# NOVE TEHNOLOGIJE

Daniela Pečnik
Dunaj

UDK 821.163.6.08 Lainšček F. :81'42:004.5/.6

# RAČUNALNIK PRI JEZIKOVNI OBDELAVI – MOŽNOSTI IN DIMENZIJE NA OSNOVI LITERARNIH BESEDIL

Pričujoči članek nakazuje nekaj možnosti, ki jih imamo pri analizi besedil s pomočjo računalnika. Izhodišče te kvantitativne analize je bil slovenski roman *Namesto koga roža cveti* avtorja Ferija Lainščka. Korpus zajema 47.976 besed (pojavnic) in 8.049 različnih leksemov. Rezultati kvantitativne analize prispevajo statistično gradivo za nadaljnja jezikovna raziskovanja, vendar je iz članka razvidno, da nam uporaba računalnika na področju jezikovnega raziskovanja sicer olajša delo, ker je možno obdelovanje večjih datotek hitrejše in zanesljivejše, a človeka, ki odloča o vprašanjih, ki presegajo kvantitativni vidik jezika, z računalnikom ne bo mogoče zamenjati.

računalniško podprta obdelava besedila, polavtomatična analiza, statistično gradivo, pogostnost, relativna pogostnost, morfološki podatki, fonološki podatki, samostalnik, predlog, medmet

The following article exemplifies the possibilities of computer-aided text analysis on a sample of Slovenian text from the novel *Namesto koga roža cveti* by Feri Lainšček. The novel has a corpus of 47,976 words (tokens) and 8,049 different lexemes. The results provide statistical material for both continuative and comparative study. The article refers to the inevitability of computer technology in linguistic research because computers make data management and handling more reliable and efficient. Despite the technical facilitation computers can provide, human intervention is still required to make decisions about questions that go beyond the quantitative aspect.

electronic word processing, semi-automatic analysis, statistical information, frequency, relative frequency, morphological information, phonological information, noun, preposition, interjection

Jezik predstavlja orodje, s katerim stopa človek v komunikacijo s soljudmi. S pomočjo jezika je človeku omogočeno izražanje mišljenja, čustev in občutkov. Dandanes, v času novih tehnologij, ki nezadržno vstopajo v naše vsakdanje življenje, je človek s svojim jezikom in jezikovnim znanjem postavljen v novo situacijo, saj – ekonomsko gledano – v prihodnosti ne bo mogel shajati brez tehničnih pripomočkov, predvsem ne brez računalnikov. Prav računalnik pa je tehnologija, ki nam lahko bistveno olajša delo tudi na področju jezikoslovja, saj z njegovo pomočjo upravljamo in obdelujemo zmeraj večje datoteke še hitreje in zanesljiveje kot doslej. Delo z računalnikom nam poleg tega omogoča hitrejšo manipulacijo, dokumentacijo in aktualizacijo podatkov.

Pričujoči članek je poskus, nakazati možnosti kombinacije »človek – računalnik« na področju jezikovnih raziskovanj ter prispevati statistično gradivo za nadaljnja jezikovna primerjalna dela.

Izhodišče za kvantitativno analizo s pomočjo računalnika je bil roman *Namesto koga roža cveti* avtorja Ferija Lainščka,[1] to se pravi literarno besedilo kot sestavni del vsake jezikovne skupnosti. V literaturi se po eni strani zrcali jezikovno stanje jezika, po drugi strani pa imajo pisatelji v literaturi možnosti za inovacije v jeziku. Poslužujejo se lahko besedišča, oblik in skladenjskih sprememb, ki odstopajo od pravil slovnice in standardnega jezika, vendar so v vsakdanjem jeziku povsem živi. S kvantitativno analizo sodobnih besedil lahko zajamemo jezikovno stanje in predvidevamo možne razvojne tendence jezika, ker lahko izhajamo iz dejstva, da vsakdanji jezik in literatura vplivata drug na drugega.

