Univerzitet u Beogradu

Akademske Master studije

Računarstvo u društvenim naukama

# Programiranje u lingvistici

# MASTER RAD

Naziv teme: Razvoj korpusa tekstova prevedenih sa japanskog jezika na srpski i engleski jezik

Mentor:

Prof. dr Ranka Stanković

Student:

Vladimir Parezanović

202/2015

Beograd,

Jul, 2018.

# Sadržaj

**Zahvalnica**

Pisanje ovog master rada ne bi bilo moguće bez moje mentorke, profesorke doktorke Ranke Stanković, koja mi je pomogla u formiranju mog projekta, pružala mi podršku i savete o pisanju rada i stručno me odvela do završetka mojih studija.

Međutim, do ovde ne bih stigao da nije mojih roditelja koji su me emotivno i finansijski podržali u ovim studijama. Reč zahvalnosti takođe dobijaju moje kolege i prijatelji koji su mi pružali podršku i savete.

V.P. U Beogradu, jul 2018.

**Sažetak**

Cilj teze „Razvoj korpusa tekstova prevedenih sa japanskog jezika na srpski i engleski jezik“ je analiza prevoda starih japanskih tekstova na srpski i engleski jezik, komparativna analiza kreiranih korpusa međusobno, uz frekvencijska poređenja sa korpusom savremenog srpskog jezika. Dodatno, dobijeni rezultati pomenute analize predstavljaju osnovu za poređenje prevoda sa japanskog na srpski i engleski jezik.

Praktičan rezultat teze su analiza korpusa sa predlogom reči za dopunu elektronskih rečnika srpskog jezika ne(pre)poznatim rečima iz korpusa prevoda sa japanskog na srpski jezik, statistička analiza vrsta reči i semantičko grupisanje korpusnih tekstova sa različitim nainima vizuelizacije rezultata. Sažeti grafički prikazu prikazuju analizu korpusa kao celine i njegovih delova, odnosno pojedinačnih tekstova ili pesama korišćenjem oblaka reči i oblaka drveta, potom u okviru stilometrijske analize drveta klastera i glavnih komponenti. Dodatno, slična analiza je urađena sa prevodima sa japanskog na engleski jezik.

Predmet istraživanja ovog master rada su korpusi prevoda japanskog teksta na srpski i engleski jezik, a korpus srpske poezije se koristi prilikom poređenja sa korpusom prevoda japanske poezije na srpski jezik.

U ovom master radu se istražuju prevodi japanskih tekstova na srpski i engleski jezik koristeći alat za korpusnu analizu *Unitex* i *Leximir* sa elektronskim rečnicima za srpski jezik, i razvojno okruženje RStudio sa jezikom R i paketom *Stylo.* Takođe, u ovom istraživanju su iz korpusa prevoda *haiku* poezije na srpski izdvojene reči i izrazi koji označavaju godišnja doba.

Rezultati koji su dobijeni pokazuju poklapanja po pitanju procentualnog odnosa vrsta reči i znakova interpunkcije u korpusima prevoda japanskog teksta na srpski i srpske poezije, a odudaranja u korpusima prevoda japanskog teksta na srpski i engleski jezik po istim kriterijumima.

Iz toga se može izvući zaključak da prevodi japanske *haiku* poezije na srpski jezik zadržavaju procentualno sličan odnos vrsta reči i znakova interpunkcije kao u srpskoj poeziji. Pored toga, zaključuje se da je rečitost originalnog teksta zadržana, a izdvojeno je 29 reči kojima treba dopuniti elektronski rečnik srpskog jezika.

**Ključne reči**: *haiku*, korpus, *kigo*, *Unitex*, R

**Abstract**

The objective of “The Development of corpora of texts translated from Japanese to Serbian and English” is the analysis of the translations of old Japanese texts to Serbian and English, mutual comparative analysis of the created corpora, as well as an analysis with a corpus of contemporary Serbian language. Furthermore, the results gained from the already mentioned analysis would be used for the comparison with the Japanese to English translations by being mutually compared.

The practical task of the thesis are corpus analysis complemented with candidates for electronic dictionary extension with unknown words from the corpus of Japanese to Serbian text translations, statistical analysis of part of speach word classes and semantic clastering of corpus texts, various types of visualizations of corpora (or parts of corpora) by using TreeCloud and TagCloud, as well as classter and principal component analysis within stylometry. Furthermore, a similar analysis is made on translations from Japanese to English.

The research subject of this master thesis are the corpora of translations from Japanese to Serbian and English, and a corpus of Serbian poetry is used for comparison with the corpus of the translation of Japanese poetry to Serbian.

This master thesis comprises research on translations of Japanese texts to Serbian and English using *Unitex* corpus analysis tool and *Leximir* with Serbian electonic morphological dictionaries, and development environment *RStudio* with language *R* and package *stylo*. Also, this thesis outlined words and expressions symbolizing the four seasons from the corpora of the Serbian translation of the haiku poetry.

The results gained show compliance (matching) according to the percentual proportions of word classes and punctuation marks between the corpora of translations of Japanese texts to Serbian and Serbian poetry, and mismatches between the corpora of translations of Japanese to Serbian and that of English by the same criteria.

From that can be concluded that translations of Japanese *haiku* poetry to Serbian keep a percentually similar proportions of classes of words and punctuation marks as in Serbian poetry. Furthermore, it is concluded that eloquence of the original text has been preserved and 29 candidate words have been suggested for expansion of the electronic dictionary of Serbian language.

**Keywords**: *haiku*, corpus, *kigo*, *Unitex*, R

**Radna biografija**

Vladimir Parezanović je rođen 19. avgusta 1992. godine u Paraćinu. Osnovnu školu „Stevan Jakovljević“ i gimnaziju završio je u Paraćinu. Diplomirao je na Filološkom fakultetu na katedri za orijentalistiku, profil japanski jezik, književnost i kultura, sa prosečnom ocenom 8,72.

Tokom studija je bio aktivan član udruženja srpsko-japanskog prijateljstva „Tagai“, koje se bavilo tradicionalnom japanskom kulturom, kao i savremenim jezikom i književnošću i asistent predsednice udruženja do 2013. godine. Postao je šegrt japanske čajne ceremonije 2011. godine.

Imao je ključnu ulogu u osnivanju udruženja građana za očuvanje starih zanata pod nazivom „Angeleon“, gde je uređivao i prevodio dokumenta i učestvovao u organizaciji događaja do juna 2017. godine.

**Izjava o akademskoj čestitosti**

Student/kinja: Vladimir Parezanović

Broj indeksa: 202/2015

Student/kinja: master akademskih studija Računarstvo u društvenim naukama

Autor maste rada pod nazivom:

Razvoj korpusa tekstova prevedenih sa japanskog jezika na srpski i engleski jezik

Potpisivanjem izjavljujem:

- da je rad isključivo rezultat mog sopstvenog istraživačkog rada;
- da sam rad i mišljenja drugih autora koje sam koristio/la u ovom radu naznačio/la ili citirao/la u skladu sa Uputstvom;
- da su svi radovi i mišljenja drugih autora navedeni u spisku literature/referenci koji su sastavni deo ovog rada i pisani u skladu sa Uputstvom;
- da sam dobio/la sve dozvole za korišćenje autorskog dela koji se upotpunosti/celosti unose u predati rad i da sam to jasno naveo/la;
- da sam svestan/na da je plagijat korišćenje tuđih radova u bilo kom obliku (kao citata, prafraza, slika, tabela, dijagrama, dizajna, planova, fotografija, filma, muzike,formula, veb sajtova, kompjuterskih programa i sl.) bez navođenja autora ili predstavljanje tuđih autorskih dela kao mojih, kažnjivo po zakonu (Zakon o autorskom i srodnim pravima, Službeni glasnik Republike Srbije, br. 104/2009,99/2011, 119/2012), kao i drugih zakona i odgovarajućih akata Univerziteta u Beogradu;
- da sam da sam svestan/na da plagijat uključuje i predstavljanje, upotrebu i distribuiranje rada predavača ili drugih studenata kao sopstvenih;
- da sam svestan/na posledica koje kod dokazanog plagijata mogu prouzrokovati napredati master rad i moj status;
- da je elektronska verzija master rada identična štampanom primerku i pristajem na njegovo objavljivanje pod uslovima propisanim aktima Univerziteta.

Beograd, _____ Potpis studenta/nje

_____

# 1 Uvod

## 1.1 Cilj rada

Teza „Razvoj korpusa tekstova prevedenih sa japanskog jezika na srpski i engleski jezik“ ima definisana dva cilja: teorijski i praktični cilj.

Teorijski cilj teze: Sprovesti analizu postojećih resursa, odnosno prevoda starih japanskih tekstova na srpski i japanski jezik, uraditi komparativnu analizu kreiranih korpusa međusobno, i sa korpusom savremenog srpskog jezika. Dodatno, dobijeni rezultati pomenute analize imaju za cilj poređenje prevoda sa japanskog na engleski jezik time što bismo ih međusobno uporedili.

Praktičan cilj teze: Analiza korpusa i dopuna elektronskih rečnika nepoznatim rečima iz korpusa prevoda sa japanskog na srpski jezik. Statistička analiza vrsta reči i semantičkih kategorija korpusnih tekstova. Vizuelizacija različitih korpusa (ili delova korpusa) korišćenjem oblaka reči i oblaka drveta. Dodatno, slična analiza će se uraditi sa prevodima sa japanskog na engleski jezik.

## 1.2 Predmet rada

Predmet ovog master rada predstavlja istraživanje i analiza korpusa tekstova prevedenih sa japanskog jezika na srpski i engleski jezik. Prikupljeni korpusi sadrže tekstove koji su prevedeni sa japanskog iz ranijih perioda, dok je korpus tekstova prevedenih na srpski jezik gotovo isključivo sačinjen od tradicionalne *haiku* poezije i *haibun* putopisa. Prevodi savremenijeg japanskog jezika nisu uključeni u korpus. Pored toga, napravljen je i korpus srpske poezije čija je svrha da se uporedi sa korpusom prevoda japanske poezije na srpski jezik i koji sadrži pesme Branka Miljkovića, Desanke Maksimović, Đure Jakšića, Jovana Dučića, Miloša Crnjanskog i Vladislava Petkovića Disa.

Istraživanje je obuhvatilo analizu postojećih korpusa japanskog jezika, ukratko su opisane njihove karakteristike, kao i karakteristike haiku poezije. Fokus istraživanja je prikupljanje, obrada i analiza prevoda sa japanskog na srpski, u okviru koje je kreiran korpus tekstova koji je u radu predstavljen. Opis korpusa treba da predstavi karakteristike srodnih tema, ali i da objasni neke potencijalne uzroke anomalija u prevođenju sa japanskog jezika na srpski i engleski jezik.

Ovaj projekat povezuje tradicionalni i savremeni svet kroz korišćenje savremene tehnologije obrađivanja tekstova i njegove primene na vekovima stare tekstove napisane na japanskom tlu na japanskom jeziku iz različitih perioda daleke prošlosti.

## 1.3 Hipoteze istraživanja

Polazna hipoteza od koje se polazi jeste da srpski prevodi, od kojih je sačinjen korpus, zadržavaju rečitost originalnog teksta. Inspiraciju za pisanje ovog master rada sam dobio u naučnim radovima *Corpus and Lexicon – Mutual Incompleteness* dr Cvetane Krstev sa Filološkog fakulteta Univerziteta u Beogradu i dr Duška Vitasa sa Matematičkog fakulteta Univerziteta u Beogradu, u kojem se istražuje nepotpunost elektronskih rečnika u odnosu na korpusnu leksiku na srpskom jeziku (Krstev i Vitas, Corpus and Lexicon - Mutual Incompleteness 2005) i *Anotacije savremenog srpskog jezika* dr Miloša Utvića sa Filološkog fakulteta Univerziteta u Beogradu, u kojem se opisuje anotacija srpskog jezičkog korpusa (Utvić 2011).

## 1.4 Metode istraživanja

U procesu izrade teze koristiće se različiti izvori podataka: teorijski i istraživački radovi o temi korpusa, anotacije, japanskog jezika i haiku poezije, izveštaji relevantnih i renomiranih istraživačkih centara koji prate ove oblasti, zvanični dokumenti sa relevantnih konferencija i slično. Elektronski rečnici i lokalne gramatike za srpski jezik u okviru alata Unitex će se koristiti za analizu teksta.

Metode istraživanja se sastoje od prikupljanja materijala, obrade prikupljenog materija koja počinje optičkim prepoznavanjem karaktera (OCR[1]), njegove obrade u alatu Unitex[2] i programskom jeziku R, uz različite vidove vizelizacije rezultata. Pretprocesiranje teksta u programu *Unitex* podrazumeva primenu elektronskih rečnika na tekst, nakon čega se mogu koristiti različiti grafovi i statističke obrade. Kvantitativna analiza prevoda sa japanskog jezika će pružiti mogućnost potpunijeg sagledavanja prevođenja sa japanskog jezika na srpski i engleski jezik. Korišćnjem paketa *stylo* u R-u će se uraditi elementarna stilistička analiza.

[1]OpticalCharacterRecognition – vizuelno prepoznavanje slova i znakova

[2]http://unitexgramlab.org/

Oblaci reči (*word cloud, tag cloud*) su postali popularni za brzi pregled sadržaja veb stranice ili nekog teksta. Relativno nova vizualizacija oblak drveta (*tree cloud*) prikazuje više informacija. Kao oblak reči, prikazuje najčešće reči teksta, gde veličina odražava frekvenciju, ali reči su raspoređene na drvetu kako bi se odrazila njihova semantička blizina prema tekstu. Ovakvi oblaci drveta pomažu da se identifikuju glavne teme dokumenta, pa čak i da se koriste za analizu teksta.

Za potrebe ovog rada će se koristiti literatura koja je mahom sa interneta i uključuje sadržaje određenih sajtova koji se bave korpusnom lingvistikom, kao i japanskim jezikom, ali će takođe uključiti određene stručne radove koji su vezani za ove teme. Prikupljanje stručne literature vezane za proučavanje *haiku*-a metodama korpusne linvistike, u većini jezika nije lak zadatak zbog oskudice resursa. Pri prikupljanju tektova za korpus, jedan od problema je i postojanje sistema za rad na korpusima koji je nazvan HAIKU, a nema veze sa haiku poezijom, što je smanjivalo preciznost pretrage.

Za prvu analizu u *Unitex*-u *kigo*[3]-a u japanskoj *haiku* poeziji su izdvojene reči koje označavaju godišnja doba i primenjene na zbirci tradicionalne *haiku* poezije *haiku* majstora Jose Busona pod nazivom „Prolećno more“, koju je na srpski jezik preveo profesor Hiroši Jamasaki-Vukelić sa Filološkog fakulteta Univerziteta u Beogradu. U ovom master radu se, pored pomenute zbirke *haiku* poezije, vrši analiza i nad tri druge zbirke tradicionalne *haiku* poezije koje je takođe preveo na srpski jezik profesor Jamasaki-Vukelić.

Prevedeni tekstovi su iz polaznog *pdf* formata, programom *ABBYY PDF Transformer+* konvertovani u mašinski čitljiv tekstualni oblik i zatim su „očišćeni“ od grešaka nastalim optičkim prepoznavanjem i uređeni koristeći *Notepad++* pre nego što bi bili pretprocesirani i dalje obrađivani u programu *Unitex.* Budući da predstavljaju komentare prevodioca kao autora zbirki pesama umesto prevoda tekstova sa japanskog jezika, predgovori i pogovori, kao i indeksi i sadržaji ovih zbirki su uklonjeni pre računarske analize tekstova.

Pre samog kreiranja korpusa od ovih prevoda sa japanskog, u ovom radu će biti objašnjeno šta je to korpus, kako se koristi i zbog čega je bitan. Takođe će se istaći neke osnovne specifičnosti japanskog jezika i njegovih korpusnih resursa.

[3]Reči koje označavaju godišnja doba.

## 1.5 Sadržaj rada

Ovaj rad je podeljen u osam celina, pri čemu prvo poglavlje daje cilj i predmet rada, kao i hipoteze i metode istraživanja. Srž rada predstavljaju poglavlja 2-5, za čim slede zaključna razmatranja, literatura i popis slika.

Drugo poglavlje obrađuje metode i alate korpusne analize teksta, gde se na početku daju osnovni pojmovi vezani za korpusnu lingvistiku, uključujući analizu konkordanci i kolokacija, kao i frekvencijsku analiza teksta. Potom sledi pregled vrsta i namena korpusa, stilometrije, alata i resursi za analizu.

U trećem poglavlju se govori o specifičnostima japanskog jezika, prvenstveno o pismima koja se koriste: kanđi i kana, kao i osnovne gramatičke i stilske karakteristike japanskog jezika, a potom se razmatra dostupnost korpusa japanskog jezika, i daju osnovne karakteristike, sadržaj i namene pojedinih korpusa.

Četvrto poglavlje donosi opis prikupljenog korpusa, u kom se prvo opisuje kolekcija haiku poezije, potom kolekcija ostalih vrsta poezije i proze, način pripreme korpusa, motive izbora i izvore, pri čemu okosnicu master rada čine četiri zbirke tradicionalne japanske haiku poezije iz perioda Edo, potom tekstovi prevedeni i na srpski i na engleski jezik. Priređena je i mala kolekcija haiku pesama domaćih autora, kao i kolekcija srpske poezije.

U petom poglavlju se daju rezultati lingvističke analize, gde se prvo obrađuje analiza prevoda sa japanskog na srpski jezik i to posebno se analizira korpus tekstova koji nije haiku, a potom se daje analiza korpusa haiku poezije. Ovo poglavlje donosi način analize korišćenjem okruženja Unitex i elektronskim morfoloških rečnika za srpski jezik, koji su omogućili prepoznavanje pojedinih vrsta reči i njihovih osnovnih (kanonskih) oblika, tako da je za svaki korpus određen broj tokena, reči, potom prepoznatih monoleksemskih i polileksemskih reči. Dalje slede poređenja srpske poezije i japanskog haiku-a na nivou znakova interpunkcije i vrsta reči, kao i poređenje haiku poezije sa korpusom proze savremenog srpskog jezika. Analiza korpusa prevoda na engleski jezik, koji osim haiku poezije sadrži i haibun i druge zbirke stare poezije, predstavlja se u ovom poglavlju na nivou celog korpusa i po vrstama tekstova. Jedan odeljak je posvećen analiza bigrama i trigrama za tri korpusa, gde se n-grai izdvajaju i grafički prikazuju po frekvencijama.

U okviru prezentacije rezultata stilometrijske analize zasnovane na paketu stylo i programskom jeziku R, je opisana priprema podataka i stilometrijska analiza, uz različite

načine vizuelizacije rezultata obrade. Peto poglavlje takođe prikazuje rezultat analize kigo-a, odnosno reči i fraza koje označavaju godišnje doba u haiku pesmama je sa jedne strane obuhvatio izdvajanje tih reči, a potom i analizu kolokacija pojedinih kigo-a. Vizuelizacija korpusa prevoda japanskih tekstova na srpski jezik oblakom reči i drveta je prikazana za pomenute korpuse. Analizira neprepoznatih reči izdvaja 29 kandidata za unos u elektronske rečnike srpskog jezika.

U šestom poglavlju se daju zaključna razmatranja i neki od daljih pravaca započetog istraživanja. Sedmo poglavlje daje pregled literature korišćene u radu,a na kraju rada, u prilogu je dat popis od 25 slika kojima su ilustrovani resursi, softveri i rezultati istraživanja sprovedenog tokom izrade ovog rada.

# 2 Metode i alati korpusne analize teksta

## 2.1 Korpusna lingvistika

Reč korpus potiče od latinske reči corpus i znači telo, ali se u savremenoj lingvistici odnosi na zbirku tekstova koje računar može da pročita i koji mogu biti pretraženi koristeći računarske metode (Ray Carey 2017). Korpus je referentni sistem zasnovan na elektronskoj zbirci tekstova koji su sastavljeni na određenom jeziku (Russian National Corpus 2017). Iako je većina dostupnih korpusa u tekstualnom obliku, postoji sve više multimodalnih korpusa, uključujući korpuse znakovnog jezika (Björkenstam 2013), kao i govora, izraza lica, položaja tela i slično (Jean-Claude Martin 2017).

Idealno, korpus je skup uzoraka jezika osmišljen da bude reprezentativan jeziku ili podjeziku kroz pažljiv odabir, a ne nasumično odabran skup podataka (Björkenstam 2013). Nacionalni korpus predstavlja taj jezik u jednom ili nekoliko stanja njegovog razvoja u raznoraznim žanrovima, stilovima, društvenim i teritorijalnim varijantama korišćenja i slično (Russian National Corpus 2017).

Korpus takođe sadrži dodatne informacije o svojstvima tekstova koji su uključeni, što se postiže anotacijama, koje su glavna odlika korpusa koja odvaja korpuse od jednostavnih zbirki odnosno biblioteka tekstova na internetu. Takve biblioteke nisu prikladne za akademski rad o prirodi jezika jer se često fokusiraju na sadržaj tekstova umesto svojstava njihovog jezika dok tvorci korpusa prepoznaju važnost književne ili naučne vrednosti tekstova, ali ih vide kao sekundarno svojstvo. Za razliku od elektronske biblioteke, nacionalni

korpus nije zbirka tekstova za koje se smatra da su sami po sebi „korisni“ ili „interesantni“, već su tekstovi u korpusu korisni i interesantni za proučavanje jezika. Takvi tekstovi mogu da uključuju ne samo velika književna dela, već i dela drugorazrednih pisaca ili transkripcije običnih razgovora (Russian National Corpus 2017).

U lingvistici i leksikografiji, korpus predstavlja telo tekstova, izgovora ili drugih primeraka koji se smatraju manje ili više reprezentativnim za jedan jezik, a obično se skladište u elektronskoj bazi podataka. Računarski korpus je velika kolekcija tekstova koje mašine mogu da pročitaju i koji može da uskladišti nekoliko stotina miliona reči čije se odlike mogu analizirati *tagovanjem*[4]i korišćenjem programa za konkordance[5] (McArthur 1992).

