FILOZOFSKI FAKULTET

SVEUČILIŠTA U ZAGREBU

ODSJEK ZA JUŽNOSLAVENSKE JEZIKE I KNJIŽEVNOSTI

KATEDRA ZA MAKEDONSKI JEZIK I KNJIŽEVNOST

Ines Cebović

SASTAVLJANJE MAKEDONSKO-HRVATSKOG KORPUSA

Diplomski rad

Mentor: dr. sc. Borislav Pavlovski

Komentor: dr. sc. Marko Tadić

Zagreb, rujan 2015.

Sadržaj

Sastavljanje makedonsko-hrvatskog paralelnog korpusa

Sažetak

U ovom su radu prikazana područja istraživanja računalne i korpusne lingvistike, pružajući definicije korpusa te opise alata za obilježavanje korpusa. Kao eksperimentalni dio sastavljen je makedonsko-hrvatski usporedni korpus, čiji je tijek sastavljanja opisan u završnom dijelu rada, zajedno s prikazom pretrage korpusa. Sastavljanje ovoga korpusa odabrano je zbog sve veće potrebe za korpusnim istraživanjima, a za makedonski jezik zasad ne postoji nikakav korpus, dok je za hrvatski jezik dostupan Hrvatski nacionalni korpus.

Ključne riječi: usporedni (paralelni) korpus, makedonski jezik, hrvatski jezik, korpus, korpusna lingvistika, Hrvatski nacionalni korpus (HNK)

Summary

In this paper the research areas of computer and corpus linguistics are presented, providing definitions of corpora and descriptions of tools used for corpus annotation. The experimental part consists of the development of a Macedonian-Croatian parallel corpus, which is described in the final part of this paper along with corpus search results. The development of this corpus was chosen because of the growing need for corpus research, and for the time being no corpus exists for the Macedonian language, while the Croatian National Corpus is available for the Croatian language.

1. Uvod

U ovom se radu opisuje sastavljanje makedonsko-hrvatskog usporednog korpusa, kao eksperimentalnog dijela diplomskog rada. Prije no što se opiše sam korpus, dan je prikaz računalne lingvistike, korpusne lingvistike kao njezine poddiscipline, te svih odlika korpusa, kao i postupka označavanja, sa svim alatima za označavanje koji danas postoje, njihovim vrstama te mogućnostima. Takav širi prikaz dan je kako bi se prikazala uloga korpusa u lingvistici i drugim znanostima, te kako bi se pokazala važnost, kao i mogućnosti, sastavljanja novih korpusa, čije značajke ovise o namjeni za koju se sastavljaju.

Potom je opisan postupak sastavljanja ovoga korpusa, čiji je cilj ponuditi istraživačima novo pomagalo pri istraživanjima, kako kontrastivnim, tako i jednojezičnim, jer je ovo jedan od pionirskih radova u korpusnoj lingvistici makedonskog jezika. Koliko nam je poznato, makedonski je zasad uključen samo u višejezični Gralis korpus koji se sastavlja na Sveučilištu u Grazu, a ovdje prezentirani korpus zasad je prvi u kojem je moguće pretraživati izvorne makedonske tekstove, bez obzira što se pri pretraživanju prikazuju i njihovi prijevodi na hrvatski jezik. Za hrvatski jezik već postoji višemilijunski Hrvatski nacionalni korpus sastavljen u Zavodu za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu, koji je potpuno besplatno u cijelosti dostupan na internetu, a studenti Sveučilišta u Zagrebu, kao i svi drugi istraživači, svakodnevno ga koriste prilikom uistinu širokog spektra istraživanja, od morfologije do semantike, preko frekvencije riječi i kolokacija, uključujući i dijakronijska istraživanja HETA potkorpusa. U samom su radu detaljnije navedene sve mogućnosti i korist upotrebe korpusnih istraživanja.

Cijeli tijek sastavljanja makedonsko-hrvatskog usporednog korpusa, odabir tekstova, kao i objašnjenje zašto su uzeti upravo ti tekstovi i upravo ti alati za sastavljanje i označavanje korpusa, opisani su u ovome radu, a na njegovu su kraju dane smjernice za daljnje razvijanje ovoga korpusa, kako u vidu povećavanja samoga makedonsko-hrvatskog korpusa, tako i sastavljanja hrvatsko-makedonskih potkorpusa, kao i popisivanje koraka za preostale razine označavanja, a time i širenje mogućnosti pretraživanja samog korpusa.

2. Računalna lingvistika

Prije definiranja korpusne lingvistike i korpusa potrebno je ponešto reći i o računalnoj lingvistici. Premda se u literaturi iznenađujuće rijetko susreće povezivanje računalne lingvistike s korpusnom, autorica ovih redaka smatra kako je potrebno definirati računalnu lingvistiku da bi se mogao odrediti predmet i doseg istraživanja korpusne lingvistike. Ponegdje se čak može naići na izjednačavanje korpusne s računalnom lingvistikom, no vjerojatno je bolja hijerarhijska podjela, pri čemu je korpusna lingvistika jedna od poddisciplina računalne.

Kako bi se računalna lingvistika mogla definirati, poslužit ćemo se riječima Tadića (1996: 603), koji kaže kako „danas istraživati prirodni jezik bez pomoći računala nije samo mukotrpno i dugotrajno nego, uslijed ljudske nemogućnosti da se u obradi zamašne jezične građe održe kriteriji i(li) koncentracija, često i paraznanstveno." Tako se razvitkom računala razvila i računalna lingvistika – grana lingvistike koja pokušava upotrebom računala olakšati i poboljšati lingvistički opis prirodnih jezika i u konačnici omogućiti izradbu sustava za strojno razumijevanje i/li generiranje ovjerenih postava prirodnoga jezika.

Računalna se lingvistika ponajprije se bavi strojnom obradom prirodnog jezika (*natural language processing*, NLP) odnosno izradom računalnih programa za obradu teksta na nekom prirodnom jeziku.

Postoje nelingvističke discipline s kojima računalna lingvistika graniči, kao što su računarstvo (što je razumljivo s obzirom da se obje discipline bave razvijanjem računalnih programa za određeno područje interesa), umjetna inteligencija, a isto tako kibernetika, kao i matematička lingvistika te djelomično i primijenjena lingvistika.

Nakon što je prikazano okruženje računalne lingvistike, može se obujmiti njezino područje istraživanja i interesa. To su ponajprije istraživanje jezika i govora, razvijanje programa za obradu istih, razvijanje programa za strojno prevođenje, razvijanje programa za računalnu upotrebu jezika i govora, bilo u vidu računalne sinteze teksta i govora, bilo kao razumijevanje teksta i govora pomoću računala. Dakako, kad je riječ o obradi govora (*speech processing*) onda je osim lingvistike u ovu aktivnost uključena i fonetika.

Budući da ne postoji literatura na ovu temu koja iscrpno obuhvaća sve domene računalne lingvistike, ovdje će se napraviti opći presjek svih područja kojima se računalna lingvistika danas bavi. Uz gore navedene to su automatsko prepoznavanje govora, komunikacija računala i čovjeka u vidu postavljanja pitanja i odgovaranja, statistička lingvistika (odnosno statistička obrada bilo kakvih lingvističkih podataka), provjera gramatike i stila, dohvaćanje informacija, modeliranje i simuliranje, uređivanje teksta, automatsko rastavljanje riječi u tekstu, provjera

pravopisa, optičko prepoznavanje znakova, generiranje teksta iz slika, ispravljanje velikih i malih slova i slično.

Za više detalja o sustavima primijene lingvistike upućujemo na *Notes on computational linguistics* (Stabler 2003).

Svi dosad navedeni programi i sustavi računalne lingvistike tiču se obrade forme jezika, no obrada značenja još je relativno slabo razvijena, iako se mnogi stručnjaci bave razvijanjem sustava za obradu značenja jezika. To je područje interesa, uz računalnu lingvistiku, umjetne inteligencije (*artificial intelligence,* AI), a kako je općepoznato da znanost još nije ušla u tajne ljudskog poimanja svijeta i strukturiranja znanja i značenja, u ovome radu neće se dalje ulaziti u pitanja zašto razvitak sustava za obradu značenja još nije moguć. Detaljnije to opisuju Gazdar i Mellish (1989: 8-9).

Nakon navođenja niza programa i sustava računalne lingvistike, može se činiti kako je ova lingvistička disciplina dugovječna, pri čemu se zaboravlja činjenicu da pojava računala datira od 2. svjetskoga rata, tako da ni ova disciplina nije starija od sedamdesetak godina. Sve većim napretkom računalnih tehnologija, i potrebe i očekivanja krajnjih korisnika sve su veća, pa se tako danas očekuje da stroj „pročita neobrađeni tekst, provjeri ispravnost, izvrši naredbe sadržane u tekstu ili da ga čak razumije dovoljno dobro da može dati razuman odgovor baziran na njegovom značenju. Ljudi žele za sebe zadržati samo konačnu odluku“ (Bolshakov i Gelbukh 2004: 16).

Pošto su prikazane domene istraživanja i interesa računalne lingvistike, te se pokušalo razgraničiti ovu disciplinu od srodnih disciplina, kao i navesti mnoge (nećemo reći sve jer se na razvitku novih radi i danas pa se lako može dogoditi da nam je u ovom radu nešto promaklo) sustave i programe za obradu jezika, potrebno je dovesti u vezu područje interesa ovoga rada s disciplinom koju smo upravo opisali.

Kako je navedeno na početku, autoričina je pretpostavka da se korpusnu lingvistiku, kao disciplinu koja se danas bavi računalnom obradom teksta, o čemu će više riječi biti u sljedećem poglavlju, može hijerarhijski podrediti računalnoj lingvistici, kao krovnoj disciplini za različite pristupe računalnoj obradi jezika, a time i govora, teksta i drugih jezičnih modaliteta. No, rijetko se dovode u vezu računalna i korpusna lingvistika, a neki, s druge strane, ove dvije discipline izjednačavaju, odnosno jedni sustave za računalnu obradu jezika (pritom se uvijek misli na računalnu obradu prirodnog jezika, NLP) ubrajaju u korpusnu lingvistiku, dok drugi korpuse i njihove alate ubrajaju u računalnu lingvistiku. U sljedećem će se poglavlju pokušati definirati domena korpusne lingvistike, zatim i korpusi i njihovi alati, kako bi se mogle razdijeliti računalna i korpusna lingvistika, kao i njihove domene, u odnosu jedna na drugu.

3. Korpusna lingvistika

Nakon definiranja računalne lingvistike, potrebno je definirati domenu korpusne lingvistike. Korpusna se lingvistika može definirati dvjema definicijama, kao proučavanje jezika na temelju podataka iz korpusa te kao razvijanje i primjena tehnika za sastavljanje digitalnih korpusa, njihovo označavanje, izvlačenje leksičkih i gramatičkih uzoraka i njihovo tumačenje ponajprije na temelju dobivenih statističkih podataka.

Korpusna lingvistika stavlja naglasak na proučavanje jezika, pri čemu se jezični podatci iz korpusa mogu dohvatiti za bilo koju jezičnu razinu: fonemsku/grafemsku, razinu riječi tj. leksičku razinu, razinu kolokacija/fraza/idioma/sintagmi, sintaktičku razinu, semantičku razinu (leksičku i rečeničnu) kao i razinu pragmatike. Druga važna dimenzija jezika kojom se korpusna lingvistika bavi jest jezična upotreba jer korpusna lingvistika proučava jezik na temelju primjera stvarne upotrebe jezika, čime se odvaja od tradicionalnih istraživanja strukture koji su do zaključaka dolazili uglavnom putem introspekcije izvornog govornika, za razliku od istraživanja upotrebe, na već postojećim rečenicama i diskursu. Sinclair (1991: 5-6) daje kritiku tradicionalnih introspektivnih gramatičara jer oni na kraju izmišljaju primjere umjesto da traže već postojeće te su njihovi primjeri bez konteksta i upitno je koliko zaista opisuju prirodni jezik.

McEnery i Wilson (2001: 25) dobro opisuju značajnost empirijskoga odnosno korpusnoga pristupa citirajući Fillmorea: „Fillmore čini se jako dobro sažima raspravu o korpusnim i ne-korpusnim lingvistima (...): 'Mislim da ne mogu postojati korpusi, koliko god opsežni, koji bi sadržavali informacije o svim područjima engleskog leksika i gramatike koje želim proučiti... ali svaki korpus koji sam imao priliku ispitati, naučio me činjenicama za koje ne mogu zamisliti da bih naučio na bilo koji drugi način. Moj zaključak je da dvije vrste lingvista trebaju jedni druge'.“

Ovdje dolazimo do prve kritike korpusa i korpusne lingvistike, ponajprije od strane Chomskoga i drugih istraživača generativne gramatike, koji tvrde da su korpusi nedovršeni, jer je jezik neprebrojiv pa nijedan konačni korpus ne može adekvatno predstavljati jezik, te bi svaki opis na temelju korpusa bio iskrivljen odnosno ništa više od pukog popisa. Druga je kritika Chomskoga da su rečenice u korpusu manjkave jer se može raditi o negramatičnim, nepristojnim, nedovršenim rečenicama i/li konstrukcijama koje u jeziku nisu ili su slabo zastupljene. Njegove su kritike zanemarene od strane korpusnih lingvista, koji su, svjesni nedostataka korpusa, nastavili razvijati programe za sastavljanje i obradu korpusa, imajući na umu da je korpus prikaz upotrebe, a ne strukture jezika, te da korpusi nipošto ne pokušavaju reprezentirati jezik u cjelini, što se vidi na temelju uzorkovanja korpusa, čime se jezik može promatrati, no ne i posve obujmiti.

Kao što je već rečeno, korpusna se lingvistika bavi istraživanjem upotrebe jezika na temelju postojećih primjera iz tekstova, a ti se tekstovi prikupljaju ovisno o potrebama istraživača. Poslužit ćemo se riječima Tadića (1996: 604): „Temelj za svako istraživanje teksta jest korpus bez obzira na to promatra li se kao jezična građa ili kao nešto drugo što se putem teksta/jezika tek ostvaruje. Za razliku od ostalih jezikoslovnih disciplina, korpusna lingvistika određena je ne toliko područjem istraživanja koliko metodološkom osnovicom na kojoj se temelji istraživanje. Stoga se korpusni pristup (ili korpusna metodologija) lako može primijeniti u različitim lingvističkim disciplinama: fonologiji, morfologiji, sintaksi, sociolingvistici, kognitivnoj lingvistici itd., i to najčešće u kombinaciji s drugim, tim disciplinama inherentnim, metodološkim postupcima. Današnji uvid u korpus ne može se ni zamisliti bez pomoći računala i svih mogućnosti koje ona pružaju pri pregledu i uređivanju građe", čime dolazimo do područja istraživanja korpusne lingvistike i pitanja: je li korpusna lingvistika grana lingvistike? Odgovor na ovo pitanje je i da i ne. Korpusna lingvistika nije grana lingvistike kao fonologija, morfologija, sintaksa ili semantika. Sve se ove discipline koncentriraju na opisivanje i objašnjavanje neke jezične razine i mogućih kombinacija jezičnih jedinica karakterističnih za tu jezičnu razinu. Korpusna je lingvistika, nasuprot tome, metodologija a ne pristup jeziku koji zahtijeva objašnjavanje ili opisivanje. Pristup na temelju korpusa može se primijeniti u mnogim aspektima lingvističkog istraživanja. Korpusna lingvistika je metodologija koja se može koristiti u gotovo svakoj lingvističkoj grani, ali ne omeđuje područje lingvistike samo po sebi.

Kao metodologija za istraživanje jezika a ne istraživanje vezano uz pojedinu jezičnu razinu, korpusna se lingvistika primjenjuje u mnogim lingvističkim pristupima, kao što su kontrastivna lingvistika, analiza diskursa, učenje jezika, semantika, sociolingvistika, teorijska lingvistika, prevođenje, stilistika, forenzička lingvistika. Kao izvor podataka za opis jezika korpusi su od velike pomoći leksikografima i gramatičarima. Danas je zapravo teško pronaći područje lingvistike u kojem se korpusno utemeljen pristup ne primjenjuje. Osim opisivanja jezičnih pojava na temelju primjera danih u tekstu, može se istraživati i same te tekstove, u smislu uspoređivanja različitih žanrova i slično.

Pošto je opisana raznovrsnost primjene korpusa u lingvistici, treba se vratiti na drugu Tadićevu (1996) tvrdnju o korpusnoj lingvistici – onu o upotrebi računala. No da bi se došlo do toga, treba se osvrnuti na povijesni razvitak korpusne lingvistike, koja je postojala i puno prije pojave računala. U svom modernom, digitalnom obliku, korpus je postojao tek od polovice 1950-ih. Osnovna ideja korištenja stvarne upotrebe računala u istraživanju jezika datira od vremena prije toga, ali problem je bio da je prikupljanje i korištenje velikih količina lingvističkih podataka u predračunalno i rano računalno doba bilo teško, gotovo nemoguće.

Zamjetni primjeri postizanja toga ostvarivali su se raspoređivanjem ogromne količine posla na veliku količinu radne snage – Kädingov Čestotni rječnik njemačkoga jezika (1897) dobar je primjer toga, pa se na tome primjeru može uočiti kako neki elementi načela korpusne lingvistike postoje već više od jednog stoljeća. Prvi jednomilijunski računalni korpus bio je Brown korpus, izrađen 1967.

Interes za računalom kod korpusnih lingvista dolazi od mogućnosti računala da pretraži, pronađe, sortira i obradi jezične podatke, bilo tekst (najčešće) ili digitalizirani govor (sve češće). Nakon višedesetljetnog sastavljanja općejezičnih korpusa za pojedine jezike, posljednja se dva desetljeća primjećuje trend sve češćeg sastavljanja višejezičnih usporednih korpusa, čime se domena korpusne lingvistike samo dodatno širi, a time se širi upotrebljivost i upotreba korpusa u drugim lingvističkim disciplinama.

Za kraj ovog poglavlja navodimo jedan citat kojim zaokružujemo definiranje područja djelovanja korpusne lingvistike, njezin povijesni razvitak te trenutno stanje, a u sljedećem poglavlju definirat ćemo korpuse, navesti vrste korpusa, a potom i alate za sastavljanje i analizu korpusa: „Korpusni pristup čine četiri glavne karakteristike: empiričan je, analizira stvarne uzorke upotrebe jezika u prirodnim tekstovima; koristi veliku i načelnu zbirku prirodnih tekstova kao svoj temelj za analizu; opsežno koristi računala u analizi; ovisi i o kvantitativnim i o kvalitativnim analitičkim postupcima“ (Bennett 2010: 7).

4. Korpus

4.1. Definicija korpusa

Nakon prikaza područja istraživanja korpusne lingvistike, vrijeme je za definiranje korpusa. Kao ni kod definicija u prethodnim poglavljima, ni oko definicije korpusa lingvisti se međusobno ne mogu u potpunosti složiti. Ovdje ćemo pokušati napraviti presjek nekih definicija kako bi se dobila gruba predodžba što je korpus u okvirima računalne i korpusne lingvistike, a zatim će biti navedene prednosti i nedostatci korpusa, kao i njihova važnost u lingvističkim istraživanjima.

