov** die Namensliste der Soldaten verlesen; **pregledati ~a** den Soldaten überprüfen, untersuchen *(medizinisch)*; **seznaniti ča** den Soldaten in Kenntnis setzen; **sporočiti**

~u dem Soldaten mitteilen; **sprejeti ča** den Soldaten aufnehmen; **usposobiti ~a** den Soldaten ausbilden; **zbrati ~e** die Soldaten sammeln; **čin ~a** Soldatendienstgrad; **dan odpusta ~a** der Entlassungstag des Soldaten; **dolžnost ~a** Funktion des Soldaten; **generacija ~ov** Einrückungsturnus; **izhod ~a** Ausgang des Soldaten; **izobrazba ~a** Bildung des Soldaten; **motivacija ~a** Motivation des Soldaten; **odgovornost ~a** Soldatenverantwortung; **odpust ~a** Entlassung des Soldaten; **počastitev pokojnega ~a** Ehrung des verstorbenen Soldaten; **počastitev ~a** Ehrung des Soldaten; **pogreb ~a** Bestattung des Soldaten; **povišanje ~a** Beförderung des Soldaten; **prevoz ~a** Soldatentransport; **prihod ~a** Einrücken des Soldaten; **prihod ~a iz rezervne sestave** Einrücken des Milizsoldaten; **prihod ~a na služenje vojaškega roka** Einrücken des Soldaten zum Grundwehrdienst; **prisega ~a** Angelobung des Soldaten; **prisotnost ~a** Anwesenheit des Soldaten; **razporeditev ~a** (Mob-) Einteilung des Soldaten; **seznam ~ov** Liste der Soldaten; **skupina ~ov** Gruppe von Soldaten; **sorodnik ~a** Angehöriger des Soldaten; **število ~ov** *(zahlenmäßige)* Stärke der Soldaten; **uradni razgovor ča** Rapport, dienstliche Aussprache des Soldaten; **usposabljanje ~a** Soldatenausbildung; **vstajanje ~a** Tagwache des Soldaten; **disciplinski postopek zoper ~a med služenjem vojaškega roka** Disziplinarverfahren gegen Präsenzdiener; **oznake činov za ~e** Soldatendienstgradabzeichen

Literatura

Militärwörterbuch Slowenisch – Deutsch (*Wehrrecht und Innerer Dienst*); *Vojaški slovar slovensko-nemški* (Vojaško pravo in Notranja služba), 2003. Wien: Sprachinstitut des Bundesheeres, LVAK.

Johann Pečnik, 2002: *Quantitative Textanalyse. Computerunterstützte Untersuchungen an ausgewählten texten – Erzählungen von Zija Dizdarević und von Ivo Andrić*. Dissertation. Universität Klagenfurt.