Nacionalni korpus stvaraju lingvisti[6] za potrebe akademskih istraživanja i proučavanje jezika. Većina glavnih svetskih jezika ima svoje korpuse. Jedan poznati primer je britanski nacionalni korpus, koji se koristi kao uzor za mnoge savremene korpuse, dok je među slovenskim jezicima vredan pomena češki nacionalni korpus, koji je sastavljen na Karlovom Univerzitetu u Pragu (Russian National Corpus 2017).

Što se tiče korpusa srpskog jezika, njegov razvoj je počeo još 1981. godine na Matematičkom institutu putem projekta Matematička i računarska lingvistika, iako je ideja o formiranju korpusa savremenog srpskog i tadašnjeg srpskohrvatskog jezika postojala još 1978. godine, kada je održana prva jugoslovenska konferencija o računarskoj obradi lingvističkih podataka[7] zahvaljujući entuzijazmu i angažovanosti Milana Šipke. Korišćen je sistem AURORA, koji je generisao konkordance i različite vrste indeksa za zadati tekst i koji se mogao uporediti sa sistemom COCOA, koji je bio vodeći sistem u to vreme. Sistem AURORA je pod okriljem projekta u periodu od 1981. do 1985. godine izneo zadovoljavajuće rezulate s obzirom na tehnološka ograničenja osamdesetih godina minulog veka: napravljena je prva zbirka tekstova u digitalnom obliku koja se sastojala od stručne literature, načinjeni su prvi eksperimenti u morfološkom generisanju srpskohrvatskog jezika, sprovedena su prva istraživanja na području korpusne lingvistike, kontakti sa vodećim evropskim istraživačima

[4]Dodavanje idenfitikujućih i klasifikujućih oznaka rečima i drugim tvorevinama.

[5] Registar reči.

[6]Specijalista za korpusku lingvistiku, discipline koja se brzo razvija.

[7]Kasnije poznatoj kao ROJP.

korpusne lingvistike su uspostavljeni. U tom periodu su uz pomoću sistema AURORA sastavljeni i obrađeni prvi paralelni korpusi kao što su srpsko-slovenački od uputstava za lekove, englesko-srpski na području informatike i srpsko-hrvatsko-slovenački na uzorcima tadašnjih saveznih zakona. Grupa za jezičke tehnologije sa Matematičkog fakulteta Univerziteta u Beogradu se uključila u projekat Evropskog saveta Jezičke industrije i time proširila krug evropskih laboratorija sa kojima je sarađivala kao što je LADL profesora Morija Grosa u oblasti razvoja leksičkih resursa u obliku sistema elektronskih rečnika, dok je sam razvoj metoda razvoja korpusa potpomogao profesor Volfgang Tojbert kroz projekat TELRI I/II Evropske Unije. Korpus srpskog jezika je postavljen na internet 2002. godine putem projekta vlade Republike Srbije Interakcija teksta i rečnika (Matematički fakultet Univerziteta u Beogradu n.d.).

Korpus može da bude korišćen na mnogo različitih načina da bi se izučavali jezici i kulture u kojima se koriste. Budući da je opširan i budući da se sastoji od različitih vrsta teksta iz različitih oblasti, on predstavlja reprezentativan uzorak jezika iz koga sve odlike pisanog jezika mogu biti proučavani (Oxford Living Dictionaries n.d.).

Sistematskom analizom podataka korpusa možemo doći do otkrića koja se potom koriste da se ažuriraju i poboljšavaju rečničke odrednice kako bi se ostvario najprecizniji mogući opis jednog jezika (Oxford Living Dictionaries n.d.).

Nove reči su najočiglednija manifestacija jezičkih promena, ali se traže i suptilnije promene u jednom jeziku, kao što su novija značenja već postojećih reči, promena u pisanju reči tokom dužeg vremenskog perioda ili čak gramatičke promene (Oxford Living Dictionaries n.d.).

Međutim, korišćenje korpusa u savremenoj leksikografiji nije samo za praćenje promena. Stotinama godina, uključujući veliki deo dvadesetog veka, leksikografi su radili sa nedovoljno podataka, a naročito ni sa čim približno korpusnim podacima, mada ponekad samo uz pomoću svoje intuicije. Čak i kada su podaci o korišćenju bili dostupni, urednici rečnika nisu imali načina da filtriraju ili urede velike količine podataka na pouzdan i efikasan način, što je postalo moguće tek nakon tehnološkog napretka kasnog dvadesetog veka, kada su računari osposobljeni da manipulišu i procesiraju veoma velike tekstove. Stoga, ogromna dobrobit korpusne leksikografije je u otkrivanju činjenica o jeziku koje nisu nove, ali nisu ranije bile primećene (Oxford Living Dictionaries n.d.).

Glavna svrha korpusa je da olakša akademska istraživanja o leksici i gramatici jezika, kao i suptilne i stalne procese promene jezika tokom relativno kratkog vremenskog perioda, recimo – tokom jednog ili dva veka. Druga svrha korpusa je da posluži kao referentna tačka za leksička, gramatička i akcentološka pitanja, kao i za istoriju jezika - da potvrdi hipoteze o jeziku kao, na primer, da odredi kako korišćenje određenog zvuka, reči ili sintakse varira. Savremene IT tehnologije čine obradu velike količine teksta značajno jednostavnijim i bržim, što stvara mogućnosti za masovnu statističku analizu tekstova. Kao rezultat toga, istraživanje jezika daje rezultate koji su pre toga mogli biti samo nagađani. U današnje vreme, istinski naučni opisi gramatičkih i akademskih rečnika moraju biti zasnovani na korpusima njihovih jezika. Korišćenje podataka korpusa, iako nije strogo neophodno, je poželjno u drugim, specijalizovanijim jezičkim istraživanjima (Russian National Corpus 2017).

Stoga, glavni korisnici nacionalnih korpusa su lingvisti raznih profila, ali su korpusi korisni i za ljude koji nisu lingvisti. Pouzdane statističke informacije o korišćenju jezika u određenom periodu ili od određenog autora bi mogli biti interesantni za istraživače književnosti, istorije i drugih humanističkih predmeta. Nacionalni korpusi su takođe korisni za predavače jezika - i maternjeg i stranog. Udžbenici iz jezika i programi za podučavanje su sve više orijentisani ka korpusima. Korpus može biti korišćen da se utvrde varijante korišćenja nepoznatih reči od strane stranaca, učenika, predavača, novinara, pisaca itd. Stoga, korpus je namenjen ljudima koji su zainteresovani u strukturu i korišćenje jezika, bez obzira na to da li je iz profesionalnog interesa ili ne (Russian National Corpus 2017).

### 2.1.1 Analiza konkordanci i kolokacija

Konkordanca je po alfabetnom redu raspoređen spisak reči koje su prisutne u tekstu ili tekstovima, obično sa citatima određenog pasusa ili sa kontekstom (Oxford Living Dictionaries 2017). Konkordanca je učinak sastavljenog korpusa koji je obrađen računarskim programom, indeks svih reči u korpusu zajedno sa njihovim najbližim lingvističkim kontekstima i informacijama o njihovoj učestalosti i lokaciji (Cobb 2018).

Linija konkordance je linija teksta koja je uzeta iz korpusa, koja može da bude sa početka, sredine ili kraja jednog od teksta i koja može biti sačinjena od jedne rečenice, jednog dela rečenice ili dela dveju rečenica. Svaki skup linija konkordance uključuje traženu reč, odnosno onu koja se proučava. Tražena reč je uvek u sredini linije konkordance, što znači da kada proučavamo reč u skupu linija konkordance možemo da vidimo njen kontekst, odnosno, reči koje se koriste pre i posle tražene reči (Haywood n.d.).

Kolokacija je kombinacija reči koje su oformljene kada se dve ili više reči ili izraza često koriste na način koji zvuči tačno, odnosno redovno korišćenje određenih reči i izraza (Cambridge Dictionary n.d.). Predlog J. Firth-a iz 1957. godine da se gleda „sa kime se reči druže“ je na nekoliko različitih načina operacionalizovan i u nekoliko različitih konteksta istražen, ali od tada mnoge naučene lekcije u izučavanju kolokacija još uvek nisu sistematski procenjene i u potpunosti primenjene u alatima koje korpusni lingvisti koriste. Ideju da se tekst u određenom polju diskursa organizuje u leksičke obrasce, što može biti vizuelizovano kao mreže reči koje se međusobno raspodeljuju je isprva predložena od strane M.K. Phillips-a 1983. godine i ta ideja ima bitne teoretske implikacije za naše razmevanje veze između leksike i teksta, kao i između teksta i čitaočevog uma (Brezina, McEnery i Wattam 2015).

Tradicionalno, predložena su tri kriterijuma za identifikovanje kolokacija, a to su rastojanje, učestalost i ekskluzivitet. Rastojanje određuje raspon oko čvorne reči[8], gde tražimo kolokacije, a taj raspon se zove kolokacioni prozor. Rastojanje kolokacije od čvorne reči može da bude najmanje jedna reč, a može da bude i četiri ili pet sa obe strane čvorne reči. Drugi kriterijum, učestalost korišćenja, je bitan pokazatelj tipičnosti u povezanosti reči tj. koje reči tipično stoje jedna pored druge.

Pored ova tri kriterijima navedena gore, S. Gries je 2013. godine istakao još tri kriterijuma koje treba uzeti u obzir: usmerenost (eng. directionality), disperzija i odnos oblik-lema (type-token) po kolokacijama. Usmerenost se odnosi na činjenicu da je snaga privlačnosti između dve reči retko kada simetrična – jedna reč može da ima jaču povezanost sa drugom rečju, ali ta druga reč može da ima jaču povezanost sa nekim drugim rečima. Ipak, mere asocijacije (kao što su z-score, t-score, MI[9], log-likelihood itd.) (Evert 2004) ne mogu da tu obuhvate razliku jer većina tih reči koje se često koriste u korpusnoj lingvistici imaju simetrične mere. Disperzija je raspoređenost čvorova i kolokacija u korpusu – koliko puta u koliko tekstova se jedna reč kolocira sa drugom. Na kraju, lematizacijom teksta, gde se oblici reči zamenjuju kanonskim oblikom (na primer: *pas*, *psa, psu*,.. se zamenjuju sa *pas*) je kriterijum koji uzima u obzir ne samo snagu date kolokacijske veze oblika reči, već i nivo pojavljivanja kolokata i u drugim gramatičkim oblicima. Vaclav Brezina, Tony McEnery i Stephen Wattam su dodali sedmi kriterijum – povezanost između zasebnih kolokacija. Kolokacije reči se ne dešavaju u

[8]Reči za koju smo zainteresovani.

[9] Mutual Information

izolovanosti, već su deo kompleksne mreže semantičkih veza koja na kraju otkriva njihovo značenje i semantičku strukturu teksta ili korpusa – jedna reč može da ne kolocira sa drugim rečima, a da je povezana sa rečju koja kolocira sa njom i drugim rečima (Brezina, McEnery i Wattam 2015).

### 2.1.2 Frekvencijska analiza teksta

Budući da je po svojoj prirodi korpusna lingvistika distributivna disciplina, smatralo se da korpusi kao takvi sadrže samo distributivnu učestalost podataka dva ili tri tipa, u zavisnosti kako neko želi da ih posmatra:

- Učestalosti okurencije[10] lingvističkih elemenata, koji mogu da budu proučavani iz dve perspektive:
  1) Koliko su učestale morfeme ili reči ili obrasci u delovima korpusa? Ova informacija može biti snabdevena u različitim oblicima lista učestalosti;
  2) Koliko jednako su morfeme ili reči ili obrasci raspodeljeni kroz korpus? Ova informacija može biti snabdevena u obliku različitih statistika disperzije;
- Učestalosti ko-okurencije [11] lingvističkih elemenata: Koliko često se lingvistički elementi poput morfema, reči, obrazaca pojavljuju pored drugih lingvističkih elemenata iz ovog sklopa ili pozicije u tekstu.

Drugim rečima, u prvom slučaju i iz perspektive puriste, sve što korpus vraća su celi brojevi veći ili jednaki nuli, naime učestalosti koliko često se nešto pojavilo u korpusu. Sve ostalo za šta su korpusni lingvisti zapravo zainteresovani mora zatim biti operacionalizovano na osnovu nekih vrsta učestalosti (Gries 2010).

Dok se stepen privrženosti ka ovom malo radikalnijim pogledom može biti diskutabilan, verovatno je pošteno reći da, zbog ovih razloga, grane lingvistike koje su koristile korpuse ili tekstualne baze podataka su oduvek bile među kvantitativno orijentisanim po disciplinama ovog polja istraživanja. Dok bi ovo obično dovelo do očekivanja da su korpusni lingvisti veoma orijentisani prema statistici, jer ipak, statistika je naučna disciplina koja nas uči kako da se obračunamo sa kvantitativnim raspodelama, ali je, nažalost, takođe pošteno reći da ne koriste sve studije zasnovane na korpusu u potpunosti dostupne statističke metode. Zapravo,

[10] Pojavljivanja (eng. *occurence*, prim. prev.)

[11] Su-pojavljivanja (eng. *co-occurence*, prim. prev.)

samo su u poslednjih nekoliko godina korpusni lingisti počeli da koriste više sofisticarinijih i obimnijih alata, od kojih su oba za baratanjem korpusnih podataka kao i za statističke analize dobijenih podataka. Međutim, ovaj trend je dovoljno nov da se disciplina nije razvila u stanje gde se takvi resursi naširoko koriste i uče, a za sada postoji samo jedan posvećen uvod u statistiku za korpusne lingviste[12] koji počinje da bude zastareo, ali i jedan uvod u korpusnu lingvistiku sa detaljnim pregledom poglavlja o statističkim metodama (Gries 2010).

## 2.2 Vrste i namena korpusa

Tekstualni korpus može biti svrstan u različite kategorije na osnovu izvora sadržaja, metapodataka, prisustva multimedija ili njegovog odnosa prema drugim korpusima. Jedan isti korpus može spadati u više od jedne kategorije ukoliko ispunjava kriterijume za više od jedne kategorije (Lexical Computing CZ s.r.o. 2017).

Jednojezični korpus je najučestaliji tip korpusa. Kao što ime kaže, sadrži tekst na samo jednom jeziku. Ovaj korpus obično ima obeležene vrste reči i koristi se od strane širokog spektra korisnika za različite zadatke, počevši od najpraktičnijih, kao što je provera pravilne upotrebe reči ili pretraga najprirodnijih kombinacija reči, pa sve do primene u nauci, kao što je prepoznavanje učestalih obrazaca ili novih trendova u jeziku (Lexical Computing CZ s.r.o. 2017).

Balansirani korpus, odnosno reprezentativni korpus, je korpus u kojem su tekstovi odabrani u predefinisanim proporcijama kako bi se ogledala određena promena ili varijacija u jeziku (Prytz 2012).

Monitorni korpusi su korpusi u kojima se novi tekstovi dodaju kako bi se nadgledala promena jezika (Prytz 2012).

Paralelni korpus se sastoji od dva jednojezična korpusa, od kojih je jedan prevod drugog. Na primer, roman i njegov prevod mogu da se iskoriste da bi se napravio paralelni korpus. Oba jezika treba da budu usklađeni, npr. odgovarajući segmenti, obično rečenice ili pasusi, treba da budu upareni. Korisnik zatim može da pretraži sve primere reči ili izraza na jednom jeziku i rezultati će biti prikazani zajedno sa odgovarajućom rečenicom na drugom jeziku. Korisnik

[12]Oakes, M. (1998).

potom može da proučava kako su pretražene reči ili izrazi prevedeni (Lexical Computing CZ s.r.o. 2017).

Višejezični korpus je veoma sličan paralelnom korpusu i često se ova dva termina koriste naizmenično. Višejezični korpus sadrži teksove na nekoliko jezika koji su prevodi istog teksta i upareni su na isti način kao paralelni korpusi. Kada se odaberu samo dva jezika, višejezični korpus funkcioniše kao paralelni korpus. Korisnik takođe može da odluči da radi na jednom jeziku i da ga koristi kao jednojezični korpus (Lexical Computing CZ s.r.o. 2017).

Komparativni korpus je sklop dva ili više jednojezična korpusa čiji tekstovi imaju istu temu, ali nisu prevodi jedni drugih, i stoga, nisu upareni. Kada korisnici pretražuju ove korpuse mogu da koriste činjenicu da korpusi imaju iste metapodatke (Lexical Computing CZ s.r.o. 2017).

Učenički korpus je korpus tekstova koji je proizveden od učenika nekog jezika. Korpus se koristi da se prouče greške i problemi tokom učenja stranog jezika (Lexical Computing CZ s.r.o. 2017).

Dijahroni korpus je korpus koji sadrži tekstove iz različitih perioda i koristi se da prouči razvoj ili promene u jeziku (Lexical Computing CZ s.r.o. 2017).

Specijalizovani korpus sadrži tekstove koji su ograničeni na jedno ili više predmeta, oblasti, tema i slično, i koristi se da se proučava kako se koristi specijalizovani jezik ili jezik struke (Lexical Computing CZ s.r.o. 2017).

Multimedijski korpus sadrži tekstove koji su potpomognuti auditivnim ili vizuelnim materijalima ili drugim vrstama multimedijskog sadržaja (Lexical Computing CZ s.r.o. 2017).

## 2.3 Stilometrija

Stilometrija, odnosno analiza prebrojivog lingvističkog sadržaja tekstova, obično je povezana sa autorskom pripisanošću ili, više senzacijski, sa jedinstvenom svrhom otkrivanja plagijata. Međutim, skorašnja istraživanja su pokazala da iste metode koje pomažu prilikom otkrivanja plagijata sa falfikikovanjem tekstova mogu da se iskoriste i u širem kontekstu književnih studija. Obrasci stilometrijskih sličnosti i razlika daju nova viđenja u vezama između različitih knjiga istog autora, knjiga različitih autora, autora različitih polova i vremenskih

perioda, prevoda istog autora ili grupe autora i time se potpomažu novi načini sagledavanja dela koja su naizgled izučavana iz svih mogućih perspektiva (University of Leipzig 2017).

Stilometrisko istraživanje pokušava da odgovori na određena pitanja, kao što su: Šta je zajedničko jeziku koji koristimo i šta je povezano sa kulturološkim kontekstima i piščevom individualnošću? Koji elementi stila su pod uticajem književnog perioda, žanra ili teme? Šta je nesvesno pripojeno od strane autora što ogleda njegovo obrazovanje, pol, veroispovest, društvene ili istorijske uslove? Koje odlike pisanog teksta mogu da odaju osobu koja ga je napisala uprkos njenim estetskim, društvenim ili istorijskim uslovima? (University of Leipzig 2017)

## 2.4 Alat za analizu

Da bi se analizirao korpus, svakako je potreban alat u obliku namenskog softvera za analizu korpusa, koji se dele po funkcijama kao što su anotaciranje ili konkordanciranje. Veliki deo ovih programa sadrži više takvih funkcija, ali to nije oduvek bio slučaj, već su programi za analizu korpusa napredovali iz generacije u generaciju.

Četiri generacije softvera za korpusnu analizu su opisane 2012. godine. Prva generacija se pojavila šezdesetih i sedamdesetih godina prošlog veka i ti alati su mogli samo da obrađuju *ASCII* skup znakova, koji je u suštini sačinjen od slova engleskog alfabeta i znakova interpunkcije, brojeva i ograničenog broja simbola, kao što su oni za matematičke jednačine, pa su stoga bili ograničeni samo na obradu korpusa engleskog jezika. Veliki deo alata je osmišljen samo za jednu funkciju, kao što je brojanje broja reči u tekstu ili stvaranje *KWIC*[13] linija konkordanci. Primeri ove generacije alata uključuju *Concordance Generator, Discon, Drexel Concordance Program, Concordance* i *CLOC*, poslednji od kojih je korišćen u poznatom projektu *COBUILD* Univerziteta u Birmingemu koji je predvodio John Sinclair. Veliki deo koncepta za alate koji su predloženi šezdesetih godina i dalje služe kao osnove za savremene alate za analizu korpusa (Anthony 2011).

Druga generacija alata za analizu korpusa je predstavljena osamdesetih i devedesetih godina prošlog veka i ona je takođe bila ograničena na obradu *ASCII* i imala ograničenu funkcionalnost, ali su ti alati imali prednost u tome da su mogli biti pokretani na ranim ličnim računarima i time dozvoljavali istraživačima da sprovode proučavanja manjih razmera, a

[13]KeyWord In Context (srp. ključna reč u kontekstu)

dozvoljavali su i nastavnicima da predstave korpusne analize u učionice za učenje jezika. Primeri softvera ove generacije uključuju *Oxford Concordance Program*, *Longman Mini Concordancer*, *Kaye concordancer* i *Micro Concord* (Anthony 2011).

Većina trenutnih alata koje koriste korpusni lingvisti se klasifikuju kao alati treće generacije. Rane verzije ovih alata su počeli da se pojavljuju kasnih devedesetih godina, ali mnogi od njih i dan-danas nastavljaju da budu razvijani i unapređeni. Glavna prednost ovih alata nad ranijim je ta što nude nekoliko funkcija, uključujući česte statističke metode, imaju poboljšanu prilagodljivost da rade sa većim korpusima, nude neki stepen podrške nekolicini jezika tako što obrađuju znakove van *ASCII* skupa znakova i uključuju interfejse koji su pogodni za korisnike sa malo iskustva u računarima. Primeri treće generacije alata uključuju *WordSmith Tools*, *MonoConc Pro* i *AntConc* (Anthony 2011).

