Metode računalne i forenzičke lingvistike za utvrđivanje autorstva

Šincek, Marijana

Master's thesis / Diplomski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:137692

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-04-25**

Filozofski fakultet
Sveučilišta u Zagrebu

Repository / Repozitorij:

ODRAZ - open repository of the University of Zagreb
Faculty of Humanities and Social Sciences

zir.nsk.hr

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

Odsjek za lingvistiku

Marijana Šincek

METODE RAČUNALNE I FORENZIČKE LINGVISTIKE ZA UTVRĐIVANJE AUTORSTVA

Diplomski rad

Mentor: dr.sc. Božo Bekavac, docent

Zagreb, 2022.

UNIVERSITY OF ZAGREB

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Department of Linguistics

Marijana Šincek

COMPUTATIONAL AND FORENSIC LINGUISTICS METHODS FOR AUTHORSHIP ATTRIBUTION

Master's Thesis

Supervisor: Dr. Božo Bekavac, Assistant Professor

Zagreb, 2022

Sažetak:

Računalna je lingvistika grana lingvistike koja se od ostalih lingvističkih grana i metodologija razlikuje upotrebom računala, koja se mogu koristiti za opis jezičnih jedinica na raznim razinama jezika (Tadić, 2003). Forenzička lingvistika predstavlja primjenu lingvistike u domeni prava i pravnih pitanja, gdje se lingvistička znanja i vještine koriste za analizu pravnih tekstova, pravnog govora, ali i za pružanje dokaza istražiteljima (Coulthard i sur., 2017).

Cilj istraživanja je ispitati mogućnosti povezivanja metoda računalne i forenzičke lingvistike te proučiti kako se one mogu koristiti za utvrđivanje autorstva tekstova koje je napisao Antun Gustav Matoš, kao i provjeriti koje metode i analize pružaju najindikativnije rezultate. U sklopu istraživanja analizirano je 26 tekstova pet autora, Antuna Gustava Matoša, Milana Begovića, Jure Turića, Josipa Kozarca i Joze Ivakića. Provedena je analiza stilometrijskih značajki teksta svih pet autora korištenjem analize varijance i Jaccardovog koeficijenta sličnosti pomoću korpusa i korpusnih alata. Rezultati istraživanja pokazali su da najlošije rezultate za utvrđivanje autorstva pružaju stilometrijske mjere dužine riječi i rečenica u tekstu, dok omjer različnica i pojavnica, koji pruža podatke o leksičkoj obogaćenosti teksta, ima veću uspješnost. Najveću uspješnost ostvarilo je mjerenje Jaccardovog koeficijenta sličnosti lema i n-grama u poznatim i provjeravanim tekstovima.

Ključne riječi: računalna lingvistika, forenzička lingvistika, metodologija, korpus, utvrđivanje autorstva

Abstract:

Computational linguistics is a branch of linguistics differentiated from other linguistic branches and methodologies by the use of computers, used for description of linguistic units on all levels of language (Tadić, 2003). Forensic linguistics denotes the application of linguistics when it comes to the law and legal matters, where linguistic knowledge and skills are used for analysis of written legal texts, legal language, as well as to provide evidence to investigators (Coulthard et al., 2017).

The aim of this research was to examine the possibilities of combining the methods of computational and forensic linguistics, how they can be used for authorship attribution of texts written by Antun Gustav Matoš, as well as to explore which methods and analyses provide most indicative results. 26 texts from five authors, Antun Gustav Matoš, Milan Begović, Jure Turić, Josip Kozarac and Joza Ivakić were included in this research. An analysis of stylometric properties of the texts by all five authors was conducted using analysis of variance and Jaccard similarity index by using corpora and corpus tools. The results of the research showed that the worst results for authorship attribution were obtained using word and sentence length, followed by a more successful type-token ratio, which provided data on the lexical richness of the texts. The most successful was the Jaccard similarity index of the lemmas and n-grams used in known and questioned text samples.

Keywords: computational linguistics, forensic linguistics, methodology, corpus, authorship attribution

Sadržaj:

1. Uvod

Kad je riječ o računalnoj lingvistici, može se uočiti dvostranost u pristupu računalnog tretmana prirodnog jezika. Ako se tom području pristupa s lingvističke strane, riječ je o računalnoj lingvistici, što se shematski može prikazati kao lingvistika + računalo, gdje je lingvistika na prvom mjestu (Tadić, 2003). Računalo je u današnje vrijeme nezaobilazno pomagalo za prikupljanje primarnih podataka koji se pohranjuju u korpuse i rječnike, a zatim se obradom dobivaju sekundarni podaci na temelju kojih se mogu izgraditi računalni modeli jezičnih (pod)sustava (Tadić, 2003).

Forenzička se lingvistika definira kao primjena lingvističkih znanja u tri domene, pisani pravni tekstovi, pravni govor i pružanje dokaza istražiteljima (Coulthard i sur., 2017). Međutim, iako se forenzična analiza teksta najčešće provodi upravo u kontekstu kaznenih djela i sudskih postupaka, metode koje koriste forenzični lingvisti mogu se koristiti i u druge svrhe, kao što je utvrđivanje autorstva književnih djela ili provjera plagijata. Uz razvoj tehnologije, u te se svrhe sve više koriste resursi i alati za računalnu obradu teksta, pogotovo korpusi.

Cilj ovog rada je ispitati mogućnosti povezivanja metoda računalne i forenzičke lingvistike te kako se one mogu koristiti za utvrđivanje autorstva teksta, provjeriti koje metode i analize računalne i forenzičke lingvistike pružaju najindikativnije rezultate te pružiti osvrt na računalnu obradu teksta u svrhu utvrđivanja autorstva. Analizirat će se ukupno 26 tekstova pet hrvatskih autora iz perioda moderne, nastalih na prijelazu 19. u 20. stoljeće, s fokusom na Antuna Gustava Matoša.

Prvi dio rada pruža pregled razvoja i nastanka forenzičke lingvistike, kao i primjere stvarnih slučajeva. Drugi dio rada bavi se utvrđivanjem autorstva, kao i dva pristupa utvrđivanju autorstva, stilističkim i stilometrijskim pristupom.

Glavni dio rada predstavlja rezultate analize 26 tekstova u svrhu utvrđivanja njihova autorstva i pregled korištenih metoda. Diseminacija rezultata pruža pogled u odnos računala i jezika pri analiziranju autorstva djela, nakon čega slijedi pregled ideja za buduće istraživanje i zaključak rada.

2. Forenzička lingvistika

Termin forenzične znanosti odnosi se na primjenu znanstvenih principa i tehnika u pravnim procesima i pružanje potpore u kriminalističkim istragama (Saferstein, 2017). Riječ je multidisciplinarnom skupu znanosti koji se sastoji od mnoštva. Iako većina ljudi na spomen riječi „forenzika“ možda prvo pomisli na primjenu prirodnih znanosti, kao što su analiza DNK ili proučavanje vlakna, polimera i drugih čestica pronađenih na mjestu zločina, društvene i humanističke znanosti ključni su dio forenzičnih znanosti. Dok prirodne znanosti pružaju istražiteljima fizičke dokaze koji se mogu koristiti za osuđivanje počinitelja, uloga društvenih i humanističkih znanosti u forenzici jest podržati istražitelje korištenjem znanja društvenih i humanističkih disciplina, kao što su psihologija, antropologija i lingvistika, u izradi profila počinitelja, žrtve i analizi dokaza.

Naziv *forenzička lingvistika* osmislio je Jan Svartvik kada je 1968. objavio analizu izjava T.J. Evansa, osuđenog i pogubljenog za ubojstvo supruge i kćeri. Jan Svartvik analizirao je četiri izjave koje je Evans dao istražiteljima nakon što je otvorena posthumna istraga kad je drugi čovjek priznao krivnju za ta ubojstva. Svartvik (1968) je u dokumentu, danas poznatijem pod nazivom „Svartvikova analiza“, otkrio stilističke razlike u izjavama te je utvrdio postojanje neslaganja u sintaktičkim obrascima Evansovih izjava. Svartvikova analiza ukazuje na to da su istražitelji vjerojatno dodali inkriminirajuće podatke u izjave, koje su zatim predstavljene kao njegove vlastite riječi i korištene kao dokazi na suđenju. U svojoj analizi Svartvik (1968) nabraja dva glavna problema, a to su to što je Evans bio gotovo pa nepismen, kao i drugi problem koji se često pojavljuje u forenzičnoj analizi teksta, a to je maleni uzorak na kojem se radi. Sveukupni korpus od četiri izjave sastoji se od 4861 riječi. Svartvik (1968) napominje da jezik korišten u Evansovim izjavama ne odgovara načinu izražavanja osobe poput Evansa i da neke od rečenica više odgovaraju govorniku koji je dobro upoznat s pravnim jezikom. Svartvikova analiza pokazuje važnost objektivnosti u analizi, što u ovom slučaju znači da značajke analize trebaju biti jasne i otvorene za vanjsku provjeru, da se mogu kvantificirati i testirati, no i da na kvantifikaciju može značajno utjecati i veličina uzorka. Često se ovaj problem rješava opširnom analizom drugih tekstova za koje je utvrđeno da ih je napisao isti autor čije autorstvo utvrđujemo, no budući da je Evans bio nepismen, to je bilo nemoguće, stoga se Svartvik oslonio na njegove ranije izjave na

suđenju. Svartvikova analiza (1968) pružila je uvid u korištenje statističkih metoda za forenzičnu analizu teksta, kao što je dužina rečenica u sve četiri izjave, ali i unutar izjava, ovisno o tematskoj cjelini, kao i količinu određenih vrsta zavisnih i nezavisnih rečenica unutar različitih cjelina. I dok se Svartvikova analiza većinom zasniva na kvantitativnim podacima, u narednim se godinama forenzička lingvistika razvila u područje u kojem se služi širokim rasponom metoda, kako kvantitativnih, tako i kvalitativnih.

Važno je napomenuti kako niti jedna znanost unutar forenzike ne može pružiti savršene dokaze. Uzrok čak 24% oslobađanja zatvorenika u SAD-u je dokazana nevinost nakon pogrešne interpretacije forenzičnih dokaza, poput uspoređivanje otiska zubi na ugrizima ili analiza DNK (The National Registry of Exonerations, n.d.). Kao glavni razlog za pogrešna osuđivanja *The Innocence Project* (n.d.) navodi to što je teško utvrditi vjerodostojnost rezultata i kako metode forenzične analize nisu dovoljno istražene, tj. rezultati nisu ispitani na dovoljno velikom broju uzoraka. Pogrešna interpretacija rezultata i neuzimanje u obzir koliko su neke pojave rijetke dovodi do velikog broja pogrešnih ili pretjerano nejasnih svjedočenja forenzičnih vještaka, što može dovesti do osuđivanja nevinih osoba. Stoga je važno što više ispitati metode i načine predstavljanja rezultata koje forenzičari, pa tako i forenzični lingvisti, koriste u analizi dokaza. Takva bi praksa ne samo značajno utjecala na živote mnogih osoba koje su osumnjičene za neko kazneno djelo, već bi bila i od značajnog doprinosa široj znanstvenoj zajednici.