Pojam korpus danas je gotovo sinonim za pojam strojno čitljiv korpus jer računalo omogućuje pretraživanje, pronalaženje, sortiranje i obradu podataka. Svaka bi se zbirka od više od jednog teksta mogla zvati korpus jer je to latinska riječ za tijelo (*corpus*), no korpus nije nasumičan skup tekstova već zbirka tekstova ili tekstovnih odsječaka ostvarenih u jednom ili više jezičnih modaliteta i/li od istih ili različitih žanrova i to u uravnoteženoj količini, u svrhu reprezentiranja jezika ili dijela jezika. Korpus je velika i strukturirana jezična baza podataka koja može sadržavati tekstove ili samo na jednome jeziku ili više njih. Glavne su odlike brižljivo sastavljenoga korpusa uzorkovanje i reprezentativnost, konačna veličina, strojno-čitljiv oblik, standardna referencija. „Kad je riječ o korpusima valja prema EAGLES (1996) jasno metodološki razlikovati: zbirku tekstova: svaki skup tekstova skupljen prema nekim kriterijima; korpus: skup jezičnih odsječaka koji su odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak; računalni korpus: korpus koji je kodiran na dosljedan i standardan način s ciljem da bude računalno pretraživ" (Tadić 2003: 28).

Već se iz ovih definicija mogu nazrijeti različite vrste korpusa, koji se mogu podijeliti na papirnate i računalne, isto tako i prema modalitetu koji koriste – pisane, govorne, video korpuse i slično, kao i prema broju jezika koje sadrže – jednojezične, dvojezične i višejezične, a sve će te vrste korpusa biti detaljnije objašnjene dalje u tekstu.

Tognini-Bonelli (2001: 2) daje još jednu nijansu u definiranju korpusa razgraničavajući tekst od korpusa, koji je zapravo skup tekstova, odabranih s ciljem da reprezentativno prikazuju dani jezik za određenu lingvističku analizu, te popisuje razlike između teksta i korpusa, gdje korpus više nije taj koji reprezentira upotrebu, *parole*, danog jezika, nego je to tekst u cjelini, a korpus kao skup više tekstova služi za reprezentaciju *langue*, jezika kao sustava, a ne upotrebe tog jezika, čime se vidi širok spektar definicije korpusa.

Sigurno bi se mogla naći još koja nijansa definicije korpusa, no smatramo da je dana gruba predodžba o tome što je korpus: korpus je, dakle, skup tekstova i/li njihovih odsječaka, govornih odsječaka ili video snimaka, u današnje vrijeme računalno pohranjenih, odabranih prema

određenim kriterijima u svrhu reprezentiranja jednog ili više jezika ili njihovih varijanata za lingvističko istraživanje strukture jezika ili jezične upotrebe u danom modalitetu i/li jezičnome varijetetu. Ovakva definicija je svakako preopširna, a takva je zbog različitih svrha za koje su pojedini korpusi sastavljeni. Donedavno je praktički svaki istraživač sastavljao vlastiti korpus kako bi mu poslužio za određeno istraživanje. Danas se, međutim, s pojavom velikih reprezentativnih općejezičnih korpusa (tzv. nacionalnih korpusa) istraživači više ne moraju baviti sastavljanjem svojih korpusa već se mogu izravno baciti na proučavanje onoga što ih zanima na temelju tako dostupne reprezentativne jezične građe. Ovime dolazimo do već spomenutih vrsta korpusa, a mogu se podijeliti s obzirom na namjenu, broj jezika, veličinu tekstova i samih korpusa te jezičnih modaliteta koji su u njima pohranjeni. Sigurno bi se mogla napraviti još koja podjela, no ovdje ćemo navesti samo nekoliko, kao i vrste korpusa s obzirom na te podjele.

4.2. Vrste korpusa

Ni kod podjele korpusa nema potpune suglasnosti među lingvistima, tako da se neke vrste izjednačavaju, ponegdje se daje podjela na više vrsta nego što bi trebalo biti i slično. Prva je nedoumica oko definiranja usporednih naspram usporedivih i prijevodnih korpusa, no većina se ipak slaže da su usporedni isto što i prijevodni jer sadrže zbirku originalnih tekstova na jednom jeziku te njihove prijevode na jedan ili više jezika, ukoliko su jednosmjerni. Ukoliko su dvosmjerni, onda postoje originali i prijevodi za svaki od jezika u tom korpusu. S druge strane, usporedivi korpusi sadrže tekstove na dva ili više jezika, no koji nisu direktni prijevodi jedni drugih nego se radi o tekstovima istog žanra ili iste tematike, a korpusi su skupljeni na temelju istih načela uzorkovanja i reprezentativnosti.

U širem smislu višejezični korpusi uključuju tekstove na dva ili više jezika, no u užem smislu moraju uključivati najmanje tri jezika jer se oni koji uključuju samo dva jezika nazivaju dvojezični korpusi. S druge strane, postoje jednojezični korpusi koji teže biti općejezični za određeni jezik, a mogu biti sinkronijski, prikazivati jezik u određenom kraćem vremenskom rasponu, ili dijakronijski pa prikazuju promjene u jeziku kroz dulji niz godina. Takvi su korpusi uglavnom vrlo veliki, od više desetaka ili stotina milijuna riječi, kako bi prikazali što je moguće cjelovitiju sliku jezika.

Bennett (2010: 13-4) daje podjelu vrsta korpusa s obzirom na njihovu namjenu – općejezični, specijalizirani, učenički, pedagoški, povijesni, usporedni, usporedivi i monitor korpusi. Specijalizirani korpusi sadrže tekstove iz neke uske domene ljudske djelatnosti, pa su iznimno korisni za npr. proučavanje specijalizirane terminologije tih područja. Učenički i pedagoški

korpusi sastavljaju se kao pomagala za učenje stranog jezika, da se utvrde greške koje učenici rade, da se utvrdi reprezentativan leksik za određeni stupanj znanja stranog jezika i slično. Povijesni su zapravo dijakronijski ili mogu biti sinkronijski, ali reprezentirati određeno povijesno razdoblje u jeziku, dok su monitor korpusi također dijakronijski jer se u njih stalno dodaju novi tekstovi, dok se oni stari mogu ili izbaciti ili zadržati, čime se veličina takvih korpusa ili stalno drži jednakom ili raste. Sve druge prethodno navedene vrste korpusa podrazumijevaju konačnost – kad se dođe do određene veličine ili kad se prikupe svi tekstovi prema određenim kriterijima, sastavljanje korpusa prestaje.

Li i dr. (2011: 9) donose malo drukčiju podjelu korpusa, s obzirom na vrstu podataka koje sadrže: „1. Heterogeni: Jezični se podatci prikupljaju neovisno o vrsti; raznovrsni jezični podatci prikupljaju se i spremaju u originalnom obliku. 2. Homogeni: Samo se podatci iste vrste prikupljaju. 3. Sustavni: Jezični se podatci prikupljaju na temelju unaprijed definiranih načela i omjera, stvarajući uravnotežen i sustavan korpus koji predstavlja jezične činjenice u određenom rasponu. (...) 4. Specijalizirani: Prikupljaju se samo podatci za određenu namjenu.“

Smatramo da smo ovom podjelom obuhvatili najčešće vrste korpusa s obzirom na njihovu namjenu i broj jezika, te možemo prijeći na druge odlike korpusa, kao što su veličina i vrsta tekstova, odnosno modalitet u kojem su primjeri za korpus prikupljeni.

4.3. Veličina korpusa

Kako se dosad moglo vidjeti, korpusi mogu biti ili konačni, s određenim brojem tekstova i bez dodavanja novih, ili nekonačni tj. monitor korpusi. Ovdje ćemo se više osvrnuti na konačne korpuse i njihovu veličinu, kao i na veličinu tekstova i/li njihovih odsječaka koji su u njima pohranjeni, a s tim u vezi i na vrstu tekstova odnosno žanrove tekstova od kojih korpusi mogu biti sastavljeni.

Uzorkovanje je prvi korak pri sastavljanju korpusa. To podrazumijeva prikupljanje uzoraka tekstova, bilo određenog žanra, određenih autora, modaliteta, jezičnoga varijeteta ili nečeg drugog. Uzorci moraju biti uravnoteženi, kako bi pružili pravu sliku jezične populacije koja se istražuje. Uzorkovanje se odvija kako bi se napravio reprezentativan prikaz željene jezične populacije jer se često radi o većoj količini tekstova koje je nemoguće prikupiti u cijelosti, bilo zbog nedostupnosti svih tekstova, manjka vremena ili zaštićenih autorskih prava jer oni koji raspolažu pravima za neki tekst ne dopuštaju korištenje teksta u cjelini. Uzorkovanje se ne provodi samo pri sastavljanju zatvorenih korpusa, kao što su korpusi djela određenog autora. No ni ovdje se svi sastavljači korpusa ne slažu pa postoji pitanje veličine uzorka, koliki uzorak teksta mora biti kako bi reprezentirao taj tekst u cjelini, kao i koji dio teksta uzeti za uzorak, s

obzirom da nisu sve značajke ravnomjerno raspršene u tekstu. Opća je praksa pokazala kako je dovoljno oko 1 000 riječi po uzorku, no neki istraživači ipak odlučuju koristiti tekstove u cjelini, ukoliko je to moguće, odnosno ukoliko su tekstovi dostupni, ili kad im je duljina manja od ciljanih 1 000 pojavnica. Kad je riječ o korpusima konačne veličine, i ovdje dolazi do problema jer je upitno koji su tekstovi reprezentativni za određeni žanr, vremenski period, jezični varijetet i slično. Zato sastavljači korpusa moraju eksplicitno navesti kriterije sastavljanja korpusa odnosno odabira reprezentativnih tekstova, broja tekstova, veličine uzoraka i svih drugih parametara koji određuju neki korpus.

Dok se nekad sastavljalo korpuse od 1 000 000 riječi, danas su računala, a time i korpusna lingvistika, toliko uznapredovala da se općejezični korpusi sastavljaju od najmanje 100 000 000 riječi (dalje u radu zadržat ćemo se na višeznačnosti *riječi* te uvesti određene varijacije s obzirom na značenje o kojem u tom trenutku govorimo), a nerijetko i od više stotina milijuna. Preporuka je da je korpus što je veći moguć, s mogućnošću rasta, a preporuka se temelji na uzorku pojavljivanja riječi u tekstu, kako je Zipf (1935) prvi istaknuo, jer se oko pola vokabulara teksta sastoji od riječi koje se javljaju samo jednom u tekstu.

Osim uzorkovanja, postoji pitanje ravnoteže, u smislu važnosti različitih dijelova u općejezičnim korpusima s obzirom na jezični modalitet. Donedavno su korpusi uvelike bili sastavljani većinski ili isključivo od pisanih tekstova, no danas se pojavljuju i tekstovi govornog zapisa, bilo u vidu transkripcije, bilo kao audio odsječci, a odnedavno se uvode i video zapisi. S druge strane, ravnoteža se tiče i ravnomjerne raspodjele žanrova, osim ako se radi o specijaliziranim korpusima poput korpusa određenog žanra ili svih djela određenog autora. Ravnoteža u govornim zapisima može se ticati i dobi, spola, podrijetla autora, što se može kontrolirati i kod pisanih tekstova. Još jedna dimenzija ravnoteže tiče se samog odabira tekstova, odnosno hoće li se odabrati značajan tekst ili autor, koji je utjecajan ili poznat, ili će se napraviti nasumičan odabir, ili će se tekstovi prilagoditi kako bi ispunili lingvističke kriterije. Najbolji je kombinirani pristup gdje se odabire iz šireg raspona vrsta tekstova.

Kad je riječ o načinu prikupljanja tekstova, video se odsječci uzimaju u originalnom obliku, a obično su popraćeni transkripcijom teksta i dodatnim formaliziranim zapisom raznih prijezičnih znakovnih sustava (geste, mimika, itd.), govorni se zapisi uz originalni zvučni zapis redovito transkribiraju, a pisani se tekstovi mogu unijeti skeniranjem, utipkavanjem, preuzimanjem s interneta ili korištenjem datoteke koja već postoji u elektronskom obliku.

Iako ne postoji univerzalna preporuka oko veličine korpusa, tekstova u njemu niti žanrova tekstova i jezičnog modaliteta u kojem su primjeri za korpus dani, u jednome se svi lingvisti

slažu – ne postoji jedinstven korpus koji bi služio svim namjenama i svaki je korpus samo približan uzorak jezičnoga varijeteta koji želimo istraživati.

Prije no što krenemo na alate za označavanje i pretragu korpusa, potrebno je navesti ulogu i važnost korpusa kako bi se moglo vidjeti zbog čega se toliki lingvisti bave korpusnom lingvistikom i sastavljaju nove i nove korpuse, iako i oni sami govore o nedostatcima korpusa, kao što su već spomenuta veličina i problemi pri uzorkovanju, nedostupnost tekstova iz određenih žanrova ili modaliteta, a kako će kasnije biti prikazano, i mnoge greške i nerazrješivi problemi pri označavanju i upotrebi korpusa.

4.4. Važnost i uloga korpusa

Kako je već navedeno, postoje različite vrste korpusa, ovisno o svrsi kojoj su namijenjeni. Postoje mnogi korpusi koji teže biti sveobuhvatni, opći, za određeni jezik, a danas postoji i tendencija napraviti takav korpus za sve veće jezike. Ovdje se ponajprije polazi od pisanih tekstova, no sve je češće prikupljanje govornih zapisa kako bi se i razgovorni stil, kao i sam govorni jezik, mogli pohraniti u korpus i kasnije biti dostupni za proučavanje. Također, razvijaju se i višejezični korpusi kao pomoć pri prevođenju, kao pomoć za razvijanje novih alata za strojno prevođenje, no i kao pomoć studentima pri učenju stranog jezika. Sve se češće sastavljaju i specijalizirani korpusi, koji mogu biti sastavljeni od tekstova određenog žanra, određenog autora ili određenog vremenskog perioda, a koji imaju točno određenu svrhu za neko istraživanje. Sve ćemo to ovdje prikazati, kako bi bilo jasno zašto lingvisti sastavljaju sve više korpusa u određene svrhe, iako se svi stalno navraćaju na već dobro poznate manjkavosti i nedostatke korpusa.

Korpusi su važni zbog pružanja empirijskih podataka, koji lingvistima omogućavaju postavljanje objektivnih tvrdnji utemeljenih na jeziku kakav je, odmičući se od tradicionalnih subjektivnih tvrdnji temeljenih na introspekciji pojedinca. Korpusi se koriste ponajprije za morfološka i sintaktička istraživanja, poput istraživanja svih oblika određene riječi, sintaktičkih uzoraka ili diskursnih struktura, kao i za istraživanje distribucije fonema, slova, interpunkcije, flektivnih i derivacijskih morfema, riječi i slično. Druga je važna upotreba korpusa u leksikografiji jer svi sadrže bogatu količinu podataka, od vrsta riječi i drugih lingvističkih oznaka do podataka o autoru, žanru, regionalnoj varijaciji i vremenu nastanka teksta.

„Korpusi su svojim trima vrstama podataka: 1. *evidencijom*: pronalaženjem ima li neke jezične jedinice u korpusu ili nema; 2. *frekvencijom*: ako je ima, brojanjem koliko se puta pojavila u korpusu; 3. *relacijom*: pronalaženjem u kakvu odnosu stoji prema drugim jezičnim jedinicama, kao i snažnim alatima za njihovo pretraživanje (djelomičnim i potpunim

konkordancijama, upitima prema vrstama riječi ili morfosintaktičkim opisima, kolokacijskim upitima itd.) u leksikografiji donijeli toliki pomak da danas ne postoji ozbiljniji leksikografski nakladnik koji se ne služi korpusima kao osnovnim sredstvom za razvitak svojih leksikografskih proizvoda“ (Tadić 2003: 31).

Podatci o autoru pisanog teksta ili govornog zapisa važan su izvor za sociolingvistiku. Specijalizirani korpusi sastavljeni od određenog žanra ili od djela određenog autora značajan su izvor podataka u stilističkim istraživanjima. I geografska komponenta može imati značaj u istraživanju dijalektologije. Također, frazeologija u korpusima može proučavati kolokacije i druge pojave jezika u željenim sekvencama.

Korpusi su uglavnom slabo zastupljeni u analizi diskursa, no mogu poslužiti kao kontrolni podatci pri proučavanju značajki određenog žanra i slično. I u psiholingvistici korpusni podatci mogu poslužiti kao kontrolni, osobito pri prepoznavanju riječi gdje daju objektivni prikaz frekvencije riječi. Isto je i s patologijom jezika jer pružaju uvid u normalnu jezičnu produkciju pa mogu služiti za usporedbu s produkcijom i procesiranjem kod patologija, kao i pri analizi razvitka jezika kod djece.

Forenzička lingvistika odnedavno koristi korpuse pri analizi vjerodostojnosti dokumenata od priznanja do pisama samoubojica, identifikaciji autorstva u akademskim okruženjima (pitanje plagijata), pismima ucjene, pismima prijetnje, čitljivosti/razumljivosti pravnog jezika, forenzičkoj fonetici (odnosno identifikaciji govornika), policijskim podatcima intervjua i ispitivanjima, jezičnim pravima etničkih manjina i diskursu sudničkog okruženja i ostalom.

Sve veća zastupljenost višejezičnih usporednih korpusa značajna je pri učenju stranog jezika jer takvi korpusi mogu poslužiti kao svojevrsni rječnici, kao i kontrolni podatci pri vježbama prevođenja za usporedbu vlastitog prijevoda s već postojećim prijevodom ili originalom.

Nakon prikaza široke lepeze lingvističkih i ne samo lingvističkih disciplina u kojima korpusi mogu poslužiti kao izvor podataka, dajemo prikaz Tognini-Bonelli (2001: 65-100) koja odlazi korak dalje pa čak lingvistička istraživanja korpusa dijeli na ona koja su temeljena na korpusima (*corpus-based approach*) u smislu da istraživač postavi hipotezu pa njezinu točnost provjerava na primjerima iz korpusa, i na ona koja su potaknuta korpusima (*corpus-driven approach*) u smislu da se korpus uzima kao prikaz jezične upotrebe pa sva istraživanja na korpusima zapravo dolaze do određenih zaključaka o jezičnoj upotrebi, bez prethodno postavljene hipoteze, odnosno u takvom se istraživanju korpus ne uzima kao dokaz ili osporavanje neke hipoteze već se zaključci donose na temelju onoga što se u korpusu nalazi, bez obzira na kojoj se jezičnoj razini istraživanje korpusa vrši. Pri tome autor korpusa mora eksplicitno navesti parametre po kojima je sastavio korpus i navesti sve detalje o tekstovima koje je u korpus unio, kako bi se

moglo zaključiti o kakvoj se jezičnoj pojavi, do koje se došlo analizom korpusa, radi. Time Tognini-Bonelli daje još veću važnost korpusnoj lingvistici i upotrebi korpusa u lingvističkim istraživanjima.