Najveće ograničenje treće generacije alata za analizu korpusa je to što imaju smetnje da barataju veoma velikim korpusima od preko stotinu miliona reči. Danas se izbacuje sve veći broj korpusa koji se automatski sastavljaju tako što se sakupljaju sa internet sajtova. Ovi korpusi mogu biti dugački nekoliko milijardi reči, a softveri treće generacije alata nisu prikladni da ih obrade. Još jedno ograničenje je to što izdavači postaju sve osetljiviji u dozvoljavanju da se njihovi podaci koriste za istraživačke svrhe. Stoga se zbirke tekstova više ne mogu sastavljati i raspodeliti za analizu korpusnim alatima na ličnim računarima. Odgovor na ova dva problema je bilo stvaranje četvrte generacije alata, kao što su*corpus.byu.edu*, *CQPweb*, *SketchEngine*, and *Wmatrix*. Ovi alati nude bolju prilagodljivost time što korpus skladište u jednoj bazi podataka na serveru mreže i pre-indeksira podatke da bi dopustili brze pretrage, a takođe nude zaštitu autorskih prava tako što sprečavaju korisnike da pregledaju ceo korpus, već korisnici moraju da pristupe korpusu kroz korisnički interfejs koji prikazuje samo mali sklop podataka korpusa u jednom trenutku. Interfejs, međutim, obično dozvoljava korisnicima da pretraže cele korpuse i proizvedu standardne rezultate iz celog korpusa, kao što su *KWIC* linije konkordansi i spiskovi učestalosti reči (Anthony 2011).

I pored prednosti četvrte generacije alata, i oni imaju nekoliko ograničenja. Prvo, mogu da budu prezahtevni ukoliko korisnik želi da sastavi mali korpus i sprovede jednostavnu analizu nad njim. Četvrta generacija alata zahteva da se podaci očiste, obrade, reformatiraju, indeksuju i konačno postave na server pre nego što analiza može da počne. Takođe, za prvobitni pristup serveru, korisnici trebaju biti registrovani za uslugu, pristati na razne licence, a ponekad i da plaćaju mesečne pretplate. Kao alternativu bi mogli da postave alat na

lični računar ili server, ali bi za to korisnik morao da kupi server (računar), podesi serverske parametre, instalira programski alat za korpuse, a zatim održava server dok traje projekat. Još jedan problem je taj što je mnogo alata četvrte generacije neprikladno za analizu korpusnih podataka osetljive prirode, kao što su interni sastanci poslovnih stranaka, prijemni ispiti za fakultete i lični dnevnici, jer kao takvi treba da se njihovi podaci okače na spoljašnji server.

Treći problem je povezan sa činjenicom da su neki alati četvrte generacije direktno povezani sa određenim korpusima (koji su obično pod autorskom zaštitom) i ne nude mogućnosti da se njima analizaju podaci drugih korpusa. Četvrti problem, koji je povezan sa trećim, je u tome što kada se stvori novi korpus koji je pod zaštitom autorskih prava, neizbežno je da se pusti kroz novo podešavanje korpusa ili alata, što kao posledicu ima eksploziju jednokratnih interfejsa na mreži sa jednim korpusom od kojih svaki ima odlike idiosinkratičkih nameštanja kontrola i operacija. Peti i konačni problem je taj što četvrta generacija alata pomućuje granice između podataka korpusa i alata koji se koristi da se korpus pregleda. Zbog načina na koji ovi alati skladište podatke korpusa u indeksiran obrazac na spoljašnjem serveru, korisnici nemaju način da promatraju sirove podatke neposredno svojim očima. Sve interakcije moraju biti putem alata koji je obično korisnički interfejs na pretraživaču mreže. Kada se ovim alatima analiziraju korpusi, istraživačima je lako da zaborave filtrirajuća dejstva alata i počnu da ga koriste na bespogovorni način (Anthony 2011).

Na osnovu navedenog istorijskog razvoja alata za analizu korpusa, jasno je da treća i četvrta generacija alata imaju svoje prednosti i mane. Očito, treća generacija alata, kao što su *AntConc*i*WordSmith Tools,* nastavljaju da budu popularni među istraživačima, dok su među četvrtom generacijom alata najpopularniji sajtovi *corpus.byu.edu, Sketch Engine* i *Wmatrix.* Ipak, treba napomenuti da je *corpus.byu.edu* jedini od navedenih alata koji nije alat opšte namene i da je tako visoko kotiran zbog činjenice da daje pristup najvećem korpusu savremenog američkog engleskog jezika (Anthony 2011).

Za analizu korpusa u ovom master radu mahom je korišćen alat *Unitex* koji pripada trećoj generaciji alata za izučavanje korpusa i predstavlja skup programa koji su razvijeni za analize tekstova na prirodnim jezicima. Negovi lingvistički resursi se sastoje od elektronskih rečnika, gramatika, leksikonsko-gramatičkih tabela *Unitex* nudi način da automatski gradi gramatike iz leksičko-gramatičkih tabela i može biti posmatran kao alat u kojem mogu da se ubace i koriste lingvistički resursi. Njegove tehničke karakteristike su njegova prenosivost, modularnost, mogućnost rada sa jezicima koji koriste posebne sisteme pisanja (kao što su

neki azijski jezici) i njegova otvorenost zahvaljujući njegovoj distribuciji otvorenog koda. Negove lingvističke karakteristike su preciznost, potpunost i uzimanje u obzir polusloženica, složenica i fraza, naročito onih koji se bave brojanjem složenica (Paumier 2003).

## 2.5 Resursi za analizu teksta (rečnici i gramatike)

Elektronski rečnici određuju proste i složene reči jednog jezika zajedno sa njegovim lemama i skupom gramatičkih kodova. Dostupnost ovih rečnika je glavna prednost kada se uporedi sa uobičajenim alatima za pretragu obrazaca jer se informacije koje sadrže mogu iskoristiti za pretragu i prepoznavanje, a time i opisivanje klasa reči koristeći veoma jednostavne obrasce. Rečnici su prikazani u DELA formatu (Gross 1989) i sastavljeni su od strane ekipa lingvista za brojne jezike (francuski, engleski, grčki, italijanski, španski, nemački, tajlandski, korejski, poljski, norveški, portugalski itd.) Gramatike koje se koriste u *Unitex*-u su predstave lingvističkih fenomena na osnovu konačnih transduktora, što je formalizam usko povezan sa konačnim automatima. Brojne studije su pokazale adekvatnost automata za lingvističke probleme na svim deskriptivnim nivoima morfologije i sintakse do fonetskih problema. Gramatike koje su napravljene u *Unitex*-u vode ovaj pristup dalje tako što koriste formalizam koji je moćniji od automata. Ove gramatike su predstavljene kao grafovi koje koristik može lako da napravi i dopuni (Paumier 2003).

Leksičko-gramatičke tabele su matrice koje opisuju svojstva nekih reči, od kojih su mnoge takve tabele konstruisane za sve proste reči u francuskom jeziku kako bi se opisale njihova značajna sintaktička svojstva. Svaka reč ima kvazi-jedinstveno ponašanje, a tabele su način da se prikaže gramatika svakog elementa u leksikonu, pa su stoga tabele dobile naziv leksičko-gramatičke tabele za ovu lingvističku teoriju (Paumier 2003).

Digitalizovani resursi srpskog jezika koji bi, za potrebe nastavnika i učenika, morali da se nađu na računaru, dele se na pet grupa:

1) Tekstovi u elektronskom obliku, koji mogu biti organizovani kao korpusi ili kao tekstuelni arhivi[14]. Njihova dostupnost podleže zakonu o autorskim pravima.

[14]Kolekcije kompletnih tekstova arhiviranih na serveru na mreži ili uskladišteni na digitalnom prostoru, koji ne moraju biti u jedinstvenom obliku.

2) *On-line* biblioteke predstavljaju drugi važan resurs, u kojima je najvažnije uspostavljanje pretraživačkih kataloga radi lakše pretrage, a čiji određeni fondovi mogu biti predstavljeni u digitalnom obliku.
3) Resurs koji je verovatno najteže prikazati i podržati informatičkim sredstvima je predstavljanje znanja o gramatičkom sistemu, čiju osnovnu teškoću verovatno predstavljaju razlike u teorijskim pristupima između različitih obrazovnih institucija, što je problem koji se obično prevazilazi kroz dokumentacione centre koji upućuju na releventan materijal složen po obrazovnim institucijama. Što se srpskog jezika tiče, postoje programi koji omogućavaju ekperimente u oblasti morfologije (automatska promena imenica, prideva i glagola, morfološka segmentacija na nivoima fleksije i derivacije itd.). Koliko god da je teško predstaviti jedan gramatički sistem u informatičkom obliku, računar omogućava da se razviju testovi poznavanja određenih nastavnih jedinica i da se automatizuju u obliku zanimljivom za učenike.
4) Rečnici su resurs koji se nasuprot gramatici smatraju podesnim za informatizaciju. Postoji osnovna podela na dve vrste rečnika, od kojih jednu čine mašinski čitljivi rečnici koji predstavljaju digitalizovanu transkripciju svojih papirnih izdanja potencijalno obogaćenu hipertekstuelnim vezama i predstavljanu kao bazu podataka, a drugu grupu čine elektronski rečnici namenjeni automatskoj transfromaciji teksta umesto ljudskom korisniku.
5) Na kraju, resursi koji podržavaju primenu pravopisa su takođe bitni i tu se iz informatičkog ugla pre svega postavlja pitanje izgleda slova, odnosno fontova, njihovog kodnog rasporeda i mogućnošću pretvaranja teksta iz jednog zapisa u drugi. Razvoj novih ćiriličnih slovnih garnitura neophodnih za prikazivanje dokumenata, kako na ekranu tako i na papiru, je aktivnost koju potpomaže i Vukova zadužbina. Univerzitet Vaterlu u Kanadi održava stranicu *sprsko-pismo.com*, gde se mogu naći iscrpne informacije o ovom pitanju u informatičkom smislu (Krstev 2000).

# 3 Specifičnosti japanskog jezika

## 3.1 O japanskom jeziku

Japanski jezik je jezik nejasnog porekla i jezičke porodice. Još uvek nije ustanovljeno da li pripada altajskoj porodici jezika, gde su turski, mongolski i drugi jezici, pošto takođe ima

sličnosti sa austronežanskom jezičkom porodicom, gde se nalaze jezici poput polinežanskog (japan-guide.com 2017).

Japanski sistem pisanja se sastoji od tri različita pisma, a to su *kanđi*[15], *hiragana* i *katakana*[16]. Japanski tekstovi se mogu pisati na dva načina: u zapadnjačkom stilu u horizontalnim redovima odozgo nadole ili u tradicionalnom japanskom stilu u vertikalnim kolonama s desna na levo. U današnje vreme se koriste oba stila podjednako (japan-guide.com 2017).

U japanskoj gramatici, faktori poput gramatičkih rodova i brojeva gotovo da ne postoje, konjugacije za glagole i prideve su jednostavni i gotovo uvek pravilni, dok se imenice ne deklinuju i ostaju u istom obliku (japan-guide.com 2017).

U poređenju sa drugim jezicima, japanski jezik nema mnogo glasova[17], ali zato ima mnogo homonima[18], zbog kojih japanski jezik upravo i koristi *kanđije* kako bi se homonimi mogli razlikovati (japan-guide.com 2017).

Japanski jezik je veoma kontekstualan i situacionalan i teško je odrediti kako reći nešto na japanskom ukoliko se ne znaju detalji društvenog konteksta – doba dana, doba godine, društvenog statusa, pola i godina govornika, sagovornika i slušaoca, kao i društvenih veza među njima (The Teaching Company 2018).

Japanski jezik je specifičan po tome što koristi različite reči i izraze u zavisnosti od hijerarhije i bliskosti govornika sa sagovornikom, pa stoga postoji nekoliko nivoa učtivosti u japanskom jeziku (japan-guide.com 2017).

U japanskom jeziku, lične zamenice bivaju rutinski izostavljane kao nepotrebne zbog tačaka referenci koje se nalaze u drugim delovima rečenice, bilo da su u obliku sufiksa ili glagola (The Teaching Company, 2018).

---

[15] Nekoliko hiljada kineskih karaktera.

[16] Dva paralelna slogovna pisma od 46 karaktera koja se zajedno zovu *kana*.

[17] Ne poznaje razlike između V i B, R i L, dok neki slogovi ne mogu da se izgovore i nemaju svoju pisanu formu u hiragani.

18 Reči koje imaju isti izgovor, ali različita značenja.

### 3.1.1 Kanđi

*Kanđi* (jap. 漢字、pismo Han dinastije) je najbrojnije od tri pisma japanskog jezika, i ono je u biti sačinjeno od kineskih znakova koji su u Japan pristigli u petom veku nove ere iz Koreje (japan-guide.com 2017).

*Kanđiji* su ideogrami – svaki simbol ima svoje značenje i odgovara određenim rečima (Slika 1). Kombinujući ove ideograme nastaje još reči. *Kanđija* u japanskom jeziku ima na desetine hiljada, od kojih je 2136 ozvaničeno kao *kanđiji* za svakodnevnu upotrebu (japan-guide.com 2017).


*Slika 1- Ideogrami japanskog jezika*

*Kanđiji* se koriste za pisanje imenica, prideva, priloga i glagola (japan-guide.com 2017).

### 3.1.2 Kana

Za razliku od kineskog jezika, u japanskom jeziku se ne može sve napisati uz pomoću ideograma. Za gramatičke završetke i reči bez odgovarajućih ideograma postoje dva slogovna pisma od kojih se svaka sastoji od 46 znakova, a to su *hiragana* i *katakana* (japan-guide.com 2017), koje su nastale oko devetog veka nove ere od *kanđija.Hiragana* i *katakana* se razlikuju po stilu – *hiragana* je obla, dok je *katakana* uglovita (japan-guide.com 2017). Na slici 2 se mogu videti prikazi obe kane u japanskom jeziku: katakana i hirakana.


*Slika 2- Slogovna slova japanskog jezika*

Oba pisma se sastoje od pet samoglasnika, a ostali slogovi su sastavljeni kombinujući ih sa suglasnikom. Međutim, jedini izuzetak je nazalno *N*. Određeni slogovi se mogu promeniti u druge dodavanjem dve crtice ili kružića u gornji desni ugao sloga. Iako se u teoriji ceo japanski jezik može ispisati *hiraganom*, ona je najčešće koristi samo kod gramatičkih završetaka glagola, imenica i prideva, kao i kod rečca i izvornih japanskih reči, dok se *katakana* koristi za pozajmljenice i strana imena koja ne mogu biti napisana ideogramima (japan-guide.com 2017).

## 3.2 Korpusi japanskog

Do negde oko 2007. godine, moglo se primetiti kako japanski jezik, iako sam po sebi bogatog rečnika, nije imao mnogo pretraživih i javno dostupnih korpusnih resursa u poređenju sa drugim svetskim jezicima (Erjavec, Kilgarriff i Srdanovic 2007). Međutim, takvo stanje korpusa japanskog jezika je u konstantnoj promeni poslednjih deset godina i danas se na svetskoj mreži može naći na desetine različitih korpusa i drugih resursa vezanih za japanski jezik, kako savremeni, tako i stari, kao i korpusi na drugim jezicima koji sadrže terminologiju iz japanskog jezika.

Mnogi od ovih korpusa se mogu naći na sajtu NINJAL[19]-a, instituta koji je osnovan davne 1948. godine kao nezavisna administrativna agencija i koji je obnovljen 2009. godine kao šesta organizacija među-univerzitetskog istraživačkog instituta korporacije „Nacionalnih Instituta za Društvene nauke“. Ovaj institut je zaslužan za sprovođenje velikog broja teorijskih i empiričkih studija kao međunarodni centar za istraživanje sa ciljem da otkrije sve aspekte japanskog jezika i produbi naše razumevanje ljudskih bića kroz proučavanje jezika, ali ima i zadatak da distribuira rezultate kolaborativnog istraživanja sa javnošću i time promoviše njihovu primenu u raznim poljima poput učenja japanskog jezika i procesovanje prirodnog jezika (National Institute for Japanese Language and Linguistics 2017).

BCCWJ[20] je javno dostupni korpus koji je stvoren u svrsi pokušaja da se razume širina savremenog japanskog jezika i koji sadrži obimne uzorke savremenih japanskih tekstova kako bi se stvorio jedinstveno uravnotežen korpus koliko je to moguće. Njegovi podaci se sastoje od više od 104 miliona reči koji potiču iz poslovnih izveštaja, blogova, knjiga, časopisa, novina, udžbenika, pravnih dokumenata i drugih vrsta izvora iz kojih su izvučeni nasumični uzorci. Morfološka analiza je sprovedena koristeći leksičke stavke iz svakog uzorka, a *tagovi* koji se odnose na strukture rečenice i precizne bibliografske informacije su obezbeđene (National Institute for Japanese Language and Linguistics 2017).

Sa druge strane, postoji i OCOJ[21], koji je dugotrajni istraživački projekat sa Univerziteta u Oksfordu sa ciljem da razvije obiman pribeležen digitalni korpus svih tekstova na starojapanskom jeziku perioda Asuka i Nara[22], koji su bili formativni period za razvoj japanske civilizacije i čiji su tekstovi od ključne važnosti za proučavanje i razumevanje porekla i razvoja japanskog jezika, pisama, književnosti, religija, istorije i kulture. Korpus je još uvek u izradi, a u planirani sadržaj su uključeni tekstovi, anotacije, prevodi i rečnik (Faculty of Oriental Studies, University of Oxford 2017).

[19]National Institute for Japanese Language and Linguistics – Nacionalni institut za japanski jezik i lingvistiku

[20]Balanced Corpus of Contemporary Written Japanese – Uravnoteženi korpus savremenog japanskog pisanog jezika

[21]Oxford Corpus of Old Japanese – Oksfordski korpus starojapanskog jezika

[22]Od šestog do osmog veka nove ere

Zatim, tu je CSJ[23], baza podataka koja sadrži veliku zbirku podataka o govornom japanskom jeziku i informacijama za korišćenje u lingvističkim istraživanjima, koju zajedničkim naporima razvijaju NINJAL, NICT[24] i Tokijski Institut za Tehnologiju. Korpus je korišćen za raznorazne istraživačke svrhe kao što su procesiranje govornog i prirodnog jezika, fonetika, psihologija, sociologija, japansko obrazovanje i dopunjavanje rečnika (Center for Corpus Development, NINJAL 2017).

NPCMJ[25] je produžetak *Keyaki Treebank*[26]-a i sintaktički i semantički pribeležen korpus pisanog i govornog savremenog japanskog jezika koji je stvoren u svrhu pretraživanja i izvlačenja japanskih reči funkcije, frazalnih struktura, klauzula i gramatičkih obrazaca iz velike količine jezičkih podataka. Radi prioritizovanja svestranosti, ovaj *treebank* korpus se ugleda na metode pribeležavanja *Penn Historical Corpus*-a (Pardeshi 2017).

## 3.3 Korpusi prevoda sa japanskog

Od korpusa koji sadrže prevode sa japanskog, najznačajniji su paralelni korpusi japanskog i engleskog jezika, kao što je, recimo, japansko-engleski pravni paralelni korpus, koji je 2014. godine sastavio Graham Neubig i postavio na internet stranicu www.phontron.com/jaen-law da bude dostupan svima. Taj korpus sadrži oko 260 000 rečenica (Neubig 2014).

JESC[27] je proizvod saradnje između Univerziteta Stanford, *Google Brain*-a i Rakuten Instituta za Tehnologiju.Stvoren je pretraživanjem interneta za filmske i televizijske titlove i njihovim poravnanjem (paralelizacijom) i predstavlja jedan od najvećih japansko-engleskih korpusa dostupnih svima, a uz to još pokriva i retko pokriven domen kolokvijalnog jezika. JESC ima za cilj da podrži istraživanje i razvoj sistema mašinskog prevođenja, izvlačenja informacija i drugih tehnika obrade jezika. Sačinjen je od 3,2 miliona paralelnih rečenica (Stanford University, Google Brain & Rakuten Institute of Technology n.d.).

---

[23]Corpus of Spontaneous Japanese – Korpus spontanog japanskog jezika

[24]National Institute of Information and Communications Technology – Nacionalni institut za informacione i komunikacijske tehnologije

[25]NINJAL Parsed Corpus of Modern Japanese – Raščlanjeni korpus savremenog japanskog jezika Nacionalnog instituta za japanski jezik i lingvistiku

[26]*Treebank*je korpus sa sintaktičkim anotacijama.

[27]Japanese-English Subtitle Corpus – Japansko-engleski korpus titlova

JST[28] je u saradnji sa NICT-om[29] sastavila ASPEC[30], koji se sastoji od japansko-engleskog korpusa od 3 miliona paralelnih rečenica (ASPEC-JE) i japansko-kineskog korpusa naučnih izvoda (apstrakata) od 680 000 paralelnih rečenica (ASPEC-JC). ASPEC-JC je jedan od dostignuća japansko-kineskog projekta za mašinsko prevođenje koji je izvođen u Japanu od 2006. do 2010. godine. Pojavila se potražnja za mašinsko prevođenje naučnih papira usled sve većeg broja objavljenih naučnih radova, a ASPEC je prvi paralelni korpus koji se usresređuje na to i teži da promoviše istraživanje mašinskog prevođenja u domenu naučnih radova (Japan Science and Technology Agency 2015).