2.1. Područja primjene forenzičke lingvistike

Unatoč tome što je riječ o relativno novoj disciplini, forenzička se lingvistika pokazala izuzetno korisnom u prikupljanju dokaza. Danas su vještačenja stručnjaka za forenzičku lingvistiku najčešća u anglofonim državama, kao i u Španjolskoj i Portugalu, te se ona u zadnjih 30 godina razvila i kao akademska disciplina (Coulthard, 2020).

Iako nosi naziv *lingvistike*, forenzička se lingvistika sastoji od mnoštva područja i nije homogena. Budući da je riječ o relativno novoj disciplini, područja forenzičke lingvistike nisu u potpunosti definirana, no postoji nekoliko pristupa forenzičnoj analizi teksta. Forenzička lingvistika može se rabiti za proučavanje pravnih tekstova, gdje istraživači proučavaju je li jezik koji se koristi u pravnim tekstovima dovoljno pristupačan svim čitateljima i načine na koji se može učiniti pristupačnijim, pogotovo ljudima koji nisu dobro upoznati s pravnim sustavom i onima čiji

materinski jezik nije isti kao jezik države u kojoj borave. Forenzička se lingvistika isto tako može koristiti i za analizu jezika korištenog na suđenjima, kao što su unakrsna ispitivanja, izlaganje dokaza, ispitivanja sumnjivaca, svjedočenja i retoriku odvjetnika u svrhu izazivanja željene reakcije, kao što je korištenje aktivnog jezika među tužiteljima i pasivnog jezika u obrani na suđenjima za silovanje (Bohner, 2001).

Forenzični lingvisti isto tako mogu pružiti podršku u mnoštvu aspekata istrage, kao što su analiza diskursa i sociolingvističko profiliranje (Fadden & Disner, 2014). Neki od primjera su analiza jezika priznanja, poput ranije spomenute Svartvikove analize (1968). Tu se ubraja i Coulthardova analiza izjave Dereka Bentleyja (Coulthard & Johnson, 2007), u kojoj je dokazano da je izjava Dereka Bentleyja, mladića osuđenog na smrt zbog ubojstva, dana pod utjecajem policajaca koji su ga ispitivali, unatoč tome što su na suđenju tvrdili da je izjavu dao svojom voljom, u obliku monologa. Coulthardova analiza narativnih obrazaca korištenih u izjavi pokazala je kako Bentleyjeva izjava nije bila monolog, već je utvrđeno da su mu policijski službenici postavljali pitanja koja su nakon toga pisali u obliku njegove vlastite izjave. Ovo otkriće dovelo je do posthumnog oslobađanja Dereka Bentleyja.

Forenzična analiza diskursa danas ima mnoštvo primjena i može se koristiti za proučavanje utjecaja konteksta na funkcionalni stil svjedoka (Saletović & Kišiček, 2012), govora mržnje (Prideaux, 2011), analizu relevantnosti tajnih snimki razgovora (Shuy, 2005), utvrđivanje autentičnosti oproštajnih pisama nakon suicida (Shapero, 2011), razlike moći kod razgovora osumnjičenih i svjedoka (Heydon, 2005), kao i proučavanje pravnih tekstova (Stygall, 2020) i jezika koji se koristi u sudnicama (Shuy, 2007).

Uz analizu diskursa, drugo područje koje se ističe u forenzičkoj lingvistici je sociolingvističko profiliranje. Ono je zasnovano na postavci da je jezična produkcija govornika pod utjecajem mnoštva društvenih faktora, kao što su dob, geografsko područje, obrazovanje ili ekonomski status. Cilj sociolingvističkog profiliranja je otkrivanje informacija o autoru ili porijeklu teksta, no ovakva se vrsta dokaza nikad nije prihvatila na sudu iz jezičnih i pravnih razloga (Perkins & Grant, 2018). Iako je nemoguće točno utvrditi tko je autor teksta prema sociolingvističkim značajkama zato što tekst može sadržavati nepotpune ili neodgovarajuće informacije ili autor može mijenjati svoj govor, ovakva vrsta profiliranja može biti vrlo korisna u

kombinaciji sa psihološkim profiliranjem i pomoći istražiteljima suziti broj sumnjivaca (Fadden & Disner, 2014). Sociolingvističko profiliranje danas je izrazito korisno u računalno posredovanoj komunikaciji (Schilling & Masters, 2015) kako bi se prepoznao utjecaj materinskog jezika (engl. *native language influence detection*) prema načinu izražavanja na stranom jeziku.

Uz razvoj informacijsko-komunikacijske tehnologije, kao i sve veću dostupnost lingvističkih alata i resursa, forenzični se lingvisti sve više okreću primjeni novih resursa i alata za utvrđivanje autorstva. Ovaj je zaokret doveo do stvaranja dvije grane, koje drugačijim pristupima pokušavaju doći do istoga cilja, utvrđivanja autorstva, o čemu će više biti rečeno u narednim odlomcima.

3. Utvrđivanje autorstva

Zadatak utvrđivanja autorstva teksta jest prepoznavanje značajki autora na temelju prethodno napisanih tekstova tog autora. Joula (2008) definira utvrđivanje autorstva kao svaki pokušaj otkrivanja značajki autora nekakve jezične građe. Pri tome napominje kako je definicija namjerno široka, kako bi se moglo uključiti i prepoznavanje govora. Prema Jouli (2008), postoje tri osnovna problema utvrđivanja autorstva:

1. Određivanje autora na određenom uzorku teksta koji pripada jednom od poznatih autora u skupu tekstova.

2. Određivanje autora na određenom uzorku teksta koji potencijalno pripada jednom od autora u skupu tekstova, no ne nužno.

3. Određivanje značajki jednog ili više autora na uzorku teksta.

Prvi problem koji Joula (2008) nabraja poznat je i kao zatvorena kategorija, dok je drugi problem otvorena kategorija. Glavna razlika je što se u prvom problemu radi sa zatvorenim skupom autora, dok je u drugom problemu potrebno razlikovati i između onih autora koji nisu uključeni u skup tekstova. Određivanje značajki jednog ili više autora poznato je i pod nazivom *stilometrija* ili *profiliranje* te je cilj odrediti značajke prema kojima se autor(i) mogu prepoznati.

Prema Coulthardu (2004), lingvisti pristupaju problemu utvrđivanja autorstva uz pretpostavku da svaka osoba ima vlastiti idiolekt. Međutim, mnogi lingvisti smatraju da je koncept

idiolekta previše apstraktan te da se već na temelju sličnosti i razlika u jeziku autora može utvrditi autorstvo (Grant, 2020). Stoga Grant (2013) predlaže da pri utvrđivanju autorstva nije nužno pokazati da se autor razlikuje od svih ostalih autora, već da je dovoljno utvrditi da razlike postoje između tog autora i ostalih relevantnih autora. U tom slučaju govori se o razlikama na razini relevantne populacije.

Joula (2008) tvrdi da riječi koje autor koristi mogu uputiti na identitet tog autora, ovisno o vremenu i kontekstu u kojem je neki tekst nastao. Neke od riječi mogu ukazati na vrijeme i prostor nastanka teksta (Johnstone, 1996), dok druge riječi mogu pokazati kojoj grupi autor pripada. Međutim, kako Joula (2008) navodi, ova vrsta analize ima dva problema. Prvi je to što je lako lažirati podatke, bilo da je riječ o osobi koja namjerno želi imitirati vokabular druge osobe, ili je riječ o lektorskim i uredničkim intervencijama u tekstu, zbog čega se čini da autor koristi oblike riječi koje ne koristi u drugim situacijama. Drugi je problem to što se neke riječi rijetko koriste i nije moguće izvući dovoljno pouzdane empirijske podatke o njihovom korištenju, pogotovo ako se radi o malom broju pojavnica koje nisu česte u korpusima.

Kao sigurniji način provjere značajki autorstva prema vokabularu, Joula (2008) navodi da je potrebno analizirati veliku količinu riječi u tekstu. Stoga je potrebno provjeriti značajke riječi korištenih u nekom tekstu, kao što su dužina, vrsta riječi ili mjerenja leksičke obogaćenosti teksta. Iako su takvi pristupi pokazali značajne rezultate u nekim slučajevima (Kruh, 1992), same po sebi nisu dostatne za donošenje pouzdanih zaključaka o autorstvu teksta. Stoga se predlaže provjera gramatičkih riječi, koje se češće koriste, a samim time i češće pojavljuju u tekstu te nisu ovisne o temi teksta.

Same po sebi, gramatičke riječi ne nose značenje, već opisuju odnose između leksičkih riječi. Stoga Joula (2008) predlaže da i sintaktičke značajke koje autor koristi u tekstovima, poput interpunkcijskih znakova, vrsta riječi i n-grama, mogu pružiti informacije na temelju kojih se može utvrditi autorstvo teksta. N-grami, koji pružaju informacije i o leksičkim, i o sintaktičkim značajkama teksta, pokazali su se kao izrazito produktivni pokazatelji autorstva teksta u forenzičkoj lingvistici (Wright, 2014), kao i u drugim primjenama u svrhu utvrđivanja autorstva, poput otkrivanja plagijata (Bosanac & Štefanec, 2011). Najuspješniji n-grami, tj. n-grami tipični za idiolekt određenog autora su, prema Wrightu (2017), oni koji se pojavljuju u testiranim

uzorcima, kao i u provjeravanom uzorku *Q* i uopće se ne pojavljuju tekstovima ostalih poznatih autora. Prema Langackeru (1988), društvene i povijesne jezične značajke te osobna iskustva utječu na to da određeni nizovi riječi postaju ukorijenjeni, stoga je potrebno provjeriti jesu li ti n-grami različiti od onih koje bi drugi autori koristili u istoj situaciji i kontekstu. Između ostaloga, moguće je analizirati i ortografske i idiosinkratske značajke u tekstu, kao što su pogreške u pisanju i razlike u kulturi koje se mogu vidjeti u tekstu. Odstupanje od dominantnog, najčešće korištenog oblika, odnosno označenost neke pojavnice, prema Chaski (2001), može biti ključni pokazatelj autorstva i može se primijeniti na sve razine lingvističke analize.

Međutim, način na koji lingvisti dolaze do podataka o značajkama teksta ovisi o pristupima koji koriste, a koji se mogu razlikovati ovisno o načinu pristupanja tekstu.

3.1. Stilometrijski i stilistički pristupi

Danas možemo govoriti o dvije metodologije utvrđivanja autorstva. S jedne strane, postoji stilistički pristup, poput ranije opisane Svartvikove analize, međutim, u zadnje se vrijeme isprofilirao i drugi, stilometrijski pristup. Dok stilistički pristup u fokus stavlja kvalitativnu analizu teksta i ne samo koje, već i kako i zašto autor koristi određene jezične značajke (Johnstone, 2000), glavni je problem stilističkog pristupa činjenica da je rezultate dobivene na takav način vrlo teško kvantificirati, a samim time, i dobiti objektivne rezultate (Grant, 2013). S druge strane, stilometrijski pristup utvrđivanju autorstva u središte stavlja automatizaciju te se provjerava niz ranije utvrđenih jezičnih značajki korištenjem statistike. Glavna zamjerka stilometrijskom pristupu utvrđivanju autorstva je to što je teško objasniti kako autori variraju koristeći lingvistička znanja, to jest, ne može se objasniti što to određenu značajku čini lingvistički relevantnom.