4.5. Podatci u korpusu

Pošto je definirano što je korpus i kakvim sve istraživanjima korpusi mogu poslužiti, potrebno je navesti informacije koje se mogu dobiti iz označenog korpusa, prije nego što se prijeđe na specifične alate i programe kojima se korpusi označuju i što samo označavanje znači. Korpusi ponajprije sadrže tri tipa informacija – metapodatke (*metadata*), tekstno označavanje (*textual markup*) i lingvističko obilježavanje (*linguistic annotation*). Metapodatci su podatci o samom tekstu, za pisane materijale to su autor, godina izdavanja i jezik na kojem je tekst napisan. Takvi podatci mogu biti kodirani u tekstu korpusa ili se mogu nalaziti u odvojenom dokumentu. Tekstno označavanje kodira informacije unutar teksta, a radi se o dijelovima koji nisu riječi same, poput označavanja gdje počinje i završava kurziv u pisanom tekstu ili kada jedan govornik počinje i završava svoj iskaz u govorenom materijalu korpusa (McEnery i Hardie 2012: 29). Za takvo su označavanje najprije razvijene univerzalne oznake poznate kao *COCOA references*, koje su mogle kodirati specifični tip tekstne informacije, poput autora, datuma i naslova. Danas se teži formaliziranijim međunarodnim standardima koji omogućuju kodiranje bilo koje vrste informacija potrebnih u strojno čitljivim tekstovima. Najšire prihvaćen je sustav *Text Encoding initiative* (TEI), no upotrebljavaju se još i *Translation Memory Exchange* (TMX) i podskup TEI-a XML *Corpus Encoding Standard* (XCES), koji će biti detaljnije objašnjeni dalje u radu.

TEI je najveći međunarodni projekt u području definiranja standarda za pripremu i razmjenu elektroničkih tekstova kako za znanstvena istraživanja, tako i za širok raspon upotreba za potrebe istraživanja s područja digitalnih humanističkih znanosti (*digital humanities*), kao i drugih oblika informacija poput slike ili zvuka. U lingvističkom smislu TEI ponajprije postavlja standarde za obilježavanje svih vrsta tekstova. TEI je nastojao definirati popis od preko 400 osobina tekstova (predstavljenih elementima) koje bi jezikoslovac ili korisnik s područja humanističkih znanosti mogao trebati. Samo je manji dio oznaka obvezatan. Proces kodiranja zamišljen je s otvorenom mogućnošću za dodavanje novih oznaka već obilježenom tekstu prema potrebama (Bekavac 2001: 54-7). TEI je krenuo od postojećega formalnoga jezika za obilježavanje struktura podataka tj. za obilježavanje dokumenta poznat kao SGML (*Standard Generalised Markup Language*), što je postao metajezik za pohranjivanje tekstova u digitalnom

obliku, a od TEI P5 inačice, tu je ulogu preuzeo XML (*Extensible Markup Language*) koji se danas sve više upotrebljava.

SGML specificira metodu za predstavljanje tekstnih podataka u ASCII dokumentima tako da se podatci mogu razmjenjivati među programima i među korisnicima bez gubitka bilo kakve informacije. Informacija u središtu nije samo prikaz znakova nego detaljna informacija o strukturi teksta. Osnovni model SGML-a je hijerarhijski. Tekstne podatke vidi kao sastavljene od elemenata različitih tipova ugrađenih jedne unutar drugih. Dan je primjer rječničkog unosa za riječ *abacus* u SGML formatu:

```
<entry>
        <headword>abacus</headword>
        <etymology>L. abacus, from Gr. abax</etymology>
        <paradigm>pl. –cuses, or –ci</paradigm>
        ...
</entry>
```

(Lawler i Dry 1998: 17)

Prednost označavanja SGML-om ili XML-om jest mogućnost njihovog jednostavnog uklanjanja ukoliko se želi dobiti izvorni tekst, a isto se tako jezičnim jedinicama označavanjem ovim jezicima za obilježavanje mogu pridodati oznake koje sadrže informaciju o njihovim gramatičkim kategorijama. Između početne i završne oznake može se staviti više riječi ili drugih znakova, što olakšava označavanje korpusa gdje se sintagma može označiti kao jedna jezična jedinica. Većina modernih softvera za konkordiranje omogućuje skrivanje oznaka prilikom gledanja konkordancije, osobito u XML-u.

Li i dr. (2011: 12-3) daju detaljniji popis metapodataka koji mogu biti pohranjeni u korpusu: Metapodatci su strukturne i standardizirane pozadinske informacije, a dijele se na opisne, strukturne i administrativne opisujući sadržaj i karakteristike svake jedinice. Opisni metapodatci opisuju sadržaj i vezu dokumenta ili izvora, kao što su bibliografski podatci. Strukturni metapodatci pružaju stvarne rezultate digitalnih arhiva za pregledavanje, pretragu i reprezentaciju, poput pregleda poglavlja knjige, ili poveznice između teksta i slika. Administrativni metapodatci pohranjuju informaciju za dugoročno upravljanje, upotrebu i pregled, poput formata dokumenta, rezolucije i prava o intelektualnom vlasništvu.

Sve dosad navedeno odnosilo se na računalno pretražive korpuse, kakvi danas isključivo i postoje. No nisu svi tekstovi *a priori* u elektronskom obliku, ponekad se za računalni korpus uzimaju tekstovi u tiskanom obliku. Danas postoje tri najčešće metode preuzimanja tekstova za korpus: prilagođavanje materijala koji su već u digitalnome obliku, preoblikovanje tekstova

optičkim prepoznavanjem pismena (*optical character recognition* tj. strojno čitanje), unos tekstova pretipkavanjem.

Kad je tekst odabran za korpus, treba odlučiti koji sve dijelovi teksta ulaze u korpus. Uglavnom se tekst čuva u najosnovnijem formatu – kao niz slova, razmaka i interpunkcijskih znakova, koji se nazivaju pismenima (*characters*), a čuva se razlika između velikih i malih slova, kurziva i slično. Brojevi stranica i paragrafa čuvaju se samo radi lakšeg kasnijeg referiranja, a ostale se informacije o izgledu teksta odbacuju. Takav neobrađeni tekst bez bilo kakvih drugih kodova najbolji je za daljnja istraživanja jer onda svaki istraživač može za određeno istraživanje pridodati svoju vrstu kodova, koja ne mora kasnije postati sastavnim dijelom korpusa.

Leech daje sedam maksima koje bi trebale biti primijenjene pri obilježavanju tekstnog korpusa. Mogu se pobrojati kao:

1. Mora biti moguće ukloniti oznake iz obilježenog korpusa i vratiti se početnom korpusu;
2. Mora biti moguće izvući oznake same za sebe iz teksta za pohranu negdje drugdje, na primjer u obliku relacijske baze podataka;
3. Obilježavanje treba biti temeljeno na uputama dostupnim krajnjem korisniku. Većina korpusa ima upute s kompletnim detaljima obilježavanja i vodičem za potpuno razumijevanje što svaki primjer obilježavanja predstavlja i zašto je određena odluka u obilježavanju donesena u slučajevima gdje je moguće više od jedne interpretacije teksta;
4. Treba biti jasno naznačeno tko je i kako napravio obilježavanje u tiskanom priručniku ili u dokumentaciji objavljenoj uz korpus. Korpus može biti obilježen ručno ili u potpunosti automatski računalnim programom, čiji rezultat mogu ili ne moraju ispraviti ljudi;
5. Krajnji korisnik mora biti svjestan da obilježavanje nije nepogrešivo nego samo potencijalno koristan alat;
6. Obilježavanje mora biti temeljeno što je više moguće na općeprihvaćenim i što neutralnijim principima;
7. Nijedno obilježavanje ne može *a priori* biti smatrano standardom. Standardi, ako postoje, proizlaze iz praktičnog dogovora (prema McEnery i Wilson 2001: 33-34).

Nešto sažetiji opis koraka u sastavljanju i obilježavanju korpusa daju O'Keeffe, McCarthy i Carter (2007: 8).

Kako je već rečeno, nekodirani je tekst samo niz pismena, a svako pisme odgovara jednoj tipki na tipkovnici. Pritom se bjelina uzima kao granica riječi, pa se svi oblici između dvije bjeline računaju kao različite riječi, poput *dječak* i *dječaci*, i *doći, dođe, došao*.

Postoji više vrsta obilježavanja, a sve će one detaljnije biti objašnjene u sljedećem poglavlju uz alate kojima se različite jezične razine obilježavaju u korpusu. Danas postoji potreba za sve više označenih podataka: isti tip oznaka za različite žanrove i različite jezike, detaljnije oznake za pojedine jezike, sravnjivanje (*alignment*) za usporedne korpuse i slično. Označavanje vrsta riječi danas je najraširenije i jedan od prvih koraka pri označavanju korpusa jer se može izvršiti računalno uz visoku preciznost bez ručne intervencije, zato što je točna vrsta riječi za bilo koju riječ predvidiva iz njezinog ko-teksta, uz minimalne informacije o jeziku (npr. najčešći sufiksi i njihove moguće vrste riječi). No, postoji nesuglasnost oko zadržavanja svih informacija za što veću korist krajnjim korisnicima i uklanjanja problematičnih razlika kako bi se automatsko označavanje učinilo točnijim. Neki su projekti označavanja znatno smanjili broj mogućih vrsta riječi u skupu oznaka.

EAGLES (*Expert Advisory Group on Language Engineering Standards*, http://www.ilc.cnr.it/EAGLES/browse.html) je napravio preporuke na temelju oznaka za vrste riječi za europske jezike. Te preporuke imaju tri razine svojstava:

- *Obavezna svojstva*, koja su najosnovnije razlike koje moraju biti označene u bilo kojem tekstu koji se označava vrstama riječi;
- *Preporučena svojstva*, koja su dodatno prepoznate gramatičke kategorije koje trebaju biti označene ukoliko je to moguće;
- *Neobavezna svojstva*, koja se mogu upotrijebiti za specifične svrhe, ali koja nisu toliko potrebna da bi bila obavezna ili preporučena.

Obavezna svojstva koja EAGLES priznaje jesu glavne vrste riječi – imenica, glagol, pridjev, zamjenica/determinativ, član, prilog, prijedlog, veznik, broj, umetanje, jedinstveno, ostalo i interpunkcija (McEnery i Wilson 2001: 52).

Korpus, kao maksimalno reprezentativan uzorak, omogućuje kvantifikaciju rezultata i usporedbu s drugim rezultatima, kao i bilo kakvo drugo znanstveno istraživanje temeljeno na podatcima. Korpus ne pruža samo kvantitativne već i kvalitativne podatke pri analizi. Razlika jest u tome da kvantitativna analiza uglavnom istražuje frekvenciju (broj pojavljivanja neke pojave u određenom kontekstu) jezičnih obilježja utvrđenih u podatcima, dok su pri kvalitativnoj analizi podatci korišteni samo za utvrđivanje i opisivanje aspekata upotrebe jezika te omogućuju primjere pojedinih pojava. Pri kvalitativnoj analizi rijetke bi pojave trebale dobiti jednaku pažnju kao i one česte jer se takvom analizom teži potpuno detaljnom opisu a ne kvantifikaciji. Tako višeznačnost koja je inherentna ljudskom jeziku, ne samo slučajno nego i namjerom govornika, može biti potpuno prepoznata u analizi: kvalitativno istraživanje ne primorava na potencijalno pogrešnu interpretaciju. Takva su istraživanja važna jer pružaju

pravu sliku o karakterističnim upotrebama, kao i o stupnju javljanja neke upotrebe unutar i između jezičnih varijeteta, što je važno ne samo za razumijevanje gramatike samog jezika nego i za proučavanje različitih jezičnih varijeteta i pri učenju jezika.

S druge strane, kvantitativna istraživanja, odnosno frekvencija, daju drugu sliku. Ljudi imaju nejasnu predodžbu o frekvenciji neke konstrukcije ili riječi. Prirodno je proučavanje podataka jedini pouzdani izvor za dokaze o svojstvima kao što je frekvencija. Korpus pruža osnovu za sustavan pristup analizi jezika jer pruža objektivnu provjeru rezultata, što se ne može reći za introspekciju. No kvantifikacija u korpusnoj lingvistici nije samo obično brojanje – postoji mnogo sofisticiranih statističkih tehnika koje omogućuju rigorozne matematičke analize kompleksnih podataka kako bi s određenim stupnjem sigurnosti pokazale da su razlike između tekstova, žanrova, jezika stvarne a ne slučajnost dobivena prilikom uzorkovanja.

Frekvencija sama po sebi ne može biti mjerilo tipičnosti – u korpusu od deset žanrova, dvije riječi mogu obje imati frekvenciju 20, no jedna od njih može se pojaviti po dva puta u svakom od deset žanrova, dok druga svih 20 pojavljivanja može imati unutar jednog žanra. Raspršivanje onda pokazuje koliko je neka riječ tipična i koliko se često pojavljuje. Frekvencija pojavljivanja ukazuje na frekvenciju upotrebe, što daje dobru osnovu za vrednovanje profila određene riječi, strukture ili iskaza u odnosu na normu. Horizontalna os konkordancije prikazuje sintagmatske uzorke, dok vertikalna os daje paradigmatsku raspoloživost odnosno izbor dostupan govorniku ili piscu u danom trenutku i unutar određenog jezičnog sustava. Bilo bi najjednostavnije pretpostaviti da je bilo koja riječ, fraza ili rečenica koja se pojavljuje u korpusu reprezentativna za jezik koji se istražuje, no to je prihvatljivo samo za potrebe čiste deskripcije. Pretpostavka da je iskaz, ako se nalazi u korpusu, po definiciji prihvatljiv i stoga bi trebao biti uključen u gramatiku jezične upotrebe, neprihvatljiv je stav iz preskriptivne perspektive zbog nekoliko razloga. Pisci i govornici ponekad namjerno ruše normalne konvencije susreta ili dokumenta kako bi prikazali grešku, ili zbog stilskog ili dramatičnog efekta, ili zbog niza drugih svakodnevnih razloga. Računalo bi pri nasumičnom odabiru primjera kad-tad naišlo na jedan od nenormalnih oblika i korisnik bi ih smatrao neprihvatljivima. Treba uvijek imati na umu da su čak i višemilijunski primjeri riječi maleni u usporedbi s količinom jezika proizvedenog i u manjim jezičnim zajednicama, stoga pojavljivanje od nule ili blizu nule može biti rezultat pogrješke pri uzorkovanju. Zato su reprezentativnost i uzorkovanje središnja pitanja pri sastavljanju korpusa. Ako se riječ pojavljuje deset puta na milijun riječi u korpusu od sto milijuna riječi, velika je šansa da će se isto dogoditi i u sljedećih sto milijuna riječi ukoliko nema velike promjene u sastavu korpusa. No, mnoge su riječi i fraze rijetke u općenitom uzorku tekstova ali vrlo česte u određenim specifičnim tekstovima. Dokaz iz opsežnog općeg korpusa

može pomoći u identifikaciji najčešćeg značenja riječi, no to se treba uzeti s oprezom. Frekvencija sama po sebi nije dovoljna. Primjerice korpusni leksikografi moraju promatrati distribuciju: pojavljuje li se riječ u više različitih tekstova ili samo u određenoj domeni ili samo kod određenog autora. Također, postoje riječi koje se u tekstu pojavljuju samo jednom, što je najčešće polovica svih riječi u tekstu, a nazivaju se *hapax legomenon* (grč. *hapax* „jednom“, *legomenon* „izgovoreno“). Korpus je reprezentativan kad zaključci temeljeni na njegovim sadržajima mogu biti generalizirani na veći hipotetični korpus.

Računalo može potražiti određenu riječ, niz riječi ili čak vrstu riječi u tekstu. Njegova sposobnost da izvuče sve primjere te riječi, obično u kontekstu, od velike je koristi lingvistu. Također može izračunati broj pojavljivanja riječi kako bi informacija o frekvenciji riječi mogla biti prikupljena. Podatci se mogu razvrstati prema nekom redu – na primjer abecednim redom riječi koje se pojavljuju zdesna ili slijeva.

Jedna od bitnih odlika korpusa jesu imena, koja su, kao i vrste riječi, postala jedna od primarnih dijelova korpusa koji se označavaju, kako bi se kasnije lakše moglo doći do potrebnih informacija pri pretraživanju korpusa. Imena (*named entities*) su ta koja u tekstovima najčešće prenose dodatne obavijesti jer izravno povezuju tekst s izvantekstnim svijetom. Uobičajena pitanja *tko? kada? što? gdje? koliko?* postavljaju se kad se želi doći do temeljnih informacija nekog događaja, a odgovori na ta pitanja u informativnim su tekstovima najčešće imena. Za konferenciju MUC-7 u postupak prepoznavanja naziva uključeno je sedam vrsta imena: osoba, organizacija, mjesto, nadnevak, vrijeme, valuta i postotak. Kasnije je tih sedam vrsta prošireno i imenima za mjere tj. izrazima koji iskazuju vrijednosti iskazane u nekim mjernim jedinicama kao i imenima za geopolitička tijela (*geo-political entities, GPE*), kao npr. *NATO, WEU* itd.

Za razliku od imena, koja olakšavaju određena pretraživanja korpusa, višeznačnost (*ambiguity*) je jedan od problema koji se već desetljećima proteže u korpusnoj lingvistici. Prirodni su jezici protkani višeznačnošću na svakoj razini opisa, od fonetske do sociolingvističke, i po tome se radikalno razlikuju od formalnih jezika. Kao korisnici prirodnih jezika, nesvjesni smo ove sveprisutne višeznačnosti – privlači nam pozornost samo u obliku marginalnih lingvističkih fenomena kao što su igre riječima ili nesporazumi.

U sljedećem poglavlju opisuje se obilježavanje korpusa, koje sve faze pri obilježavanju postoje, koje se sve razine i kako mogu obilježiti, kakvi alati za obilježavanje postoje i slično.

5. Obilježavanje korpusa

5.1. Obilježavanje

Prije no što se prikažu postojeći alati za obilježavanje korpusa, dajemo sažet prikaz koraka pri obilježavanju Lawlera i Dryja (1998: 240-2):

1. korak: *Analiza podataka* Prije nego analiziramo podatke, moramo ih imati u nekom računalno čitljivom obliku. Danas imamo korpuse koji obuhvaćaju širok raspon tipova podataka (govor, tekst, multimedijski dokumenti) i više jezika. Zajednica također ima koristi od dobro organiziranih napora u skupljanju podataka, prikazano primjerom MUC-a i uslugama i proizvodima Lingvističkog podatkovnog konzorcija (*Linguistic Data Consortium*) ili Europske agencije za jezične resurse (*European Language Resources Agency*).