Pored navedenih korpusa, nađena je jedna korpusna analiza *haiku* poezije i njenih prevoda na engleski jezik. Naime, reč je o „A crow on a bare branch: A comparison of Matsuo Basho's *haiku Kare-eda-ni...* and its English translations", u kojem istraživač, Elin Sütiste, upoređuje Bašoov originalnu *haiku* pesmu *Kare-eda-ni...* i njenih 32 prevoda na engleski jezik, koji su pravljeni između 1899. i 2000. godine. U svom radu, istraživač objašnjava kakve sve varijacije postoje među prevodima, ali navodi i karakteristike kakve postoje u originalu, kao što su *kiređi*[31], nedostatak naslova i znakova interpunkcije, *kigo* i odstupanje od slogovnog formata 5-7-5. Istraživač u svom radu izdvaja numerisane prevode koji štrče u poređenju jedni sa drugim, kao što su oni koji sadrže naslove, znakove interpunkcije, one koje su verni prevodu po broju slogova i one koji su verni ustaljenom slogovnom formatu *haiku* pesme. Izdvojena su tri prevoda kojima su dati naslovi, četiri prevoda koji sadrže sveukupno 19 slogova, jedan od kojih zadržava slogovni format 5-9-5 iz originala, četiri prevoda koji sadrže samo 10 slogova, jedan prevod koji sadrži čak 31 slog, dva koji održavaju 17 slogova kao tipična *haiku* pesma, jedan koji ima isti slogovni obrazac kao original, jedan prevod koristi svojstveni *kiređi,* sedam koji ne koriste nijedan znak interpunkcije, osam koji koriste crticu, šest koji koriste dve tačke, četiri koji koriste tačka-zarez, dva koja koriste zarez, dva koja koriste tri tačke i tri koja koriste tačku (Sütiste 2001). Ovi radovi su poslužili kao inspiracija za neke od sprovedenh analiza u ovom radu.

---

[28]Japan Science and Technology Agency – Agencija za Nauku i Tehnologiju Japana

[29]National Institute of Information and Communications Technology – Nacionalni Institut za Informacije i Komunikacionu Tehnologiju

[30]Asian Scientific Paper Excerpt Corpus – Azijski korpus apstrakata naučnih radova

[31] Rečca koja označava prekid rečenice.

# 4 Opis prikupljenog korpusa

## *4.1* Kolekcija *haiku* poezije

Od prikupljenog korpusa se nalaze četiri zbirke tradicionalne japanske haiku poezije iz perioda Edo[32], kada je u izolovanom Japanu nastala *haiku,* odnosno *haikai,* poezija (Sütiste 2001). Ovaj materijal u vidu datoteka u *pdf* formatu, su skenirane stranice literature koja mi je bila neophodna za osnovne akademske studije iz japanskog jezika, kulture i književnosti, a sa japanskog iz perioda Edo ih je preveo profesor Hiroši Jamasaki-Vukelić, jedan od najistaknutijih prevodilaca sa japanskog na srpski i sa srpskog na japanski i jedna od veoma retkih osoba na našim prostorima sa poznavanjem starojapanskog jezika[33].

U planu je bila veća kolekcija korpusa koji je takođe sačinjen od prevoda sa japanskog na srpski u *pdf* formatu, ali su samo ove četiri zbirke uspešno prošle kroz OCR za dalju digitalnu analizu njihovih podataka. Zbirke nose nazive „Uska staza ka dalekom severu“, „Vrapčeva priča“, „Prolećno more“ i „Svenulo polje“.

Najobimnije i najpoznatije od navedenih dela je „Uska staza ka dalekom severu“ jednog od najvećih *haiku* majstora Macuo Bašoa i ono po čemu se razlikuje od drugih zbirki koje su izabrane za analizu jeste to što ona predstavlja književni oblik *haibun*[34] i jednu vrstu putopisa[35], a *haiku* pesme u njemu su u drugom planu i ima ih manje nego u ostale tri zbirke – pedeset, a napisali su ih sam Bašo i njegov učenik Sora (p. H.-V. Macuo Bašo 2012).

„Svenulo polje“ je, sa druge strane, zbirka izabrane haiku poezije Macuo Bašoa u pravom smislu te reči i sadrži osamdeset i jednu *haiku* pesmu koja je raspoređena u pet ciklusa, koji

---

[32]1600.-1868, takođe poznat i kao period Tokugava. (Bleed, i dr. 2003)

[33]Profesor Jamasaki-Vukelić je, zajedno sa profesorima Danijelom Vasić, Dalimirom Kličkovićem i Divnom Glumac (Tomić), preveo najstariji sačuvani japanski zapis, *Kođiki,* sa starojapanskog na srpski jezik (goodreads 2017).

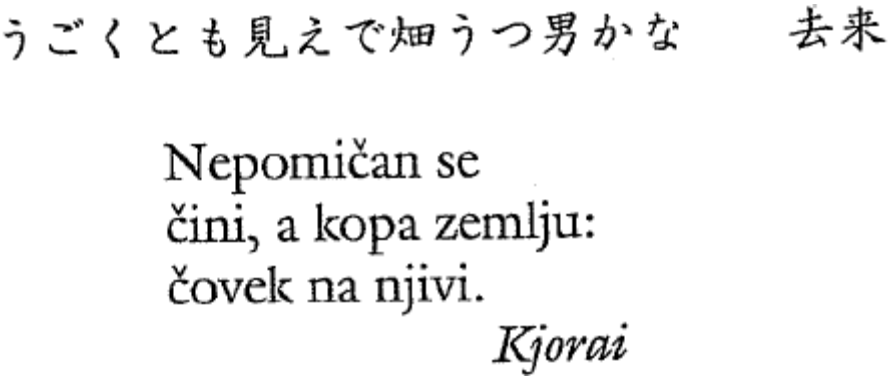
[34]Esejistički tekst sa *hoku* pesmom (Bašo, Jamasaki-Vukelić i Mitrović 2008).

[35]*Usku stuzu ku Dalekom severu* nije putopis u smislu dokumentarnog zapisa, već je to sračunato rekonstruisan sled događaja. Jednostavno rečeno, to je - fikcija. Kako je, poređenjem sa Sorinim *Dnevnikom,* otkrivenim 1933. godine, nedvosmisleno utvrđeno, Bašo je izostavio jedne, izmenio druge i izmislio treće događaje (Bašo i Jamasaki-Vukelić, Uska staza ka dalekom severu 2012).

su sačinjeni od godišnjih doba i Nove godine kao posebnog perioda u jednoj godini (p. H.-V. Macuo Bašo 2008).

„Prolećno more“ je slična vrsta zbirke kao „Svenulo polje“ po tome što se u njemu takođe nalazi osamdeset i jedna *haiku* pesma koja je, opet, raspoređena u istih pet ciklusa, ali je ovaj put reč o pesmama *haiku* majstora Jose Busona, koji se smatra najvećim *haiku* pesnikom posle Bašoa. Ova zbirka je takođe bila predmet računarske analize u mom prethodnom radu (Josa Buson 1999).

„Vrapčeva priča“ je, pak, kompilacija sačinjena od osamdeset i jedne *haiku* pesme trideset i troje *haiku* pesnika iz perioda Edo osim Macuo Bašoa i Jose Busona, od kojih je najpoznatiji i najzastupljeniji Kikaku (Jamasaki-Vukelić 2011). Na slici 3 je prikazan primer haiku pesme iz zbirke „Vrapčeva priča“.


うごくとも見えで畑うつ男かな　　去来

Nepomičan se
čini, a kopa zemlju:
čovek na njivi.

*Kjorai*

*Slika 3 Primer haiku pesme iz zbirke „Vrapčeva priča“*

Ovi korpusi imaju tri zajedničke karakteristike:

- Svaki od njih je preveden sa japanskog jezika iz perioda Edo;
- Svaki od njih je preveden od strane profesora Hirošija Jamasakija-Vukelića[36];
- Svaki od njih sadrži tradicionalnu *haiku* poeziju.

Pored ovih zbirki *haiku* poezije prevedenih na srpski jezik, u korpus su dodate dve zbirke tradicionalne *haiku* poezije prevedenih na engleski jezik, jedna je zbirka *haiku* pesama Macuo Bašoa (Barnhill 2004), a druga je zbirka *haiku* pesama više autora (Yasuda 2001).

Da bi se razumele pojedinosti vezane za *haiku* poeziju, potrebno je objasniti njenu formu.

[36] „Prolećno more“ i „Svenulo polje“ su takođe prevođeni u saradnji sa Srbom Mitrovićem.

*Haiku*, odnosno komični stih, je tradicionalna japanska poetska forma, najkraća od svih, sa svega sedamnaest slogova raščlanjenih u tri stiha, odnosno na tri člana - 5/7/5. Zbog toga se za *haiku* može reći daje „trostih" ili „tročlani stih". Izvesno odstupanje, višak ih manjak slogova, dopušteno je kada za to ima opravdanje. Pored te relativno strogo određene forme, za haiku je karakteristično da sadrži *kigo*, tj. reč ili izraz koji određuje godišnje doba u pesmi, kao na primer, trešnjev cvet (proleće), setva pirinča (leto), zadušnica (jesen) ili svenulo polje (zima). *Kigo* je nosilac senzibiliteta prema godišnjim dobima koji je negovan kroz vekove i izvor je višeslojnih asocijacija koje dele pesnik i čitalac. Haiku bez znaka godišnjeg doba je moguć, ali je prava retkost. Na kraju, u *haiku* pesmi treba da postoji rečca prekida - *kiređi*. On je tu da bi se *haiku* razlikovao od prvog dela duže poetske forme *vaka* (5/7/ 5/7/7) iz koje je nastao. Taj prekid ne mora da se pojavi samo na kraju pesme. On može da bude iza prvog ili drugog, kao i usred dragog ili trećeg stiha i lomi pesmu na dva dela te čini njenu struktura složenom. U prevodu Hirošija Jamasakija-Vukelića i Srbe Mitrovića „lom" u strukturi pesme se prikazuje tačkom, dvema tačkama, ili nekim dragim znakom interpunkcije (p. H.-V. Macuo Bašo 2008).

U vreme pisanja ovih tradicionalnih *haiku* pesama, japanski jezik nije poznavao znake interpunkcije. To je za savremenog čitaoca veliki problem, jer ne može lako da sazna gde je kraj rečenice, gde počinje a gde se završava citat ili upravni govor itd. Zato su kasniji redaktori uneli u tekst tačke, zapete, znake navoda, kao i podelu na pasuse i odeljke, i to po sopstvenom tumačenju. To je dovelo do izvesnih razlika među izdanjima (p. H.-V. Macuo Bašo 2012).

*Haiku* vuče koren iz *vaka* pesništva. *Vaka*[37] ili *tanka*[38] je reprezentativna poetska forma sa trideset jednim slogom (577/5/7/7). U zbirci pesama *Kokin vakašu* (oko 905. godine) postoji rubrika *haikaika*[39]. U kasnijem periodu, odvajanjem prvog i dragog dela *vaka* pesme, nastaje kolektivna pesnička igra *renga*[40], u kojoj prvi učesnik daje početne stihove *hoku* (5/7/5) a drugi sledeća dva (7/7) nadovezujući ih na prethodne i tako redom. Prva zbirka *renga* pesama

[37]Japanska pesma

[38]Kratka pesma

[39]Komična pesma

[40]Niska pesama

pojavila se 1357. godine. Pošto se za uspešno nadovezivanje na prethodne stihove tražila, između ostalog, i duhovitost, neke od tih niski svrstane su u komične - *haikai no renga*, iz kojih se početkom šesnaestog veka razvija niska komičnih stihova - *haikai no rengu* kao poseban književni oblik. Kasnije će se reč *haikai* upotrebljavati gotovo isključivo za ovu vrstu književnosti (p. H.-V. Macuo Bašo 2008).

*Haiku* pesme tradicionalno nemaju svoje samostalne naslove, ali mnogi prevodi dodaju haiku pesmama naslove kako bi ih približili tekstove pesama ciljnoj publici. Takođe, zbog relativne jednostavnosti fonetskog i slogovnog sistema, kao i nedostatka naglašavanja reči u japanskom jeziku, u japanskoj poeziji ne postoji rima (Sütiste 2001).

Na početku, *haiku* je bio naziv za sve stihove u niski komičnih stihova. Ali veliki reformator tradicionalne japanske poezije Šiki (1867-1902) insistirao je da se početni stihovi *hoku* (5/7/5) potpuno osamostale pod nazivom *haiku*, priznajući samo njemu literarnu vrednost (p. H.-V. Macuo Bašo 2008).

## 4.2 Kolekcija ostalih vrsta poezije i proze

Od ostalih vrsta tekstova u ovom korpusu treba izdvojiti pre svega putopisni prozni tekst – *haibun*, koji prati *haiku* poeziju Macuo Bašoa i koji je preveden i na srpski (p. H.-V. Macuo Bašo 2008) i na engleski jezik (Barnhill 2004).

Što se ostalih vrsta poezije tiče, pre svega treba spomenuti *vaka* odnosno *tanka* poezijuu dve najstarije zbirke japanske poezije *Man'jošu*(Donald Keene 2000) i *Kokinšu*(Henkenius 1984)[41]koje su prevedene na engleski jezik, kao i zbirka pesama monaha-pesnika Rjokana (Yuasa 1981) iz poznog perioda Edo, koja, pored *vaka* poezije, sadrži tek mali broj *hokku* pesama i malo putopisnog proznog teksta.

## 4.3 Kolekcija srpskog

Za potrebe poređenja korpusa prevoda sa srpskog na japanski, potrebna je određena kolekcija srpskog jezika. Za tu svrhu su nađene *haiku* pesme domaćih autora sa *WordPress* stranice *Beleg* (Lujak n.d.), kao i poezija velikih srpskih pesnika kao što su Branko Miljković, Desanka Maksimović, Đura Jakšić, Jovan Dučić, Miloš Crnjanski i Vladislav Petković Dis, čije su pesme za korpus uzete sa stranice internet stranice *Poezija noći* (Poezija noći n.d.).

[41]Prvobitno poznata kao *Kokin Vakašu*.

# 5 Rezultati lingvističke analize

## 5.1 Analize prevoda na srpski jezik

Budući da je izvorna ideja bila da se skeniraju prevodi japanskog teksta na srpski jezik, ali je sačuvani korpus uglavnom sačinjen od haiku poezije, napravljene su dve analize – jedna kojom se analizira celokupan korpus prevoda japanskog teksta na srpski i druga koja se bavi analizom isključivo haiku pesama, gde su prozni tekst i drugi oblici poezije izostavljeni. Prva analiza bi se bavila nekim generalnim karakteristikama kao što su izdvajanje reči koje rečnik ne prepoznaje, brojeva određenih vrsta reči i njihov procenat u celom tekstu, dok bi se druga takođe bavila znacima interpunkcije, odnosno njihovom zasebnom učestalošću u srpskom prevodu tradicionalne japanske haiku poezije kako bi se uporedili sa istraživanjem koju je sprovela Elin Sütiste nad raznim prevodima Bašoove poezije na engleski jezik.

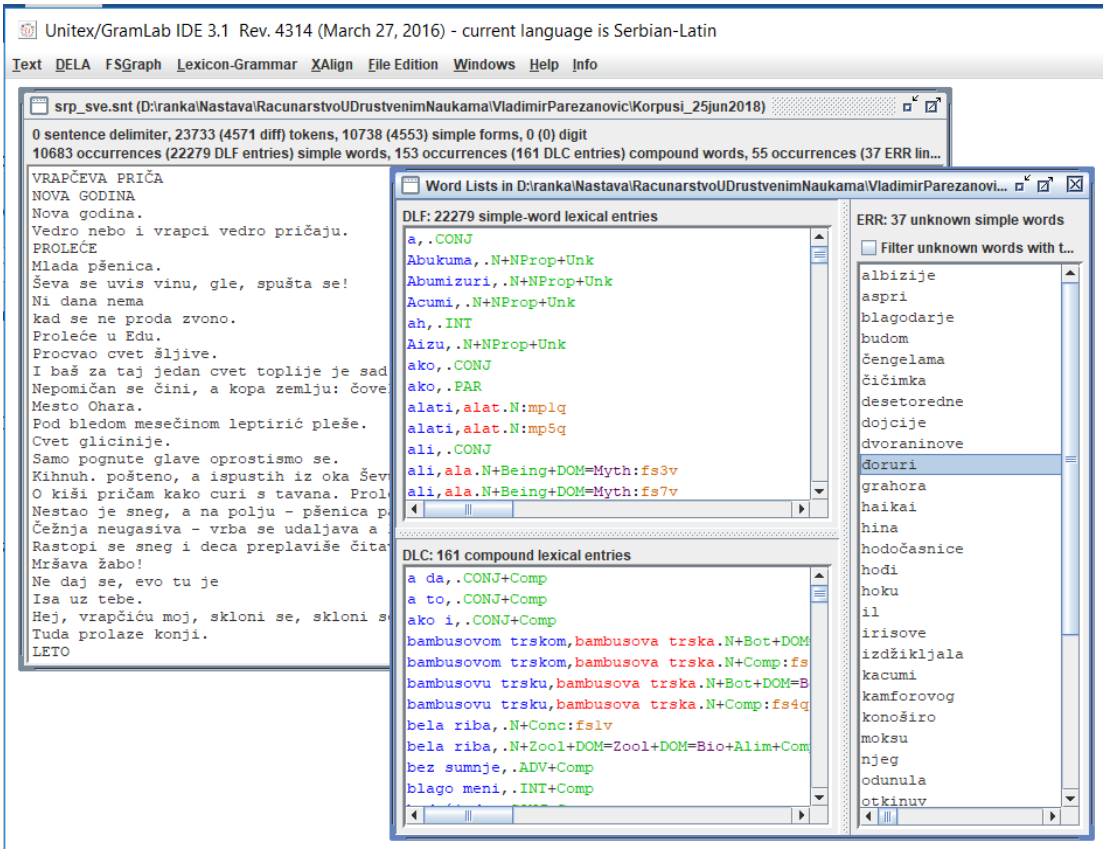
### *5.1.1* Analiza teksta koji nije *haiku*

Pretprocesiranjem elektronskim rečnicima celokupnog korpusa japanskog prevoda na srpski jezik, tekst je obeležen sa 673 oznake za kraj rečenice, pronađeno je 25616 tokena, od toga 4604 različitih, od čega je 10789 oblika prostih reči (4575 različitih), 634 cifara (10 različitih). Imajući u vidu jedan oblik može da ima različite gramatičke interpretacije u elektronskom rečniku, na primer:

oblasti,oblast.A+DOM=Geol+FLX=A6:adms1g
oblasti,oblast.A+DOM=Geol+FLX=A6:adms4q
oblasti,oblast.A+DOM=Geol+FLX=A6:aems5g
oblasti,oblast.A+DOM=Geol+FLX=A6:aemp1g
oblasti,oblast.A+DOM=Geol+FLX=A6:aemp5g

to sistem izveštava da ima 22306 oblika prostih (monoleksemskih) reči povezanih sa lemama, odnosnokanonskim oblicima reči ili tipova reči kako se navodi u nekim istraživanjima. Pojavljivanja 79 složenih (polileksemskih) reči ima 229 interpretacija u elektronskom rečniku.

Na slici 4 je prikazano radno okruženje Unitex sa korpusom haiku poezije gde se može videti osnovna statistika primene elektronskih rečnika na tekst. U okviru na desnoj strani slike se prikazuju prepoznate monleksemske reči, polileksemske reči i neprepoznate reči.




*Slika 4 - Korpus haiku poezije obrađen u Unitex-u*

Daljom analizom je utvrđeno da se prepoznaje 4151 imenicu, što pokriva oko 17% teksta, 2970 glagola (12% teksta), 1511 prideva (6% teksta), 938 priloga (4% teksta).

Da bi se uz pomoću programa *Unitex* našle ove četiri vrste reči, korišćena je opcija *Locate Pattern*[42]i pretražen je u *Regular Expression*[43] leksičkim maskama: <N> za imenice, <V> za glagole, <A> za prideve i <ADV> za priloge. Ipak, postoje greške u pretrazi koje treba napomenuti. Naime, prilikom pretrage imenica u korpusu, među rezultatima se 91 put pojavljuje „su“ kao imenica, onoliko puta koliko se pojavljuje kao oblik glagola „jesam“.

Zatim, red je da se dalje analiziraju i izdvoje nepoznate reči. To je postignuto jednostavnom pretragom <!DIC> za nepoznate reči u korpusu. Dobijeni rezultati se razlikuju od početnih i dobija se da ima sveukupno 37 nepoznatih reči, odnosno reči (tokena) kojih nema u rečniku. Navešću ovde tokene koji nisu prepoznati u rečniku, a pritom nisu vlastite imenice, i podvući ih u svojim delovima rečenice, osim ukoliko su deo haiku stihova, kada se navodi ceo haiku.

[42] Srp. - Pronaći obrazac

[43] Srp. – Uobičajeni izraz

*„U Kisagati.*
*Cvet albizije na kiši*
*ko setna Seiši.“*
*„Ledena bura*
*na škrgu duva: riba*
*na čengelama.“*
*„Roj komaraca,*
*tamo gde sa čičimka*
*opada cveće.“*
*„Da kucnem tu gde*
*nema cveća dojcije.*
*Vrata u mraku."*
*„U rukavima*
*dvoraninove halje*
*miču se svici.“*
*„Pod istim krovom*
*spavaše i bludnice.*
*Grahor i mesec.“*
*„Četvrti dan smo proveli sastavljajući venac haikai stihova u glavnom hramu.“*
*„Baciše tri niske aspri za piće...“*
*„...i njeno blagodarje preplavljuje svaki kutak sveta...“*
*„Pred budom Amidom u mojoj kolibi...“*
*„...od tog šaša prave desetoredne prostirke...“*
*„I koliba mi*
*sada promeni gazdu.*
*Dom za hina lutke.“*
*„Tu behu još i hoku pesme Sanpua i Đokušija.“*
*„Oblak cvetova.*
*Zvoni li u Uenu,*
*il' Asakusi?“*
*„Iz njegovog korena, na zemlji su izdžikljala dva odvojena stabla...“*
*„...zamenim vrpcu na šeširu i zapalim moksu na golenicama...“*
*„S njeg berem vodeni peršun.“*
*„Studena bura -*
*odunula kamenčić;*
*pade na zvono.“*
*„Rukom dotaknuh,*
*ne otkinuv a prođoh*
*- slezova ruža.“*
*„Kakva je radost*
*pregazit letnju reku!*
*Sandale u ruci.“*
*„Za oči zeleno, kukavica u gori, rani tunj prugavac.“*
*„Ona kao da u sebi nosi neku setu, pojačanu pustošju...“*
*„Oj, te žalosti!*
*Pod ratnikovim šlemom*
*sad peva cvrčak.“*
*„Brzo li teče*
*skupljajuć majsku kišu -*
*reka Mogami.“*
*„I majska kiša*
*ostavlja u suvoti*
*paviljon Svetlost.“*

*„Na tržici još*
*širi se neki zadah.*
*Sja letnji mesec.“*
*„...a Hara Anteki mi pokloni svoju vaka pesmu o ostrvlju u Borovom zalivu.“*
*„Osnovao ga je zen učitelj Dogen.“*

Kao što se može videti iz priloženih primera, kada se izuzmu vlastite imenice, naročito one na stranom jeziku, jeste da tuđica iz japanskog jezika, u stvari, nema u velikom broju u ovim prevodima. Reči kojih nema u rečniku su podvučene. Japanske reči su *hina*, *hoku*, *haikai*, *vaka* i *zen*. Njihov mali broj pokazuje tendenciju korišćenja srpskih reči gde god je moguće. Nepoznatih reči koje nisu vlastite imenice u ovom korpusu ima 26, od kojih se neke ne bi našle tu da nisu određene vrste krnjih oblika glagola, zamenica ili rečci, kao što su „skupljajuć“, „pregazit“, „il'“ i „njeg“. Razlog za ovakve krnje oblike inače prepoznatljivih reči u prevodima tradicionalne haiku poezije jeste najverovatnije da bi se zadržala slogovna forma tipična za haiku pesme, jer bi se dodavanjem nedostajućih samoglasnika dodao i slog koji bi narušio tu formu. Takvih reči nema u proznom delu teksta.