Najvažnejši postopek pri takšnih analizah je lematizacija, ker imamo pri tem postopku največ možnosti manipulacije s podatki. Lematizacija je postopek, ki ni popolnoma avtomatiziran in nam omogoča dodajanje različnih kvalifikatorjev, odvisno od zastavljenega cilja. Podatkovna baza je povezana s slovarjem, kjer se shranujejo že lematizirani leksemi. Če začnemo obdelovati novo besedilo, računalnik sprva poseže po že shranjenih podatkih, lematizirati moramo nato le še lekseme, ki jih računalnik nima v slovarju. Če je v slovarju shranjenih več variant, se mora uporabnik odločiti za eno izmed njih (primer: *da* = 3. oseba ednine glagola *dati*; *da* = medmet, *da* = veznik).

Statistični rezultati so različne narave. Po eni strani nam odgovarjajo na vprašanje, za kakšno vrsto besedila pri analizi gre, po drugi strani pa podatki zajemajo vsebinske in jezikovne značilnosti. Kar zadeva statistične rezultate vsebinske narave, je treba pripomniti, da so odvisni od cilja, ki si ga moramo zastaviti že pred analizo, ker nas šele kvalifikatorji, ki jih pri lematizaciji dodajamo leksemom, privedejo do rezultatov zastavljenega cilja analize. Brez dodajanja kakršnihkoli kvalifikatorjev, to se pravi avtomatično, pridemo do rezultatov, ki zajemajo jezikovne značilnosti besedila (glasoslovje, oblikoslovje, skladnja in besedišče).

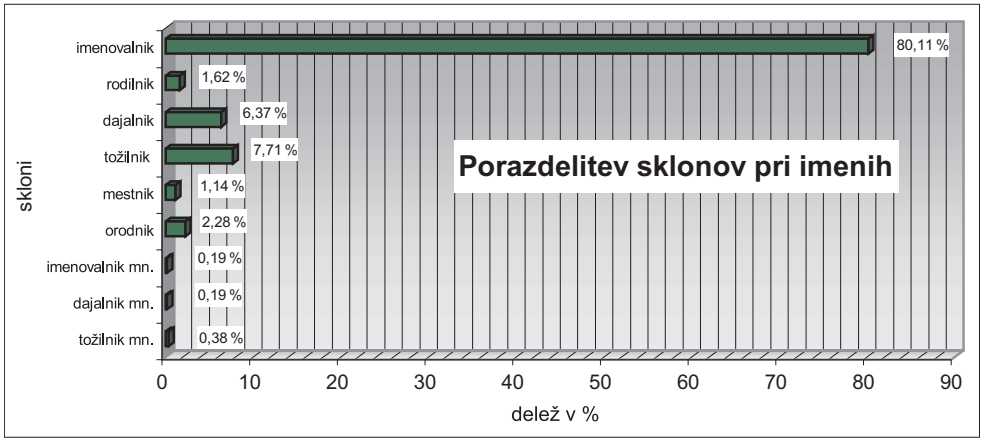
Za kvantitativno analizo je roman *Namesto koga roža cveti* s svojim korpusom 47.976 besed pojavnic primeren, ker predstavlja dokaj pregledno besedilo. Osnova tem 47.976 pojavnicam je 8.049 besednih oblik, te besedne oblike pa pripadajo 4.438 različnim (besednim) leksemom (lemam).[2] Če primerjamo med seboj lekseme in besedne oblike, je povprečna pogostnost leksemov v besedilu 1,81, medtem ko se vsak leksem pojavi v nasprotju s pojavnicami povprečno 10,81-krat. Povprečna frekvenca 10,81 je v primerjavi z že obstoječimi tovrstnimi analizami relativno visoka. Kot razlago navajam tri vzroke: pogostnost je v neposredni povezavi z vrsto

[1] Feri Lainšček, *Namesto koga roža cveti*, Ljubljana: Prešernova družba, 1991.