Analiza treba biti vođena glavnim zadatkom, bilo to transkripcija, prevođenje, pretraga, sažimanje, ažuriranje baze podataka ili upiti za bazu podataka. Naš je cilj pronaći sustavno i robusno preslikavanje od inputa do outputa. Podatci pružaju input; struktura zadatka pruža output, oblik outputa je označeni oblik inputa (npr. dokument proširen vrstama riječi za svaku riječ).

2. korak: *Postavljanje hipoteze postupka* Bazirano na našoj analizi podataka, postavljamo hipotezu postupka koji će omogućiti preslikavanje od inputa do outputa. Važno je da se taj postupak može primijeniti u računalnom programu. Postoje mnogi pristupi, od neuronskih mreža preko stohastičkih modela do sustava baziranih na pravilima. Neki od njih koriste eksplicitna pravila koja je stvorio čovjek, neki koriste strojno učenje, a neki su bazirani na statističkim procesima.

3. korak: *Provjera postupka* Korpusno bazirana metodologija koristi podatke za dvije različite namjene: potaknuti analizu i omogućiti mjerilo za provjeru analize. Ta metoda zahtijeva pažljivo definiranje evaluacijskog mjerila koje daje rezultat, tako da možemo usporediti strategije ili skupove pravila. Važno je da koristimo nove podatke (odnosno podatke koji se ne koriste u fazi postavljanja hipoteze) za evaluaciju kako bismo osigurali stvaranje sustava koji je robustan u odnosu na vrste podataka kojima će sustav morati rukovati. Ako to ne uradimo, riskiramo izradu strategije koja će biti pretjerano specifična s obzirom na podatke koje smo koristili za obuku. Točnost evaluacije najviše ovisi o točnosti u određivanju „ispravnog“ outputa u 90% slučajeva, nemoguće je razviti sustav koji će biti 95% točan jer se ljudi ne mogu dogovoriti što to točno znači.

4. korak: *Iteracija* Jednom kada evaluiramo naš pristup, možemo koristiti standardne metode za poboljšanje rezultata, kao što su pristup sustavnog strojnog učenja, iterativno ispravljanje

grešaka ili regresivno testiranje. Tijekom iteracije možemo ponovno promotriti bilo koji od prethodnih koraka. Možda ćemo morati poboljšati skup oznaka, postupak – ili čak evaluaciju.

Obilježavanje je još u počecima i samo je mali dio mogućih oznaka jasno definiran i spreman za upotrebu. Za svako je obilježavanje potrebno specificirati konačni popis oznaka imajući na umu svaki mogući kontekst. Što je opis lingvističke pojave precizniji, to je veća mogućnost pojavljivanja raspršenih podataka. Idealno bi obilježavanje trebalo biti jasno i bez višeznačnosti te bi trebalo biti jednostavno za razumijevanje nekome bez opsežnog lingvističkog znanja. „Označavanje (*tagging*) je proces pridruživanja oznaka (*tags*) iz skupa ili popisa oznaka dijelovima teksta (pojavnica, rečenica i sl.) koji su delimitirane jezične jedinice“ (Bekavac 2001: 18).

Neobilježeni se korpus prvo stavlja u alate za obilježavanje i obilježava se, stvarajući inačicu teksta s oznakama kodiranima u XML-u ili drugom metajeziku. Takav se obilježeni korpus pohranjuje u posebnom alatu kako bi korisnik mogao pristupiti pretraživanju i dobiti potrebne rezultate. Ova se dva koraka najčešće izvode odvojeno. Alati za pretraživanje korpusa postali su vrlo jednostavni za korištenje. S druge strane, programi za obilježavanje korpusa uglavnom zahtijevaju napredno poznavanje računala kako bi se instalirali i koristili pa su nepristupačni većini lingvista (McEnery i Hardie 2012: 33).

Program kazuje računalu kako analizirati korpus, koji tekst treba koristiti kao input, koja lingvistička svojstva treba analizirati i kako ih prepoznati, te koju vrstu outputa treba proizvesti.

Označeni se korpusi razlikuju po količini informacija koje sadrže o riječima. Svi označivači fokusirani su ponajprije na podatke o vrsti riječi, no različiti označivači uključuju različitu količinu gramatičkih podataka, a neki također uključuju semantičke i sintaktičke podatke.

Ovdje treba napraviti distinkciju među nekim terminima. Nerijetko se na ovome polju nalaze termini obilježavanje (*annotation*) i označavanje (*tagging*). Valja napomenuti kako se ova dva termina koriste za različite vrste dodavanja oznaka, pri čemu se označavanje najčešće odnosi na označavanje vrsta riječi (*POS tagging*), dok je obilježavanje nadređen pojam i može se odnositi na neke druge vrste, poput dodavanja gramatičkih oznaka, semantičkih uloga, kao i na obilježavanje strukture teksta tj. na nelingvističku razinu obilježavanja. Označavanje može u korpus dodavati oznake za vrste riječi, podvrste glagola, tipove zavisnih rečenica i slično.

Označavanje riječi tj. svake pojedine pojavnice u korpusu smatra se najnižom standardnom procedurom koja prethodi svim ostalim analizama. Cilj je ove procedure da svakoj riječi u korpusu pridruži najprikladniju morfosintaksnu kategoriju što znači da jedna riječ u danoj rečenici može imati samo jednu oznaku. Ako se pak dogodi da je jednoj riječi moguće pripisati više oznaka, radi se o riječi koja može imati više značenja, ali i više funkcija u rečenici. Takve

riječi u hrvatskome jeziku najčešće imaju i različit izgovor, no zbog nekorištenja oznaka za naglasak u pisanju hrvatskih tekstova, tu razliku ne možemo iskoristiti za ovakvu analizu (Vučković 2009: 30).

„5 svojstava koje svaki označivač mora imati:

– robusnost – sustav se ne sruši ako naiđe na riječ koja je označivaču nepoznata, negramatična;

– efikasnost – vrijeme obrade raste linearno s porastom količine teksta;

– točnost – pokušati svakoj riječi pridružiti točnu POS oznaku;

– mogućnost podešavanja – označivač se može podešavati različitim lingvističkim pomoćnim oznakama za različit korpus;

– ponovna iskoristivost (*reusable*) – lako prilagodljiv novom korpusu, novom skupu oznaka i novom jeziku“ (Isto: 33).

Označeni korpusi doveli su do treniranja stohastičkih komponenti za obradu prirodnog jezika što može rezultirati znatnim poboljšanjima za parsanje i razrješavanje višeznačnosti riječi. Ovakav je uspjeh potaknuo razvitak sve veće raznovrsnosti korpusa s bogatijim i raznolikijim obilježavanjem, na primjer obilježavanje automatskog izvlačenja sadržaja (oznake za imena, vlastite riječi, semantičke odnose i događaje), semantičko obilježavanje, obilježavanje semantičkih uloga, i pragmatičko obilježavanje, poput vremenskih odnosa. Unatoč odabranoj metodi, dobro označeni korpusi važan su izvor za testiranje i poboljšanje jezičnog modela. Kod dovoljno raznolikog i velikog korpusa gdje je točna interpretacija označena, ako gramatičar zaboravi neki gramatički fenomen pri izradi pravila, testni korpus će ga podsjetiti na previd.

Većina algoritama za označavanje pripada jednoj od klasa:

- Označivači bazirani na pravilima – uglavnom uključuju velike baze podataka ručno ispisanih pravila o razrješavanju višeznačnosti.
- Stohastički označivači – višeznačnost razrješavaju upotrebom korpusa za treniranje za procjenu vjerojatnosti dane riječi s obzirom na danu oznaku u danom kontekstu.
- Označivač baziran na transformacijama ili Brillov označivač – dijeli značajke s obje klase označivača. Temelji se na pravilima koja određuju kad višeznačna riječ treba imati određenu oznaku, no sadrži i komponentu strojnog učenja – pravila su automatski preuzeta iz prethodno označenog korpusa za treniranje.

Alati za strojnu obradu jezika mogu se bazirati na korpusnim podatcima, koji se koriste radi treniranja nekog modela jezika koji sustav sadrži. Treniranje se može napraviti na neoznačenom korpusu ili na ručno obilježenom korpusu. Takvo ručno obilježavanje može imati barem dvije funkcije – omogućava modelu kojeg je program razvio da bude što ispravniji, te takvi podatci

mogu biti korisni za evaluaciju testiranja za takve programe. Omogućavajući označivaču vrsta riječi da označi tekst koji je prethodno već označen znatno je brže i automatiziranije – računalo može preko ručnih oznaka ocijeniti svoju učinkovitost, umjesto oslanjanja na ljude za ispitivanje i ocjenjivanje outputa. Za višeznačne riječi, većina označivača koristi probabilističke informacije, koje se temelje na prethodno točno označenim korpusima.

„Dobro bi kodiran korpus trebao biti:

• višestruko uporabiv (*reusable*), potencijalno uporabiv u više istraživačkih projekata i za više namjena,

• proširljiv (*extensible*), u smislu mogućnosti daljnjega nadograđivanja postojećega korpusa“ (Bekavac 2001: 54).

5.2. Opojavničenje

Elektronski je tekst u stvari slijed pismena. Prije bilo kakve obrade teksta, tekst se mora rastaviti na jezične jedinice kao što su riječi, interpunkcija, brojevi, alfanumerički znakovi i drugo. Taj se proces naziva opojavničenje. Termin *riječ* je višeznačan: riječ iz vokabulara jezika može se pojaviti više puta u tekstu, ali i dalje je samo individualna riječ jezika. Tako postoji razlikovanje između riječi vokabulara i oblika riječi i višestrukih pojavljivanja ovih riječi u tekstu koje se nazivaju pojavnicama (*token*). Zato se proces rastavljanja pojavnica u tekstu naziva opojavničenje. Iako je razlika između različnica (*type*) i pojavnica važna, najčešće se za oboje koristi *riječ* gdje je god u tekstu jednoznačnost implicirana. Pojavnica je svako pojedinačno pojavljivanje *riječi* u korpusu, pa bi se pod pojmom milijunski korpus podrazumijevao korpus od milijun pojavnica. Različnica je jedinstveni lik (najčešće grafijski) pojavnice iz korpusa.

Pri obilježavanju se određenim objektima pridodaju oznake za početak i kraj, a takav se objekt (niz pismena) naziva element. Jednostavan element može izgledati ovako:

<riječ>Ivan</riječ> (Bekavac 2001: 31).

Izgled većine označivača značajno je slična: Opojavničenje – tekstni ulazni podatci podijeljeni su na pojavnice prikladne za daljnju analizu: interpunkcija, jedinice riječi i granice iskaza. Provjera višeznačnosti uključuje leksikon i alat za pretpostavljanje (*guesser*) za pojavnice koje se ne nalaze u leksikonu. U najjednostavnijem obliku, leksikon može biti popis oblika riječi i njihovih mogućih vrsta (*POS*). Ekonomičnije je rješenje bazirano na modelima s konačnim brojem stanja (*finite-state models),* na primjer dvorazinska morfologija, gdje se lingvističke generalizacije (o fleksiji i derivaciji) mogu odgovarajuće modelirati računalnim modelom morfologije nekoga jezika. Alat za pretpostavljanje analizira preostale pojavnice.

Dizajn takvog alata najčešće je baziran na onome što je poznato o leksikonu. Na primjer, ako je poznato da leksikon sadrži sve zatvorene klase vrsta riječi kao što su zamjenice i članovi, alat za pretpostavljanje sa sigurnošću može ponuditi samo otvorenu klasu vrsta riječi (npr. imenice ili glagole). Upotrijebljeni s tumačem, leksikon i alat za pretpostavljanje sastavljaju leksički (ili morfološki) analizator koji pruža sve razumne analize kao alternativu za svaku pojavnicu (Mitkov 2003: 221).

„Segmentacija teksta na rečenice (*sentence segmentation, sentence boundary disambiguation)* u mnogim je slučajevima prvi korak za brojna područja strojne obrade jezika kao što je npr. označavanje vrsta riječi *(POS tagging)*, sintaktički parsing, sravnjivanje rečenica usporednoga korpusa ili pak za određivanje čitljivosti teksta. Segmentacija se rečenice obavlja ubacivanjem jedinstvenih nizova pismena, tj. graničnih oznaka na početak, odnosno na završetak rečenica u tekstu (u suvremenim shemama za obilježavanje teksta to su nizovi <S> i </S>)“ (Bekavac 2001: 20).

Vrsta informacije na kojoj je opojavničitelj utemeljen može se razlikovati – opojavničitelj temeljen na riječima može tražiti potencijalne pojavnice u leksikonu; neki detektori rečenica koriste informacije o vrstama riječi, čime zahtijevaju označene ulazne podatke. Drugi sustavi izvlače informacije o frekvenciji koje se mogu upotrijebiti za odluke u višeznačnim slučajevima.

Djelovanje opojavničitelja unekoliko graniči s djelovanjem parsera. Dok opojavničitelj rastavlja tekst na riječi i sintagme, parser rastavlja na rečenice. Pritom oba alata mogu imati dodatne funkcije, poput obilježavanja riječi, odnosno dodavanja sintaktičkih kategorija riječima u tekstu. Sljedeće poglavlje opisuje djelovanje i vrste parsera koji postoje.

5.3. Parsanje

Rastavljanje na rečenice važan je dio razvijanja mnogih aplikacija za obradu teksta – sintaktičko parsanje, dohvaćanje informacija, strojno prevođenje, sravnjivanje teksta, sažimanje dokumenta i sl. U većini je slučajeva rastavljanje jednostavno – točka, upitnik ili uskličnik označavaju granicu rečenice, no ima slučajeva kada je točka dio kratice pa tako ne ukazuje na granicu rečenice. Kratica sama po sebi može biti zadnja pojavnica u rečenici pri čemu je njezina točka dio kratice i granica rečenice. Rastavljanje rečenice tako može predstavljati neočekivane teškoće koje se moraju riješiti. Rastavljanje na rečenice ili segmentacija na rečenice zahtijeva analizu lokalnog konteksta oko točke i drugih interpunkcija koje bi mogle ukazivati na kraj rečenice. Ne označava svaka bjelina granicu među riječima. Sintagme (*multi-word expressions, MWE*) se sastoje od niza dvije ili više jedinica odvojenih

bjelinama, a zbog njihovog visokog stupnja leksikalizacije, cijeli se niz može smatrati i samo jednom pojavnicom.

„Naziv *parsanje* (*parsing*) je pojednostavljeni naziv koji se koristi u računalnoj znanosti i lingvistici umjesto formalnijega i preciznijega naziva *sintaktička* (ili *sintaksna*) *analiza*." (Agić 2012: 6). Termin parsanje odnosi se na proces automatske analize dane rečenice, promatrane kao slijed riječi, kako bi se utvrdile sve moguće osnovne sintaktičke strukture. Sintaktička je analiza ili parsanje teksta pisanoga prirodnim jezikom raščlamba rečenica toga teksta od razine rečenice do razine riječi, u skladu s prethodno zadanim okvirom za sintaktički opis toga jezika (Isto: 8-9). Čim su osnovne morfosintaktičke kategorije identificirane u tekstu, moguće je međusobno dovesti te kategorije u sintaktičke odnose višeg nivoa.

Sintaksna struktura rečenice označava način na koji su riječi u rečenici međusobno povezane, kako se riječi grupiraju u skupine, koje riječi opisuju druge, koje su riječi od centralne važnosti u rečenici, koje veze postoje između skupina. Procesom parsanja izvode se strukturna svojstva rečenice i daje se sintaksni prikaz kojim se pridružuje sintaksno ime svakoj osnovnoj vrsti strukture. U slučajevima višeznačnosti, sintaksni opis može uključivati popis više mogućih sintaksnih prikaza (Vučković 2009: 48).

Tablično parsanje prikuplja alternativne analize u tablicu, organizira ih i procjenjuje, što se koristi vrlo uspješno u analizi sintaktički višeznačnih rečenica.

Tehnike parsanja prirodnog jezika koriste gramatiku za dodjeljivanje sintaktičke analize nizu riječi. Razina detaljnosti ovisi o zadatku obrade jezika koji se izvodi i pristupu zadatku koji se obavlja – na primjer, odluka o anafori može tražiti samo identifikaciju granica osnovnih fraza, dok obrada upita nad bazom podataka može tražiti detaljno parsanje.

Tri su osnovne tehnike parsanja:

– silazna ili top-down metoda – orijentirana je prema cilju, kreće od početnog simbola *S* koji je obavezni korijen za sve rečenice, pokušavajući doći do listova stabla uz pomoć postojeće gramatike; upravljana je krajnjim ciljem ili hipotezom (*goal driven*),

– uzlazna ili bottom-up metoda – orijentirana je prema podatcima tj. prema listovima stabla od kojih pokušava doći do korijena stabla; kreće se od riječi u rečenici i njihovih leksičkih kategorija koje se spajaju u skupove, sve dok se ne dođe do konačnog neterminala *S* tj. oznake za rečenicu; upravljana je podatcima (*data driven*),

– kombinirana metoda - metoda koja se koristi i silaznom i uzlaznom metodom parsanja istovremeno.

Za više detalja o metodama parsanja v. Jurafsky i Martin 2000.

Parser je računalni program sposoban za analizu sintaktičke strukture rečenica. Takvi se programi usredotočuju na rješavanje gramatičke višeznačnosti i razrješuju točna grupiranja jedinica. Minimalno mora identificirati riječi u rečenici, dodijeliti ispravne sintaktičke opise tim riječima, grupirati te riječi u jedinice višeg stupnja (uglavnom sintagme i surečenice) koje identificiraju glavne sintaktičke sastavnice rečenice, te imenovati te sastavnice.

„Parseri prirodnoga jezika obično obrađuju tekst u dvije faze. U prvoj fazi opojavničavatelji, morfološki analizatori, prevode niz znakova u niz riječi dok u drugoj fazi sintaksni analizator ili parser prevodi niz riječi u parsanu rečenicu, tj. u niz parsanih rečenica“ (Isto: 37).

Postoje parseri koji dodaju sintaktičku analizu korpusu, identificiraju subjekte, glagole i objekte, kao i kompleksnije sintaktičke informacije, semantička svojstva, i prozodijska svojstva za govorne korpuse.

Još jedan od načina na koji možemo promatrati proces sintaksne analize je pretraživanje parsera kroz šumu mogućih stabala u potrazi za najboljim parsnim stablom ulazne rečenice. (Isto: 47). „Parser treba biti moćan i fleksibilan, a četiri osnovna svojstva koja bi trebao zadovoljiti su:

– robusnost – za svaku rečenicu u tekstu treba dati najmanje jednu analizu;

– uklanjanje višeznačnosti – za svaku rečenicu u tekstu, treba dati najviše jednu analizu;

– točnost – svaka analiza koju ponudi treba biti točna u što je moguće većem broju;

– efikasnost – za svaku analizu treba koristiti što je moguće manje vremena i računalne memorije“ (Isto: 51).