Međutim, ono što objedinjuje ova dva pristupa je fokus na leksik (Wright, 2014). Unatoč tome, korisnici stilometrijskih pristupa utvrđivanju autorstva stilističkom pristupu zamjeraju manjak baza podataka na temelju kojih donose odluke, što dovodi do subjektivnosti (Chaski, 2005), dok Grant (2013) tvrdi da taj pristup previše ovisi o subjektivnom znanju osobe koja provodi analizu, a zamjera mu i nedostatak mogućnosti ponovnog korištenja u drugim slučajevima, budući da se analiza radi od slučaja do slučaja (Nini & Grant, 2013). Dok stilistički pristup analizi teksta počinje od samog teksta i lingvist traži značajke u tekstu, stilometrijski pristup zasnovan je na nizu ranije utvrđenih značajki koje se traže u tekstu, što povećava mogućnost ponavljanja i

generalizacije rezultata (Wright, 2014). Unatoč prividnoj uspješnosti stilometrijskog pristupa, glavni je problem nedostatna utemeljenost u ranijim lingvističkim znanjima, što može utjecati na rezultate forenzične analize teksta (Grant, 2008). Stoga danas zagovaratelji stilometrijskog, ali i stilističkog pristupa, sve veći naglasak stavljaju na suradnju između dvije metodologije, u svrhu dobivanja što boljih rezultata, koji su istovremeno pouzdani, ali i utemeljeni u lingvističkim teorijama i znanjima (Wright, 2014).

Kako bi se postigao spoj između te dvije metodologije, lingvisti se okreću značajkama koje jasno ocrtavaju jezik autora. Neka od istraživanja provedenih u svrhu objedinjavanja metoda su Grantovo (2013) istraživanje, gdje koristi kvantitativne metode za utvrđivanje stilističkih značajki u SMS porukama, korištenje analize varijance za uspoređivanje čestotnosti određenih značajki (Nini & Grant, 2013), kao i Wrightova (2014, 2017) istraživanja o primjeni Jaccardovog koeficijenta na n-grame. Upravo Wright (2014) tvrdi da ovakva vrsta pristupa vodi do metoda koje su utemeljene u teoriji, ali i na pouzdanim i ponovljivim statističkim pristupima, koji pružaju rezultate koji se mogu objasniti koristeći ranija lingvistička znanja. Kao jedan od glavnih alata korištenih u ovu svrhu istaknuli su se korpusi, koji dopuštaju spajanje i unaprjeđivanje kvantitativnih i kvalitativnih rezultata (McEnery & Wilson, 2001). Stoga će se u nastavku rada predstaviti primjer utvrđivanja autorstva utemeljen na podacima dobivenim iz korpusa, koji će se prvo analizirati kvantitativno, a nakon toga će uslijediti daljnja analiza dobivenih rezultata, kako bi se pobliže proučile jezične karakteristike koje mogu poslužiti za utvrđivanje autorstva tekstova koje je napisao Antun Gustav Matoš.

4. Cilj, istraživačka pitanja i hipoteze

Cilj istraživanja je provjera značajki autorstva koje razlikuju pisca Augusta Gustava Matoša od četiriju ostalih autora uključenih u korpus tekstova na temelju kojih se radi analiza, kao i provjera koje metode pružaju najindikativnije rezultate.

Istraživački fokus ovog rada bio je provjeriti kako prosječna dužina riječi, prosječna dužina rečenica, omjer različnica i pojavnica, leme i n-grami pronađeni u korpusu tekstova mogu pomoći pri utvrđivanju autorstva teksta. U tu su svrhu postavljene sljedeće hipoteze temeljem pretpostavke da će Matoš kao pisac koristiti bogatiji jezični inventar (Kaštelan, 1956) od ostalih autora čija se djela analiziraju:

H1. Prosječna dužina riječi u tekstovima A.G. Matoša bit će značajno veća od ostalih autora.

H2. Prosječna dužina rečenica u tekstovima A.G. Matoša bit će značajno veća od ostalih autora.

H3. Omjer različnica i pojavnica u tekstovima A.G. Matoša bit će značajno veći od ostalih autora.

H4. Jaccardov koeficijent sličnosti lema u tekstovima A.G. Matoša bit će značajno veći nego u tekstovima ostalih autora.

H5. Jaccardov koeficijent sličnosti n-grama u tekstovima A.G. Matoša bit će značajno veći nego u tekstovima ostalih autora.

5. Metoda

U svrhu istraživanja prikupljeno je 26 tekstova pet hrvatskih autora nastalih krajem 19. i početkom 20. stoljeća. Tekstovi su javno dostupni u digitalnom formatu na portalu e-Lektire i dostupni su svima koji imaju elektronički identitet u sustavu AAI@EduHr, kao i onima koji imaju korisnički račun na portalu e-Lektire.

Kriteriji za izbor tekstova korištenih u istraživanju bili su da je riječ o pripovijetkama ili novelama, s brojem pojavnica većim od 1000 i da su dostupni u digitalnom formatu zbog lakše mogućnosti sastavljanja korpusa, budući da u slučaju već digitalno dostupnih tekstova nije potrebno prethodno digitaliziranje kao priprema za sastavljanje korpusa. Razlog izbora tekstova koji sadrže preko 1000 pojavnica jest to što mjerenja leksičke obogaćenosti teksta variraju ovisno o duljini toga teksta (Daller i sur., 2007). Istraživanja o utjecaju duljine teksta na leksičku obogaćenost teksta pokazala su da nakon 1000 pojavnica leksička obogaćenost teksta postaje stabilna neovisno o duljini teksta, dok u tekstovima s manje od 1000 pojavnica leksička obogaćenost teksta eksponencijalno raste ovisno o duljini teksta (Shi & Lei, 2022).

U korpus Pet autora uključeni su sljedeći tekstovi, koje će u istraživanju poslužiti kao primjeri poznatih tekstova autora, nabrojani u Tablici 1.

Tablica 1.

25 tekstova korištenih za izradu korpusa Pet autora

Antun Gustav Matoš	**Milan Begović**	**Ivan Kozarac**	**Joza Ivakić**	**Jure Turić**
Balkon	*Dva bijela hljeba*	*Gnjili...*	*Didak i baka*	*Njihova ljubav*
Cvijet sa raskršća	*Krzno od sibirske vjeverice*	*Moji ljudi*	*Gospodična Jelica*	*Prosci*
Ugasnulo svjetlo	*Kvartet*	*Stara rana*	*Mlada žena*	*Srce*
Lijepa Jelena	*Nerotkinja*	*Sudoperka*	*Otrov*	*Tko je kriv?*
Put u ništa	*Posljednji posjet*	*U nagonu*	*Sestra*	*U mraku*

Uz 25 nabrojanih djela, u istraživanje je uključen i tekst *Camao* A. G. Matoša, koji će u ovom slučaju poslužiti kao *Q* (provjeravani uzorak, engl. *questioned sample*), tj. tekst čije se autorstvo utvrđuje.

5.1. Sastavljanje korpusa

Prvi korak sastavljanja korpusa koji se sastoji od ranije nabrojanih djela bio je čišćenje teksta. Uklonjene su sve natuknice, kao i zvjezdice, brojevi i drugi znakovi koji se odnose na natuknice, spojnice, zaglavlja, te brojevi stranica i poglavlja iz teksta kako ne bi utjecali na analizu i doveli do pogrešnog prepoznavanja pojavnica unutar korpusa. Nakon toga, tekstovi su spremljeni u .txt formatu. Koristeći *Sketch Engine*, koji dopušta korisnicima stvaranje korpusa do milijun pojavnica, 25 tekstova uvezeni su i prikupljeni u korpus nazvan *Pet autora*. Korpus se sastoji od

ukupno 140,537 pojavnica i 115,803 riječi. Zatim smo stvorili pet potkorpusa, nazvanih po prezimenima autora čija su djela uključena u korpus.

Slika 1.

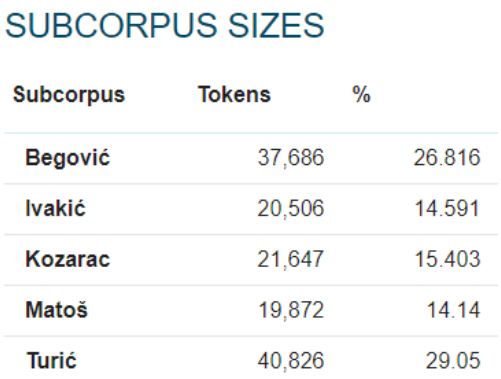
Informacije o korpusu Pet autora.

COUNTS

Tokens	140,537
Words	115,803
Sentences	7,406
Documents	25

Slika 2.

Informacije o potkorpusima Begović, Ivakić, Kozarac, Matoš i Turić.



SUBCORPUS SIZES

Subcorpus	Tokens	%
Begović	37,686	26.816
Ivakić	20,506	14.591
Kozarac	21,647	15.403
Matoš	19,872	14.14
Turić	40,826	29.05

Pri tome, važno je napomenuti kako je upravo potkorpus Antuna Gustava Matoša onaj s najmanjim brojem pojavnica. Iako su potkorpusi različitih veličina, ovakva situacija česta je u forenzičnoj analizi teksta, koja se nekad provodi na tekstovima različitih veličina (Cotterill, 2010). Također, radi se o korpusu s manjim brojem pojavnica nego što je to uobičajeno, budući da i forenzični lingvisti često moraju raditi na manjim tekstovima u svojim analizama (Cotterill, 2010). Stoga heterogenost broja pojavnica, kao i manji korpus, mogu dati više uvida o tome kako se metode korištene u istraživanju mogu koristiti i u drugim forenzičnim analizama teksta uz pomoć korpusa te ukazati na to koji su neki od mogućih izazova rada na tekstu ovakve veličine.

Uz korpus *Pet autora*, izrađen je i korpus *Camao*, koji je u sebi sadržavao samo tekst *Camao*. Podaci o tom korpusu mogu se vidjeti na Slici 3.

Slika 3.

Informacije o korpusu Camao.

5.2. Utvrđivanje obilježja teksta

Kako bi se provjerila obilježja Antuna Gustava Matoša koja bi mogla pomoći pri utvrđivanju autorstva njegovih djela, provedena je preliminarna analiza obilježja koja mogu biti karakteristična za tog autora. Pri tome su korištene dvije statističke metode, ANOVA test i Jaccardov koeficijent sličnosti.

5.2.1. Prikupljanje podataka

Kao prvi korak određivanja značajki autora provedena je provjera osnovnih stilometrijskih značajki teksta, kao što su dužina riječi i rečenica te omjer različnica i pojavnica unutar svakog djela. U tu svrhu korišten je paket *WordSmith Tools*, uz opciju *WordList* koja korisnicima omogućava izvlačenje statističkih podataka o korpusu.