[2] *Pojavnica* (token) je ena pojavitev besedne oblike v besedilu/korpusu. *Besedna oblika* (word form) je beseda, kakor je zapisana črka za črko med presledki oz. ena od oblik v oblikoslovni paradigmi. *Lema* (lemma) je osnovna oblika besede kot del leksikona besednih oblik. (Op. ur.: Za pomoč se zahvaljujemo S. Kreku. V nadaljevanju namesto izraza lema puščamo avtoričin izraz leksem.)

besedila in z njegovim obsegom. Dodatno je odvisna od posebnosti, ki jih moramo pripisati avtorju, ker je od avtorja odvisno, kakšno besedišče izbere ter kako in kolikokrat ga v svojem besedilu zares tudi realizira. Iz analize je razvidno, da je povprečna pogostnost leksemov večja, čim daljše je besedilo.

Glede na zaželeni cilj in uporabljene kvalifikatorje pri lematizaciji nam kvantitativni rezultati odgovarjajo na vprašanja, povezana s temo, figurami, prostorom, časom, družbenimi sloji itd. Za ponazoritev si natančneje oglejmo figure, ki jih najdemo v romanu *Namesto koga roža cveti*. Pri branju besedila bralec neposredno karakterizira nastopajoče osebe kot glavne junake, stranske osebe in osebe, ki nimajo aktivne vloge. Pogostnostna razdelitev oseb nato potrdi bralčevo porazdelitev ali tudi ne. V obdelanem besedilu se pojavi junak Halgato 222-krat, sledita mu Pišti (189-krat) in Bumbaš (177-krat). Vse ostale osebe se z imenom pojavijo v besedilu manj kot 67-krat, več kot polovica celo manj kot 10-krat. Oseb, ki ne igrajo aktivne vloge, je 38,30 %. Porazdelitev imen po sklonih je pokazala, da je 80,11 % imen v imenovalniku, kar kaže na dinamiko besedila, saj nam imenovalnik pove, kdo kaj dela, po drugi strani pa je to treba pripisati pogostemu prememu govoru.




Najbolj pogoste osebe, Halgato, Pišti in Bumbaš, imajo v besedilu celo dve različni imeni oz. tri. Tako je glavni junak Halgato poleg Mežikaš, kar izraža eno njegovih značilnosti, poimenovan tudi Šanji. Veliko stranskih oseb ima dvojna imena, vendar pisatelj dvojnega poimenovanja ne uporablja dosledno. To pomeni, da tako prvi kot tudi drugi del imena lahko odpade.

Književni čas iz romana ni natančno razviden. Razen letnice 1932, ki pa ni v neposredni povezavi z glavnim dogajanjem, ni nobenih drugih letnic. Enako kot književni čas, tudi književni prostor ni natančno določen. Poimenovanja krajev Lacki roma (72-krat), Velika vejs (20-krat) in Mesto (28-krat) so izmišljena, vendar najdemo v romanu tudi resnična zemeljepisna imena (Amerika – 6-krat, Avstralija – 1-krat, Brazilija – 3-krat, Kanada – 1-krat in Jugoslavija – 1-krat).

Besedišče je del jezika, ki je podvržen stalnim spremembam. Besedišče ni nič stalnega, je dinamično in se po eni strani spreminja zaradi zgodovinskih preobratov, po drugi strani pa spremembe zahtevajo gospodarski, tehnični, politični in družbeni razvoj. Nove besede se sprejemajo v besedišče, druge izginjajo iz vsakdanje rabe. Besede pa tudi spreminjajo svoj pomen in svojo frekvenčnost. V ožjem pomenu predstavlja besedišče knjižni jezik, v širšem pomenu mu lahko prištevamo tudi regionalno in narečno besedišče. Ker je besedišče orodje, s katerim pisatelj ustvarja, postane prav besedišče zanj značilno.