Cilj parsera je identificiranje točne sintaktičke analize među svim mogućim analizama rečenice. Duljinom rečenice eksponencijalno raste broj mogućih analiza. Zadaci parsera:

– razdioba rečenice na svoje sastavne skupove, podskupove i leksičke kategorije,

– označavanje sastavnica,

– izgradnja hijerarhijskog prikaza njihovih struktura,

– mapiranje nizova u njihove strukture,

– dubinska pretraga obavlja jednu po jednu hipotezu, dok površinska pretraga paralelno obavlja hipoteze,

– spremanje međurazinskih rezultata.

Većina označivača koristi rječnike koji popisuju kategorije kojima određena riječ može pripadati. Neke su riječi jednoznačne pa mogu jednostavno biti prepoznate. Druge su riječi višeznačne. Rječnici također mogu identificirati ustaljene izraze, a mogu imati popis riječi koje poprimaju određene gramatičke uzorke (npr. glagoli ili imenice koji kontroliraju dopune).

Zadovoljavajući parseri u automatskim sveobuhvatnim parsanjima rijetko prelaze granicu od 60% i najčešće se zaustavljaju na točnosti od 30-40%. Usporedbe radi, označivači vrsta riječi još su 1970-ih postizali točnost od 77%. Uzimajući u obzir takvu usporedbu, može se zaključiti da je problem parsanja znatno kompleksniji, a to je i za očekivati s obzirom da je kompleksnost mogućih kombinacija jezičnih jedinica na sintaktičkoj razini znatno veća od kompleksnosti na morfološkoj razini.

Parsna stabla korisna su u sustavima obrade riječi pri provjeri gramatike jer rečenica koja ne može biti parsirana može imati gramatičku grešku. Parsanje je važno na srednjoj razini reprezentacije za semantičku analizu, što ima važnu ulogu u strojnom prevođenju, odgovaranju na pitanja i dohvaćanju informacija.

Lawler i Dry (1998: 175) sintaktičko označavanje odnosno parsanje korpusa dijele na nekoliko koraka:

- Testiranje fonoloških pravila – primjenjuje fonološka pravila na korpusne podatke (fonološka analiza).
- Morfološko parsanje – program za parsanje riječi na njihove sastavne morfeme neprocjenjiv je za jezike s kompleksnom morfološkom strukturom.
- Sintaktičko parsanje – može se koristiti ne samo za primjer i testiranje analize nego i za praktične zadatke poput sintaktičkog označavanja teksta.
- Interlinearna analiza teksta – moguće interlinearno obilježavanje uključuje fonološku reprezentaciju, objašnjenja morfema, objašnjenja riječi i sintaktičke kategorije.

Pošto je obavljeno opojavničenje te parsanje, treba prijeći na sljedeći korak u obilježavanju korpusa, a to je lematizacija, koja je usko vezana uz označavanje vrsta riječi, kao što će biti prikazano u sljedećim poglavljima.

5.4. Lematizacija

Lematizacija (*lemmatisation*) je svođenje pojavnica iz korpusa na njihove natukničke oblike, tj. svođenje različitih pojavnica (članova iste paradigme) na zajedničku lemu (Bekavac 2001: 27).

Lematizacija je usko povezana s identifikacijom vrsta riječi. Uključuje redukciju riječi u korpusu prema njihovim odgovarajućim leksemima – glavnim riječima koje bi netko potražio ako traži riječ u rječniku. Tako bi, primjerice, oblici *udara, udario* i *udarile* svi bili reducirani na leksem *udariti*. Ti oblici čine lemu leksema *udariti*. Lematizacija se jednako primjenjuje na morfološki nepravilne oblike.

Lematizacija je važan postupak u istraživanju temeljenom na korpusu. U istraživanjima vokabulara i leksikografiji omogućuje istraživaču izvlačenje i proučavanje svih varijanata određenog leksema bez stavljanja svih mogućih varijanata u input, i izvlačenje informacija o frekvenciji i distribuciji leksema.

Leksikoni lema smanjuju redundantnost. Lema je kanonski oblik – uglavnom osnovni oblik – uzet kao reprezentativan za sve oblike paradigme (Mitkov 2003: 38).

Važno je pitanje odluke o fizičkom obliku koji lema treba imati. Tradicionalno, osnova, ili nepromijenjen oblik korišten je čak i kad je takav oblik teško ili nemoguće naći. No postoje mnoge alternative, primjerice oblici koji se najčešće upotrebljavaju mogli bi se uzeti za lemu, a prvi rezultati računala mogu omogućiti dobar temelj u planiranju novih metoda pristupa oblicima riječi nekog jezika.

5.5. Označavanje vrsta riječi

Vrste su riječi u lingvistici poznate još od Dionizija Tračanina (oko 100. pr. Kr.) koji je razlikovao osam vrsta riječi koristeći uglavnom formalne kriterije: imenice, glagoli, participi, članovi, zamjenice, prepozicije, prilozi, veznici. Najbolji kriterij za vrste riječi je gramatički (a ne semantički): (1) sintaktička distribucija, (2) sintaktička funkcija i (3) morfološke i sintaktičke klase kojima različite vrste riječi mogu pripadati.

Označavanje vrsta riječi (*POS tagging*) je proces u kojem je svakoj pojavnici u korpusu dodijeljena odgovarajuća vrsta riječi. Dobivši opojavničeni ulazni tekst, označivač određuje moguće vrste riječi za svaku pojavnicu, provjeravajući ih u leksikonu. Ako je pojavnica višeznačna između dvije ili više vrsta riječi, označivač mora odrediti točnu vrstu riječi prema danom kontekstu (razrješavanje višeznačnosti). Ako je pojavnica nepoznata, odnosno ako se ne nalazi u leksikonu, označivač mora pretpostaviti njezinu vrstu riječi. Informacija potrebna za razrješavanje višeznačnosti može se prikupiti uvidom u ko-tekst višeznačne pojavnice i iz značajki pojavnice same, kao što je frekvencija pojavljivanja s određenom vrstom riječi. Oznake se uglavnom primjenjuju i na interpunkcijske znakove. Postoji više shema označavanja vrsta riječi, temeljenih na popisima od 40 do 2000 oznaka.

Značajke većine označivača su:

- komponenta temeljena na pravilima – može se upotrijebiti za identifikaciju struktura koje slijede pravilne nizove,
- morfološki analizator – koristi određene morfološke karakteristike kako bi pomogao analizirati riječi koje nisu pronađene u rječniku.

Svaka se riječ u označivaču interpretira sa svim mogućim lingvističkim interpretacijama, a sustav prvo pokušava vidjeti je li svaka riječ prisutna u strojno čitljivom leksikonu koji je dostupan. Takvi leksikoni najčešće imaju oblik <riječ><vrsta riječi 1, ... vrsta riječi n>. Ako je riječ prisutna u leksikonu, sustav onda dodjeljuje riječi cijeli popis vrsta riječi s kojima može biti povezana. Informacije o vrstama riječi koristan su oblik obilježavanja koji može biti uveden u tekst s visokim stupnjem automatizma.

Razliku između označivačâ čini: (1) informativnost i specifičnost popisa oznaka i (2) stupanj točnosti dodjeljivanja oznaka, gdje se u obzir mora uzeti da točnost od 90% može biti ostvarena preko odabiranja najčešće vrste riječi za danu pojavnicu u višeznačnim slučajevima, dok je za nepoznate riječi točnost znatno niža.

Jezici uglavnom imaju relativno mali broj zatvorenih vrsta riječi, koje su često vrlo frekventne, uglavnom funkcionalne riječi, dok otvorene vrste riječi uglavnom uključuju različite tipove imenica, glagola, pridjeva. Sustavi za razlikovanje osnovnih vrsta riječi prošireni su dodatnim informacijama, kao što su lice i broj i u tome slučaju uključivanja obavijesti o dodatnim gramatičkim kategorijama više ne govorimo o označavanju vrsta riječi (*POS-tagging*) već o morfosintaktičkome označavanju (*MSD-tagging*).

Produkcija jedne riječi (ili vrste riječi) utječe na vjerojatnost druge riječi (ili vrste riječi) koja ju slijedi, kao dio koherentne strukture. Takva je struktura potrebna za stohastičke procese koji se računalno modeliraju kao tranzicijska matrica kako bi djelovali učinkovito. Označivači vrsta riječi koriste strukturu jezika pri razrješavanju zadataka s višeznačnim vrstama riječi jer riječi nisu međusobno neovisne.

Značajnost je vrsta riječi u pružanju informacija o riječi i njezinim susjedima. Na primjer znanje o tome je li riječ posvojna ili lična zamjenica može nam reći koje će se riječi prije naći u njezinoj blizini.

Najjednostavniji algoritam za nepoznate riječi pretpostavlja da je svaka nepoznata riječ višeznačna između svih mogućih oznaka, s jednakom vjerojatnosti. Označivač se onda oslanja na kontekstualne trigrame (tj. dvije riječi ispred i promatranu riječ, riječ ispred promatrane riječi i riječ iza nje, promatranu riječ i dvije riječi iza nje) kako bi predložio najvjerojatniju oznaku. Nešto kompleksniji algoritam pretpostavlja da je distribucija vjerojatnosti oznaka nepoznatih riječi slična distribuciji oznaka riječi koje se pojavljuju samo jednom u nizu (*hapax legomenon*). Na primjer, nepoznate riječi i *hapax legomenon* slične su jer su oboje najčešće imenice ili iz neke druge otvorene klase riječi, a ne mogu biti iz zatvorene klase.

Glavni izvor nepoznatih oblika riječi (u korpusu) je nestandardna ortografija, pa ne očekujemo da će takvi oblici biti dio leksikona standardnog označivača vrsta riječi. Druga važna

kategorija riječi koje se ne nalaze u leksikonu označivača tehnički su pojmovi, kratice ili riječi stranih jezika, te u neformalnim tekstovima nerijetko i emotikoni. Imena, osobito strana, greške označavanja i drugi izvori zajedno tvore ostatak nepoznatih riječi.

Korpusi koji su označeni vrstama riječi vrlo su korisni za lingvističko istraživanje, na primjer za pronalaženje u velikim korpusima instancija i/li frekvencije pojedinih konstrukcija, koje se sastoje od karakterističnih nizova vrsta riječi (tzv. *POS-patterns*), te kao brz, automatski i pouzdan izvor za izgradnju leksikona. Bez korpusâ ovi bi se leksikoni (obično veličine do stotinu tisuća jedinica) morali izgrađivati ručno. Takvi su leksikoni toliko veliki jer je broj riječi u prirodnom jeziku vrlo velik pa što je veći leksikon veće su šanse identifikacije riječi i pridruživanja odgovarajuće vrste riječi. Kad promatrana riječ nije prepoznata, na scenu dolazi morfološki analizator. Morfološka analiza koja se provodi nije prava morfološka analiza, nego je više pogađanje temeljeno na učestalim završetcima riječi. Neki završetci riječi na koje su sustavi osjetljivi čak i ne čine prave morfeme. Morfološka se jedinica može prizvati nekoliko puta, dok pokušava stvoriti različite riječi uklanjanjem različitih završetaka. Kod morfoloških analizatora često postoji interakcija leksikona i morfologije u automatskoj analizi teksta.

Pošto je pojavnica analizirana pomoću leksikona/morfološkog procesora i sve su njezine okoline identificirane, zadatak pridruživanja jedinstvenih oznaka vrsta riječi riječima daleko je od završetka. Leksikon i morfološka sastavnica samo ukazuju na niz vrsta riječi koje se mogu pridružiti nekoj riječi. Možemo znati da *žene* može biti i imenica i glagol, ali ne znamo što je u trenutnom kontekstu. Postupak razrješenja višestrukih mogućih interpretacija zove se razobličenje (*disambiguation*) i može se temeljiti na pravilima ili na statistici tj. vjerojatnosti da neka oznaka vrste riječi slijedi drugu u prepoznatoj okolini riječi i oznaka.

Obrada na razini morfologije potrebna je za flektivno bogate jezike koji zahtijevaju da se ista riječ u tekstovima pojavljuje u različitim oblicima. Kad govornik takvoga jezika želi referirati na sve oblike u kojima se neka riječ može pojaviti, onda odabere jedan od njih i rabi ga kao predstavnika svih ostalih. Takav se oblik naziva lema i upravo je on taj koji se najčešće koristi kao upitna riječ u tražilicama. Intuitivno, govornici flektivno bogatih jezika najčešće očekuju pronalaženje i onih dokumenata u kojima se lema pojavljuje i u ostalim oblicima (Tadić 2003: 61).

Na razini leksičke semantike moguće je označavanje smisla temeljeno na postojećem inventaru smisla ili skupu semantičkih klasa. Da bi to bilo moguće, potrebno je lematizirati svaku pojavnicu u korpusu i potom je označiti smislom ili klasom koja je tom pojavnicom (i njezinom lemom) prenesena.

„Povezivanje ovako razrađene semantičke mreže sa sustavima za pretraživanje teksta rezultira gotovo neslućenim mogućnostima crpljenja podataka iz teksta. Za sada su te mogućnosti otvorene istraživačima s npr. područja korpusne lingvistike koji mogu u pojedinim sustavima za pretraživanje korpusa postavljati složene upite u kojima se kombiniraju podatak o vrsti riječi i njezinu značenju. Tako je u CQP sustavu za pretraživanje korpusa moguće postaviti upit:

"kill.*" []? [pos="N.*" & ishuman(word)]

kojim se traži engleska riječ *kill* u bilo kojem obliku kad poslije nje slijedi imenica koja označuje ljudsko biće. Podatak o vrijednosti atributa *ishuman* za riječ *word* iz korpusa priskrbljuje se automatskim uvidom u WordNet za okolinsku riječ" (Tadić 2003: 71-2).

Tako je moguće u pretraživanje korpusa uključiti i semantičku komponentu, iako se zasad označavanje korpusa više temelji na morfološkom i sintaktičkom označavanju, kako je prikazano, a uspostavljanje semantičkih odnosa tek je za veći broj jezika u začetcima. Upravo je postojanje WordNeta za neki jezik dobar temelj za označavanje korpusa na razini leksičke semantike. Pošto je korpus označen, treba prikazati kako se pretražuje i kako ta pretraga izgleda.

5.6. Konkordancija

Nesumnjivo jedan od najvažnijih dostupnih alata korpusnim lingvistima jest alat za konkordanciju (*concordancer*), koji omogućava pretraživanje korpusa i izvlačenje određenog niza pismena bilo koje duljine – riječ, dio riječi, ili sintagma. U konkordancijama se tražena riječ ili niz riječi naziva stožernica i uglavnom se prikazuju u sredini konkordancijskoga retka uz tekst prije i poslije svake stožernice.

Stvaranje konkordancije teksta iziskuje kombinaciju slaganja i pretraživanja. Dobar konkordancijski program mora riješiti problem različnica i pojavnica: ako se želi konkordirati leme, treba uzeti u obzir njihove flektivne oblike. To je moguće ukoliko je korpus prethodno lematiziran i program za konkordiranje dopušta pretragu po lemama.

Ko-tekst ili okolina je u konkordancijama specificiran kao *x* riječi slijeva i *y* riječi zdesna od stožernice. Najčešće se uzima po pet riječi sa svake strane, no to se može prilagoditi prema potrebi.

Odnedavno su konkordancijski alati za analizu korpusnih podataka mrežno dostupni i mogu biti vrlo korisni za istraživanje frekvencije riječi, kombinacije riječi, a čak i za određene morfološke karakteristike. U gramatički označenim korpusima konkordancijski alati mogu se koristiti za istraživanje gramatičkih vrsta riječi.

„Unesena pojavnica za pretraživanje nalazi se između lijeve i desne okoline, a ona se u kontekstu konkordancija naziva stožernica (*headword*). Uobičajeno je da su stožernice u konkordancijama razvrstane onim redoslijedom kojim se pojavljuju u tekstu, međutim moguće ih je razvrstati po brojnim parametrima. Desno-redana konkordancija je poredana abecednim redoslijedom stožernice i pojavnica koje slijede nakon nje. Takvo redanje može poslužiti za analizu onih slučajeva gdje stožernica otvara neku frazu, ili je sadržana u njoj“ (Bekavac 2001: 74-5).

Kad je tekst vrlo dug, popis riječi će isto tako biti dug, a konkordancije iznimno dugačke. Nisu sve informacije potrebne svaki put pa se izbor može napraviti:

- Prema frekvenciji. Ako iz popisa izostavimo oblike riječi koji se pojavljuju samo jednom, smanjujemo popis za otprilike polovicu. Uobičajeno je razlikovanje između gramatičkih i leksičkih jedinica na ovaj način.
- Prema obliku. Moguće je specificirati riječi prema abecednom sastavu ili prema slovima u njima ili prema kombinaciji oba. Specifikacije poput ovih mogu se konstruirati za omogućivanje istraživaču odabira nekoliko klasa riječi.

Uzmimo u obzir faktore koji utječu na oblik i korisnost konkordancija:

1. Je li konkordancija selektivna ili iscrpna. Sposobnost iscrpnosti jedna je od temeljnih značajki konkordancija, jer može prikazati sve dostupne informacije, te je jasno superiornija od popisa odabranih citata gdje nema striktnih pravila o odabiru. Trenutno, jedina potreba za odabirom je u slučaju najčešćih riječi u vrlo velikim tekstovima. Uzorak pojavljivanja riječi u tekstu znači da u većim tekstovima postoje riječi koje se pojavljuju prečesto i one koje se ne pojavljuju dovoljno često da bi njihovo ponašanje bilo točno istraženo.
2. Duljina citata. Gotovo univerzalni oblik konkordancija je KWIC, gdje je duljina citata određena širinom stranice na računalnom zaslonu, s ključnom riječi u sredini. Ovaj je oblik dosta koristan, ali za istraživanje nekih riječi nije prikladan, pa se drugi oblici moraju smisliti. Trenutno opseg konkordancijskih formata raste. Često je korisno krenuti s jednostavnom KWIC konkordancijom i zatim se prebaciti na veći kontekst ili rečenični kontekst za pobliže istraživanje.
3. Raspoređivanje potvrda. Gdje ima na desetke, stotine ili tisuće potvrda oblika riječi treba razmotriti kako mogu biti ispisani za daljnja istraživanja. Najjednostavnija metoda je poredak tekstova, ali za neke potrebe može biti korisno ispisivanje abecednim redom riječi koje slijede ili prethode ključnoj riječi (Sinclair 1991: 42-43).

Konkordancijsko pretraživanje, s obzirom da se može pretraživati više riječi odjednom ili čak više vrsta riječi, a u prikazu se daje i ko-tekst tražene riječi, omogućava istraživanje odnosa

riječi, kao što su kolokacija, veza riječi nešto slabija nego što je to u sintagmi jer obje riječi barem donekle zadržavaju vlastiti smisao, te koligacija, veza između različitih vrsta riječi. Kolokacija i koligacija zajedno tvore jedinstvene smislene nizove ili dijelove jezika koji su pohranjeni u memoriji govornika i koji daju građu za sintagme. Dijelovi su spremni za upotrebu u svakom trenutku i ne iziskuju ponovno sastavljanje svaki put kad se koriste. Biber, Conrad i Reppen (2000: 262-8) daju cijeli niz testova za izračun kolokacija na temelju korpusnog pretraživanja, primjerice izračun jesu li dvije ili više riječi kolokacijski povezane na temelju njihove frekvencije u korpusu kao pojedinačne riječi naspram frekvencije kada se pojavljuju zajedno.