### 5.1.2 Analiza *haiku* poezije

Druga analiza, koja se bavi znacima interpunkcije je jednostavnija za sprovođenje i ima za svrhu da se uporedi sa istraživanjem koju je Elin Sütiste sprovela nad prevodima „*Kare-eda-ni...*“ i da izdvoji brojeve znaka interpunkcije u prevodima *haiku* pesama koje je radio profesor Hiroši Jamasaki-Vukelić.

Jednostavna analiza koja ima za svrhu da izdvoji znake interpunkcije je postignuta pretragom svih tokena leksičkom maskom <TOKEN>. Od znakova interpunkcije dobijeni su sledeći rezultati:

- 789 tačaka;
- 1017 zareza;
- 20 znakova pitanja;
- 40 tačka-zareza;
- 97 dvotačaka;
- 53 crtice.

Drugih znakova interpunkcije koji postoje u ustraživanju Elin Sütiste nema u prevodima Hirošija Jamasakija-Vukelića. U poređenju sa prevodima koje je prikupila Elin Sütiste, u kojima sedam ne koriste nijedan znak interpunkcije, osam koriste crticu, šest koriste dve tačke, četiri koriste tačka-zarez, dva koriste zarez, dva koriste tri tačke i tri koriste tačku, može se primetiti da prevodi profesora Hirošija Jamasakija-Vukelića češće koriste određene

znakove interpunkcije u odnosu na to koliko su frekventno koristili prevodioci jedne određene *haiku* pesme – „*Kare-eda-ni*...“. Ipak, reč je o stilovima prevoda sa japanskog, ali su jezici na koje su prevodi pravljeni različiti – reč je o srpskom i o engleskom, tako da bi svakako morala da postoje određena prilagođavanja u prevodima shodna jezicima na koje se prevodi, ali u kojima se manje ili više žrtvuje određena forma prilikom prevođenja.

### 5.1.3 Poređenje srpske poezije i japanskog *haiku*-a

Za potrebe poređenja korpusa prevoda japanskog jezika na srpski jezik sa korpusom srpskog jezika, pripremljen je korpus srpske poezije koji sadrži, pored pesama nekoliko velikih imena srpske poetske scene, i *haiku* poeziju napisanu na srpskom i jedan *haibun.*

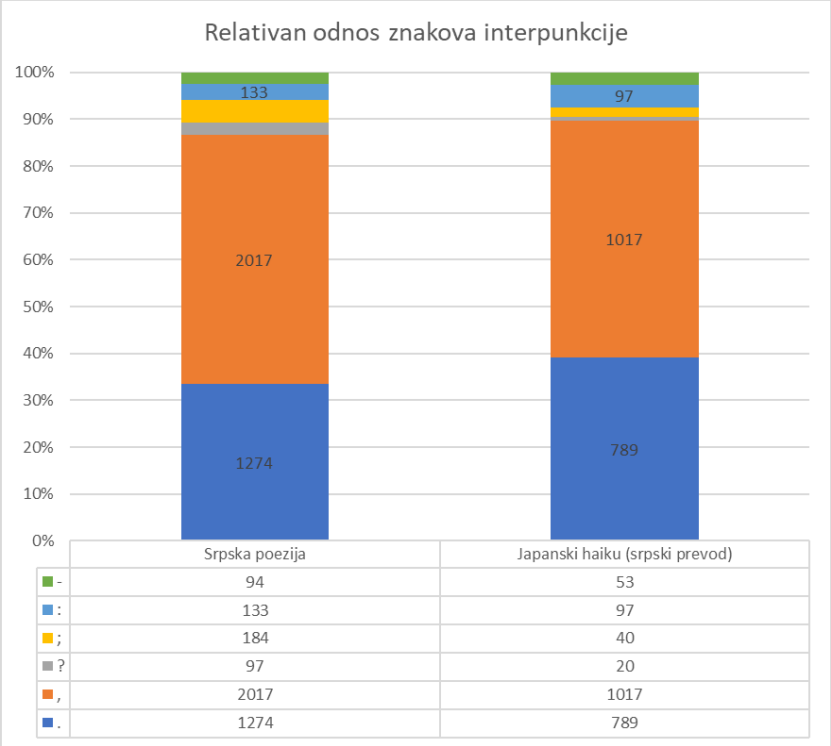
Analizom korpusa autohtone srpske poezije je utvrđeno da se prepoznaje 7087 imenica, što pokriva oko 16,4% teksta, 5320 glagola (12% teksta), 2904 prideva (6,7% teksta), 2032 priloga (4,7% teksta).

Kada se uporede procenti ove četiri vrste reči između korpusa prevoda japanskog teksta na srpski jezik i korpusa autohtone srpske poezije, može se primetiti određeno poklapanje. Naime, u korpusu prevoda, imenice pokrivaju oko 17% teksta, a u autohtonom korpusu 16,4%; pridevi u korpusu prevoda pokrivaju 6% teksta, a u autohtonom korpusu 6,7%; u korpusu prevoda, prilozi sačinjavaju 4% teksta, a u autohtonom korpusu 4,7%; u oba korpusa glagoli sačinjavaju 12% teksta.

Jednostavnu analizu koja ima za svrhu da izdvoji znake interpunkcije koja je postignuta pretragom svih tokena leksičkom maskom <TOKEN> je ponovo iskorišćena. Od znakova interpunkcije dobijeni su sledeći rezultati:

- 1274 tačaka;
- 2017 zareza;
- 97 znakova pitanja;
- 184 tačka-zareza;
- 133 dvotačaka;
- 94 crtice.

Na slici 5 se vidi relativan odnos znakova interpunkcije u korpusima srpske poezije i prevoda japanske haiku poezije na srpski jezik, iz koga se mogu uočiti sličnosti u procentualnim odnosima znakova interpunkcije ova dva korpusa poezije na srpskom jeziku, sa izuzetkom povećane učestalosti znaka pitanja i tačka-zareza u korpusu srpske poezije.




*Slika 5 - Relativan odnos znakova interpunkcije u korpusima srpske poezije i prevoda japanske haiku poezije na srpski jezik*

### 5.1.4 Poređenje *haiku* poezije sa korpusom proze

Izdvajanje ključnih reči korpusa haiku poezije je urađeno korišćenjem statističkih mera kojima se izražavaju razlike frekvencije reči u odnosu na referentni korpus, u ovom slučaju korpus Srpskog jezika od 23 miliona reči (Utvić 2011). Korišćena je mera „ključnosti“

$$\text{Keyness} = \frac{RelativFrek_{Haiku}+1}{RelativFrek_{SrpKor}+1}$$

Relativna frekvencija je frekvencija reči „na milion reči“ i računa se kao količnik apsolutne frekvencije i broja reči u korpusu, pomnoženo sa milion.

Ova mera omogućava pronalaženje reči koje predstavljaju fokus teksta, odnosno pojavljuju se značajno češće nego u standardnom jeziku, kao i reči čije je pojavljivanje ređe nego u standardnom jeziku. Korišćenjem alata Leximir i elektronskih rečnika za srpski jezik (Stanković, et al. 2016) su izračunate apsolutne frekvencije za korpus *haiku* (oznaka AFr), korpus srpskog jezika (AFd), odgovarajuće relativne frekvencije RFr i RFd i keyness. Na slikama 6, 7 i 8 se redom daju primeri za imenice, prideve i glagole koje su najfrekventnije u svakom od korpusa.

| lema | POS | keyness | RFr | RFd | AFr | AI |
|---|---|---|---|---|---|---|
| taka | N | 1696.1 | 1.1 | 3538.8 | 24 | 38 |
| svetilište | N | 295.7 | 5.9 | 2048.8 | 131 | 22 |
| divlja ruža | N | 274.2 | 0.4 | 372.5 | 8 | 4 |
| glicinija | N | 268.3 | 0.0 | 279.4 | 1 | 3 |
| grahorica | N | 268.3 | 0.0 | 279.4 | 1 | 3 |
| koliba | N | 222.3 | 7.4 | 1862.5 | 163 | 20 |
| šaš | N | 212.9 | 0.3 | 279.4 | 7 | 3 |
| cvrčak | N | 201.7 | 1.3 | 465.6 | 29 | 5 |
| halja | N | 193.0 | 0.5 | 279.4 | 10 | 3 |
| rusomača | N | 187.3 | 0.0 | 186.3 | 0 | 2 |
| zaseok | N | 187.3 | 0.0 | 186.3 | 0 | 2 |
| uliv | N | 187.3 | 0.0 | 186.3 | 0 | 2 |
| riža | N | 179.2 | 0.0 | 186.3 | 1 | 2 |
| hrizantema | N | 175.2 | 1.1 | 372.5 | 25 | 4 |
| kreket | N | 171.6 | 0.1 | 186.3 | 2 | 2 |
| svilara | N | 171.6 | 0.1 | 186.3 | 2 | 2 |
| čajdžinica | N | 171.6 | 0.6 | 279.4 | 14 | 3 |
| slavuj | N | 171.5 | 1.7 | 465.6 | 38 | 5 |

| lema | POS | keyness | RFr | RFd | AFr | AI |
|---|---|---|---|---|---|---|
| primer | N | 0.3 | 374.8 | 93.1 | 8274 | 1 |
| trenutak | N | 0.2 | 385.4 | 93.1 | 8508 | 1 |
| govor | N | 0.2 | 402.3 | 93.1 | 8883 | 1 |
| škola | N | 0.2 | 404.6 | 93.1 | 8933 | 1 |
| igra | N | 0.2 | 416.0 | 93.1 | 9185 | 1 |
| nega | N | 0.2 | 837.7 | 186.3 | 18495 | 2 |
| deo | N | 0.2 | 1273.6 | 279.4 | 28119 | 3 |
| pravo | N | 0.2 | 856.5 | 186.3 | 18910 | 2 |
| vek | N | 0.2 | 467.5 | 93.1 | 10321 | 1 |
| Gor | N | 0.2 | 492.0 | 93.1 | 10862 | 1 |
| snaga | N | 0.2 | 519.9 | 93.1 | 11478 | 1 |
| posao | N | 0.1 | 871.2 | 93.1 | 19234 | 1 |
| rada | N | 0.1 | 1058.0 | 93.1 | 23359 | 1 |
| vlada | N | 0.1 | 1071.2 | 93.1 | 23651 | 1 |

*Slika 6 – Najčešće imenice u prevodu japanske haiku poezije (levo) i srpske proze (desno)*

Analiza ključnosti imenica pokazuje da se u haiku poeziji ćešće javljaju reči taka, svetilište, divlja ruža, glicinija, grahorica,… nego u korpusu srpskog jezika, dok se za imenice kao snaga, posao, vlada može reći da se vrlo retko javljaju u haiku poeziji, bar na osnovu resursa sa kojima je analiza rađena.

Kada su u pitanju pridevi, haiku poeziju karakterišu: bambusov, topal, gorski, koren, pirinčan, opevan, borov, trešnjev,… dok se pridevi jasan, vojni, brz, važan,… javljaju samo jednom u korpusu haiku.

| lema | POS | keyness | RFr | RFd | AFr | AI |
|---|---|---|---|---|---|---|
| bambusov | A | 368.9 | 0.8 | 651.9 | 17 | 7 |
| topal | A | 280.4 | 0.0 | 279.4 | 0 | 3 |
| gorski | A | 232.8 | 1.4 | 558.8 | 31 | 6 |
| koren | A | 167.0 | 0.7 | 279.4 | 15 | 3 |
| pirinčan | A | 167.0 | 0.7 | 279.4 | 15 | 3 |
| opevan | A | 158.3 | 1.4 | 372.5 | 30 | 4 |
| borov | A | 143.5 | 4.8 | 838.1 | 107 | 9 |
| gospodarev | A | 133.0 | 0.4 | 186.3 | 9 | 2 |
| setveni | A | 133.0 | 0.4 | 186.3 | 9 | 2 |
| trešnjev | A | 114.6 | 0.6 | 186.3 | 14 | 2 |
| trščan | A | 111.5 | 0.7 | 186.3 | 15 | 2 |
| rascvao | A | 94.1 | 0.0 | 93.1 | 0 | 1 |
| padni | A | 94.1 | 0.0 | 93.1 | 0 | 1 |
| poredan | A | 94.1 | 0.0 | 93.1 | 0 | 1 |
| suzni | A | 93.8 | 1.0 | 186.3 | 22 | 2 |
| majski | A | 92.6 | 8.1 | 838.1 | 178 | 9 |
| slezov | A | 90.1 | 0.0 | 93.1 | 1 | 1 |
| šljivin | A | 90.1 | 0.0 | 93.1 | 1 | 1 |
| kalemljen | A | 90.1 | 0.0 | 93.1 | 1 | 1 |
| suzan | A | 89.7 | 1.1 | 186.3 | 24 | 2 |
| obrastao | A | 89.5 | 5.3 | 558.8 | 116 | 6 |
| zarudeo | A | 86.3 | 0.1 | 93.1 | 2 | 1 |
| ribarski | A | 85.1 | 4.5 | 465.6 | 99 | 5 |
| zemljopisni | A | 79.7 | 0.2 | 93.1 | 4 | 1 |
| mrzan | A | 79.7 | 0.2 | 93.1 | 4 | 1 |
| lučki | A | 79.3 | 2.5 | 279.4 | 56 | 3 |
| božanstven | A | 77.9 | 1.4 | 186.3 | 31 | 2 |

| lema | POS | keyness | RFr | RFd | AFr | AI |
|---|---|---|---|---|---|---|
| običan | A | 0.5 | 189.8 | 93.1 | 4191 | 1 |
| malen | A | 0.5 | 575.1 | 279.4 | 12698 | 3 |
| tačan | A | 0.5 | 195.9 | 93.1 | 4326 | 1 |
| nedavni | A | 0.5 | 198.6 | 93.1 | 4385 | 1 |
| ostao | A | 0.5 | 804.4 | 372.5 | 17759 | 4 |
| čest | A | 0.5 | 412.0 | 186.3 | 9097 | 2 |
| verovatan | A | 0.5 | 206.7 | 93.1 | 4564 | 1 |
| uspeo | A | 0.4 | 214.2 | 93.1 | 4730 | 1 |
| izuzetan | A | 0.4 | 222.8 | 93.1 | 4920 | 1 |
| uspešan | A | 0.4 | 224.9 | 93.1 | 4966 | 1 |
| budući | A | 0.4 | 230.1 | 93.1 | 5080 | 1 |
| viši | A | 0.4 | 1789.0 | 651.9 | 39498 | 7 |
| siguran | A | 0.4 | 259.9 | 93.1 | 5739 | 1 |
| prethodan | A | 0.3 | 271.1 | 93.1 | 5985 | 1 |
| mnogi | A | 0.3 | 1081.0 | 372.5 | 23867 | 4 |
| poseban | A | 0.3 | 586.1 | 186.3 | 12941 | 2 |
| poslednji | A | 0.3 | 588.3 | 186.3 | 12989 | 2 |
| jasan | A | 0.3 | 306.3 | 93.1 | 6763 | 1 |
| vojni | A | 0.3 | 310.2 | 93.1 | 6848 | 1 |
| brz | A | 0.3 | 345.5 | 93.1 | 7627 | 1 |
| važan | A | 0.2 | 408.6 | 93.1 | 9022 | 1 |
| dobar | A | 0.2 | 1692.9 | 279.4 | 37376 | 3 |
| rad | A | 0.1 | 1431.8 | 186.3 | 31612 | 2 |
| nov | A | 0.1 | 1783.0 | 186.3 | 39365 | 2 |

*Slika 7 - Najčešći pridevi u prevodu japanske haiku poezije (levo) i srpske proze (desno)*

Glagoli: besiti, miti, liti, razvedriti, cvasti, konačiti,… karakterišu haiku poeziju, dok su u njoj retki glagoli kao postojati, raditi, smatrati, trebati, uspeti,…

| lema | POS | keyness | RFr | RFd | AFr | Al |
|---|---|---|---|---|---|---|
| besiti | V | 789.4 | 0.2 | 931.3 | 4 | 10 |
| miti | V | 483.8 | 15.2 | 7822.7 | 335 | 84 |
| liti | V | 453.2 | 4.8 | 2607.6 | 105 | 28 |
| odorati | V | 373.5 | 0.0 | 372.5 | 0 | 4 |
| letati | V | 187.3 | 0.0 | 186.3 | 0 | 2 |
| razvedriti | V | 187.1 | 1.0 | 372.5 | 22 | 4 |
| mesti | V | 170.1 | 17.1 | 3073.2 | 377 | 33 |
| svenuti | V | 128.9 | 0.5 | 186.3 | 10 | 2 |
| opevati | V | 128.6 | 3.4 | 558.8 | 74 | 6 |
| cvasti | V | 103.2 | 0.8 | 186.3 | 18 | 2 |
| konačiti | V | 103.2 | 0.8 | 186.3 | 18 | 2 |
| presti | V | 101.7 | 2.7 | 372.5 | 59 | 4 |
| snevati | V | 100.6 | 0.9 | 186.3 | 19 | 2 |
| pomokriti | V | 98.2 | 0.9 | 186.3 | 20 | 2 |
| koketovati | V | 94.1 | 0.0 | 93.1 | 0 | 1 |

| lema | POS | keyness | RFr | RFd | AFr | Al |
|---|---|---|---|---|---|---|
| podneti | V | 0.4 | 235.5 | 93.1 | 5200 | 1 |
| prihvatiti | V | 0.4 | 247.4 | 93.1 | 5462 | 1 |
| imati | V | 0.4 | 2717.1 | 1024.4 | 59989 | 11 |
| brojati | V | 0.3 | 540.8 | 186.3 | 11940 | 2 |
| završiti | V | 0.3 | 277.3 | 93.1 | 6123 | 1 |
| održati | V | 0.3 | 325.8 | 93.1 | 7193 | 1 |
| dodati | V | 0.3 | 326.5 | 93.1 | 7208 | 1 |
| dobiti | V | 0.3 | 684.1 | 186.3 | 15103 | 2 |
| uspeti | V | 0.3 | 347.4 | 93.1 | 7669 | 1 |
| otvoriti | V | 0.3 | 367.8 | 93.1 | 8120 | 1 |
| trebati | V | 0.3 | 1467.8 | 372.5 | 32406 | 4 |
| smatrati | V | 0.2 | 498.9 | 93.1 | 11014 | 1 |
| raditi | V | 0.2 | 1028.3 | 186.3 | 22702 | 2 |
| postojati | V | 0.1 | 638.2 | 93.1 | 14090 | 1 |

*Slika 8 - Najčešći glagoli u prevodu japanske haiku poezije (levo) i srpske proze (desno)*

## 5.2 Analize prevoda na engleski jezik

Za potrebe poređenja korpusa prevoda japanskog jezika na srpski jezik sa drugim korpusom sačinjen od prevoda japanskog jezika na engleski jezik, koji sadrži, pored *haiku*-a i *haibun*-a, dve najstarije zbirke japanske poezije *Man'jošu* i *Kokinšu*[44]i pesme koje je sastavio zen monah Rjokan.

Analizom korpusa prevoda japanskog tekstana engleski jezikje utvrđeno da se prepoznaje 29926 imenica, što pokriva oko 26% teksta, 22338 glagola (18,8% teksta), 10953 prideva (9% teksta), 8379 priloga (7% teksta). Ovi procenti se ne poklapaju kada se uporede sa oba korpusa na srpskom jeziku.