Svi su pročišćeni tekstovi spremljeni u *WordSmith Tools* u *WordList* formatu (.lst), te analizirani funkcijom *WordList*. Kao glavni jezik, koji se koristi kao zadani pri pokretanju, postavljen je hrvatski. Dobiveni su podaci o frekvenciji pojavnica, kao i statistički podaci o svakom od tekstova, poput aritmetičke sredine broja riječi u svakoj rečenici, gdje se dužina rečenice mjeri u broju riječi koje se u njoj nalaze, prosječna dužina riječi mjerena u broju znakova, te obični i standardizirani omjer pojavnica i različnica unutar svakog teksta (ORP), kao i standardna devijacija mjera dužina riječi i rečenica. Alat *WordSmith* pruža podatke o omjeru različnica i pojavnica (ORP) i o standardiziranom omjeru različnica i pojavnica (SORP). Za razliku

od ORP-a, koji analizira cijeli tekst, standardizirani omjer različnica i pojavnica analizira svaki *n* broj riječi, a zatim nastavlja računati iznova za svaki *n* broj riječi nakon toga. Ako je *n*=1000, program će izračunati ORP za prvih 1000 riječi, zatim krenuti ispočetka za idućih 1000, i tako sve do kraja teksta, te na kraju analize daje rezultate o prosječnom omjeru pojavnica i različnica *n* riječi svakog od komada teksta. Budući da su ranija istraživanja (Shi & Lei, 2022) pokazala da nakon 1000 pojavnica leksička obogaćenost teksta postaje stabilna, za daljnje provjere izabran je omjer različnica i pojavnica, a ne standardizirani omjer različnica i pojavnica.

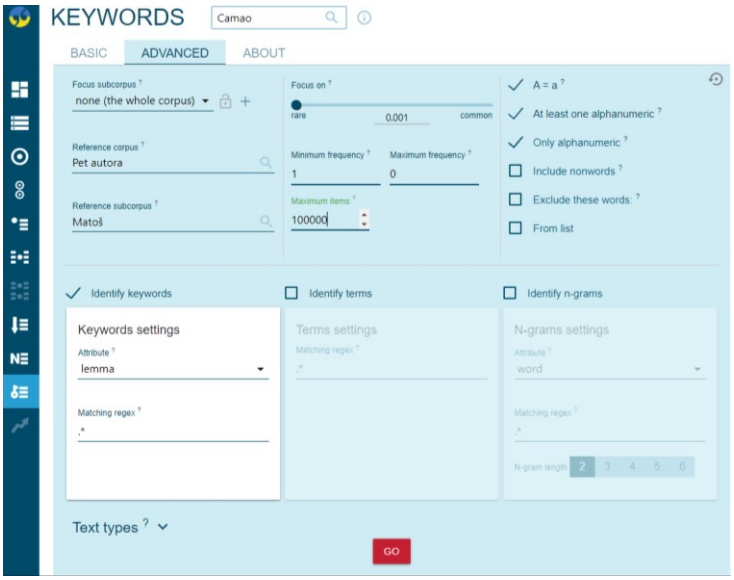
Nakon što su dobiveni podaci o stilometrijskim značajkama teksta, potrebno je dobiti podatke o zajedničkim lemama i n-gramima u poznatim tekstovima svakog od autora *K* tekstova (poznati tekstovi, engl. *known sample*) kako bi se provjerila vjerodostojnost metode. Prvi korak ove analize bila je priprema uzoraka 5%, 10%, 15%, 20% i 25% pet tekstova Antuna Gustava Matoša. Kako bi svi tekstovi bili jednako zastupljeni, od svakog teksta napravljen je uzorak odgovarajuće veličine, tako da se ukupni uzorak od 5% Matoševih teksta sastoji od 5% teksta svakog djela, 10% uzorka od 10% svakog djela i tako dalje. Analizirani uzorci izabrani su nasumično, a dijelovi teksta uključeni u uzorke različitih veličina mijenjali su se. Nakon toga izrađeni su uzorci 95%, 90%, 85%, 80% i 75% teksta za svakog od autora. Pri tome je važno napomenuti kako uzorci nisu bili jednake veličine, budući da je broj pojavnica svakog od autora drugačiji. Svi su uzorci spremljeni u dva korpusa. Manji uzorci Matoševih tekstova spremljeni su u korpus pod nazivom *Q tekstovi*, budući da će u svrhu ove provjere simulirati provjeravane tekstove, dok su ostali uzorci spremljeni u korpus *K tekstovi* jer predstavljaju poznate tekstove autora. Zatim su svi manji uzorci uspoređeni s pet uzoraka odgovarajuće veličine svakog od autora, tako da je 5% Matoševih tekstova uspoređeno s uzorcima od 95% tekstova Antuna Gustava Matoša, zatim 95% tekstova Milana Begovića, Joze Ivakića, Ivana Kozarca i Jure Turića, a nakon toga ista je provjera provedena i s 10% Matoševih i 90% tekstova svih autora i tako dalje.

Ovakva vrsta usporedbe pružit će mogućnost dodatne provjere vjerodostojnosti ove metode, kao i postoji li mogući utjecaj veličine uzorka na rezultate. Nakon provjere na poznatim uzorcima teksta, slijedi usporedba lema i n-grama poznatih tekstova svih autora s *Camaom*. U tu svrhu, svaki od potkorpusa unutar korpusa *Pet autora* uspoređen je s provjeravanim uzorkom *Camao*.

Za dobivanje podataka o zajedničkim lemama i n-gramima, korištena je funkcija *Identify keywords* u *Sketch Engineu*, koja korisnicima dopušta usporedbu dva korpusa i pruža podatke o frekvenciji termina, višerječnih izraza i n-grama. Prikaz ovakve vrste pretraživanja za leme u provjeravanom tekstu *Camao* i Matoševim poznatim tekstovima može se vidjeti na Slici 3.

Slika 4.

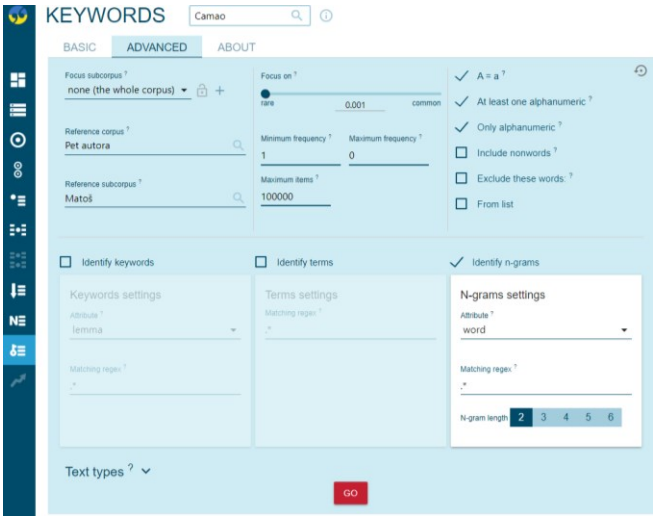
Usporedba lema korištenih u provjeravanom tekstu Camao i Matoševim poznatim tekstovima funkcijom Identify keywords.



Nakon što su prikupljeni podaci o lemama u tekstu, ista je provjera provedena i za svaki od bigrama, trigrama, 4-grama i 5-grama u tekstu. 6-grami nisu uključeni u analizu budući da niti jedan od provjeravanih uzoraka nije imao zajedničke 6-grame. Primjer pretraživanja zajedničkih bigrama u provjeravanom tekstu *Camao* i Matoševim poznatim tekstovima prikazan je na Slici 5.

Slika 5.

Usporedba bigrama korištenih u provjeravanom tekstu Camao i Matoševim poznatim tekstovima funkcijom Identify keywords.



5.2.2. Analiza varijance

Kako bi se provjerilo postoje li razlike u dužini riječi i rečenica te omjera različnica i pojavnica, provedene su tri jednosmjerne analize varijance (ANOVA) za svaki od statističkih podataka dobivenih funkcijom *WordList* da bi se utvrdilo postoje li statistički značajne razlike među autorima. Za provedbu ovih analiza korišten je SPSS i aritmetičke sredine podataka o dužini riječi i rečenica, kao i o omjeru različnica i pojavnica iskazanima u postotcima. Pri tome je prosječna dužina riječi i prosječna dužina rečenice te omjer različnica i pojavnica za pojedini tekst tretiran kao podatak (X). Nakon toga, provedene su dvije analize varijance za podatke o dužini riječi i rečenica kako bi se provjerilo postoje li odstupanja unutar autora.

Budući da nisu dobiveni podaci o dužini svake riječi i svake rečenice u tekstovima, već samo aritmetička sredina i standardne devijacije dužine riječi i dužine rečenica, ANOVA je računata ručno prilagođenom formulom za ovakve situacije.

Prvi korak bio je provesti ANOVA-u za svih pet poznatih djela jednoga autora, a ako razlika F-omjera bude značajna, potrebno je provesti i post-hoc analizu kako bi se utvrdilo koji se tekstovi razlikuju od ostalih u skupini. Na početku je potrebno postaviti dvije hipoteze, H_0, prema kojoj se djela unutar grupe neće razlikovati međusobno, kao i alternativnu hipotezu H_1, da će se

barem jedan tekst razlikovati od ostalih tekstova unutar grupe. Budući da su poznati podaci o prosječnom broju znakova (za dužinu riječi) ili riječi (za dužinu rečenice) M_x, kao i o pripadajućoj standardnoj devijaciji, SD_x, utvrđen je zbroj kvadrata odstupanja između grupa, ili SSB, prema ovoj formuli:

$$SSB = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

Zbroj kvadrata odstupanja između grupa računa se kao umnožak broja N, što u ovom slučaju predstavlja broj riječi u tekstu, s M_x (dužina riječi) i prosjekom svih M-ova (aritmetičkom sredinom dužina riječi), na kvadrat:

$$SSB = N(M_x - M)^2$$

Nakon toga, određuje se zbroj kvadrata odstupanja unutar grupa, tj. varijanca, koristeći standardnu devijaciju:

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Računa se kao kvadrat standardne devijacije svakog od tekstova, pomnožen s brojem N-1:

$$SSW = (SD)^2 N_x - 1$$

Nakon toga slijedi izračun ukupnog zbroja kvadrata odstupanja, to jest, zbroj kvadrata odstupanja između grupa i kvadrata odstupanja unutar grupa:

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

$$SST = SSB + SSW$$

Nakon što smo potrebni podaci o zbrojevima kvadrata dobiveni, napravljena je tablica u kojoj su navedeni zbrojevi, stupnjevi slobode, sredine kvadrata i F-omjer:

Tablica 2.