5.7. Alati za označavanje korpusa

Tadić (2003: 83) daje iscrpan popis alata koji se mogu upotrijebiti pri označavanju korpusa, a kako su u ovom radu bili opisani samo najuobičajeniji alati, dajemo prikaz njegovog popisa: „Neki su od alata još uvijek na razini akademskih prototipova dok su drugi gotovo dosegli status zaokruženih komercijalnih proizvoda. Može ih se podijeliti u nekoliko skupina:

- pretvornici (npr. 2XML koji pretvara dokumente iz RTF ili HTML zapisa u XML zapis)
- alati za obilježavanje dokumenata (specijalizirani urednici za SGML/XML obilježavanje jezičnih resursa, SGML/XML parseri i validatori itd.)
- alati za konkordancije (pretražuju korpuse i izručuju rezultate najčešće u obliku konkordancija)
- alati za statističku analizu jezičnih resursa
- označivači i lematizatori (označuju tekst na morfološkoj ili morfosintaktičkoj razini i/li lematiziraju tekst)
- banke sintaktičkih stabala (sintaktički analizirani korpusi)
- semantičke mreže (jezični resursi s eksplicitno kodiranim semantičkim odnosima)
- alati za sravnjivanje (uspostavljaju eksplicitne veze između odsječaka izvornoga teksta i njihovih prijevoda u dvo- ili višejezičnim tekstovima)
- sustavi za strojno i strojno potpomognuto prevođenje (pružaju potpuno/djelomično strojno prevođenje ili uspostavu i uporabu prijevodnih memorija u procesu prevođenja).“

5.8. Usporedni korpusi

Prije no što se navedu točni alati korišteni pri označavanju makedonsko-hrvatskog usporednog korpusa koji se u ovom radu opisuje, treba navesti alate za označavanje usporednih

korpusa. Usporedni se korpusi označavaju na jednak način kao i drugi korpusi, odnosno alatima koji su već opisani, a ovdje se daje prikaz alata koji se koristi isključivo pri označavanju usporednih korpusa.

Usporedni bi korpusi trebali biti stvarno usporedni, tako da se softver za njihovo sravnjivanje potpuno oslanja na neposredno podudaranje rečenicu po rečenicu između dva teksta. No idealni usporedni korpus je gotovo nemoguće naći. Trebao bi biti jednako tečan s obzirom na izvorni i ciljni jezik, a istovremeno pružati što dosljedniji prijevod, gdje je svaka zasebna rečenica sravnjena s odgovarajućom rečenicom prijevoda, što je zapravo u neskladu jedno s drugim.

„Sravnjivanjem se definiraju eksplicitne veze među odsječcima tekstova usporednoga korpusa. U biti, sravnjivanje (*alignment*) je povezivanje elemenata (rečenica, fraza ili riječi) koje su uzajamni prijevodi. Danas se sravnjivanje rečenica može izvesti automatski s visokim stupnjem točnosti. Alat koji obavlja sravnjivanje zove se program za sravnjivanje (*aligner*)“ (Bekavac 2001: 99).

Tipične primjene usporednih korpusa uključuju treniranje prevoditelja, dvojezičnu leksikografiju i strojno prevođenje (O’Keeffe, McCarthy i Carter 2007: 19).

Mogućnost sravnjivanja usporednih korpusa razgraničava takve korpuse od usporedivih, gdje nema potpunih prijevoda rečenica pa se takvi korpusi ne mogu sravniti.

6. Makedonsko-hrvatski usporedni korpus

Makedonsko-hrvatski usporedni korpus je jednosmjeran, tako da je korpus sastavljen od tekstova na makedonskom jeziku i njihovih prijevoda na hrvatski jezik. S obzirom da je granica suvremenosti za Hrvatski nacionalni korpus (dalje HNK) postavljena na 1990. godinu, tako da se prikupljaju tekstovi nastali 1990. i suvremeniji, isto je učinjeno i za ovaj usporedni korpus. To znači da su tekstovi zapravo odabirani s obzirom na njihove dostupne prijevode na hrvatski jezik, nastale od 1990. naovamo. Starost makedonskih originala nije uzimana u obzir jer je njihova suvremenost postavljena na 1945. godinu, od izlaska prve makedonske gramatike, a sva su djela nastala nakon te godine. Dakle, odabir je učinjen prema književnim djelima čiji su prijevodi na hrvatski bili dostupni. U izbor su ušla sva prozna književna djela, od romana, preko drama, do kratkih priča i novela. Poezija je izostavljena zbog poetskih odlika koje dopuštaju prijevodna odstupanja u korist rime i drugih poetskih figura, a nauštrb semantički i kontekstno odgovarajućih prijevodnih ekvivalenata pa bi takvi prijevodi na kraju napravili više štete nego koristi pri sastavljanju usporednog korpusa, osobito u ovom slučaju kad se radi o prvom uopće sastavljanju usporednog korpusa makedonskog i hrvatskog jezika, pa funkcionalnost ima prednost nad količinom. Upravo zbog funkcionalnosti, kao i manjka vremena i resursa, odabrana je književnost, jer se do novinskih i administrativnih usporednih tekstova u ovom trenutku nije moglo doći. Tekstovi su uzimani u cijelosti, osim pojedinih antologija ili zbirki gdje su uzimane pripovijetke i novele koje su bile dostupne na oba jezika, a ne cijele zbirke. Autori i prevoditelji pristali su na korištenje tekstova, no nikakav služben ugovor nije bio potpisan jer se u konkordancijskom prikazu neće moći dobiti više od jedne rečenice, pa se ne krše autorska prava o eksponiranju cijelih književnih djela.

Djela koja su ušla u ovaj korpus su (popisana na hrvatskom):

- Andonovski, Venko (2011). *Pupak svijeta*. Zagreb: Algoritam
- Andonovski, Venko (2015). *Genetika pasa*. Zagreb: Društvo hrvatskih književnika
- Andonovski, Venko. *Granica*. Rukopis ustupio prevoditelj dr. sc. Borislav Pavlovski
- Andonovski, Venko. *Svetica tame*. Rukopis ustupio prevoditelj dr. sc. Borislav Pavlovski
- Dimovski Vlado (2014). *O vuku i ježu.* U: *Makedonske priče za djecu*, Zagreb: Vijeće makedonske nacionalne manjine Grada Zagreba, str. 57-60.
- Isakovski, Igor (2012). *Pješčani sat*. Zagreb: Meandarmedia
- Jankovski, Andrea (2014). *Točno u proljeće*. U: Balkan Express 07, Zagreb: Klub studenata južne slavistike A-302, str. 218-229.
- Lafazanovski, Ermis (2010). *Hrapeško*. Zagreb: Europapress holding: Novi Liber

- Nikolova, Olivera (2014). *Zoki Poki*. Zagreb: Klub studenata južne slavistike A-302
- Osmanli, Tomislav (2012). *21*. Zagreb: Sandorf
- Sazdova, Vinka (2011). *Posljednji čaj*. Zagreb: V.B.Z.
- Smilevski, Goce (2007). *Razgovor sa Spinozom*. Zagreb: Edicije Božičević
- Smilevski, Goce (2013). *Sestra Sigmunda Freuda*. Zaprešić: Fraktura
- Starova, Luan (2000). *Vrijeme koza*. Zagreb: Znanje
- Prozni dijelovi zbirke *Revija malih književnosti – Makedonija* (2013). Zagreb: Udruga za promicanje kultura 'Kulturtreger'

Hrvatski dio korpusa ima 509 455 pojavnica, a makedonski 531 936 pojavnica. Makedonski dio očekivano ima više pojavnica, točnije 4,33% više, zbog svoje analitičnosti.

6.1. Označavanje usporednog korpusa

Pri označavanju ovog usporednog korpusa zasad je napravljeno samo sravnjivanje rečenica. Sva daljnja označavanja, osobito makedonskog dijela, ne mogu se izvršiti dok ne postoji gotov označivač vrsta riječi za makedonski, osim ako se svaku pojavnicu označuje ručno, što bi bio dugotrajan i mukotrpan posao. Sravnjivanje omogućuje prikaz usporednih rečenica na oba jezika, što je prvi korak upotrebe korpusa jer se barem može dobiti kontekst traženih pojavnica, kao i njihovi prijevodi na drugi jezik. Pretraživanje se može izvršiti na oba jezika, što je dodatna korist istraživačima koji bi se ovim korpusom mogli služiti, jer mogu naći prijevodne ekvivalente i s hrvatskog na makedonski, i obratno. Iako je korpus makedonsko-hrvatski, odnosno sadrži makedonske originale i hrvatske prijevode, u korpusu se hrvatski jezik smatra polazišnim jer za njega postoje dostupni alati za svaki korak označavanja, pa tako može biti detaljnije obrađen nego makedonski dio, za koji zapravo ne postoji ništa, s obzirom da zasad ne postoji Makedonski nacionalni korpus niti bilo kakav korpus makedonskog jezika dostupan na internetu, pa se ne može doći do bilo kakvih alata za označavanje makedonskog dijela usporednog korpusa, ako bilo kakvi alati negdje i postoje. Bilo je pokušaja izgradnje alata za označavanje korpusâ pri Makedonskoj akademiji znanosti i umjetnosti, no zbog nesuglasica lingvista oko nekih gramatičkih kategorija, zasad nije završen ni označivač vrsta riječi. Kako za hrvatski jezik postoji višemilijunski Hrvatski nacionalni korpus, tako postoje alati za sve korake označavanja, no za ovaj korpus i ograničenost vremena nije bilo moguće doći do svih alata kojima bi se mogao označiti hrvatski dio usporednog korpusa.

U sljedećem koraku trebalo bi se napraviti označavanje vrsta riječi, za koje je olakotna okolnost što hrvatski za označavanje koristi MTE oznake, koje postoje i za makedonski jezik.

Izvorna autorska djela konvertirana su u obični tekstni format kako bi se uklonile suvišne informacije poput formatiranja teksta (boje, fontovi i sl.), slika i drugih stvari koje nisu potrebne za korpus.

U prikupljenim je tekstovima trebalo provjeriti usporednost rečenica jer se moglo dogoditi da nisu sve rečenice istog redoslijeda u originalu i prijevodu, da je prijevod na zahtjev izdavača skraćen i slično. To se provjeravalo ručno u .doc dokumentima jer trenutno ne postoji alat za oba jezika koji bi to učinio automatski.

Zatim je tekst pretvoren u .txt dokumente kako bi se specificirao UNICODE UTF-8 format koji podržava sva pismena oba jezika, odnosno pisma, a u uređivaču teksta ručno je provjereno jesu li sve rečenice u odgovarajućim ulomcima jer alat za sravnjivanje radi na razini ulomaka i ne može prepoznati postoji li neka rečenica ako je slučajno u drugome ulomku (V. Prilog 1).

Kako ne postoji alat za sravnjivanje posebno prilagođen za oba jezika, na internetu je pronađen besplatno dostupan alat NOVA Text Aligner (http://www.supernova-soft.com/wpsite/products/text-aligner/) koji podržava UNICODE formate, a segmentaciju na rečenice radi putem interpunkcijskih znakova (V. Prilog 2). S obzirom da je već ručno provjeren poredak rečenica, nije bilo straha od prevelikih pogrešaka, samo su se trebale ručno provjeriti i ispraviti novonastale greške, uglavnom zbog netreniranja ovog alata na danim jezicima, pri čemu je svaka točka uzimana za kraj rečenice, što u hrvatskom jeziku nije uvijek slučaj jer se, primjerice, i redni brojevi pišu s točkom. Upravo je to u hrvatskom dijelu bio problem jer su datumi rastavljani na zasebne rečenice, što se trebalo naknadno ispraviti. U makedonskom dijelu to nije bio problem jer se, unatoč pisanju rednih brojeva s točkom, datumi pišu bez točke jer mjesec upućuje na znamenku za dan pa se podrazumijeva da je ona redni broj, a godine su ionako sve četveroznamenkaste pa se točka ne stavlja jer se zna o čemu se radi, osobito ako slijedi iza riječi za mjesec.

Druga greška pri segmentaciji javljala se kod tri točke koje ponekad nisu označavale granicu rečenice nego zastajanje u govoru pri upravom govoru u tekstu, što se, osim prema smislu teksta, moglo uočiti zbog malih slova sljedećeg dijela teksta. U tekstovima je bilo vrlo malo kratica s točkom, no i one su predstavljale problem, jer alat nije istreniran za dane jezike pa u njega nisu ubačene kratice kao standardni elementi s točkom.

S druge strane, segmentacija na rečenice nije izvršena kada rečenica završava navodnicima, što se često javljalo pri upravnom govoru u tekstovima, pa je sljedeća rečenica slijedila nakon nje, a ne kao zasebna rečenica. To se javljalo u oba jezika jer alat očito navodnike ne prepoznaje kao granicu rečenica, tako da se to moralo razdvojiti ručno.

Drugih grešaka nije bilo. Alat je iznenađujuće točno segmentirao rečenice, s obzirom da nije istreniran za makedonski i hrvatski jezik pa nije prepoznavao datume i kratice kao ustaljene izraze koji završavaju točkom a ne moraju označavati granicu rečenice.

Sravnjivanjem je završeno predviđeno označavanje korpusa jer su rečenice oba jezika međusobno uparene, tako da se pretraživanje može odvijati na razini rečenice, što znači da se prilikom pretrage određene riječi ili dijela riječi (poput korijena ili afiksa) u traženom jeziku prikazuje tražena riječ ili dio riječi, a u onom drugom jeziku prikazuje se cijela rečenica koja je uparena s onom iz traženog jezika gdje se tražena riječ nalazi. Riječi međusobno nisu povezane, no pretpostavka je da je korpus i u ovoj fazi upotrebljiv za razna istraživanja jer odmah nudi i kontekst na oba jezika, a istraživač koji će se njime služiti ionako bi trebao znati oba jezika pa bi lako mogao prepoznati traženu riječ i u prijevodu, ne samo u traženom jeziku. Pretraživanje se može napraviti za oba jezika, pa premda je u ovom korpusu makedonski originalni jezik, a hrvatski prijevodni, pri pretraživanju tekstovi korpusa mogu poslužiti i obratno, barem do neke mjere.

NOVA Text Aligner ima mogućnost izvoza teksta u TMX formatu.

Za svrhu izrade usporednog korpusa dva su najčešća standarda koja se primjenjuju, a koja smo već spomenuli u dijelu o podatcima u korpusu:

- TEI (*Text Encoding Initiative*);
- TMX (*Translation Memory Exchange*) i
- XCES (*XML Corpus Encoding Standard*) kao podskup TEI-ja.

Sva su tri standarda hijerarhijski orijentirana, od zaglavlja s metapodatcima o nazivu dokumenta, autoru, pravima, izdavaču, godini itd., preko razina za naslove, podnaslove, rečenice. Ukratko su prikazane prednosti i nedostatci s obzirom na izradu usporednog korpusa:

	TEI	**TMX**	**XCES**
Dokumentacija	Jako dobro dokumentirano	Dobro dokumentirano	Slabo dokumentirano
Jednostavnost pri kodiranju usporednog korpusa	Nije jednostavno	Jako jednostavno	Nije jednostavno
Mogućnost sravnjivanja na razini rečenica	Slaba mogućnost	Jako dobra mogućnost	Slaba mogućnost
Mogućnost sravnjivanja na razini riječi	Jako dobra mogućnost	Nije moguće	Jako dobra mogućnost

Dostupnost alata za izradu i održavanje	Dobra dostupnost	Jako dobra dostupnost	Dobra dostupnost
Dostupnost kodiranih korpusa	Dobra dostupnost	Jako dobra dostupnost	Loša dostupnost

Konačna odluka o tome koji standard koristiti ovisi o primjeni. Na primjer:

- Ako će se korpus koristiti za računalne programe prevođenja, TMX je najbolji odabir.
- Ako je potrebno sravniti nekoliko XML dokumenata na razini rečenice, praktičnije je koristiti TEI ili XCES jer se može stvoriti zasebna datoteka u kojoj se samo definiraju sravnjenja dijelova iz XML dokumenata.
- Pri izradi neoznačenog korpusa može se odabrati bilo koji od tri standarda, ali je TMX najjednostavniji.

Odluka također ovisi o dostupnosti alata i dokumentacije, kao što je istaknuto u tablici – tako primjerice TMX ima najbolju dostupnost alata i dokumentacije, TEI ima jako dobru dokumentaciju, ali slabije dostupne alate osobito za izradu usporednih korpusa, dok je XCES ponešto zapušten i nedostatak dobre dokumentacije ovaj standard čini trenutačno manje korisnim.

Kao što je već navedeno, odabran je TMX standard zbog svoje jednostavnosti, dostupnosti alata i dokumentacije te ispunjava sve potrebe za izradu neoznačenog usporednog korpusa (V. Prilog 3). Za nastavak obrade podataka korpusa i označavanje riječi, što je izvan dosega ovog rada, bit će potrebno korištenje nekog drugog standarda, TEI ili XCES standarda.

Programski jezik koji je odabran za izradu programskog rješenja za dohvaćanje i adaptaciju sravnjenih tekstova je Python (https://www.python.org/), koji je odabran zbog dostupnosti raznih biblioteka za obradu teksta i web poslužitelja, kao i opsežne dokumentacije.

Kako bi se podatci pretrage mogli vizualizirati i kako bi bili dostupni putem interneta, bilo je potrebno odabrati platformu koja podržava odabrani programski jezik, koja je jednostavna za upotrebu i besplatna. Odabrana je platforma Heroku (https://www.heroku.com/).

Pomoću Python biblioteka za izradu poslužitelja, upravljanje istim i posluživanje web formata izrađen je program koji kombinira prethodno navedene programe za čitanje TMX datoteka i konkordanciju kako bi podatke dobivene pretragom prikazao u prikladnom obliku. Dizajn stranice s kojom korisnik ima interakciju je s naglaskom na podatcima pa tako postoji polje za unos izraza, mogućnost odabira jezika koji se pretražuje, gumb za početak pretrage i dva vodoravno postavljena okna u kojima se prikazuju podatci pretrage (V. Prilog 4 i 5). Rezultati pretrage upareni su s drugim jezikom na razini rečenice kako bi osnovna ideja i cilj

ovog rada bili ostvareni – izrada neoznačenog sravnjenog makedonsko-hrvatskog usporednog korpusa.[1]

7. Zaključak

U ovom se radu dao sažet prikaz područja istraživanja računalne lingvistike kao lingvističke discipline te ju se smjestilo u odnos s drugim znanostima i (ne)lingvističkim disciplinama koje su joj srodne. Zatim se prikazala domena korpusne lingvistike te su definirane odlike korpusa, kao i njihova podjela s obzirom na vrstu, broj jezika, veličinu, namjenu i jezični modalitet. Potom su prikazani alati za obilježavanje korpusa, prema mogućem redoslijedu, kao i alat specifičan za obilježavanje usporednih korpusa.