Leksičkom maskom <TOKEN> smo izdvojili sve tokene i našli sledeće brojeve znakova interpunkcije:

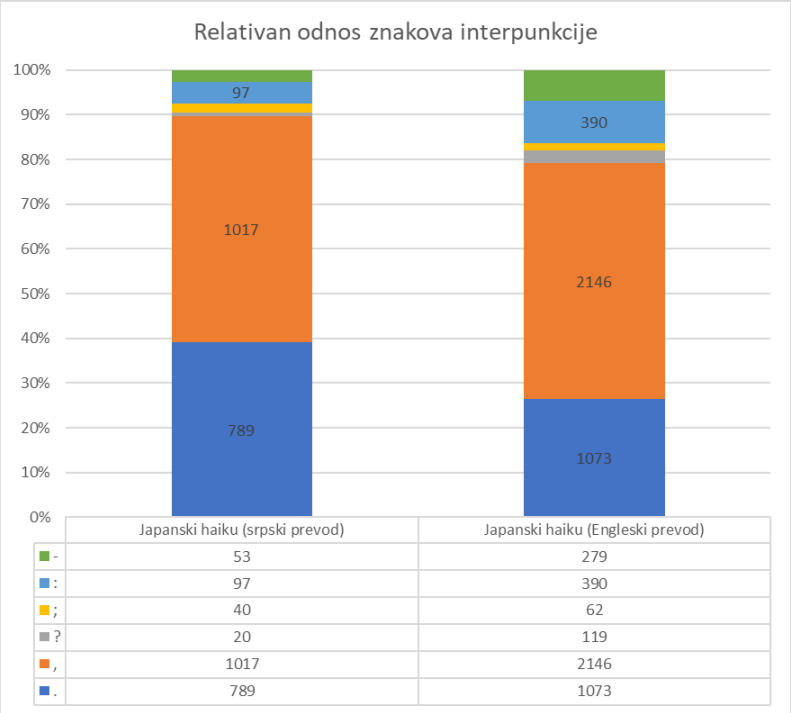
- 1073 tačaka;
- 2146 zareza;
- 119 znakova pitanja;
- 62 tačka-zareza;
- 390 dvotačaka;
- 279 crtice.

Treba imati na umu da je korpus prevoda japanskog jezika na engleski jezik ispao veći zbog postojanja većeg sadržaja i duže tradicije prevođenja na engleski jezik, ali i uzeti u obzir

[44] Poznat i kao *Kokin Vakašu*.

faktor da je engleski, iako jezik u indo-evropskoj jezičkoj porodici, pripada drugoj grupi u odnosu na srpski jezik (Antić n.d.) i stoga ne bi trebalo da bude čudno da je sastav vrsta reči drugačiji.

Na slici 9 je prikazan relativan odnos znakova interpunkcije u korpusima prevoda japanske haiku poezije na srpski i engleski jezik, iz koga se mogu uočiti sličnosti u procentu tačaka i tačka-zareza u oba korpusa, ali primetno različiti odnosi kod ostalih znakova interpunkcije u oba korpusa.




*Slika 9 - Relativan odnos znakova interpunkcije u korpusima prevoda japanske haiku poezije na srpski i engleski jezik*

## 5.3 Analiza n-grama

U računarskoj lingvistici se pod n-gramom podrazumeva susedna sekvenca n stavki iz datog uzorka teksta ili govora. Stavke mogu biti fonemi, slogovi, slova, reči ili bazni parovi prema aplikaciji. N-grami se obično sakupljaju iz tekstualnog ili govornog korpusa. Koristeći numeričke prefikse, n-gram veličine 1 naziva se "unigram"; veličine 2 je "bigram" (ili, manje uobičajeno, "digram"); veličina 3 je "trigram". Model n-gram je tip jezičkog modela zasnovanog na verovatnoćama za predviđanje sledeće stavka u određenoj sekvenci. Modeli n-grama sada se široko koriste u računskoj lingvistici pri statističkoj obradi prirodnog jezika. Dobru stranu n-gram modela (i algoritama koji ih koriste) predstavljaju jednostavnost i

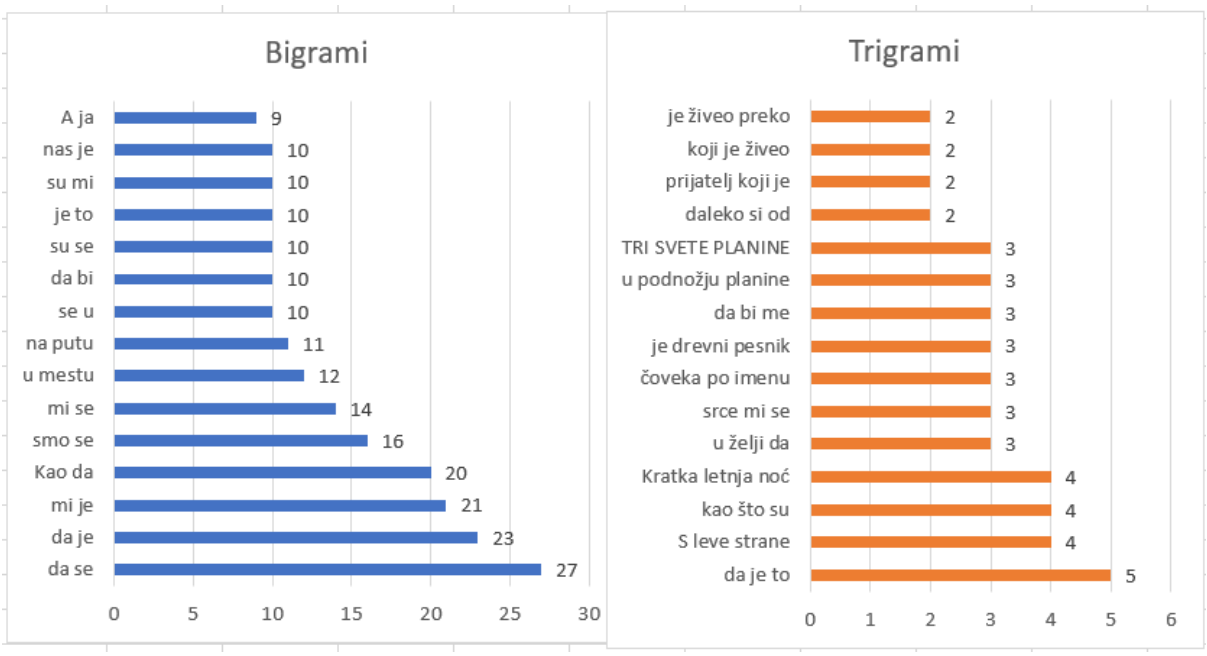
skalabilnost - sa većim n, ali je preciznost obično na strani metoda zasnovanih na leksičkim resursima i pravilima.

Unitex omogućava različite vrste statističkih analiza:

- Kolokacije prema z-vrednost (eng. z-score): prethodna reč, plus dodatne informacije (broj pojavljivanja kolokata u prepoznatom kontekstui u celom korpusu, odnosno z-vrednost kolokacije).
- Kolokacije prema frekvencijama: prikazuju tokene koji se pojavljuju zajedno u prepoznatom kontekstu
- Konteks prema frekvencijama: prikazuje prepoznavanjesa levim i desnim kontekstom sa brojem pojavljivanja kombinacije: prepoznata niska + kontekst.

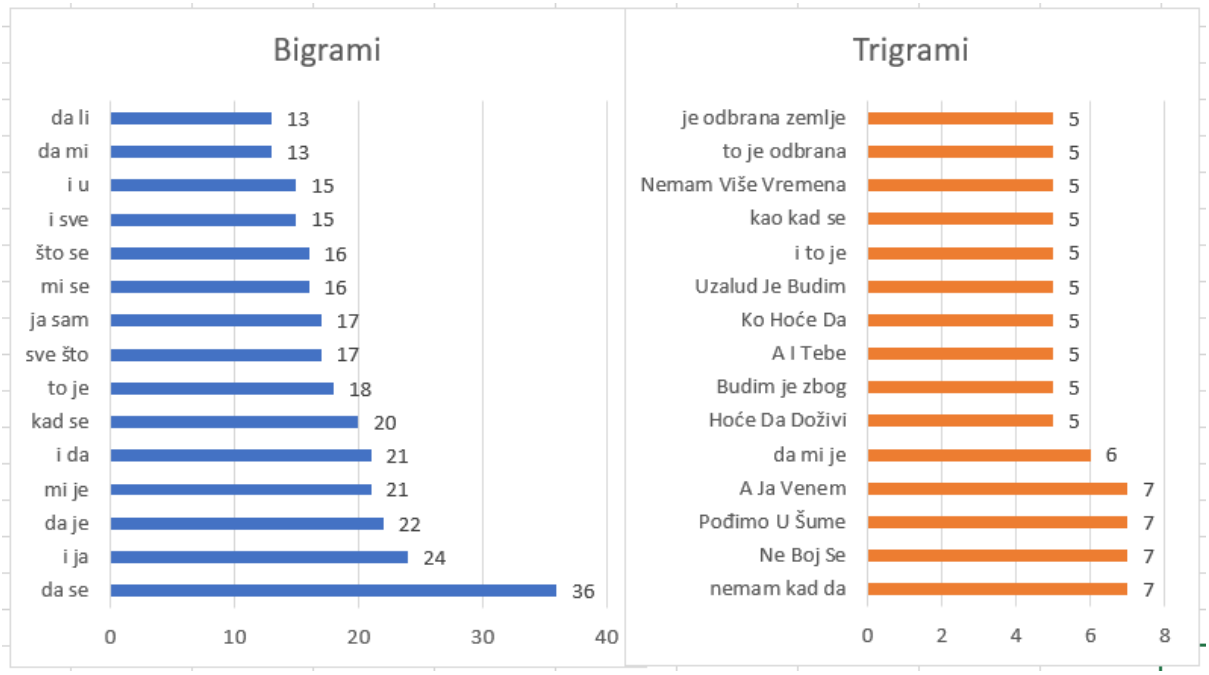
Kako bi pronašli bigrame i trigrame u tri korpusa opisana u 5.1 – 5.3, analizaje podrazumevala pretraguteksta leksičkim maskama<MOT><MOT> za bigrame, odnosno<MOT><MOT><MOT> za trigrame.

Analiza bigrama korpusa prevoda japanskog jezika na srpski je dala 8626 rezultata (81% teksta), a trigrama 6636 rezultata (77% teksta). Najfrekventniji bigrami i trigrami mogu videti na slici 10.




*Slika 10 –Najfrekventniji bigrami i trigrami korpusa prevoda japanskog teksta na srpski*

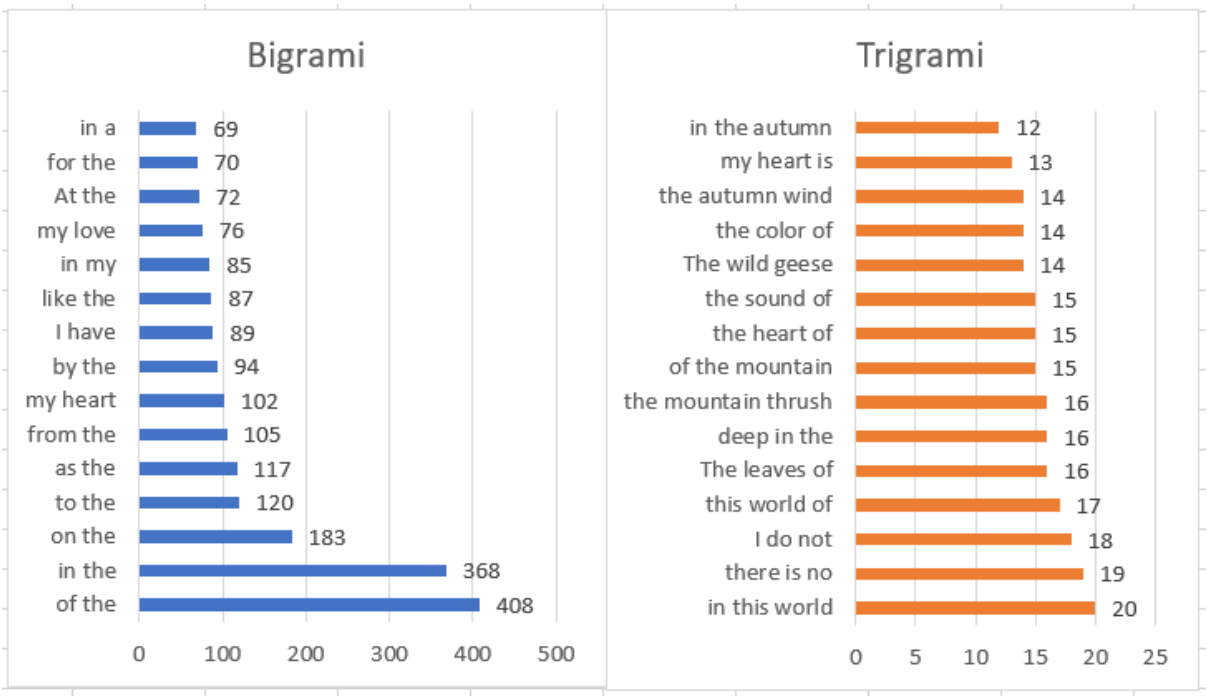
Analiza bigrama korpusa autohtone srpske poezije je dala 15631 rezultat (80,2% teksta), a trigrama 12196 rezultata (76,8% teksta).Najfrekventniji bigrami i trigrami mogu videti na slici 11.




*Slika 11 - Najfrekventniji bigrami i trigrami korpusa autohtone srpske poezije*

Bigrami *da se*, *da je*, *mi je*sumeđu najfrekventnijim u oba korpusa, dok kod najfrekventnijih trigrama nema poklapanja, što je posledica verovatno skromne veličine korpusa.

Analiza bigrama korpusa prevoda japanskog jezika na engleski je dala 51413 rezultata (91% teksta), a trigrama 46378 rezultata (90% teksta). Najfrekventniji bigrami i trigrami mogu videti na slici 12.Najfrekventniji bigrami su *of the*, *in the*, *on the*, dok kod trigrama je *in this world*, *there is no*, *I do not.*




*Slika 12 - Najfrekventniji bigrami i trigrami korpusa prevoda japanskog teksta na engleski*

## 5.4 Rezultat stilometrijske analize

Za analizu lingvističkih podataka se može koristiti i programski jezik R i razvojno okruženje *RStudio*, u okviru kog je moguće korišćenje različitih paketa poput *tm, nlp, corpora, stylo* i drugih. U ovom radu je korišćen Paket *stylo*, koji je namenjen za analize iz oblasti računarske stilistike i omogućava pripremu podataka i stilometričku analizu, od učitavanja teksta do vizuelizacije rezultata (Eder 2013). Osnovna funkcija istoimenog paketa je *stylo()*, u kojoj su implementirane različite metode istraživanja podataka pomoću kojih se vrše procene o sličnostima i razlikama između tekstova i obavljaju različite klasifikacije. Osim uobičajenog rada iz komandne linije, *stylo* paket ima grafički korisnički interfejs za unos parametara, čime olakšava korišćenje: automatsko učitavanje korpusa, njegovo pretprocesiranje i različite stilometrijske analize, od multivarijacione statistike do vizuelizacije stilističkih sličnosti između tekstova.

Paket `stilo` obezbeđuje jednostavne implementacije raznih utvrđenih analiza u polju računarske stilistike, uključujući netradicionalno autorstvo, klasifikaciju prema žanru, razvoj stila (stilometriju), itd. Dodatno s dostupne i metode nadgledanog mašinskog učenja funkcijom *classify()* (Delta, metoda potpornih vektora, naivni Bajes, k-najbližih suseda, …). Funkcija *rolling.delta()* analizira zajednički rad i pokušava da odredi autentičnost fragmenata izvučenih iz njih. Funkcija *rolling.classify()* pruža fleksibilniji interfejs za sekvencijalnu

klasifikaciju zajedničkih radova. Funkcija *oppose()* vrši kontrastivnu analizu između dva skupa tekstova: između ostalog, on generiše spisak reči koje jedan ili više autora preferira i izbegava u odnosu na tekstove drugog autora (ili više njih).

Istraživanje teksta ovom funkcijom obično se izvodi u sledećem redosledu:

1. pronalaženje najfrekventnijih reči za ceo korpus
2. pronalaženje najfrekventnijih reči za pojedinačne tekstove i njihovih frekvencija pojavljivanja
3. normalizovanje frekvencija i kreiranje konačne liste reči (upisuje se u datoteku wordlist.txt)
4. primena različitih statističkih procedura (klaster analiza, faktorske analize, analize osnovnih komponenti), kako bi se grupisali tekstovi i procenila njihova sličnost
5. grafička reprezentacija izračunatih rastojanja tj. udaljenosti (upisuje u datoteku results.txt)

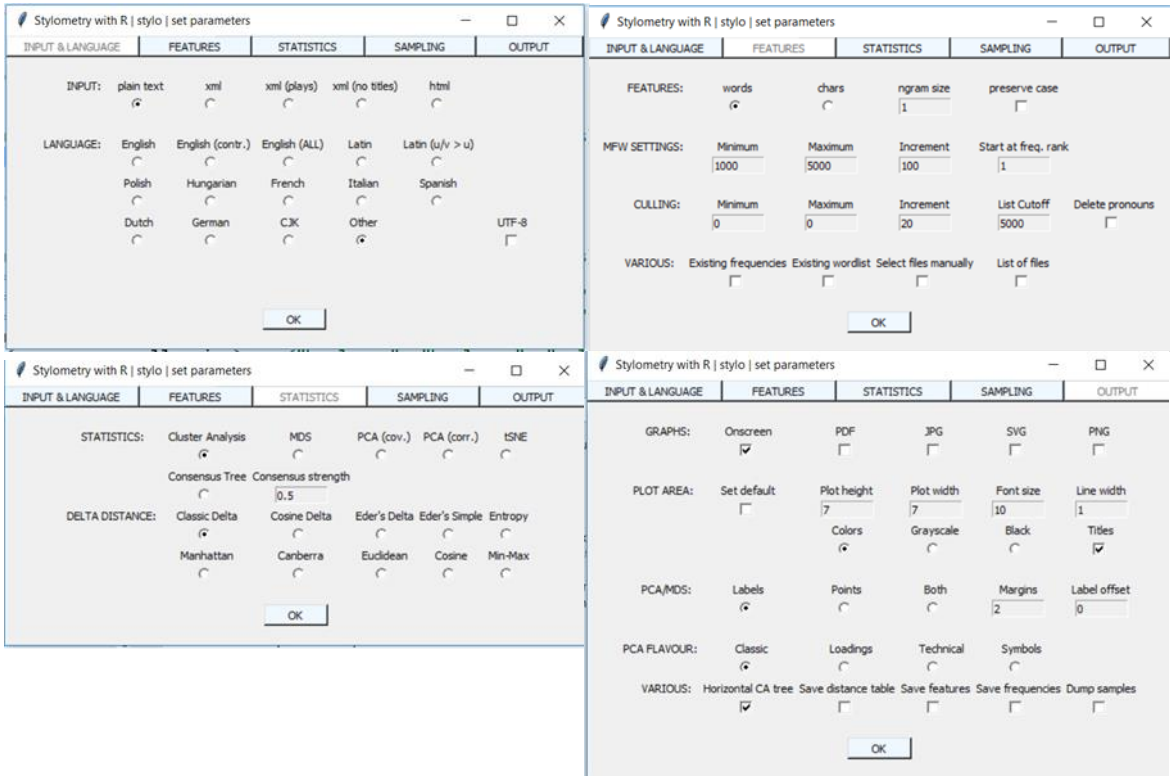
Za potrebe stilističke analize je kreiran korpus izabranih pesama koji se sastoji od tekstualnih datoteka gde svaka datoteka sadrži po jednu pesmu. Svaki projekat zahteva svoj poseban direktorijum koje se ujedno koristi i kao naslov grafikona, u našem slučaju (npr. Pesme). U ovaj direktorijum će biti upisani svi rezultati izvršavanja. Tekstovi koji se analiziraju treba su smešteni u direktorijum corpus unutar radnog direktorijuma.

Posebnu pažnju treba obratiti na imenovanje datoteka, kako bi funkcija stylo () mogla da vrši upoređivanja teksta. Imena treba da slede sledeću sintaksu kategorija_naslov.txt, pri čemu je kategorija ono u odnosu na šta se posmatraju razlike. Na primer, ukoliko želimo da vidimo da li postoje razlike u stilu između muških i ženskih pisaca, kao kategoriju možemo da koristimo oznake M (za muški rod, eng. male) i F (za ženski rod, eng. female). Tada bi nazivi mogli da budu M-Conrad-Lord-Jim.txt, M-Joyce- Dubliners.txt, F-Woolf-Night-and-day.txt, F-Woolf-Waves.txt .... Informacije sadržane u ovako formiranim naslovima koriste se od strane više funkcija, kao na primer za dodeljivanje boja na grafikonima i sl.

Moguće je da datoteke budu i u drugim formatima, poput HTML ili XML. U svakom slučaju, važno je da sve datoteke jednog korpusa budu u istom formatu.

Jedna od analiza je uključila ispitivanje razlika u odnosu na žanr teksta, svakoj datoteci sa pesmom ćemo na početak naziva dodati prefiks HA koji govori da li je u pitanju haiku pesma, BM da je pesma Branka Miljkovića, DM Desanke Maksimović i DJ za Đuru Jakšića.

Funkcija *stylo()* ima veliki broj parametara kojima se definiše način njenog izvršavanja. Ukoliko je pozovemo bez parametara, otvara se grafički korisnički interfejs za unos parametara (slika 13). Korisnik može za definiše neki od ugrađenih jezika ili da izabere obradu bez podrške za specifičan jezik, što je rađeno u ovom istraživanju. Analiza može da se bira po rečima i karakterima, pri čemu smo mi radili analizu po rečima. Za tip statistike se takođe mogu izabrati različite opcije, na primer klaster, MDS (višedimenzionalno skaliranje) ili PCA (Principal Components Analysis) zasnovan na matrici kovarijansi. Rastojanje koje se računa između dokumenata može biti bazirano na različitim formulama, pri čemu je u ovom radu uglavnom korišćeno klasično delta rastojanje.