Če si v obravnavanem romanu natančneje ogledamo jezik, ki ga lahko zasledimo v premem govoru, nas to privede do zaključka, da gre za jezik nižjega socialnega sloja. Feri Lainšček je skušal s pomočjo stilizacije predstaviti bralcu jezik, ki je živ med tistimi, ki se niso učili knjižnega jezika in govorijo, kakor pač vedo in znajo. Zato ni čudno, da je moral seči po besedišču, ki je sicer v pogovoru živo, vendar ne ustreza knjižni normi in je zaradi tega v slovarju[3] zaznamovano z različnimi kvalifikatorji. Celotni korpus vsebuje 679 zaznamovanih pojavnic, ki pripadajo 263 različnim leksemom. Zaznamovane lekseme lahko porazdelimo v šest skupin: v skupino slabšalnih, pogovornih, starinskih in vulgarnih, v skupino regionalizmov in skupino razno.[4] Analiza je pokazala, da je samostalnik tista besedna vrsta, ki najpogosteje odstopa od knjižnega jezika, čeprav je pri porazdelitvi pogostnosti besednih vrst s 14,84 % šele na četrtem mestu. Rezultati torej kažejo na to, da so samostalniki kot izrazi za predmetnost govorcem najbližji in da zaradi tega dejstva pri govorcu s pomočjo samostalnika najlažje pride do konotacij.

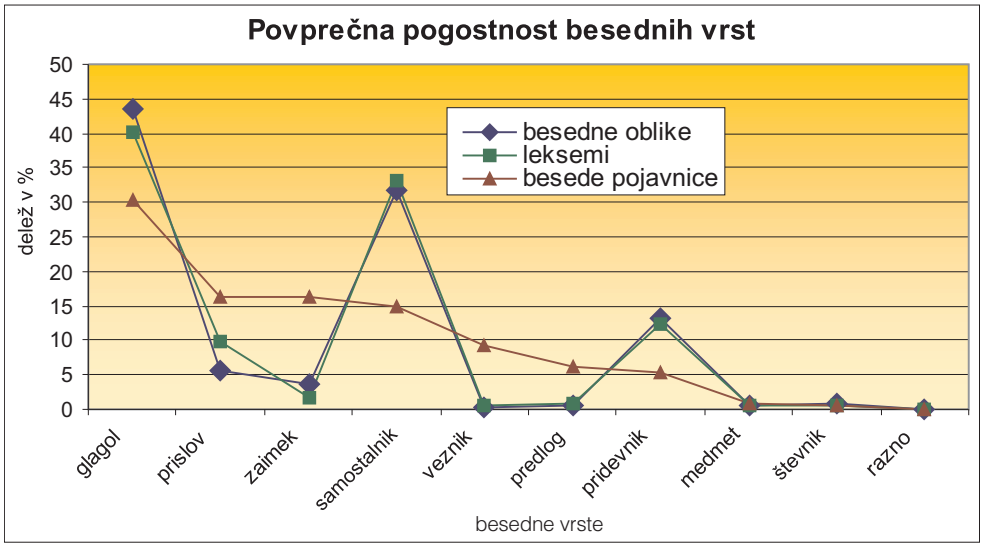
Za roman je značilen tudi položaj žensk, ki se kaže po eni strani v njihovi funkciji v družbi in po drugi strani pri izbiri besed za karakterizacijo tega spola. Ženske izvajajo le okrog 18 % poklicev, ki se pojavljajo v besedilu in celo ti poklici so izbrani stereotipno: kuharica, točajka, vzgojiteljica itd. Če primerjamo pogostnost poimenovanj za ženske in moške osebe, je poimenovanj za ženske osebe le okrog 1/5 moških.