Na samom kraju rada opisan je tijek sastavljanja i obilježavanja makedonsko-hrvatskog usporednog korpusa, koji je sastavljen kao eksperimentalni dio ovog diplomskog rada. Korpus je jednosmjeran, s književnim tekstovima na makedonskom jeziku i njihovim prijevodima na hrvatski jezik. Kako je ovo jedan od prvih pokušaja sastavljanja korpusa za makedonski jezik, većina je koraka u obradi ovoga korpusa učinjena ručno ili ručno provjerena, dok za hrvatski jezik postoji višemilijunski jednojezični Hrvatski nacionalni korpus, kao i niz višejezičnih usporednih korpusa čija je jedna od komponenti i hrvatski jezik. U korpusu se zasad provelo samo sravnjivanje rečenica, a razvijen je softver za pretragu pismena na oba jezika.

Daljnja obrada korpusa uključit će označavanje vrsta riječi prema MTE oznakama i razvitak softvera za pretragu MSD oznaka i lema, na oba jezika, kao i moguće proširenje korpusa u vidu dodavanja hrvatsko-makedonskog *potkorpusa* i proširivanja makedonsko-hrvatskog korpusa, kako novim književnim djelima, tako i drugim žanrovima, poput novinskih i administrativnih tekstova, do kojih se pri sastavljanju ovog korpusa nije moglo doći.

[1] Autor navedenih programskih rješenja, Ilija Ćosić, zadržava sva autorska prava i izričito zabranjuje dijeljenje, dupliciranje, izmjenu ili korištenje programskih rješenja bez prethodnog dopuštenja.

8. Literatura

1. Agić. Ž. (2012). *Pristupi ovisnom parsanju hrvatskih tekstova.* Zagreb: Filozofski fakultet Sveučilišta u Zagrebu. http://darhiv.ffzg.unizg.hr/2337/ (Pristupljeno 6. kolovoza 2015.)

2. Bekavac. B. (2001). *Primjena računalnojezikoslovnih alata na hrvatske korpuse*. Zagreb: Filozofski fakultet Sveučilišta u Zagrebu. http://darhiv.ffzg.unizg.hr/2360/ (Pristupljeno 6. kolovoza 2015.)

3. Biber. C., Condrad. S., Reppen. R. (2000). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press

4. Bolshakov. I., Gelbukh. A. (2004). *Computational Linguistics: Models. Resources. Applications*. Mexico: Instituto Politécnico nacional

5. Brown. K., Miller. J. (2013). *The Cambridge Dictionary of Linguistics*. Cambridge: Cambridge University Press

6. Clark. A., Fox. C., Lappin. S. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. Chichester, West Sussex: Wiley-Blackwell.

7. Gazdar. G., Mellish. C. (1989). *Natural Language Processing in Prolog: An Introduction to Computational Linguistics*. Wokingham: Addison-Wesley Publishing Company

8. Jurafsky. D., Martin. J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics. and Speech Recognition*. New Jersey: Prentice Hall

9. Kennedy. G. (1998). *An Introduction to Corpus Linguistics*. London: Longman

10. Lawler. J. M., Dry. H. A. (1998). *Using Computers in Linguistics: A Practical Guide*. London: Routledge

11. Li, P. i dr. (2011). *A Starter's Guide to Linguistic Corpora Building*. Taipei: Taiwan e-Learning and Digital Archives Program. Taiwan Digital Archives Expansion Project

12. McEnery. T., Hardie. A. (2012). *Corpus Linguistics*. Cambridge: Cambridge University Press

13. McEnery. T., Wilson. A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press

14. Mitkov. R. (2003). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press

15. O'Keeffe. A., McCarthy. M., Carter. R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press

16. Schäfer. R., Bildhauer. F. (2013). *Web Corpus Construction*. San Rafael. CA: Morgan and Claypool Publishers

17. Sinclair. J. (1991). *Corpus. Concordance. Collocation*. Oxford: Oxford University Press

18. Stabler. E. (2003). *Notes on Computational Linguistics*. http://www.linguistics.ucla.edu/people/stabler/185-03.pdf (Pristupljeno 16. srpnja 2015.)

19. Tadić. M. (1996). *Računalna obradba hrvatskoga i nacionalni korpus*. U: Suvremena lingvistika. br. 41/42. str. 603-611. http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=38525 (Pristupljeno 2. svibnja 2015)

20. Tadić, M. (1997). *Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive.* U: Suvremena lingvistika, br. 43/44. str. 387-394. http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=37485 (Pristupljeno 2. svibnja 2015.)

21. Tadić. M. (2003). *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex libris

22. Tadić, Marko (2009) *New version of the Croatian National Corpus*. U: Hlaváčková, Dana ; Horák, Aleš ; Osolsobě, Klara ; Rychlý, Pavel (ur.) After Half a Century of Slavonic Natural Language Processing. Masaryk University, Brno, str. 199-205.

23. Tognini-Bonelli. E. (2001).. *Corpus Linguistics at Work*. Amsterdam: J. Benjamins

24. Vučković. K. (2009). *Model parsera za hrvatski jezik.* Zagreb: Filozofski fakultet Sveučilišta u Zagrebu. http://darhiv.ffzg.unizg.hr/2343/ (Pristupljeno 6. kolovoza 2015.)

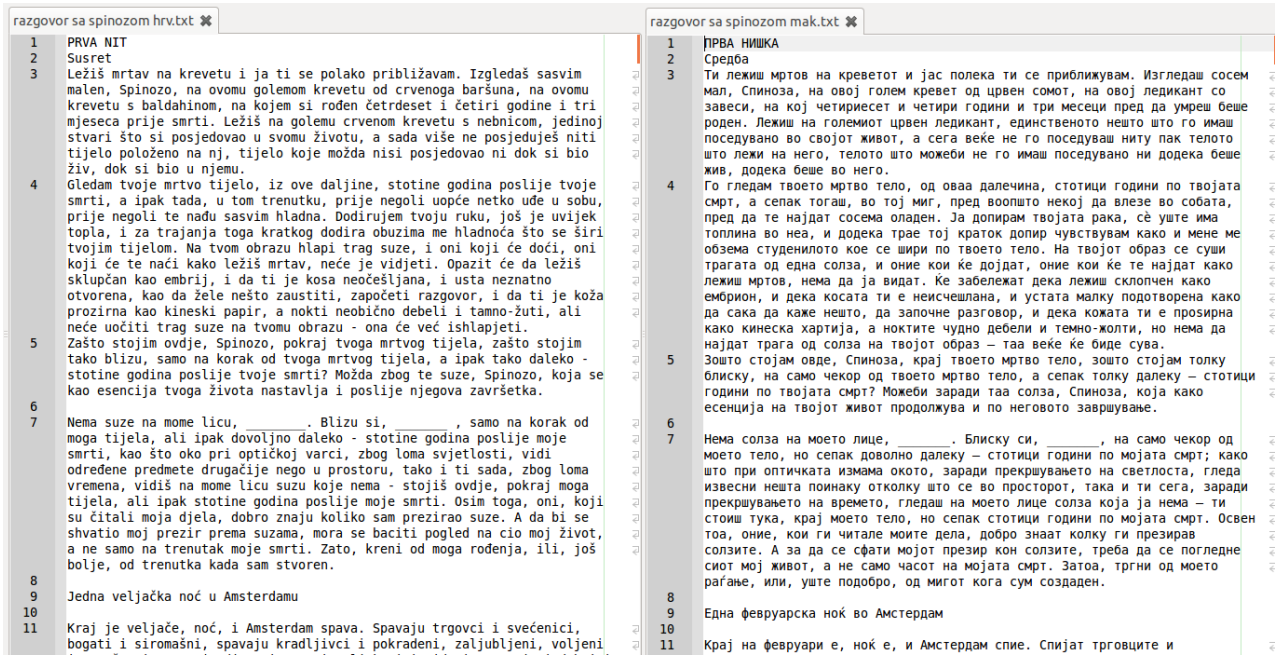
25. Dublin Core Metadata Element Set. Version 1.1. (n.d.). http://dublincore.org/documents/dces (Pristupljeno 7. kolovoza 2015.)

26. Browse the EAGLES Guidelines. (n.d.). http://www.ilc.cnr.it/EAGLES/browse.html (Pristupljeno 8. kolovoza 2015.)

27. Supernova Software (Supernova Software). http://www.supernova-soft.com/wpsite/products/text-aligner/ (Pristupljeno 10. rujna 2015.)

28. Welcome to Python.org (Python.org). https://www.python.org/ (Pristupljeno 16. rujna 2015.)

9. Prilozi



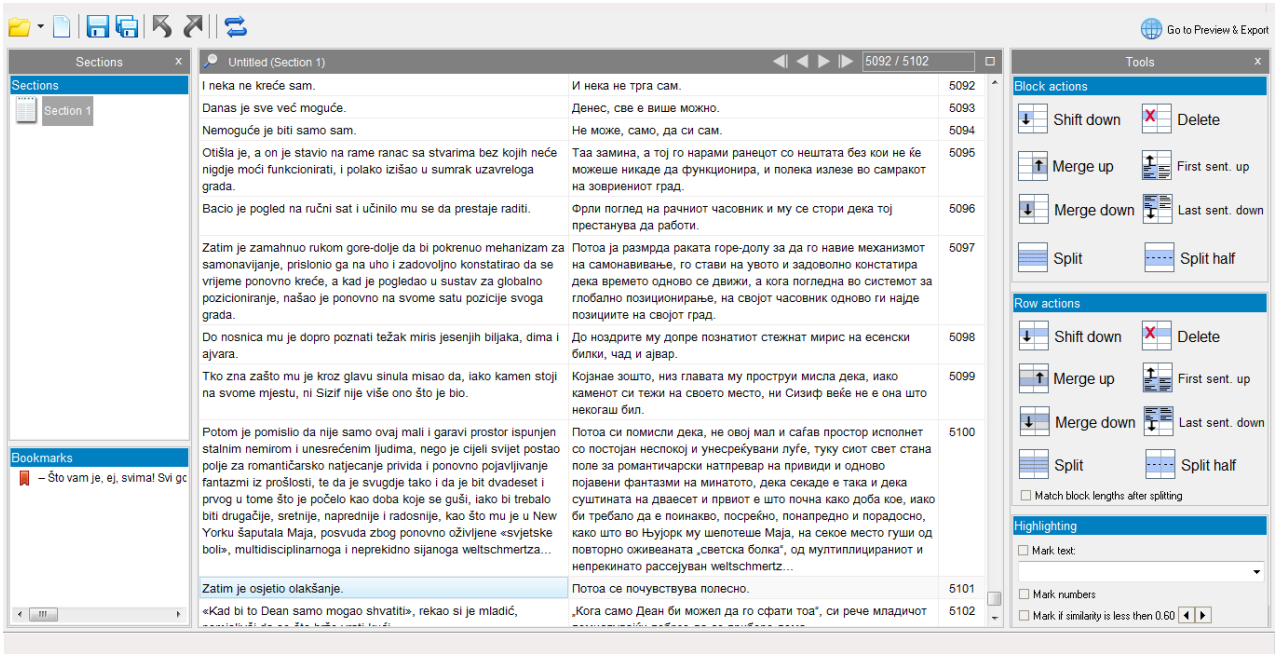
razgovor sa spinozom hrv.txt

```
1  PRVA NIT
2  Susret
3  Ležiš mrtav na krevetu i ja ti se polako približavam. Izgledaš sasvim
   malen, Spinozo, na ovomu golemom krevetu od crvenoga baršuna, na ovomu
   krevetu s baldahinom, na kojem si rođen četrdeset i četiri godine i tri
   mjeseca prije smrti. Ležiš na golemu crvenom krevetu s nebnicom, jedinoj
   stvari što si posjedovao u svomu životu, a sada više ne posjeduješ niti
   tijelo položeno na nj, tijelo koje možda nisi posjedovao ni dok si bio
   živ, dok si bio u njemu.
4  Gledam tvoje mrtvo tijelo, iz ove daljine, stotine godina poslije tvoje
   smrti, a ipak tada, u tom trenutku, prije negoli uopće netko ude u sobu,
   prije negoli te nađu sasvim hladna. Dodirujem tvoju ruku, još je uvijek
   topla, i za trajanja toga kratkog dodira obuzima me hladnoća što se širi
   tvojim tijelom. Na tvom obrazu hlapi trag suze, i oni koji će doći, oni
   koji će te naći kako ležiš mrtav, neće je vidjeti. Opazit će da ležiš
   sklupčan kao embrij, i da ti je kosa neočešljana, i usta neznatno
   otvorena, kao da žele nešto zaustiti, započeti razgovor, i da ti je koža
   prozirna kao kineski papir, a nokti neobično debeli i tamno-žuti, ali
   neće uočiti trag suze na tvomu obrazu - ona će već ishlapjeti.
5  Zašto stojim ovdje, Spinozo, pokraj tvoga mrtvog tijela, zašto stojim
   tako blizu, samo na korak od tvoga mrtvog tijela, a ipak tako daleko -
   stotine godina poslije tvoje smrti? Možda zbog te suze, Spinozo, koja se
   kao esencija tvoga života nastavlja i poslije njegova završetka.
6
7  Nema suze na mome licu, _______. Blizu si, _______ , samo na korak od
   moga tijela, ali ipak dovoljno daleko - stotine godina poslije moje
   smrti, kao što oko pri optičkoj varci, zbog loma svjetlosti, vidi
   određene predmete drugačije nego u prostoru, tako i ti sada, zbog loma
   vremena, vidiš na mome licu suzu koje nema - stojiš ovdje, pokraj moga
   tijela, ali ipak stotine godina poslije moje smrti. Osim toga, oni, koji
   su čitali moja djela, dobro znaju koliko sam prezirao suze. A da bi se
   shvatio moj prezir prema suzama, mora se baciti pogled na cio moj život,
   a ne samo na trenutak moje smrti. Zato, kreni od moga rođenja, ili, još
   bolje, od trenutka kada sam stvoren.
8
9  Jedna veljačka noć u Amsterdamu
10
11 Kraj je veljače, noć, i Amsterdam spava. Spavaju trgovci i svećenici,
   bogati i siromašni, spavaju kradljivci i pokradeni, zaljubljeni, voljeni
```

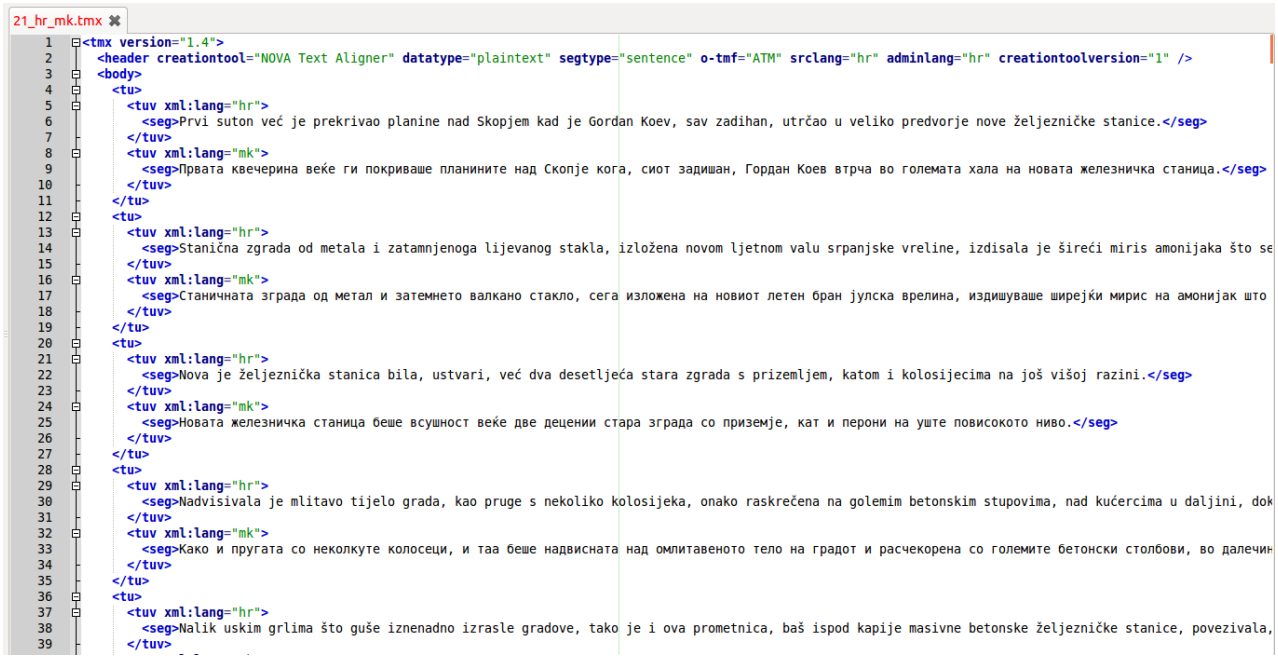
razgovor sa spinozom mak.txt

```
1  ПРВА НИШКА
2  Средба
3  Ти лежиш мртов на креветот и јас полека ти се приближувам. Изгледаш сосем
   мал, Спиноза, на овој голем кревет од црвен сомот, на овој ледикант со
   завеси, на кој четириесет и четири години и три месеци пред да умреш беше
   роден. Лежиш на големиот црвен ледикант, единственото нешто што го имаш
   поседувано во својот живот, а сега веќе не го поседуваш ниту пак телото
   што лежи на него, телото што можеби не го имаш поседувано ни додека беше
   жив, додека беше во него.
4  Го гледам твоето мртво тело, од оваа далечина, стотици години по твојата
   смрт, а сепак тогаш, во тој миг, пред воопшто некој да влезе во собата,
   пред да те најдат сосема оладен. Ја допирам твојата рака, сè уште има
   топлина во неа, и додека трае тој краток допир чувствувам како и мене ме
   обзема студенилото кое се шири по твоето тело. На твојот образ се суши
   трагата од една солза, и оние кои ќе дојдат, оние кои ќе те најдат како
   лежиш мртов, нема да ја видат. Ќе забележат дека лежиш склопчен како
   ембрион, и дека коса ти е неисчешлана, и устата малку подотворена како
   да сака да каже нешто, да започне разговор, и дека кожата ти е прозирна
   како кинеска хартија, а ноктите чудно дебели и темно-жолти, но нема да
   најдат трага од солза на твојот образ – таа веќе ќе биде сува.
5  Зошто стојам овде, Спиноза, крај твоето мртво тело, зошто стојам толку
   блиску, на само чекор од твоето мртво тело, а сепак толку далеку – стотици
   години по твојата смрт? Можеби заради таа солза, Спиноза, која како
   есенција на твојот живот продолжува и по неговото завршување.
6
7  Нема солза на моето лице, _______. Блиску си, _______, на само чекор од
   моето тело, но сепак доволно далеку – стотици години по мојата смрт; како
   што при оптичката измама окото, заради прекршувањето на светлоста, гледа
   извесни нешта поинаку отколку што се во просторот, така и ти сега, заради
   прекршувањето на времето, гледаш на моето лице солза која ја нема – ти
   стоиш тука, крај моето тело, но сепак стотици години по мојата смрт. Освен
   тоа, оние, кои ги читале моите дела, добро знаат колку ги презирав
   солзите. А за да се сфати мојот презир кон солзите, треба да се погледне
   сиот мој живот, а не само часот на мојата смрт. Затоа, тргни од моето
   раѓање, или, уште подобро, од мигот кога сум создаден.
8
9  Една февруарска ноќ во Амстердам
10
11 Крај на февруари е, ноќ е, и Амстердам спие. Спијат трговците и
```

Prilog 1. – Prikaz sravnjenih ulomaka za roman *Razgovor sa Spinozom* gdje svaki broj označava novi ulomak.