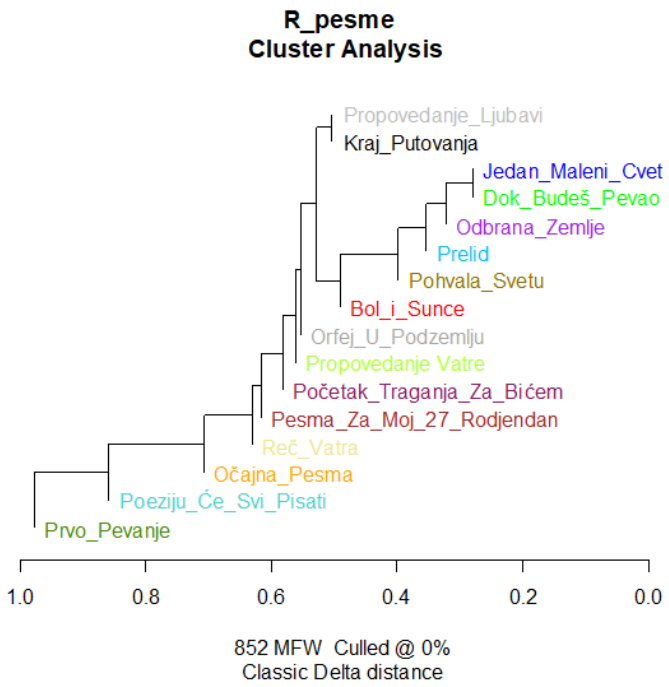



*Slika 13 Grafički korisnički interfejs funkcije stylo()*

Druga opcija korišćenja *stylo()* funkcije je direktno prosleđivanje parametara, pri čemu se naredba izvršava i na ekranu se pojavljuje rezultat izvršavanja.

```
#        izvršava naredbu i crta grafikon
stylo(gui = F, mfw.min = 1000, mfw.max = 5000)
```

Na slici 14 je prikazan rezultat klasifikacije pesama korpusa poezije Branka Miljkovića.

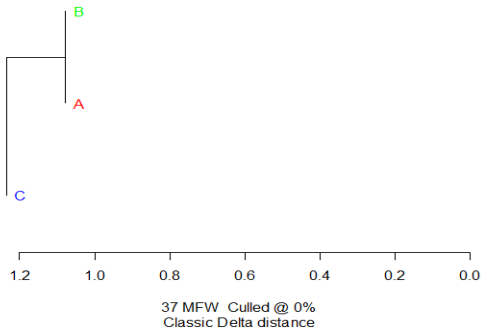



*Slika 14 - Klaster analiza*

Moguće je koristiti stylo() funkciju i sa prosleđivanjem tokenizovanih vektora teksta umesto čistog teksta, kao u sledećem jednostavnom primeru:

```
# korišćenje postojećeg korpusa (liste tokenizovanog teksta):
txt1 = c("Moja","ljubav","puna","drugih","je","deo","zore","koju","budim")
txt2 = c("Ne","povredite","zemlju","ne","dirajte","vazduh","ne","učinite","nikakvo",
         "zlo","vodi","ne", "posvađajte","me","sa","vatrom")
txt3 = c("Postoji","jedna","topla","obala","breg","zelenila","i","jedna","Beatriča",
          "ali", "su","tri","čeljusti","tri","makaze","i","tri","noža")
custom.txt.collection = list(txt1, txt2, txt3)
names(custom.txt.collection) = c("A", "B", "C")
stylo(parsed.corpus = custom.txt.collection)
```

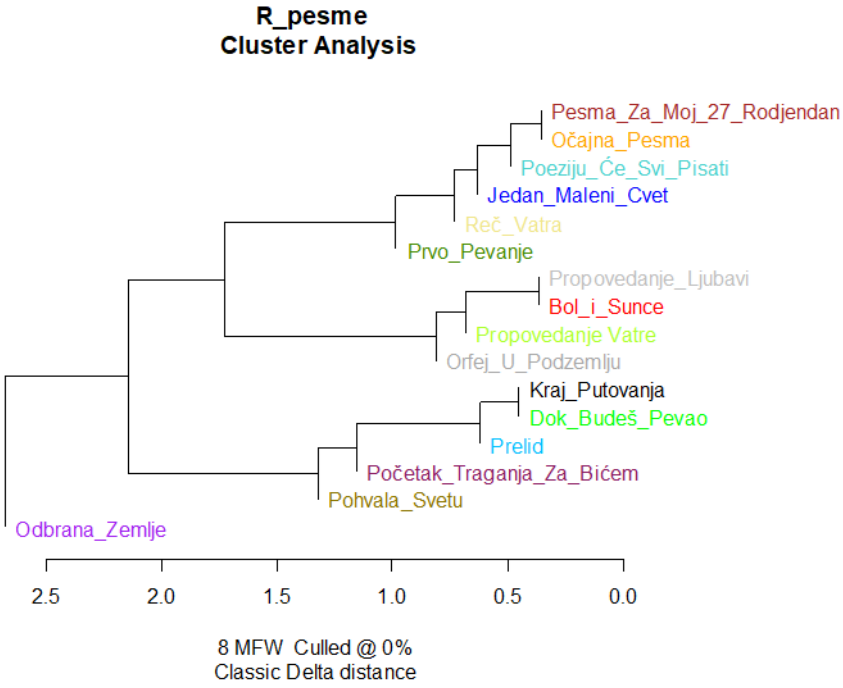
Na slici 15 je prikazana klasterizacija prethodnog primera:




*Slika 15 – Primer klasterizacije*

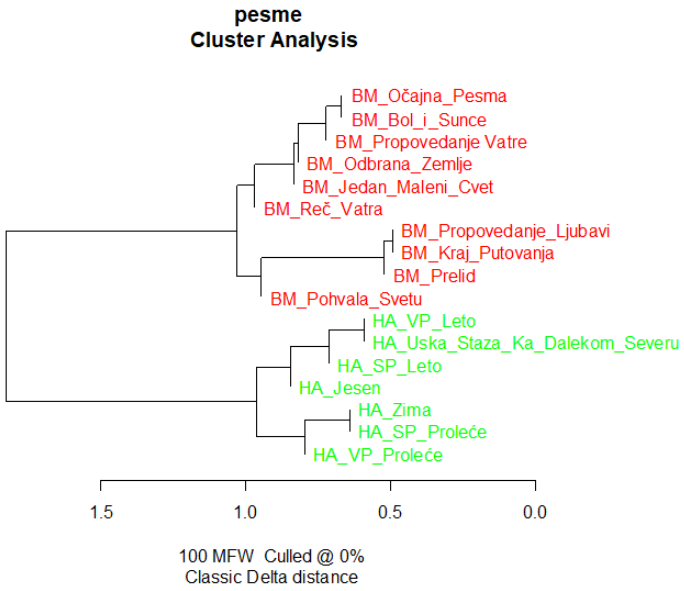
Još jedna opcija je da se koristi zadati skup fičera (reči, n-grama) za analizupesama.

```
my.selection.of.function.words = c("je", "da", "ne", "se", "i", "ako","ili", "od", "do")
stylo(features = my.selection.of.function.words)
```

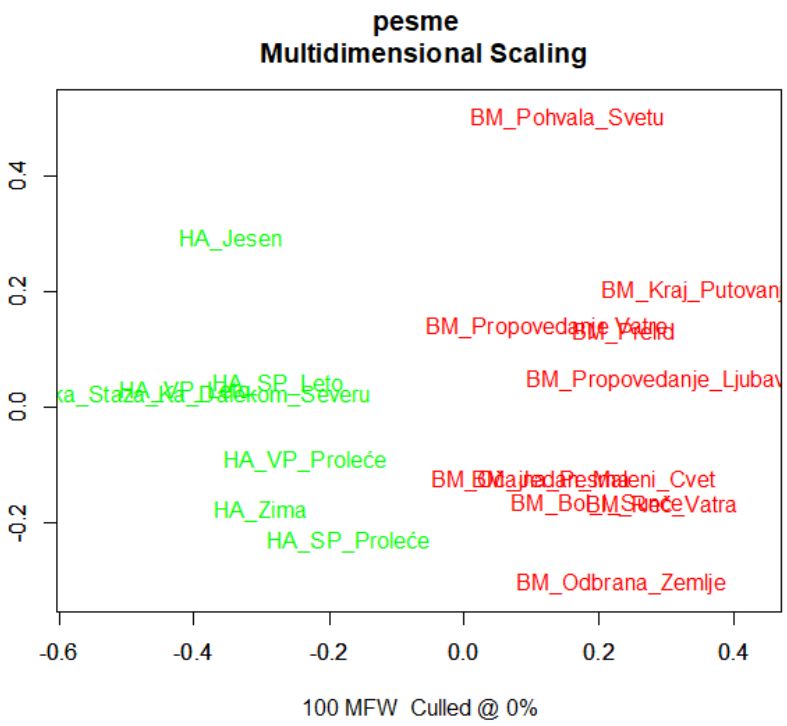



*Slika 16 - Klasifikacija na osnovu skupa fičera*

Drugi eksperiment je uključio 10 pesama Branka Miljkovića i 5 *haiku* pesama i tri je uključio i pesme Đure Jakšića i Desanke Maksimović, što se može videti na slikama 16, 17 i 18.

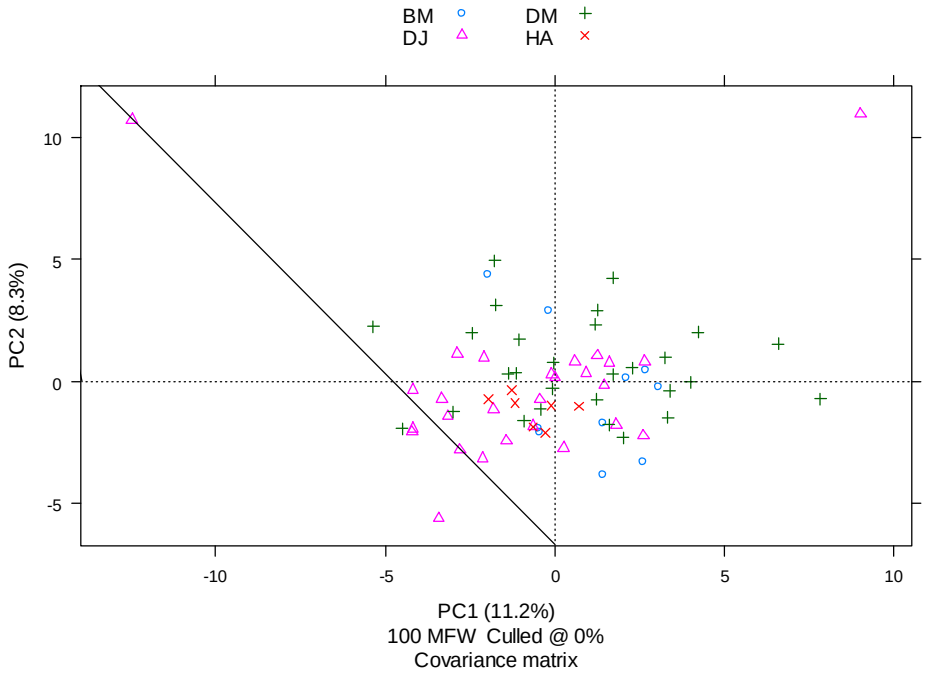



*Slika 17 – Klaster analiza haiku pesama (zeleno) i srpske poezije (crveno)*




*Slika 18 – Višedimenzionalno skaliranje haiku pesama (zeleno) i srpeke poezije (crveno)*

Na slici 19 se može videti klasifikacija dokumenata uz pomoću simbola. BM predstavlja Branka Miljkovića, DM Desanku Maksimović, DJ Đuru Jakšića, a HA predstavlja japansku *haiku* poeziju.




*Slika 19 – Simbolima predstavljena klasifikacija dokumenata*

## 5.5 Rezultat analize *kigo*-a

Na početku istraživanja je pravljena analiza zbirke „Prolećno more“ gde se korišćenjem programa Unitex proučavao *kigo*[45]u toj zbirci *haiku* pesama. I pored početnih problema korišćenja softvera i podešavanja resursa za srpski jezik, napravljene su dve analize teksta koje su pretraživale neke od češćih *kigo* reči.

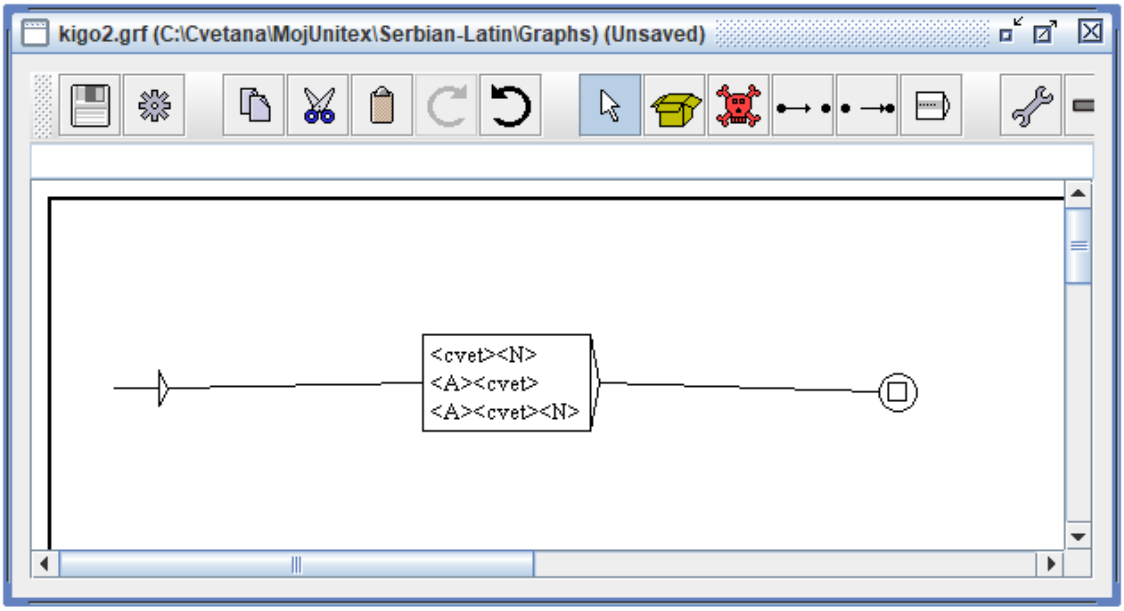
Prva analiza nije bila toliko precizna, jer se sastojala od pretrage *kigo*-a uz pomoću četiri različita grafa celog teksta zbirke, pa se na taj način *kigo* reči nisu razlikovale od istih reči koje ne predstavljaju *kigo* svog reprezentativnog godišnjeg doba. Stoga je, za potrebe druge analize, tekst podeljen na četiri dela, gde je svaki deo predstavljao jedan ciklus izuzev Nove godine, koja je izostavljena jer se sastojala od samo jedne pesme.

Druga analiza, koja se sastojala od četiri manje analize, donela je bolje i preciznije rezultate s obzirom na tehničke nedostatke koji su onemogućavali lakšu pretragu reči. Rezultati druge analize su izdvojili sedam *kigo* za proleće, šest za leto, osam za jesen i deset za zimu, a takođe su pokazali da su određeni *kigo*-i prisutni u dva godišnja doba: cvet, kiša i more. Pored toga, izdvojene su tri *kigo* fraze: vrba bez lišća, cvet šljive i cvet trešnje.

Korpus prevoda japanskog teksta na srpski jezik se, još jednom, sa punom verzijom rečnika biti podvrgnut analizi određenih *kigo*-a. Od *kigo*-a, biće pretraživane fraze „vrba bez lišća“, „cvet šljive“ i „cvet trešnje“ i reči „kiša“, „konj“, „mesec“, „more“, „nebo“, „noć“, „Sunce“, „školjka“, „zora“, „živica“, „gusenica“, „pljusak“, „mrav“, „oblak“, „riba“, „ruža“, „trava“, „veče“, „vrabac“, „jutro“, „kit“, „magla“, „orkan“, „ptica“, „suton“, „vetar“, „vrba“, „guska“, „hladnoća“, „koren“ i „miš“ (Reichhold 2000).

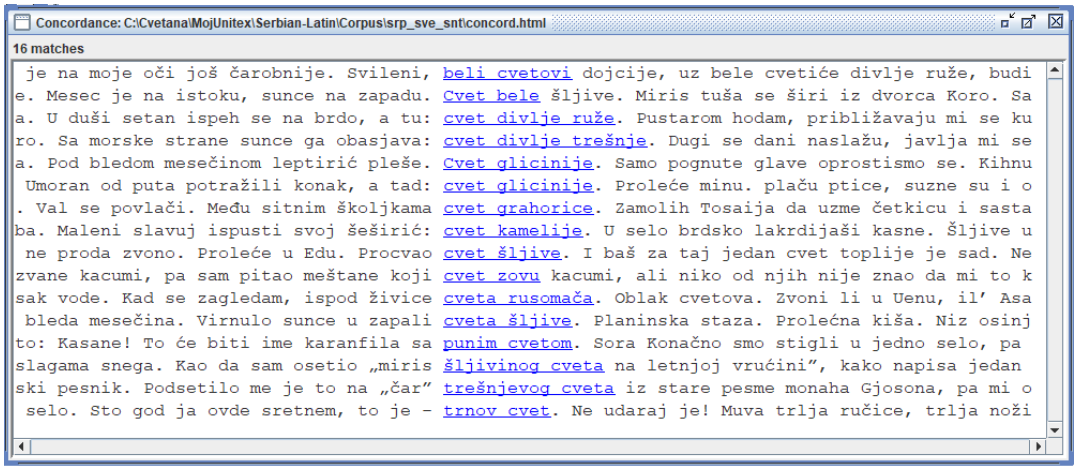
Da bi se uz pomoću programa Unitex našle ove reči i izrazi, korišćeni su grafovi kao što je onaj prikazan na slici 16, čija je funkcija da pretraži korpus prevoda *haiku* poezije za fraze koje su vezane za cvet, kao što su *kigo*-i „cvet šljive“ i „cvet trešnje“, dok se na slici 17 mogu videti konkordance koje su rezultat grafa iz slike 16.

[45] Reči i fraze koje označavaju godišnje doba u *haiku* pesmi (Jamasaki-Vukelić 2011).




*Slika 20 – Graf kojim se pretražuje korpus da se nađu kigo fraze vezane za cvet*

```
#Unigraph
SIZE 1188 840
FONT Times New Roman:  10
OFONT Arial Unicode MS:B 12
BCOLOR 16777215
FCOLOR 0
ACOLOR 13487565
SCOLOR 16711680
CCOLOR 255
DBOXES y
DFRAME y
DDATE y
DFILE y
DDIR n
DRIG n
DRST n
FITS 100
PORIENT L
#
3
"<E>" 56 116 1 2
"" 368 116 0
"<cvet><N>+<A><cvet>+<A><cvet><N>" 186 116 1 1
```




*Slika 21 – Prikaz konkordanci kao rezultat grafa iz Slike 16*

## 5.6 Vizuelizacija oblakom reči i drveta

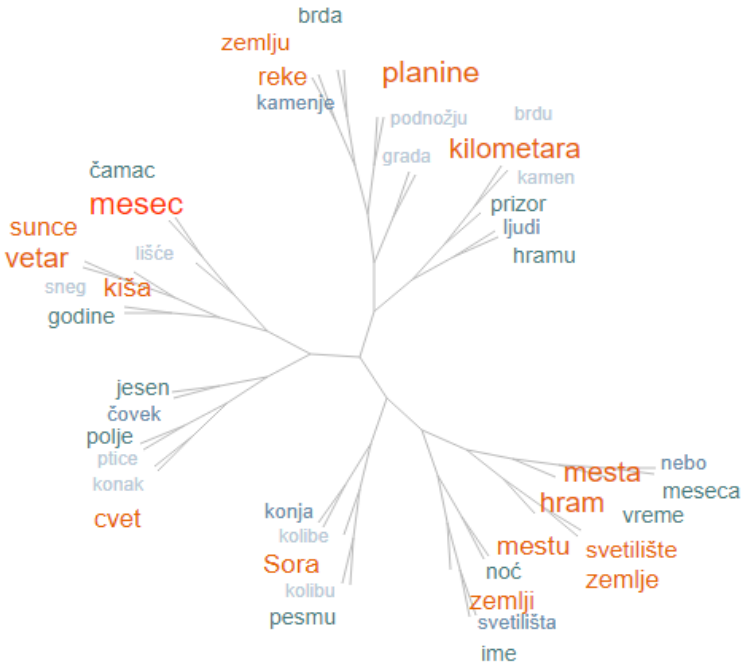
Oblak reči (engl tag cloud) je stekao veliku popularnost na internetu za brz i izražajan prikaz sadržaja sadržaja sajta, veb strane ili nekog dokumenta. Druga opcija koja postoji je alat TagCloud, kod kojepostoji više raznih sajtova i varijanti, kao što je TagCrowd (Steinbock n.d.). Na slici 21 se može videti primer vizuelizacije uz pomoću oblaka reči.


*Slika 22 - Vizuelizacija korpusa prevoda japanskih tekstova na engleski jezik oblakom reči*

Drugi način vizuelizacije teksta, oblak drveta (eng. TreeCloud) prikazuje više informacija. Naime, osim najfrekventnijih reči gde veličina odražava frekvenciju, reči su uređene kao drvo sa semantičkom sličnošću izvedenom iz teksta. Rastojanje između dve reči je predstavljeno dužinom putanje između njih. Ovakvi oblaci drveta pomažu da se identifikuju glavne teme dokumenta, pa čak i da se koriste za analizu teksta. Vizuelizacija drvetom povezanih reči se dobija korišćenjem alata TreeCloud[46] pri čemu se bira jezik srpski za srpske tekstove i engleski za engleski. Alat ima implementirane metode za procenu kvaliteta dobijenog oblaka

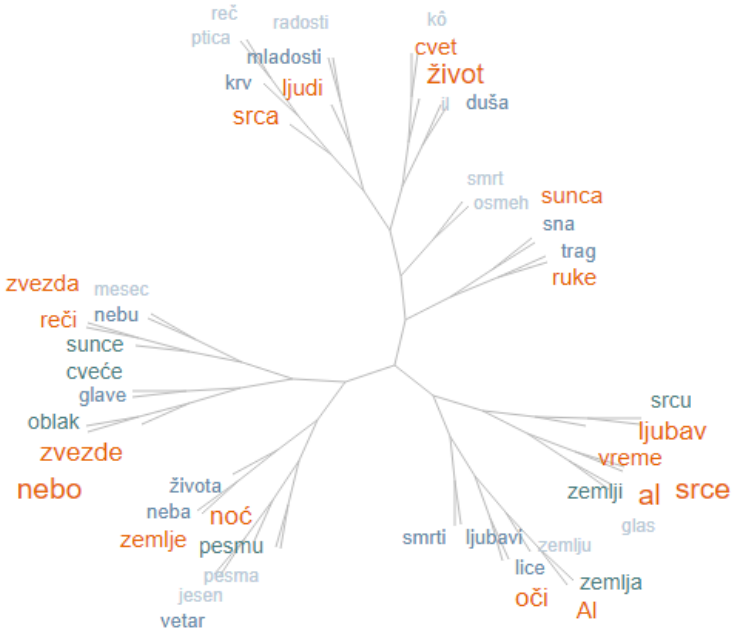
[46] http://treecloud.univ-mlv.fr/cgi-bin/NuageArbore_10_EN.cgi

drveta i nekoliko ključnih koraka njegove izgradnje. Na slikama 23, 24 i 25 su prikazani primeri vizuelizacija uz pomoću oblaka drveta (Gambette i Véronis 2009).