Pri lematizaciji pojavnic je bila dodelitev besednih oblik ustreznim osnovnim oblikam in porazdelitev v besedne vrste izvedena na podlagi *Slovarja slovenskega knjižnega jezika*. Besedne vrste niso razdeljene po sintaktičnih kriterijih, temveč po tradicionalni razvrstitvi, kot jo najdemo v šolski slovnici. Deležnike (končnice na *-l*, *-n/-t*, *-č*), ki imajo sicer značilnosti pridevnika, a so izpeljani iz glagolov, sem dosledno pripisala glagolom. Za obravnavani roman dobimo naslednjo porazdelitev leksemov v besedne vrste: glagol 30,40 %, prislov 16,33 %, zaimek 16,30 %, samostalnik 14,84 %, veznik 9,26 %, predlog 6,25 %, pridevnik 5,25 %, medmet 0,75 %,

[3] *Slovar slovenskega knjižnega jezika*, elektronska izdaja, verzija 1.0, Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, DZS, Amebis d. o. o., 1993–1997.

[4] Skupina »razno« zajema lekseme, ki jih ni v slovarju niti jih avtor sam ni mogel klasificirati. Dodatno so v tej skupini besede, ki so sicer po svoji osnovni obliki knjižne, vendar se v besedilu pojavijo v nenormativni obliki (npr. nooože, škaaarje itd.). Tudi namenilnik, ki se v romanu pojavi 10-krat, sem uvrstila v to skupino.

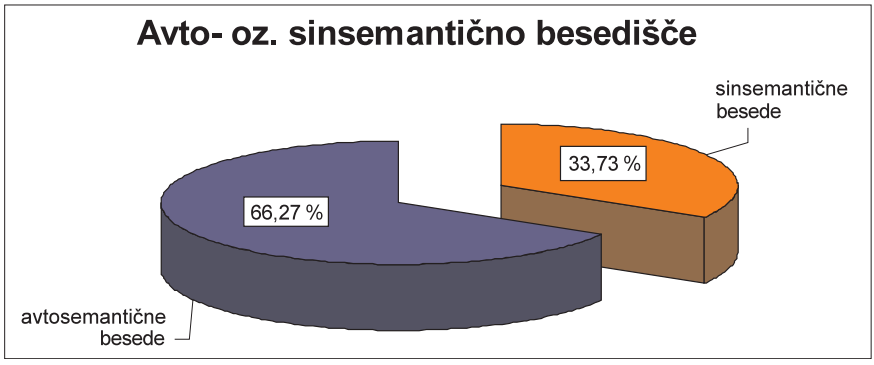
številnik 0,60 % in razno 0,02 %. Porazdelitev besednih vrst potrjuje, da je obravnavano besedilo literarno in ne strokovno, da prevladuje dinamičnost pred statičnim in konkretnim elementom. Merljivost namreč zaradi majhnega števila številnikov, razmeroma nizkega števila samostalnikov in pridevnikov ter zaradi velikega števila prislovov izključujem. Besedna vrsta, ki besedilo prav tako poživi, so medmeti. Medmeti so z 0,75% na videz le redko zastopana besedna vrsta, vendar v primerjavi z drugimi analizami iz literature pri Lainščku relativno pogosta.




Če govorimo o pogostnosti besednih vrst, ne smemo spregledati, da ni nujno, da so besede (pojavnice) najbolj frekventnih besednih vrst tudi relativno najbolj pogoste. Čim manjše je namreč število besed ene besedne vrste, tem večja je njihova relativna pogostnost. Iz analize je torej razvidno, da moramo razen zaimkov pripisati tri relativno najbolj pogoste besede besednim vrstam, ki so nepregibne in nepolnopomenske (sinsemantične). Če ponazorimo, bi to pomenilo, da se vsak posamezni veznik v romanu pojavi v povprečju 177,68-krat, medtem ko vsak glagol najdemo povprečno le 8,2-krat.