Prikaz tablice sa zbrojevima kvadrata, stupnjevima slobode, sredinama kvadrata i F-omjerom

Izvor varijabilnost	**Zbrojevi kvadrata**	**Stupnjevi slobode**	**Sredine kvadrata**	**F-omjer**
Između grupa	*SSB*	$k-1$	$MSB = SSB/(k-1)$	MSB/MSW
Unutar grupa	*SSW*	$n-k$	$MSW = SSW/(n-k)$	
Ukupno	*SST*	$n-1$		

Ako je utvrđeno da se tekstovi međusobno razlikuju usporedbom dobivenog F-omjera s kritičnim F-omjerom za te stupnjeve slobode, provodi se post-hoc analiza korištenjem Tukeyjevog testa. Razlog izbora ovog testa je to što je robusniji na nejednake brojeve slučajeva u grupama, zbog čega je formula prilagođena. Za provođenje Tukeyjevog testa potrebno je postaviti nultu hipotezu, H_0, koja glasi:

$$\mu_B = \mu_A$$

Isto tako, postavlja se i alternativna hipoteza H_1:

$$\mu_B \neq \mu_A$$

Oznake A i B odnose se na bilo koji par grupa. Budući da se u grupama nalazi nejednak broj riječi, u ovom se slučaju Tukeyjev test računa prema ovoj formuli:

$$q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{\frac{MS_{Error}}{n_i} + \frac{MS_{Error}}{n_j}}{2}}}$$

Prvi par grupa provjeravanih u Tukeyjevom testu predstavljaju dva teksta, onaj s najvećim, kao i onaj s najmanjim M-om, to jest, tekstovi s najkraćim i najdužim riječima ili rečenicama. Pod $\bar{X}$ se uvrštavaju aritmetičke sredine dužine riječi ta dva teksta, a zatim se pod MS_{Error} uvrštavaju sredine kvadrata unutar grupe. Naposljetku, uvrštavaju se i brojevi znakova (za dužinu riječi) ili riječi (za dužinu rečenica). Nakon što su dobiveni rezultati o iznosu q, uspoređuje se s tablicom kritičkih omjera $q_{critical}$ za dobivene stupnjeve slobode. Ako je q manji od $q_{critical}$, prihvatit će se hipoteza H_0, a ako je q veći od $q_{critical}$, odbacuje se H_0 i prihvaća alternativnu hipotezu H_1. U slučaju prihvaćanja alternativne hipoteze H_1, potrebno je nastaviti provjeru na sljedećoj grupi tekstova, onom s najdužim riječima i drugom najmanjem tekstu prema provjeravanoj dužini. Postupak je potrebno nastaviti sve dok se ne provjere hipoteze za svaki od parova.

5.2.3. Jaccardov koeficijent sličnosti

Kako bi se provjerilo koji tekstovi dijele najviše zajedničkih lema i n-grama, korišten je Jaccardov koeficijent sličnosti, koji je već i ranije korišten u forenzičnim analizama teksta (Wright, 2017; Joula, 2013; Grant, 2013). Jaccardov koeficijent sličnosti pokazuje u koliko se slučajeva određeni elementi pojavljuju ili ne pojavljuju u određenom skupu. Formula se može raspisati i ovako:

$$J(A,B) = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}}$$

Vrijednost Jaccardovog koeficijenta kreće se od 0 do 1, gdje 0 znači da ne postoji nikakva sličnost između dva skupa, dok 1 znači da su skupovi identični. Skup A koji ostvari najveći Jaccardov koeficijent sličnosti pripisat će se onom skupu s kojim ga je ostvario, što u slučaju ove provjere znači da se autorstvo provjeravanih tekstova pripisuje autoru poznatih tekstova za koje je Jaccardov koeficijent sličnosti najveći.

6. Rezultati

U središtu prvog dijela predstavljanja rezultata nalaze se rezultati analize varijance dužina riječi i rečenica, kao i omjera različnica i pojavnica, koji pružaju podatke o leksičkoj obogaćenosti teksta. Nakon toga slijede rezultati provjere Jaccardovog koeficijenta sličnosti za leme i n-grame korištene u poznatim djelima iz korpusa *Pet autora*, kao i rezultati usporedbe s tekstom *Camao*.

6.1. Analiza varijance za usporedbu dužina riječi i dužina rečenica autora

Kako bi se provela analiza dobivenih statističkih podataka o tekstu, kao i provjerilo postoje li u uzorcima tekstovi koje se međusobno razlikuju iako je poznato da ih je napisao isti autor, provedena je analiza varijance i post-hoc analiza postupkom koji je opisan u prijašnjem dijelu rada. Budući da SPSS ne dopušta provođenje ovakve vrste analize varijance u slučajevima kad su dostupni samo podaci o aritmetičkoj sredini i standardnoj devijaciji, a ne i pojedinačni podaci, provjera je provedena ručno na dobivenim statističkim podacima o dužinama riječi, kao i o dužinama rečenica za svaki od poznatih tekstova. Dobiveni podaci o dužinama riječi i dužinama rečenica u tekstu mogu se pronaći u Prilogu.

6.1.1. Dužina riječi

Za tekstove Antuna Gustava Matoša ($F=7,18$; df: 4,17064; $p<0,01$), postoji statistički značajna razlika, uz razinu rizika od jedan posto, između *Cvijet sa raskršća* i Matoševa ostala četiri djela, tako da je autor u *Cvijetu sa raskršća* koristio statistički značajno kraće riječi nego u ostalim tekstovima.

Djela Milana Begovića ($F=48,64$; df: 4,32455; $p<0,01$) također imaju statistički značajnu razliku. *Nerotkinja* ima kraće riječi nego ostala četiri teksta. *Dva bijela hljeba* i *Posljednji posjet* međusobno se ne razlikuju po dužini riječi, ali imaju kraće riječi nego *Krzna od sibirske vjeverice* i *Posljednjeg posjeta*, koji se isto tako razlikuju međusobno. Begovićeve se djela mogu podijeliti u tri skupine. Prvu skupinu čini *Nerotkinja*, čija je prosječna dužina riječi 3,75 znakova, zatim *Dva bijela hljeba* i *Posljednji posjet*, čije riječi u prosjeku sadrže četiri znaka i *Krzno od sibirske vjeverice* i *Posljednji posjet*, čije se riječi sastoje od prosječno 4,25 znakova. Iako postoje statistički značajne uz razinu rizika od jedan posto, jasno je kako je i dalje radi u razlici od jednog znaka.

I za pet tekstova Jure Turića ($F=18,93;df:4,36160$; $p<0,01$) utvrđena je statistički značajna razlika. Njegovi se tekstovi mogu podijeliti u dvije skupine prema dužini riječi. Prva skupina, *Prosci* i *U mraku* sadrže statistički značajno duže riječi od druge skupine, koja se sastoji od *Njihova ljubav*, *Tko je kriv* i *Srce*. I u ovom slučaju postoji statistički značajna razlika uz razinu rizika od jedan posto, no riječ je u razlici dužine od četvrtine znaka.

Djela Ivana Kozarca ($F=8,43;df: 4,18028$; $p<0,01$) statistički se značajno razlikuju. Post-hoc analize pokazale su razlike između *Moji ljudi* i *Stare rane*, *Moji ljudi* i *Gnjili...* te *Moji ljudi* i *Sudoperka*. Pri tome, *Moji ljudi* ima duže riječi nego ostala tri teksta. Također, *U nagonu* ima duže riječi samo u usporedbi sa *Starom ranom*. Za ostale je usporedbe prihvaćena nulta hipoteza, to jest, ti se parovi tekstova ne razlikuju statistički značajno po dužini riječi.

Naposljetku, tekstovi Joze Ivakića ($F=45,70;df: 4,18408$; $p<0,01$) također se statistički značajno razlikuju. Prema post-hoc analizi, utvrđene su razlike između tekstova *Gospodična Jelica* i *Otrov*, *Gospodična Jelica* i *Sestra*, kao i *Gospodična Jelica* i *Didak i baka*, gdje *Gospodična Jelica* ima duže riječi od ostala tri djela. Za ostale članove grupe prihvaćena je nulta hipoteza i ti se parovi tekstova ne razlikuju po dužini riječi na statistički značajan način.

Budući da su ovom metodom utvrđene razlike unutar poznatih djela autora, nismo nastavili analizu i uz *Camao* jer se metoda pokazala nevjerodostojnom u slučaju dužina riječi ovih autora.

6.1.2. Dužine rečenica

Kao i za dužine riječi, provedena je ANOVA na temelju podataka o broju rečenica, prosječnoj dužini rečenica i pripadajućoj standardnoj devijaciji.

U tekstovima Antuna Gustava Matoša ($F=9,01;df:4, 1223$; $p<0,01$), postoji statistički značajna razlika, uz razinu rizika od jedan posto, između dužine rečenica u tekstovima *Lijepa Jelena* i *Cvijet sa raskršća*, *Lijepa Jelena* i *Put u ništa*, kao i za *Lijepa Jelena* i *Balkon*. Post-hoc analiza pokazala je kako autor u djelu *Lijepa Jelena* koristi statistički značajno duže rečenice u odnosu na ostale tekstove, osim za *Ugasnulo svjetlo*. Autor u tekstu *Lijepa Jelena* koristi u prosjeku pet riječi više nego u rečenicama *Cvijeta s raskršća*, dok u odnosu na ostale tekstove koristi između dvije i četiri riječi više.

Djela Milana Begovića (F=14,83;df:4, 2362; p<0,01) isto tako imaju statistički značajnu razliku u dužini rečenica te se prema dužini mogu podijeliti u dvije skupine. Prva skupina sastoji se od *Krzna od sibirske vjeverice*, *Kvarteta* i *Dva bijela hljeba*, dok se druga sastoji od *Posljednjeg posjeta* i *Nerotkinje*. Ove se dvije skupine međusobno razlikuju tako da prva skupina sadrži statistički značajno duže rečenice, dok druga skupina ima kraće rečenice.

Jure Turić (F=12,18;df:4, 2292; p<0,01) isto se tako razlikuje unutar svojih tekstova prema dužini rečenice. Post-hoc analiza pokazala je statistički značajnu razliku u dvije skupine tekstova. Skupina *Tko je kriv* i *Njihova ljubav* sadrži statistički značajno duže rečenice od skupine *U mraku* i *Prosci*, dok *Srce* sadrži statistički značajno duže rečenice samo od *U mraku*.

Ivan Kozarac (F=12,52;df:4, 1341; p<0,01) ima dva teksta koji se statistički značajno razlikuju od ostalih prema dužini rečenica. *U nagonu* razlikuje se od *Sudoperke*, *Moji ljudi* i *Stare rane*, dok se ne razlikuje od *Gnjili…*, koja se zauzvrat razlikuje samo od *Sudoperke*. U slučaju ostalih tekstova potvrđena je nulta hipoteza.

Joza Ivakić (F=14,35;df:4, 1316; p<0,01) isto tako varira unutar broju rečenica u tekstovima. *Mlada žena* statistički se značajno razlikuje se od *Otrova*, *Gospodične Jelice* i *Sestre*, dok se *Didak i baka*, *Sestra* i *Gospodična Jelica* razlikuje samo od *Otrova*.

Budući da su ovom metodom utvrđene razlike unutar poznatih djela autora, analiza nije nastavljena za *Camao* jer se i ova metoda pokazala nevjerodostojnom u slučaju ovih autora.

6.2. Jednosmjerna analiza varijance

U Tablici 3 prikazani su deskriptivni podaci za varijable koje su praćene u istraživanju.

Tablica 3.