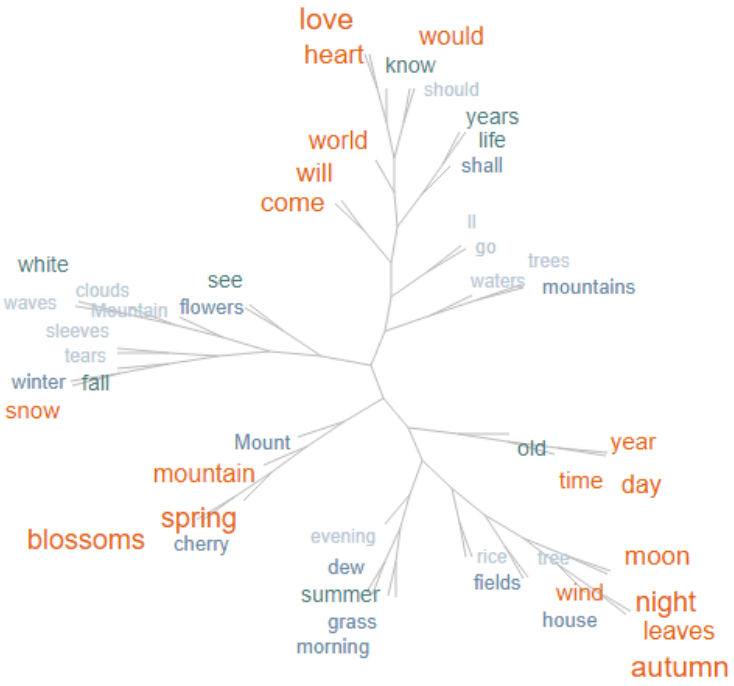

*Slika 23 – Vizuelizacija korpusa prevoda japanskih tekstova na srpski jezik oblakom drveta*

Na slici 23 se mogu videti reči i njihove grupacije. Može se, recimo, videti kako je reč „hram“ grupisana u istoj grupi kao reči „svetilište“, „zemlje“, „„mesta“, „nebo“, „meseca“ i „vreme“. Reči „hram“ i „mesta su podebljana jer se ponavljaju češće“ u toj grupaciji reči.




*Slika 24 - Vizuelizacija korpusa autohtone srpske poezije oblakom drveta*

Na slici 24 se može videti još reči i njihovih grupacija. Može se videti kako je reč „nebo“ grupisana u istoj grupi kao reči „zvezde“, „oblak“, „„glave“, „cveće“, „reči“ i „sunce“ i kako su u toj grupaciji reči „zvezde“, „zvezda“ „reči“ i „nebo“ podebljane u odnosu na ostale..




*Slika 25 - Vizuelizacija korpusa prevoda japanskih tekstova na engleski jezik oblakom drveta*

Na slici 25 se mogu videti reči i njihove grupacije iz korpusa prevoda na engleski jezik, gde reči „love“, „heart“, „world“, „will“, „come“ i „would“ formiraju jednu veliku grupu frekventnih reči.

## 5.7 Predlozi zadopunu elektronskih rečnika

Od ukupno 10789 reči u korpusu prevoda sa japanskog na srpski, 37 reči nije prepoznato, od čega 29 su reči kojima bi mogli da se dopune elektronski rečnici srpskog jezika. Kao priprema za unos u rečnik, uz prepoznati oblik u tekstu je dat kanonski oblik, odnosno lema, vrsta reči, za imenice rod i definicija, opciono sa izvorom iz kog je definicija preuzeta. Oznake su N za imenicu, A za pridev, V za glagol, m za muški rod, f za ženski i n za srednji.

*albizije – albicija, N, f –vrsta biljke, kineska mimoza (Rasadnik Mihalek n.d.)*
*aspri – apsra, N, f – sitan srebrni novac (Ćirković n.d.)*
*blagodarje- blagoslov, N, n*
*budom – buda, N, m*
*čengelama – čengele, N, f– vešala za obradu svinje prilikom klanja (Sučić n.d.)*
*čičimka – čičimak, N, f – vrsta biljke, žižula, kineska datula (Dulist 2014)*
*desetoredne – desetoredna, A, – pridev, od deset redova*
*dojcije – dojcija, N, f – vrsta biljke(Bertović, i dr. 1997)*
*dvoraninove – dvoraninov, A*

*grahora – grahor, N, m – vrsta biljke, burčak, kukolj (Zirojević 2011)*
*haikai - haiku, N, m*
*hina – hina, N, f -vrsta lutke koja se poklanja devojčicama za Dan devojčica (3. mart)(Šikošek 2016)*
*hodočasnice – hodočasnica, N, f – ženski hodočasnik*
*hoku, hoku, N, m – početni stihovi u renga poeziji od kojih je nastao haiku*
*irisove – iris, N, m – vrsta cveta*
*izdžikljala – izdžikljati, V, – naglo izrasti (Novo Miloševo n.d.)*
*kacumi – kacumi, N, m- vrsta biljke*
*kamforovog– kamfor, N, m– vrsta drveta*
*konoširo – konoširo, N,m - vrsta ribe*
*moksu – moksa, N, f –vrsta tamjana (Mandić 2016)*
*odunula – odunuti, V, - oduvati*
*prugavac–prugavac, N, m - vrsta ribe, tunj prugavac (EUR-LEX 1992)*
*ratnikovim– ratnikov, A*
*sutre – sutra, N, f–hinduistički i budistički sveti tekstovi*
*tržici – tržica–? (nije pronađena definicija)*
*vaka – tanka, japanska pesnička forma*
*zen–tradicija mahajanskog budizma u Japanu (Perić 2012)*
*suvoti – suvota - suvoća*
*pustošju – pustoš*

Određeni broj reči se javlja u krnjem obliku ili nestandardnim fleksijama, pa navodimo i njih:

*il – il' - krnji oblik veznika „ili“*
*otkinuv – otkinuv' – krnji oblik glagolskog prideva prošlog „otkinuvši“*
*skupljajuć – skupljajuć' – krnji oblik glagolskog prideva sadašnjeg „skupljajući“*
*pregazit – pregazit' – krnji oblik glagola „pregaziti“*
*njeg – njeg' – krnji oblik zamenice „njega“*

Reči kao što su „albicija“, „aspra“, „blagodarje“, „dvoranin“, „grahor“, „hodočasnica“, „iris“, „izdžikljati“, „kamfor“, „moksa“, „odunuti“, „prugavac“, „pustoš“, „ratnik“, „sutra“, „suvota“ i „*zen*“ se ne nalaze u elektronskom rečniku srpskog jezika i stoga treba dopuniti rečnik tim rečima.

# 6 Zaključak

Na osnovu istraživanja koja su sprovedena u ovom radu, iznosimo zaključak da je polazna hipoteza tačna u odnosu na prevode sa japanskog na srpski profesora Hirošija Jamasakija-Vukelića, a to je da prevodi sa japanskog na srpski zadržavaju rečitost.

Razlozi za zaključivanje u korist polazne hipoteze se ogledaju u broju nepoznatih lema, ali i u njihovim oblicima. Pored toga što su većinski sačinjeni od vlastitih imenica kao što su lična imena i lokacije, među nepoznatim lemama se nalazi 29 reči, od kojih ima samo pet tuđica iz japanskog jezika, što pokazuje tendenciju da se što više koriste domaće reči, sa time da postoji određen broj reči koje se frekventnije koriste u hrvatskom jeziku. Korišćenje takvih reči su verovatno imale za cilj zadržavanje forme *haiku* stihova, ali i jezika iz drugog perioda književnosti. Takođe, neke reči su okrnjene, ali isključivo u *haiku* poeziji, odnosno njenom sprskom prevodu, što pokazuje i tendenciju da se održi tradicionalna slogovna struktura *haiku* pesme, jer bi se punim i pravilnim oblikom tih reči narušila ta slogovna struktura.

Što se tiče poređenja korpusa prevoda japanske poezije na srpski i korpusa srpske poezije, mogu se primetiti podudaranja u procentualnom odnosu vrsta reči, kao i većinska podudaranja u procentualnom odnosu znakova interpunkcije, sa izuzetkom znaka pitanja i tačka-zareza, iz čega se može zaključiti da se u srpskoj poeziji znak pitanja i tačka-zarez češće koriste nego u prevodima japanske *haiku* poezije.

Sa druge strane, prilikom poređenja korpusa prevoda japanske *haiku* poezije na srpski i engleski jezik, može se primetiti da se procenti vrsta reči ne poklapaju, kao ni procenti znakova interpunkcije, sa izuzetkom tačke i tačka-zareza, iz čega se može zaključiti da srpski i engleski jezik imaju značajno drugačije sastave svojih rečenica po pitanju vrsta reči i znakova interpunkcije, verovatno jer pripadaju različitim grupama indo-evropske jezičke porodice.

Na kraju, elektronski rečnik srpskog jezika treba dopuniti sa 29 reči koje nisu prepoznate prilikom analize korpusa prevoda japanskog teksta na srpski jezik.

Za dalju analizu korpusa prevoda japanskih tekstova na druge jezike potrebno je analizirati japanske tekstove u njihovom izvornom jeziku koristeći računarske resurse japanskog jezika, kako bi mogli da se uporede sa svojim prevodima. Jedna od daljih aktivnosti može biti i paralelizacija prevoda na srpski ingleski jezik haiku poezije i drugih tekstova prevedenih na oba jezika, naravno uključujući i paralelizaciju sa izvornim tekstom na japanskom.

# 7 Literatura


1. Anthony, Laurence. *A critical look at software tools.* Tokyo: Waseda University, 2011.
2. Antić, Dejan. *Indoevropski jezici danas.* http://www.dml.rs/index.php/lat/tekstovi-lat/istorijska-lingvistika-lat/111-indoevropski-danas-lat (accessed 06 23, 2018).
3. Barnhill, David Landis. *Basho's Haiku - Selected Poems of Matsuo Basho.* New York: State University of New York Press, Albany, 2004.
4. Bašo, Macuo, and Hiroši Jamasaki-Vukelić. *Uska staza ka dalekom severu.* Beograd: Tanesi, 2012.
5. Bašo, Macuo, Hiroši Jamasaki-Vukelić, and Srba Mitrović. *Svenulo polje.* Beograd: Rad, 2008.
6. Bertović, Stjepan, Milan Generalović, Josip Karavla, and Jakob Martinović. *Priroda i parkovni objekti u općini Rijeka.* Rijeka: Izvorni Znanstveni Članci, 1997.
7. Björkenstam, Kristina Nilsson. *What is a corpus and why are corpora important tools?* Stockholm: Stockholm University, 2013.
8. Bleed, Peter J., et al. *Istorija Japana Iz Kodanšine Ilustrovane enciklopedije Japana.* Beograd: Zavod za udžbenike, 2003.
9. Brezina, Vaclav, Tony McEnery, and Stephen Wattam. *Collocations in context - A new perspective on collocation networks.* Lancaster: John Benjamins Publishing Company, 2015.
10. Buson, Josa, Hiroši Jamasaki-Vukelić, and Srba Mitrović. *Prolećno more.* Beograd: Rad, 1999.
11. Cambridge Dictionary. *collocation.* https://dictionary.cambridge.org/dictionary/english/collocation (accessed December 23, 2017).
12. Carey, Ray, Kaisa Pietikäinen, and Netta Hirvensalo. *What's a corpus?* 20 November 2017. https://elfaproject.wordpress.com/whats-a-corpus/.
13. Center for Corpus Development, NINJAL. *Overview.* 20 November 2017. http://pj.ninjal.ac.jp/corpus_center/csj/en/.
14. Ćirković, Đorđe V. *Šimanovci.* http://www.simanovci.rs/ckalj/sremacki-recnik/?lang=lat (accessed June 20, 2018).
15. Cobb, Tom. *Chapter 2 - Corpus & Concordance in Linguistics & Language Learning.* 15 January 2018. https://www.lextutor.ca/cv/webthesis/Thesis2.html.
16. *Computational Approaches to Collocations*http://www.collocations.de/AM/
17. Dulist. *UPOZNAJMO ČIČIMAK Koliko znate o ovom ukusnom voću?* 10 September 2014. https://www.dulist.hr/upoznajmo-cicimak-koliko-znate-o-ovom-ukusnom-vocu/194354/ (accessed June 20, 2018).
18. Eder, Maciej. *Mind your corpus: systematic errors in authorship attribution.* Literary and Linguistic Computing, Oxford: Oxford University Press, 2013.

19. Erjavec, Tomaz, Adam Kilgarriff, and Irena Srdanovic. "A large public-access corpus for Japanese language." *Inaugural Workshop on Computational Japanese Studies.* Ikaho, Japan: Adam Kilgarriff Publications, 2007. 1.

20. EUR-LEX. *UREDBA VIJEĆA (EEZ) br. 1536/92.* 9 June 1992. https://eur-lex.europa.eu/legal-content/HR/TXT/?uri=celex%3A31992R1536 (accessed June 20, 2018).

21. Faculty of Oriental Studies, University of Oxford. *The Oxford Corpus of Old Japanese.* 20 November 2017. http://vsarpj.orinst.ox.ac.uk/corpus/index.html.

22. Gambette, Philippe, and Jean Véronis. "Visualising a Text with a Tree Cloud In Locarek-Junge H. and Weihs C., editors." *Classification as a Tool of Research.* Dresden: IFCS, 17 March 2009.

23. goodreads. *Kođiki, Zapisi o drevnim događajima.* 20 November 2017. https://www.goodreads.com/book/show/27150528-ko-iki-zapisi-o-drevnim-doga-ajima.

24. Gross, Maurice. *The Use of Finite Automata in the Lexical Representation of Natural Language.* Paris: University of Paris, 1989.

25. Haywood, Sandra. *Using Concordance Lines.* https://www.nottingham.ac.uk/alzsh3/acvocab/concordances.htm (accessed January 15, 2018).

26. Jamasaki-Vukelić, Hiroši. *Vrapčeva priča.* Beograd: Tanesi, 2011.

27. Japan Science and Technology Agency. *ASPEC (Asian Scientific Paper Excerpt Corpus).* 26 February 2015. http://orchid.kuee.kyoto-u.ac.jp/ASPEC/ (accessed June 20, 2018).

28. japan-guide.com. *Hiragana.* 20 November 2017. https://www.japan-guide.com/e/e2047.html.

29. —. *Japanese Language.* 20 November 2017. https://www.japan-guide.com/e/e621.html.

30. —. *Kanji.* 20 November 2017. https://www.japan-guide.com/e/e2046.html.

31. Keene, Donald, Masayuki Miyata, Makoto Ooka, and Ian Hideo Levy. *Love Songs from the Man'Yoshu.* Tokyo - New York - London: Kodansha International, 2000.

32. Krstev, Cvetana. *Srpski jezik u informatičkom okruženju.* Beograd: Književnost i jezik, 2000.

33. Krstev, Cvetana, and Duško Vitas. *Corpus and Lexicon - Mutual Incompleteness.* Birmingham: The University of Birmingham Press, 2005.

34. Lexical Computing CZ s.r.o. *Corpus types.* 23 December 2017. https://www.sketchengine.co.uk/user-guide/user-manual/corpora/corpus-types/.

35. Lujak, Tamara. *Beleg.* https://belegbg.wordpress.com/haiku-2/ (accessed June 15, 2018).

36. Mandić, Ira. *Moksibustija - liječenje toplinom.* 5 May 2016. https://www.sensa.hr/clanci/tretmani/moksibustija-lijecenje-toplinom?page=2 (accessed June 20, 2018).

37. Martin, Jean-Claude, Patrizia Paggio, Michael Kipp, and Dirk Heylen. *Multimodal-Corpora.org.* 20 November 2017. http://www.multimodal-corpora.org.

38. Matematički fakultet Univerziteta u Beogradu. *Kratka istorija.* http://www.korpus.matf.bg.ac.rs/prezentacija/istorija.html (accessed Novembar 20, 2017).

39. McArthur, Tom. *The Oxford Companion to the English Language, ed.* Oxford: Oxford University Press, 1992.

40. National Institute for Japanese Language and Linguistics. *General Description.* 20 November 2017. https://www.ninjal.ac.jp/english/info/aboutus/.

41. —. *Introduction to the BCCWJ.* 20 November 2017. http://pj.ninjal.ac.jp/corpus_center/bccwj/en/.

42. Neubig, Graham. *Japanese-English Legal Parallel Corpus.* 23 July 2014. http://www.phontron.com/jaen-law/ (accessed 06 20, 2018).

43. Novo Miloševo. *Prvi "novoMiloševačko - srbski tolmač" ( rečnik... )...* https://www.novomilosevo.devbin.org/recnik.html#I (accessed June 20, 2018).

44. Oxford Living Dictionaries. *concordance.* 23 December 2017. https://en.oxforddictionaries.com/definition/concordance.

45. —. *Using the Corpus.* https://en.oxforddictionaries.com/explore/using-the-corpus (accessed November 20, 2017).

46. Pajić, Vesna. *Unitex 3.0 uputstvo za upotrebu.* Beograd, Centralna Srbija, 25 maj 2016.

47. Pardeshi, Prashant. *About this project.* 20 November 2017. http://npcmj.ninjal.ac.jp/?lang=en.

48. Paumier, Sébastien. *UNITEX 2.1 User Manual.* Munich: Ludwig-Maximilians-Universität, 2003.

49. Perić, Vladimir. *Zen budizam.* 14 April 2012. http://afirmator.org/vladimir-peric-zen-budizam/ (accessed June 20, 2018).

50. Poezija noći. *Poezija noći - Ljubavna poezija, najlepše pesme.* https://www.poezijanoci.com/domaca-poezija.html (accessed June 15, 2018).

51. Rasadnik Mihalek. *Rasadnik Mihalek.* http://rasadnikmihalek.com/?p=2133 (accessed June 20, 2018).

52. Rasplica, Laurel Rodd, and Mary Catherine Henkenius. *Kokinshu - A Collection of Poems Ancient and Modern.* Princeton, New Jersey: Princeton University Press, 1984.

53. Reichhold, Jane. *A Dictionary of Haiku Classified by Season Words with Traditional and Modern Methods.* 2000. https://www.ahapoetry.com/aadoh/adofinde.htm (accessed June 23, 2018).

54. Russian National Corpus. *what is the Corpus?* 20 November 2017. http://www.ruscorpora.ru/en/corpora-intro.html.

55. Šikošek, Majda. *Lutke Japana.* 29 June 2016. http://lepotazivota.rs/lutke-japana/ (accessed June 20, 2018).

56. St. Gries, Stefan. *Useful statistics for corpus linguistics.* Santa Barbara: University of California, 2010.

57. Stanford University, Google Brain & Rakuten Institute of Technology. *JESC - Japanese-English Subtitle Corpus.* https://nlp.stanford.edu/projects/jesc/ (accessed 06 20, 2018).

58. Stanković, Ranka, Cvetana Krstev, Ivan Obradović, and Olivera Kitanović. *Rule-based Automatic Multi-Word Term Extraction and Lemmatization.* Portorož, Slovenia: ELRA, 2016.

59. Steinbock, Daniel. *TagCrowd.* https://tagcrowd.com (accessed June 23, 2018).

60. Sučić, Miroslav. *Kupres.de.* http://www.kupres.de/rjecnik/rjecnik.htm (accessed June 20, 2018).

61. Sütiste, Elin. *A Crow on a bare branch: A Comparison of Matsuo Basho's haiku "Kare-eda-ni..." and its English translations.* Tartu: Studia Humaniora Tartuensia, 2001.

62. The Teaching Company. *The Japanese Language: Context Is Everything.* 10 May 2018. https://www.thegreatcoursesdaily.com/aspects-of-the-japanese-language/.

63. University of Leipzig. *Stylometry: Computer-Assisted Analysis of Literary Texts.* 23 December 2017. http://www.culingtec.uni-leipzig.de/ESU_C_T/node/389.

64. Utvić, Miloš. *Anotacija savremenog srpskog jezika.* Beograd: INFOteka, 2011.

65. W3-Corpora project. *Corpus Linguistics.* 21 November 2017. https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction2.html.

66. Wynne, Martin, and Ylva Berglund Prytz. *Types of corpora and some famous (English) examples.* Oxford: OUCS Hilary, 2012.

67. Yasuda, Kenneth. *The Japanese Haiku - Its Essential Nature and History.* Boston - Rutland, Vermont - Tokyo: Tuttle Publishing, 2001.

68. Yuasa, Nobuyuki. *The Zen Poet of Ryokan.* Princeton, New Jersey: Princeton University Press, 1981.

69. Zirojević, Olga. *Prilozi za Orijentalnu Filologiju.* Sarajevo: Orijentalni institut Univerziteta u Sarajevu, 2011.

# 8 Prilog: popis slika