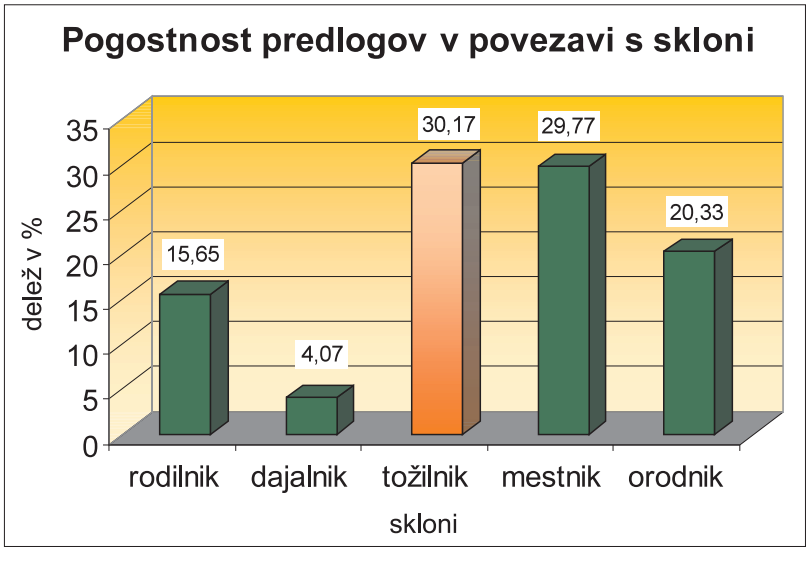
Razen glagola *biti* in zaimkov kot *on* in *ta* pripada 10 napogostejših leksemov nepregibnim besednim vrstam. Skupaj zajema teh 10 leksemov 30,76 % tega korpusa s 47.976 besedami pojavnicami.

Besedišča ne členimo le po besednih vrstah, pač pa lahko tudi glede na to, ali nosijo besede pomen ali izvajajo samo določeno funkcijo v stavku. Od teh 47.976 besed pojavnic jih pripada 31.795 avtosemantičnemu, ostalih 16.181 sinsemantičnemu besedišču. Razmerje avtosemantičnega besedišča proti sinsemantičnemu je 2 : 1.




Samostalnikov je v celotnem korpusu 7.118 ali 14,84 %. Če razčlenimo samostalnike po spolu, opazimo, da prevladuje z 49,61 % moški spol pred ženskim spolom z 39,45 %. Najmanj zastopan je srednji spol z 10,94 %, kar bi ustrezalo 779 pojavnicam in 203 leksemom. Posebej je treba izpostaviti besede ženskega spola s končnico na *-ost*, ki jih uvrščamo v 2. sklanjatev in zajemajo kar 34,45 %.

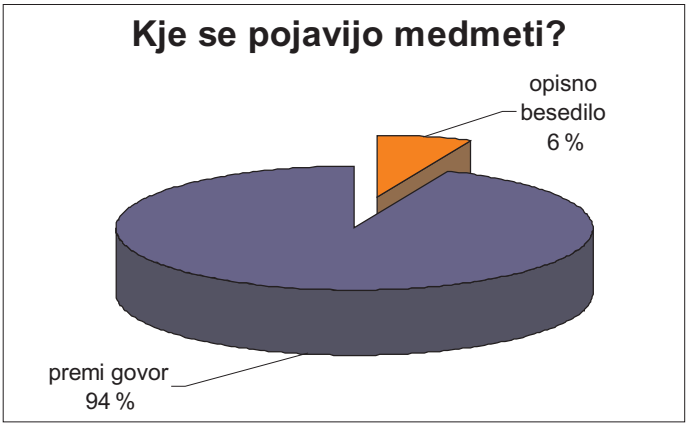
Predlogov je v romanu 36 različnih, uporabljeni pa so v pogostnosti 6,25 %. Rezultati analize so pokazali, da je predlogov, ki zahtevajo tožilnik, največ, in sicer 30,17 %. Z 29,77 % jim sledi skupina predlogov, ki zahtevajo mestnik. Najslabše zastopana skupina so predlogi z dajalnikom, saj jih je le 4,07 %. Približno 15 % predlogov stoji pred zaimki, vsem drugim sledi samostalnik. Število predlogov je v neposredni povezavi s samostalniki in zaimki. To se pravi, da je število predlogov tem večje, čim več je samostalnikov in zaimkov. V romanu so predlogi *v, z/s, na, za* in *po* najbolj pogosti in zajemajo skupaj 68,86 % vseh predlogov.