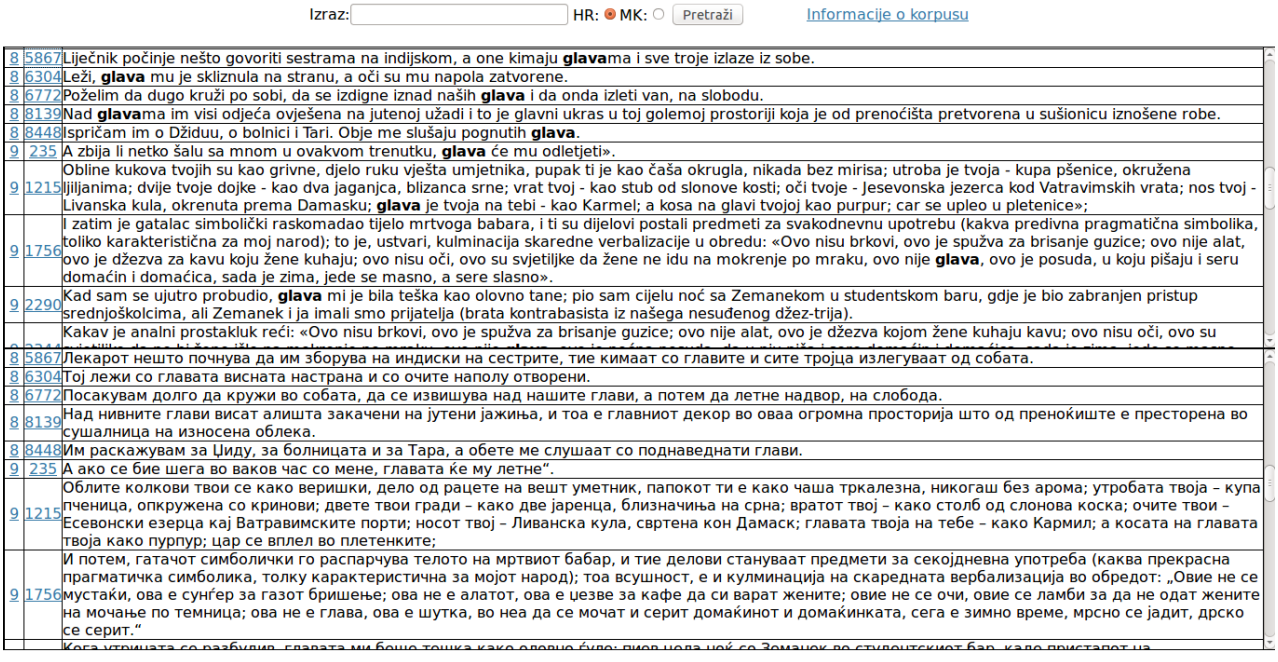


Prilog 2. – Prikaz sravnjenih rečenica u NOVA Text Aligneru za kraj romana *21*.



```
21_hr_mk.tmx
<tmx version="1.4">
  <header creationtool="NOVA Text Aligner" datatype="plaintext" segtype="sentence" o-tmf="ATM" srclang="hr" adminlang="hr" creationtoolversion="1" />
  <body>
    <tu>
      <tuv xml:lang="hr">
        <seg>Prvi suton već je prekrivao planine nad Skopjem kad je Gordan Koev, sav zadihan, utrčao u veliko predvorje nove željezničke stanice.</seg>
      </tuv>
      <tuv xml:lang="mk">
        <seg>Првата квечерина веќе ги покриваше планините над Скопје кога, сиот задишан, Гордан Коев втрча во големата хала на новата железничка станица.</seg>
      </tuv>
    </tu>
    <tu>
      <tuv xml:lang="hr">
        <seg>Stanična zgrada od metala i zatamnjenoga lijevanog stakla, izložena novom ljetnom valu srpanjske vreline, izdisala je širећi miris amonijaka što se
      </tuv>
      <tuv xml:lang="mk">
        <seg>Станичната зграда од метал и затемнето валкано стакло, сега изложена на новиот летен бран јулска врелина, издишуваше ширејќи мирис на амонијак што
      </tuv>
    </tu>
    <tu>
      <tuv xml:lang="hr">
        <seg>Nova je željeznička stanica bila, ustvari, već dva desetljeća stara zgrada s prizemljem, katom i kolosijecima na još višoj razini.</seg>
      </tuv>
      <tuv xml:lang="mk">
        <seg>Новата железничка станица беше всушност веќе две децении стара зграда со приземје, кат и перони на уште повисокото ниво.</seg>
      </tuv>
    </tu>
    <tu>
      <tuv xml:lang="hr">
        <seg>Nadvisivala je mlitavo tijelo grada, kao pruge s nekoliko kolosijeka, onako raskrečena na golemim betonskim stupovima, nad kućercima u daljini, dok
      </tuv>
      <tuv xml:lang="mk">
        <seg>Како и пругата со неколкуте колосеци, и таа беше надвисната над омлитавеното тело на градот и расчекорена со големите бетонски столбови, во далечин
      </tuv>
    </tu>
    <tu>
      <tuv xml:lang="hr">
        <seg>Nalik uskim grlima što guše iznenadno izrasle gradove, tako je i ova prometnica, baš ispod kapije masivne betonske željezničke stanice, povezivala,
      </tuv>
```

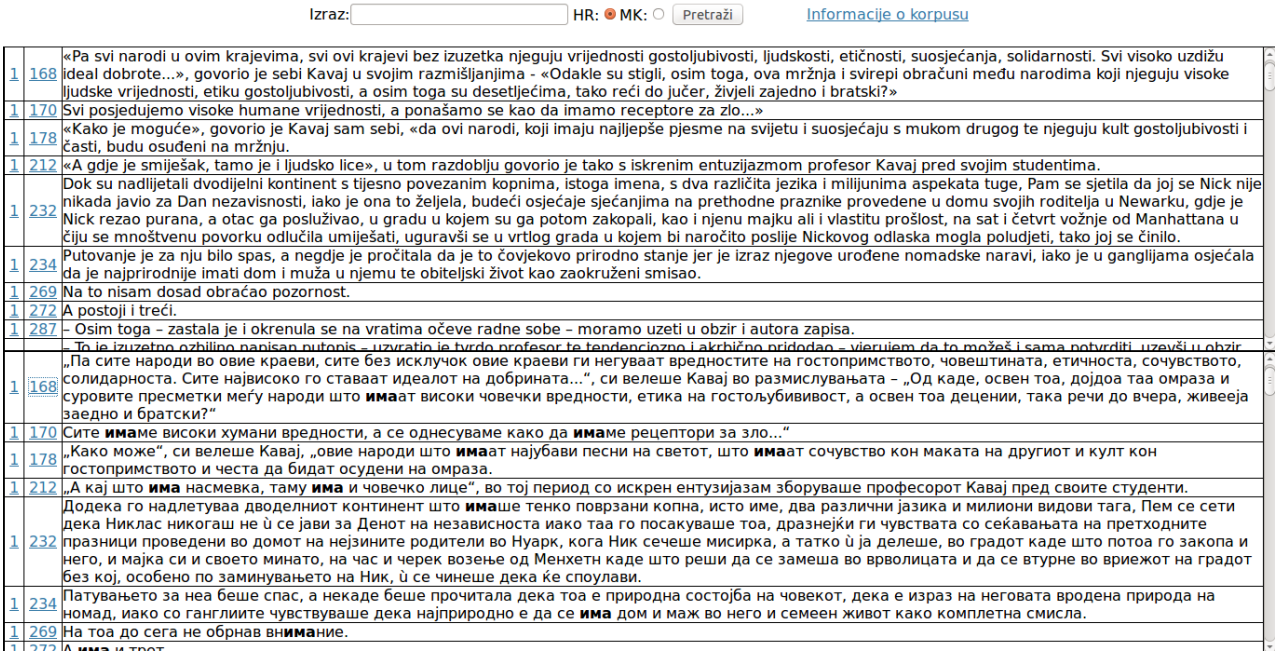
Prilog 3. – Prikaz TMX formata za početak romana *21*. Zatamnjene su makedonska originalna rečenica i njezin sravnjen prijevod na hrvatski.



Izraz: HR: MK: Pretraži Informacije o korpusu

8	5867	Liječnik počinje nešto govoriti sestrama na indijskom, a one kimaju **glava**ma i sve troje izlaze iz sobe.
8	6304	Leži, **glava** mu je skliznula na stranu, a oči su mu napola zatvorene.
8	6772	Poželim da dugo kruži po sobi, da se izdigne iznad naših **glava** i da onda izleti van, na slobodu.
8	8139	Nad **glava**ma im visi odjeća ovješena na jutenoj užadi i to je glavni ukras u toj golemoj prostoriji koja je od prenoćišta pretvorena u sušionicu iznošene robe.
8	8448	Ispričam im o Džiduu, o bolnici i Tari. Obje me slušaju pognutih **glava**.
9	235	A zbija li netko šalu sa mnom u ovakvom trenutku, **glava** će mu odletjeti».
9	1215	Obline kukova tvojih su kao grivne, djelo ruku vješta umjetnika, pupak ti je kao čaša okrugla, nikada bez mirisa; utroba je tvoja - kupa pšenice, okružena ljiljanima; dvije tvoje dojke - kao dva jaganjca, blizanca srne; vrat tvoj - kao stub od slonove kosti; oči tvoje - Jesevonska jezerca kod Vatravimskih vrata; nos tvoj - Livanska kula, okrenuta prema Damasku; **glava** je tvoja na tebi - kao Karmel; a kosa na glavi tvojoj kao purpur; car se upleo u pletenice»;
9	1756	I zatim je gatalac simbolički raskomadao tijelo mrtvoga babara, i ti su dijelovi postali predmeti za svakodnevnu upotrebu (kakva predivna pragmatična simbolika, toliko karakteristična za moj narod); to je, ustvari, kulminacija skaredne verbalizacije u obredu: «Ovo nisu brkovi, ovo je spužva za brisanje guzice; ovo nije alat, ovo je džezva za kavu koju žene kuhaju; ovo nisu oči, ovo su svjetiljke da žene ne idu na mokrenje po mraku, ovo nije **glava**, ovo je posuda, u koju pišaju i seru domaćin i domaćica, sada je zima, jede se masno, a sere slasno».
9	2290	Kad sam se ujutro probudio, **glava** mi je bila teška kao olovno tane; pio sam cijelu noć sa Zemanekom u studentskom baru, gdje je bio zabranjen pristup srednjoškolcima, ali Zemanek i ja imali smo prijatelja (brata kontrabasista iz našega nesuđenog džez-trija).
		Kakav je analni prostakluk rekli: «Ovo nisu brkovi, ovo je spužva za brisanje guzice; ovo nije alat, ovo je džezva kojom žene kuhaju kavu; ovo nisu oči, ovo su
8	5867	Лекарот нешто почнува да им зборува на индиски на сестрите, тие кимаат со главите и сите тројца излегуваат од собата.
8	6304	Тој лежи со главата висната настрана и со очите наполу отворени.
8	6772	Посакувам долго да кружи во собата, да се извишува над нашите глави, а потем да летне надвор, на слобода.
8	8139	Над нивните глави висат алишта закачени на јутени јажиња, и тоа е главниот декор во оваа огромна просторија што од преноќиште е претворена во сушалница на износена облека.
8	8448	Им раскажувам за Џиду, за болницата и за Тара, а обете ме слушаат со поднаведнати глави.
9	235	А ако се бие шега во ваков час со мене, главата ќе му летне“.
9	1215	Облите колкови твои се како веришки, дело од рацете на вешт уметник, папокот ти е како чаша тркалезна, никогаш без арома; утробата твоја – купа пченица, опкружена со кринови; двете твои гради – како две јаренца, близначиња на срна; вратот твој – како столб од слонова коска; очите твои – Есевонски езерца кај Ватравимските порти; носот твој – Ливанска кула, свртена кон Дамаск; главата твоја на тебе – како Кармил; а косата на главата твоја како пурпур; цар се вплел во плетенките;
9	1756	И потем, гатачот симболички го распарчува телото на мртвиот бабар, и тие делови стануваат предмети за секојдневна употреба (каква прекрасна прагматичка симболика, толку карактеристична за мојот народ); тоа всушност, е и кулминација на скаредната вербализација во обредот: „Овие не се мустаќи, ова е сунѓер за газот бришење; ова не е алатот, ова е џезве за кафе да си варат жените; овие не се очи, овие се ламби за да не одат жените на мочање по темница; ова не е глава, ова е шутка, во неа да се мочат и серит домаќинот и домаќинката, сега е зимно време, мрсно се јадит, дрско се серит.“
		Кога утринава се разбудив, главата ми беше тешка како оловно ѓуле: пиев цела ноќ со Земанек во студентскиот бар, кадe пристапот на

Prilog 4. – Pretraga „glava“ na hrvatskom jeziku – s obzirom da korpus nije POS označen, prilikom pretrage softver promatra skup pismena pa se za „glava“ daju rezultati poput „glavama“ i sl. U donjem dijelu prikazane su odgovarajuće rečenice na makedonskom jeziku, a prema brojevima u drugom stupcu slijeva prikazani su parovi rečenica. Prvi stupac slijeva označava tekstove u kojima se prikazane rečenice nalaze.


Izraz: HR: MK: Pretraži Informacije o korpusu

1	168	«Pa svi narodi u ovim krajevima, svi ovi krajevi bez izuzetka njeguju vrijednosti gostoljubivosti, ljudskosti, etičnosti, suosjećanja, solidarnosti. Svi visoko uzdižu ideal dobrote...», govorio je sebi Kavaj u svojim razmišljanjima - «Odakle su stigli, osim toga, ova mržnja i svirepi obračuni među narodima koji njeguju visoke ljudske vrijednosti, etiku gostoljubivosti, a osim toga su desetljećima, tako reći do jučer, živjeli zajedno i bratski?»
1	170	Svi posjedujemo visoke humane vrijednosti, a ponašamo se kao da imamo receptore za zlo...»
1	178	«Kako je moguće», govorio je Kavaj sam sebi, «da ovi narodi, koji imaju najljepše pjesme na svijetu i suosjećaju s mukom drugog te njeguju kult gostoljubivosti i časti, budu osuđeni na mržnju.
1	212	«A gdje je smiješak, tamo je i ljudsko lice», u tom razdoblju govorio je tako s iskrenim entuzijazmom profesor Kavaj pred svojim studentima.
1	232	Dok su nadlijetali dvodijelni kontinent s tijesno povezanim kopnima, istoga imena, s dva različita jezika i milijunima aspekata tuge, Pam se sjetila da joj se Nick nije nikada javio za Dan nezavisnosti, iako je ona to željela, budeći osjećaje sjećanjima na prethodne praznike provedene u domu svojih roditelja u Newarku, gdje je Nick rezao purana, a otac ga posluživao, u gradu u kojem su ga potom zakopali, kao i njenu majku ali i vlastitu prošlost, na sat i četvrt vožnje od Manhattana u čiju se mnoštvenu povorku odlučila umiješati, uguravši se u vrtlog grada u kojem bi naročito poslije Nickovog odlaska mogla poludjeti, tako joj se činilo.
1	234	Putovanje je za nju bilo spas, a negdje je pročitala da je to čovjekovo prirodno stanje jer je izraz njegove urođene nomadske naravi, iako je u ganglijama osjećala da je najprirodnije imati dom i muža u njemu te obiteljski život kao zaokruženi smisao.
1	269	Na to nisam dosad obraćao pozornost.
1	272	A postoji i treći.
1	287	– Osim toga – zastala je i okrenula se na vratima očeve radne sobe – moramo uzeti u obzir i autora zapisa.
		– To je izuzetno ozbiljno napisan putopis – uzvratio je tvrdo profesor te tendenciozno i akribično pridodao – vjerujem da to možeš i sama potvrditi, uzevši u obzir
1	168	„Па сите народи во овие краеви, сите без исклучок овие краеви ги негуваат вредностите на гостопримството, човештината, етичноста, сочувството, солидарноста. Сите највисоко го стават идеалот на добрината...“, си велеше Кавај во размислувањата – „Од каде, освен тоа, дојдоа таа омраза и суровите пресметки меѓу народи што **има**ат високи човечки вредности, етика на гостољубививост, а освен тоа децении, така речи до вчера, живееја заедно и братски?“
1	170	Сите **има**ме високи хумани вредности, а се однесуваме како да **има**ме рецептори за зло...“
1	178	„Како може“, си велеше Кавај, „овие народи што **има**ат најубави песни на светот, што **има**ат сочувство кон маката на другиот и култ кон гостопримството и честа да бидат осудени на омраза.
1	212	„А кај што **има** насмевка, таму **има** и човечко лице“, во тој период со искрен ентузијазам зборуваше професорот Кавај пред своите студенти.
1	232	Додека го надлетуваа дводелниот континент што **има**ше тенко поврзани копна, исто име, два различни јазика и милиони видови тага, Пем се сети дека Никлас никогаш не ѝ се јави за Денот на независноста иако таа го посакуваше тоа, дразнејќи ги чувствата со сеќавањата на претходните празници проведени во домот на нејзините родители во Нуарк, кога Ник сечеше мисирка, а татко ѝ ја делеше, во градот каде што потоа го закопа и него, и мајка си и своето минато, на час и черек возење од Менхетн каде што реши да се замеша во врволицата и да се втурне во вриежот на градот без кој, особено по заминувањето на Ник, ѝ се чинеше дека ќе споулави.
1	234	Патувањето за неа беше спас, а некаде беше прочитала дека тоа е природна состојба на човекот, дека е израз на неговата вродена природа на номад, иако со гангliите чувствуваше дека најприродно е да се **има** дом и маж во него и семеен живот како комплетна смисла.
1	269	На тоа до сега не обрнав вн**има**ние.
1	272	А **има** и трет

Prilog 5. – Pretraga „има“ na makedonskom jeziku – bez označavanja vrsta riječi daje rezultate sa svim riječima koje sadrže „има“ poput „внимание“ i „плима“. Ovdje su prikazani rezultati gdje je „има“ dio paradigme glagola „има“ na makedonskom jeziku, zato su sada donji rezultati podebljani, a gore se, u prozoru za hrvatski jezik, prikazuju odgovarajuće rečenice za makedonske rezultate.