Deskriptivni podaci za varijable pet autora.

	Dužina rečenica		**ORP**		**Dužina riječi**	
Autori	**M**	**SD**	**M**	**SD**	**M**	**SD**
Matoš	14,21	2,20	50,90	2,48	4,44	0,13
Kozarac	14,04	2,36	41,17	4,91	4,01	0,11

Turić	16,32	1,86	31,16	4,96	3,83	0,11
Begović	13,30	2,19	36,72	5,83	4,05	0,24
Ivakić	13,18	2,52	35,07	4,35	3,69	0,24
Prosjek	14,21	2,35	39,00	8,10	4,01	0,31

Kako bi se provjerilo razlikuju li se djela pet autora po dužini riječi, dužini rečenica i odnosa različnica i pojavnica (ORP), provedene su tri jednosmjerne analize varijance u SPSS-u.

Testovi homogenosti varijance nisu statistički značajni niti za jednu od tri provedene analize te je zbog toga opravdano provoditi analize. Analize su pokazale da su F-omjeri za dužinu riječi ($F=7,43$; $p<0.01$) i ORP ($F=11,59$; $p<0.01$) statistički značajni, dok je F-omjer za dužinu rečenica neznačajan ($F=1,41$; $p>0.10$).

Provedene su post-hoc analize, Scheffeov test i Boferronijev test, za dužinu riječi i ORP kako bi se utvrdilo koji autori se razlikuju od drugih autora. Za dužinu riječi utvrđene su statistički značajne razlika uz razinu rizika od 1% između Matoša i Turića te Matoša i Ivakića, a na razini rizika od 5% između Matoša i Kozarca i Matoša i Begovića. Razlika Begovića i Ivakića značajna je uz razinu rizika od 10% prema Schefferovom testu, a prema Bonferronijevom testu, razlika je značajna uz razinu rizika od 5%. Uvidom u Tablicu 3. i usporedbom pripadajućih aritmetičkih sredina utvrđeno je da je Matoš koristio statistički značajno dulje riječi nego četiri preostala autora.

Analize omjera različnica i pojavnica pokazale su da se Matoš razlikuje od Kozarca na razini rizika od 10% (po Scheffeovom testu), na razini rizika od 5% po Bonferronijevom testu, a od ostalih autora Matoš se razlikuje na razini rizika od 1%. Kozarac se razlikuje od Turića uz razinu rizika od 5%. Matoš ima prosječni ORP statistički značajno veći nego ostali autori, a Kozarac ima prosječni ORP statistički značajno veći nego Turić.

6.3. Omjer različnica i pojavnica

Uvidom u analizu omjera različnica i pojavnica dobivenu alatom *WordSmith Tools*, utvrđeno je da omjer različnica i pojavnica u *Camau* iznosi 45,68%. SPSS ne dopušta daljnje provođenje jednosmjerne analize varijance na samo jednom podatku, a zbog nedostatka informacija o standardnoj devijaciji u tekstu, daljnji nastavak analize varijance nije moguć.

Međutim, uvidom u ranije dobivene rezultate, prema kojima se Matoš razlikuje od svih ostalih autora tako da koristi statistički značajno više različnica od ostalih autora, dobiveni podaci za *Camao* najbliži su upravo rezultatima dobivenima za Matoševe poznate tekstove, budući da je omjer različnica i pojavnica u *Camau* veći od ORP-a ostalih autora.

6.4. Jaccardov koeficijent sličnosti

Provedene su dvije analize za provjeru razlika u n-gramima i lemama poznatih autora te mogu li se tekstovi *Q* prepoznati koristeći Jaccardov koeficijent sličnosti. Cilj prve analize bio je utvrditi može li se autor prepoznati prema lemama koje koristi, a cilj druge analize bio je provjeriti može li se autor uspješno prepoznati koristeći potvrđene n-grame autora u usporedbi s n-gramima korištenima u nepoznatim tekstovima *Q*.

6.5. Leme

Prije početka usporedbe lema iz provjeravanog teksta *Camao* s lemama iz poznatih tekstova, provedeno je probno mjerenje na uzorcima iz poznatih tekstova Antuna Gustava Matoša. U tu svrhu napravljeni su uzorci od 5%, 10%, 15%, 20% i 25% tekstova Antuna Gustava Matoša, koji su spremljeni u korpuse u *Sketch Engineu*. Zatim je ostatak Matoševih tekstova, kao i uzorci tekstova ostalih autora podijeljen u korpuse od 95%, 90%, 85%, 80% i 75% tekstova tih autora. Svaki od uzoraka Matoševih tekstova uspoređen je sa svojim parovima tekstova, tako da je uzorak od 5% uspoređen s 95% poznatih tekstova, uzorak od 10% s 90% poznatih tekstova i tako dalje.

Uz 80% uspješnosti, leme iz 10%, 15%, 20% i 25% uzoraka Matoševih tekstova uspješno su pripisane ostatku Matoševih tekstova. Međutim, 5% uzoraka teksta je pogrešno pripisano 95% teksta Joze Ivakića. Rezultati usporedbe mogu se vidjeti u Tablici 4.

Tablica 4.

Rezultati provjere Jaccardovog koeficijenta sličnosti na poznatim tekstovima.

Veličina uzorka 1	**Veličina uzorka 2**	**Antun Gustav Matoš (%)**	**Milan Begović (%)**	**Joza Ivakić (%)**	**Josip Kozarac (%)**	**Jure Turić (%)**
5%	**95%**	7,22%	5,07%	7,39%	5,77%	5,59%

10%	**90%**	11,40%	8,22%	10,64%	8,93%	8,59%
15%	**85%**	14,07%	9,97%	12,49%	10,85%	10,64%
20%	**80%**	15,43%	11,49%	13,40%	11,90%	11,99%
25%	**75%**	16,76%	12,65%	13,97%	12,92%	12,87%

Razlog pogrešnog pripisivanja 5% uzoraka Matoševih poznatih tekstova Jozi Ivakiću leži u tome da Joza Ivakić u svojim poznatim tekstovima ima najmanji broj lema od svih autora, što upućuje na nisku razinu leksičke raznolikosti u tekstu. Iako 5% teksta i 95% teksta Antuna Gustava Matoša dijele 326 zajedničkih lema, a 5% Matoševih tekstova i 95% Ivakićevih tekstova samo 231 lemu, broj lema u tih 95% teksta iznosi 2639 lema, dok je u 95% Matoševih poznatih tekstova 4029 lema. Rijetko koja od lema iz 95% Ivakićevih tekstova može se smatrati stilski označenom. S druge strane, u 95% Matoševih tekstova mogu se pronaći stilski označene leme koje dijeli s 5% teksta, kao što su „naočar“, „nesrećan“ ili „pariski“. Budući da se u ovom slučaju radi o vrlo maloj razlici, a zajedničke leme nisu stilski obilježene, analiza je nastavljena.

U drugom dijelu analize, koristeći *Sketch Engine* i *Identify keywords*, dobiveni su podaci o svim lemama u tekstu. Nakon toga, poznati tekstovi svih pet autora uspoređeni su s Matoševim tekstom *Camao* koji je poslužio kao nepoznati tekst *Q* u svrhu ove analize. Usporedbom svakog od potkorpusa s tekstom *Q* dobiveni su podaci o zajedničkim lemama u poznatim tekstovima svakog od autora, kao i o lemama u tekstu *Q*. Kao tekst *Q* poslužila je Matoševo djelo *Camao*. U usporedbi s korištenim lemama poznatih autora, *Camao* je najsličniji poznatim tekstovima Antuna Gustava Matoša, s kojima dijeli 17,18%, to jest, 1110 lema. Nakon Matoša, 14,88% sličnosti s *Camaom* dijele leme u tekstovima Joze Ivakića, 14,53% leme Jure Turića, 14,07% leme Ivana Kozarca, a 13,94% zajedničkih lema su između teksta *Camao* i lema koje koristi Milan Begović.

Uvidom u dobivene rezultate usporedbe može se uočiti kako leme koje tekst *Camao* i ostalih pet analiziranih tekstova A.G. Matoša nisu samo pojavnice koje se mogu često čuti u govoru, već među njima postoje i pojavnice koje mogu pružiti više uvida u informacije o autorstvu, kao i o samom autoru. Neki od primjera pojavnica koje se pojavljuju samo u tekstu *Camao* i u uzorku poznatih tekstova Antuna Gustava Matoša nalaze se u Tablici 5.

Tablica 5.

Primjeri pojavnica u tekstu Camao i poznatom uzorku A. G. Matoša.

Lema	**Primjer iz teksta *Q* (*Camao*)**	**Primjer iz poznatih tekstova**	**Poznati tekstovi u kojima se pojavljuju**
sljepoočica **Frekvencija (*Camao*): 1** **Frekvencija (poznati tekstovi): 4**	„Zabolilo ga u sljepočicama.“ (Matoš, 1900, str. 47)	*Balkon*: „Hitajući pred veče spram dragog vrta, okrene zlehuda moja slutnja noćno nebo u olovnu lubanju, a vjetar me lupa i lupa od jedne metalne sljepočice u drugu.“ (Matoš, 1909, str. 44)	*Balkon* *Lijepa Jelena* *Put u ništa*
nješto **Frekvencija (*Camao*): 3** **Frekvencija (poznati tekstovi): 12**	„No Kamenski nije mrzio; bijaše to nješto drugo.“ (Matoš, 1900, str. 44)	*Pur u ništa*: „Ja je dakle Ništa i treba ga realizovati, valja dakle Ništa pretvoriti u Nješto.“ (Matoš, 1909, str. 92)	*Put u ništa* *Ugasnulo svjetlo* *Lijepa Jelena*
alaj **Frekvencija (*Camao*): 1** **Frekvencija (poznati tekstovi): 2**	„Alaj si mi oslabio, uvedrio, lijepi viteže, idući na daleka proštenja!“ (Matoš, 1900, str. 48)	*Cvijet sa raskršća*: „Bože, alaj se uplaših, jer malo te ne padoh preko vas!“ (Matoš, 1909, str. 99)	*Cvijet sa raskršća* *Put u ništa*

Tablica 5. - nastavak

Primjeri pojavnica u tekstu Camao i poznatom uzorku A. G. Matoša.

atelijer **Frekvencija (*Camao*): 4** **Frekvencija (poznati tekstovi): 4**	**„Moj gusar je u Parizu, hara na burzi i vucara se po atelijerima i iza kulisa.“ (Matoš, 1900, str. 48)**	***Lijepa Jelena*: „I zlovoljan, otrovan padah na ležište u mom visokom atelijeru navrh Ménilmontanta, dok zorom ne usnuh, snivajući o ženi sa licem divnim od duše i očima velikim od ljubavi.“ (Matoš, 1909, str. 23)**	***Lijepa Jelena***
naočar **Frekvencija (*Camao*): 3** **Frekvencija (poznati tekstovi): 2**	„Samotno šumsko veče, plahi i pogureni Š. u crnim naočarima sastaje se sa crnim, vampirskim očima i strasnim usnama koje su namijenjene drugomu.“ (Matoš, 1900, str. 45)	*Lijepa Jelena*: „Kola se jedva počeše micati, a ona namještaše koprenu i naočare.“ (Matoš, 1909, str. 26)	*Lijepa Jelena* *Ugasnulo svjetlo*

Kako bi se dodatno provjerilo mogu li dobivene leme poslužiti kao indikatori autorstva, provedena je usporedba parova sinonima u korpusu hrWaC. Rezultati se nalaze u Tablici 6.