Natančnejša analiza je veljala tudi medmetom, ki so v besedilu *Namesto koga roža cveti* razmeroma pogosti. V korpusu zasledimo 42 različnih medmetov, ki so razdeljeni v tri skupine: razpoloženjski, onomatopoetski in velelni medmeti. Skupina razpoloženjskih, ki izražajo človeško občutenje, je z 52,5 % najmočnejša, najmanj zastopani so onomatopoetski medmeti. Za okrepitev sporočilnosti medmetov pisatelj uporablja tri različne možnosti:

1. dodajanje različnega števila samoglasnikov ali soglasnikov: *beee, eee, jaaa, psss* (tip 1 = a + V(C) + ... + nV(C);[5]
2. ponavljanje istega medmeta: *Eh, eh!, Ja, ja!* (tip 2 = a + a + ... + an);
3. nizanje različnih medmetov: *Eh, ja!, Hm, ja-, O, ja!, O, ne!* (tip 3 = a + b + ... + n).

Ojačevanje medmetov se pojavi v besedilu 57-krat, prevladuje tip 2 z 28-kratno ponovitvijo.




Analiza sodobnega slovenskega romana *Namesto koga roža cveti* je pokazala, da nam računalnik na področju jezikovnega raziskovanja na vsak način olajša delo, kljub temu pa človeka, ki odloča o vprašanjih, ki presegajo kvantitativni vidik jezika, ni mogoče in ga tudi v bližji prihodnosti ne bo mogoče zamenjati z računalnikom, ker računalnik še naprej ostaja to kar je: stroj, ki zna računati, ne pa misliti.

## Literatura

Gerhard Neweklowsky, 1986: Zur Paradigmatik in Trubars Katechismus 1550. *Obdobja 6. 16. stoletje v slovenskem jeziku, književnosti in kulturi*. Ur. B. Pogorelec, J. Koruza. Ljubjana: Filozofska fakulteta. 307–317.

[5] Pojasnilo: V = samoglasnik, C = soglasnik; a, b = različni medmeti, n = poljubno število.

Gerhard Neweklowsky, 1988: Zur Häufigkeit morphologischer Kategorien in slowenischen Prosatexten. *Obdobja 8. Sodobni slovenski jezik, književnost in kultura*. Ur. B. Paternu, F. Jakopin, P. Weiss. Ljubjana, Filozofska fakulteta. 337–349.

Daniela Pečnik, 2002: *Versuch einer computerunterstützten quantitativen Textanalyse am Beispiel des slowenischen Romans Namesto koga roža cveti*. Diplomarbeit. Wien.

Johann Pečnik, 1994: *Quantitative Textanalysen – Ausgewählte Erzählungen von Ivo Andrić*. Diplomarbeit. Klagenfurt.

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

*Slovar slovenskega knjižnega jezika*, elektronska izdaja, verzija 1.0, 1993–1997. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, DZS, Amebis d. o. o.

Stane Suhadolnik, Marija Ježič, 1962: Plasti in pogostnost leksike. *Jezik in slovstvo* 8/1–2. 45–49.

Stane Suhadolnik, Marija Ježič, 1962: Plasti in pogostnost leksike. *Jezik in slovstvo* 8/3. 73–78.

*Besedišče slovenskega jezika z oblikoslovnimi podatki. Po gradivu za slovar slovenskega knjižnega jezika. Zbrane besede, ki niso bile sprejete v SSKJ*, 1998. Ur. I. Šircelj Žnidaršič. Ljubljana: ZRC SAZU.