Tablica 6.

Usporedba označenih i neoznačenih parova u korpusu hrWaC.

Prvi oblik	**Relativna frekvencija (na milijun pojavnica)**	**Drugi oblik**	**Relativna frekvencija (na milijun pojavnica)**
sljepoočica	0,1	sljepoočnica	0,78
nješto	0,13	nešto	766,21
alaj	0,38	ala	5,48
atelijer	0,98	atelje	2,89
naočar	0,19	naočale	19,48

Dobiveni rezultati pokazuju kako se leme pronađene u Matoševim tekstovima koriste rjeđe te da bi zbog toga mogle poslužiti kao indikatori autorstva, budući da dobiveni oblici predstavljaju označene oblike te je njihova frekvencija manja od frekvencije neoznačenih oblika.

6.6. N-grami

Da bi se provjerila primjena na n-grame, koji su se, prema Wrightu (2017) pokazali kao vrlo uspješni kandidati za utvrđivanje autorstva, prvo je provedena probna verzija mjerenja na poznatim tekstovima Antuna Gustava Matoša. Izabrani su uzorci 5%, 10%, 15%, 20% i 25% tekstova Antuna Gustava Matoša i zatim su spremljeni u korpuse u *Sketch Engineu*, kao nepoznati, tj. *Q* tekstovi. Nakon toga, preostali tekstovi Antuna Gustava Matoša, kao i tekstovi ostalih autora, spremljeni su u korpuse koji su sadržavali 95%, 90%, 85%, 80% i 75% tekstova tih autora, koji su poslužili kao poznati tekstovi *K*. U provjeru su uključeni bigrami, trigrami, 4-grami i 5-grami.

Ova je metoda uz 80% uspješnosti pripisala Matoševe n-grame točnom autoru. Svi se rezultati ove provjere nalaze u Prilogu.

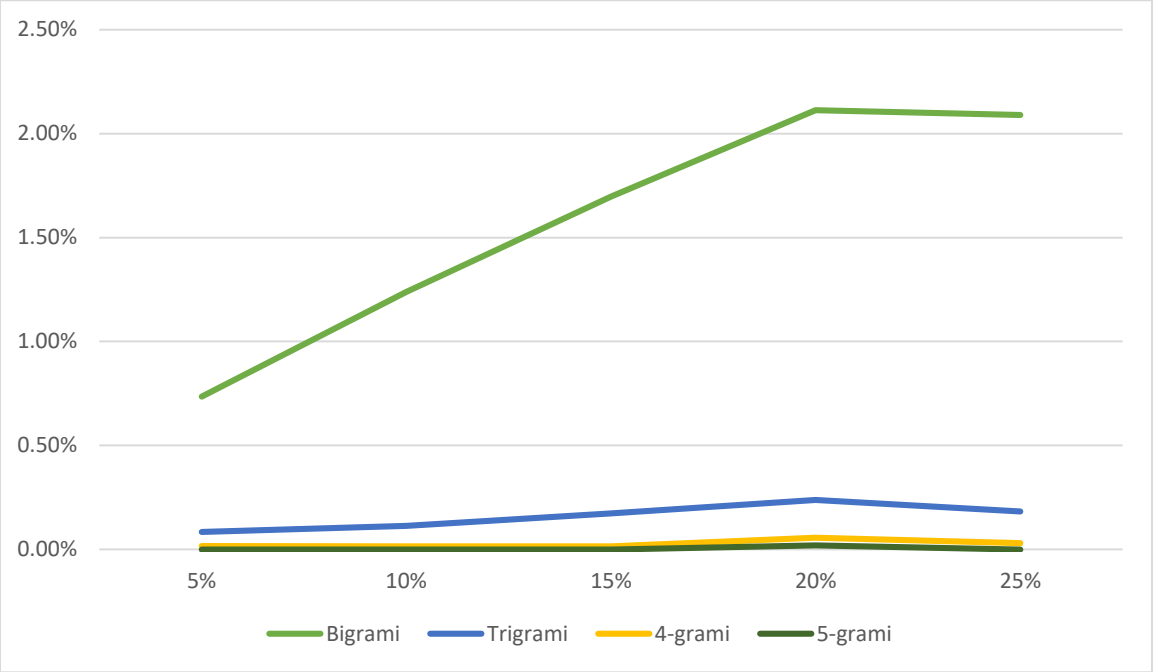
Bigrami, trigrami i 4-grami imali su 100%-nu uspješnost u prepoznavanju sličnosti između Matoševih tekstova, a jedini neuspješni bili su 5-grami. Iako su svi tekstovi sadržavali 5-grame, samo je u usporedbi 20% uzorka s 80% tekstova uspješno prepoznat jedan 5-gram, koji je točno pripisan Matošu. U niti jednom drugom tekstu nije uspješno prepoznat niti jedan zajednički 5-gram ni za kojega autora, to jest, Jaccardov koeficijent sličnosti za 5-grame iznosio je 0. Ukoliko

bi se 5-grami isključili iz ovog mjerenja, uspješnost prepoznavanja iznosila bi 100% za cijelu provjeru.

Rezultati prepoznavanja mogu se provjeriti na Slici 6.

Slika 6.

Rezultati provjere Jaccardovog koeficijenta sličnosti na n-gramima poznatih tekstova za Matoša.



Koristeći 5% Matoševih tekstova u usporedbi s 95% tekstova ostalih autora, bigrami su uspješno identificirali Matoša kao najvjerojatnijeg autora 5% tekstova. Matoševi tekstovi sadržavali su 0,73% bigrama, tj. 74 bigrama korištenih u 95% ostalih tekstova, dok su ostali autori dijelili manji udio bigrama. Trigrami u 5% Matoševih tekstova isto su tako uspješno pripisani Matošu uz 0,08% sličnosti, to jest, sedam zajedničkih trigrama, dok su 4-grami pripisani uz 0,02% sličnosti, to jest, imali su jedan zajednički 4-gram.

Na temelju 10% Matoševih tekstova, bigrami su uspješno pripisani Matošu, uz 1,24% ili 131 zajednički bigram, dok su trigrami uz 0,11% i 10 istih trigrama isto tako uspješno pripisani Matošu. Jedan zajednički 4-gram i 0,01% sličnosti pripisani su Matošu.

Bigrami u 15% Matoševih tekstova isto su tako uspješno pripisani Matošu, uz 1,69% i 188 zajedničkih bigrama. Trigrami su pripisani Matošu uz 0,17% sličnosti sa 16 zajedničkih trigrama, a dva su teksta ponovno dijelila jedan zajednički 4-gram uz 0,01% sličnosti.

U 20% tekstova, bigrami su ponovno uspješno pripisani Matošu uz 2,11% sličnosti i 235 zajedničkih bigrama, kao i trigrami s 22 zajednička trigrama, odnosno 0,24%, dok su ovaj put u tekstu dijelili čak četiri 4-grama uz 0,06% sličnosti i jedan zajednički 5-gram uz 0,02% sličnosti. Upravo je usporedba 20% i 80% sličnosti teksta pružila najviše Jaccardove koeficijente, dok je usporedba 25% i 75% teksta polučila nešto niže koeficijente. Tekstovi su dijelili 221 bigram, to jest, 2,09% bigrama, 16 trigrama odnosno 0,18% zajedničkih trigrama, dok su se u tekstu pojavila i dva zajednička 4-grama, to jest, ova dva teksta imali su Jaccardov koeficijent u iznosu od 0,03%. Razlog ovoj razlici leži u tome da su uzorci nakon podjele u 25%, odnosno 75% Matoševih tekstova prebačeni u isti tekst.

Uz 80% uspješnosti, ova se metoda pokazala kao granična metoda za prepoznavanje autorstva, no bigrami, trigrami i 4-grami su se pokazali kao vrlo uspješni pokazatelji autorstva, budući da su u svakom od uzoraka uspješno pripisani Matošu.

Provjera n-grama nastavljena je i na tekstu *Camao*, to jest, provjeravanom tekstu ili tekstu *Q*. Uz 40% uspješnosti, bigrami i trigrami iz teksta *Camao* uspješno su pripisani Matošu, no 4-grami i 5-grami nisu, budući da niti jedan autor nije imao iste 4-grame i 5-grame kao u *Camau*. Poznati Matoševi tekstovi sadržavali su 2,75% istih bigrama kao u *Camau* te su dijelili sveukupno 290 bigrama, kao i 14 zajedničkih trigrama, odnosno 0,17%, u usporedbi s Matoševim tekstovima. Svi rezultati provjere nalaze se u Prilogu.

Naravno, važno je napomenuti kako nisu svi n-grami otkriveni u tekstu potencijalni pokazatelji autorstva, pogotovo u slučajevima n-grama kao što su „kao da je“ ili „da se“, koji se pojavljuju u gotovo svim poznatim, ali i u provjeravanom tekstu. Međutim, otkriveno je nekoliko n-grama koji se mogu posebno istaknuti, a navedeni su u Tablici 7.

Tablica 7.

Primjeri bigrama u tekstu Camao i poznatim tekstovima.

Bigram	**Primjer iz teksta *Q* (*Camao*)**	**Primjer iz poznatih tekstova**	**Poznati tekstovi u kojima se pojavljuju**
ne bijaše **Frekvencija (*Camao*): 1** **Frekvencija (poznati tekstovi): 10**	„To zapravo i ne bijaše salon, nego kraljevski atelijer, pun divnih perzijskih sagova, pušaka beduinaka i sabalja šamijanaka, etrurskih i helenskih posuda, emalja i bižua.“ (Matoš, 1900, str. 51)	*Lijepa Jelena*: „Ne bijaše kao ostale dame pretovarena nakitima.“ (Matoš, 1909, str. 24)	*Balkon* *Cvijet sa raskršća* *Lijepa Jelena* *Ugasnulo svjetlo* *Nerotkinja* (1)
veli ona **Frekvencija (*Camao*): 1** **Frekvencija (poznati tekstovi): 3**	„Gospodine, vi ste i više nego smiješni – veli ona spokojno.“ (Matoš, 1900, str. 53)	*Cvijet sa raskršća*: „Jeste li se probudili? - veli ona naprečac glasom u kojem bijaše sjena tuge koja ne boli. - Pst, Fido!“ (Matoš, 1909, str. 98)	*Balkon* *Cvijet sa raskršća*
mi štogod **Frekvencija (*Camao*): 1** **Frekvencija (poznati tekstovi): 2**	„Pričaj, pričaj mi štogod – izgubljena i opet nađena ljubavi!“ (Matoš, 1900, str. 48)	*Balkon*: „Odsvirajte mi štogod da se sjetim.“ (Matoš, 1909, str. 41)	*Balkon* *Put u ništa*

Tablica 7. - nastavak

Primjeri bigrama u tekstu Camao i poznatim tekstovima.

u Parizu **Frekvencija (*Camao*): 4** **Frekvencija (poznati tekstovi): 4**	**„U Parizu naiđem na Morgue, žalosnu kuću gdje izlažu unesrećene neznanike ne bi ih kogod upoznao.“ (Matoš, 1900, str. 50)**	***Put u ništa*: „Nije tragično živjeti u Parizu bez hrane i krova.“ (Matoš, 1909, str. 88)**	***Lijepa Jelena*** ***Put u ništa***
i rekne **Frekvencija (*Camao*): 2** **Frekvencija (poznati tekstovi): 2**	„Poznati ga F. uzme ispod miške, odvede svojoj kući i rekne: – Odsada sam vam ja skrbnikom.“ (Matoš, 1900, str. 43)	*Put u ništa*: „Orlović se tek prezirno nasmjehne, sneveseli i rekne : – Našto riječi... riječi... riječi, moj brajane?“ (Matoš, 1909, str. 90)	*Balkon* *Put u ništa*

Navedeni bigrami prikazuju tendenciju autora da koristi imperfekt, ali i aorist u svom tekstu. Uz to, neki od primjera sadrže i ranije spomenute stilistički označene izraze, kao što su „veli“, „rekne“ i „štogod“, a s druge strane, česta pojava bigrama „u Parizu“, koji se ne pojavljuje u drugim poznatim tekstovima i koja ukazuje na to da autor radnju svojih djela smješta u Francusku, isto tako može poslužiti kao dobar pokazatelj autorstva, pogotovo u kombinaciji s činjenicom da se u mnogim poznatim djelima, ali i u *Camau*, pojavljuju izrazi na stranim jezicima, pogotovo na francuskom.

Tablica 8.

Primjer trigrama u tekstu Camao i poznatim tekstovima.

Trigram	**Primjer iz teksta *Q* (*Camao*)**	**Primjer iz poznatih tekstova**	**Poznati tekstovi u kojima se pojavljuje**
malo te ne **Frekvencija (*Camao*): 2** **Frekvencija (poznati tekstovi): 3**	„Prozaično ga živovanje modernog čovjeka, koji ga malo te ne pregazi na pločniku mrtva od gladi, najprije zanimaše, docnije bi mu na užas i najzad mu se gadijaše.“ (Matoš, 1900, str. 44)	*Cvijet sa raskršća*: „Bože, alaj se uplaših, jer malo te ne padoh preko vas!“ (Matoš, 1909, str. 99) *Ugasnulo svjetlo*: „Spasu me dvije gospođe, i moram im pričati o njemačkoj književnosti, o mom universitetu, o njemačkoj vjernosti i o njemačkoj ljubavi. – Vi odista ne nalikujete njemačkom učenjaku – rekne mi mlađa, ljepša, domahujući mi lepezom opojni miris sa razgolićenog vrata i sa bijele dojke. – Da ste jurist, hajde de, ali arheolog, ô mon Dieu! – Vi nas, njemačke profesore,	*Cvijet sa raskršća* *Lijepa Jelena* *Ugasnulo svjetlo*

	jamačno sudite kao "Fliegende Blätter" – reknem i malo te ne prasnem u smijeh.“ (Matoš, 190, str. 9)

U slučaju trigrama, posebno treba istaknuti trigram „malo te ne“, koji autor sustavno koristi umjesto izraza „maltene“ kao jedan od potencijalnih indikatora autorstva teksta.

7. Diskusija

Rezultati istraživanja pokazali su uspješnost različitih vrsta pristupa analizi teksta na temelju tekstova Antuna Gustava Matoša i četiri ostala autora iz razdoblja moderne. Početno je za svakog autora provjeravano razlikuju li se njegovih pet tekstova međusobno po dužini riječi i dužini rečenica, a bilo je očekivano da se neće razlikovati. Utvrđeno je kako postoje statistički značajne razlike između dužine riječi i dužine rečenice za svakoga od autora i barem jedan od tekstova u skupini statistički se značajno razlikuje od ostalih tekstova istog autora. Premda se u ovom slučaju radi o razlici od pola ili jednog znaka, to jest, pola ili jedne riječi, razlike su svejedno bile statistički značajne. Značajni F-omjeri u ovoj analizi varijance pokazuju da postoje unutarnje varijacije kod autora te dolazi do preklapanja distribucija variranja različitih autora. Ako i postoje razlike između različitih autora, zbog velikog variranja unutar djela pojedinih autora, one mogu ostati prikrivene. Djelo jednog autora može se svojim obilježjima naći u okviru raspona variranja obilježja djela drugog autora. Iako je nulta hipoteza, da se dužina riječi i rečenica istog autora ne razlikuje u različitim tekstovima, a koja je preduvjet za razmatranje razlike između autora, odbačena, nastavljeno je i s analizama u kojima se uspoređuju djela različitih autora kako bi se odgovorilo na postavljene hipoteze istraživanja.

Kako bi se usporedili rezultati statističke analize uz pomoć alata *WordSmith Tools* korišten je SPSS i analiza varijance za nezavisne podatke. Podaci koji su analizirani bili su prosječne dužine riječi i rečenica, kao i omjer različnica i pojavnica za svaki od poznatih tekstova grupiranih prema autorima. Utvrđeno je kako Antun Gustav Matoš koristi statistički značajno duže riječi od ostalih autora. Isto tako, utvrđeno je da prema postotku omjera različnica i pojavnica Matoš ima statistički značajno veći omjer različnica i pojavnica u odnosu na ostale autore, to jest, da u svojim tekstovima

koristi raznolikije riječi. Važno je za napomenuti kako se neki od autora razlikuju samo uz razinu rizika od 5% ili 10%, što vodi do pitanja je li uopće riječ o uočljivoj razlici između autora.

Isto tako, analiza nije pokazala da postoje statističke značajne razlike u dužinama rečenica svakoga od autora, iako je razlika u najmanjoj i najvećoj aritmetičkoj sredini za dužinu rečenice određenog autora tri riječi. U slučaju usporedbe dužine rečenica, došlo je do statističke pogreške tipa dva, to jest, statistički neznačajne stvarne razlike dovode do pogrešnog zaključivanja tipa dva te je grupa tekstova različitih autora svrstana u istu skupinu.

Prvi dio analize uspješno je istaknuo problem stilističkog naprema stilometrijskog pristupa utvrđivanju autorstva. Naime, dok se u stilometrijskom pristupu vrlo često koriste dužine riječi ili rečenica kao ukazatelji autorstva, dubinska provjera takvih mjera pokazala je kako u nekim slučajevima postoje prevelike varijacije unutar samoga autora da bi te mjere bile jezično i statistički pouzdane. Isto tako, potrebno je napomenuti kako postoje slučajevi gdje su forenzični lingvisti uspješno pokazali autorstvo temeljem dužine rečenica (Grieve, 2007), no razlike utvrđene u tim slučajevima bile su puno veće nego razlike koje su uočene u slučaju ovih pet autora. Naravno, rezultati ove provjere ne dokazuju da je ova metoda u potpunosti neupotrebljiva u forenzično lingvističkom kontekstu, već da, kako bi se utvrdile razlike između autora koje su ne samo statistički, već i jezično značajne, potrebna je puno veća razlika između tih omjera, jer u suprotnome može doći do pogrešnog zaključivanja.

Hipoteza H1, da će prosječna dužina riječi u tekstovima Antuna Gustava Matoša biti značajno veća u odnosu na ostale autore, potvrđena je u kontekstu aritmetičkih sredina, no provjere variranja unutar samog autora pokazale su kako je riječ o nepouzdanoj metodi. Hipoteza H2 nije potvrđena jer nije utvrđena statistički značajna razlika u dužinama rečenica Antuna Gustava Matoša u odnosu na ostale autore niti u jednoj od dvije provedene analize varijance. Unatoč ranijim uspješnim primjenama u drugim istraživanjima, u slučaju ovih pet autora, ova se mjera pokazala kao nedovoljno pouzdana za donošenje zaključaka o autorstvu.

S druge strane, hipoteza H3, prema kojoj će omjer različnica i pojavnica u tekstovima Antuna Gustava Matoša biti značajno veća u odnosu na ostale autore, potvrđena je, no uz dozu opreza. Budući da u slučaju omjera različnica i pojavnica ne postoji standardna devijacija, jer se mjerenje radi na cijelom tekstu, nije postojala mogućnost dodatnih provjera, premda je zanimljivo za istaknuti kako je upravo mjera omjera različnica i pojavnica najviše utemeljena u ranijim

lingvističkim saznanjima. Naime, omjer različnica i pojavnica koristi se kao jedna od mjera za provjeru u forenzičnim analizama plagijata. Mjere leksičke obogaćenosti teksta mogu se koristiti u usporedbi dva teksta, a što je razlika različnica i pojavnica dva teksta veća, to su veće šanse da je tekst s višim omjerom plagijat (Olsson, 2010).

Drugi dio analize zasnovan je na Jaccardovom koeficijentu sličnosti. Za analizu je korišten *Sketch Engine* i njegova funkcija *Identify keywords* kako bi se izvukli podaci o podudaranju u lemama i n-gramima poznatih i provjeravanih tekstova.

U analizi su uspoređene leme korištene u 25 tekstova pet poznatih autora s lemama nepoznatog teksta, za što je u ovom slučaju poslužila Matošev *Camao*. U prvom dijelu analize provedena je usporedba na uzorku od 5%, 10%, 15%, 20% i 25% Matoševih poznatih tekstova, koji su uspoređeni s uzorcima od 95%, 90%, 85%, 80% i 75% poznatih tekstova svih pet autora te je 5% Matoševih poznatih tekstova pripisano 95% Ivakićevih tekstova. Potencijalni razlog za inicijalno nepodudaranje leži u velikoj razlici u omjeru pojavnica i različnica Ivakića i Matoša. Naime, Ivakić je imao najmanji broj lema u 95% tekstova, dok je Matoš imao najveći omjer različnica i pojavnica. Upravo je taj omjer ukazatelj Matoševa bogatog vokabulara, stoga ovakva vrsta mjerenja, zasnovana na usporedbama skupova, bez dubljeg poniranja u same leme, nije dovoljno pouzdana, pogotovo ako se radi na uzorku manjem od 1000 pojavnica. Međutim, uz 80% uspješnosti i nakon dodatne provjere pogrešnog pripisivanja 5% uzoraka Matoševih tekstova 95% teksta Joze Ivakića, utvrđeno je da je metoda vjerodostojna, stoga je provjera nastavljena i na tekstu *Camao*. Nakon što su leme u *Camau* točno pripisane ostatku Matoševih tekstova, utvrđeno je kako hipoteza H4, prema kojoj su leme Antuna Gustava Matoša najsličnije onima u nepoznatim tekstovima, potvrđena. Daljnja analiza zajedničkih lema u korpusu hrWaC pokazala je i odlike Matoševog jezika, pogotovo dijalektalizama, ali i korištenje izraza iz stranih jezika, što može poslužiti u daljnjoj analizi stilističkih značajki teksta.

Posljednja hipoteza H5, da će n-grami u provjeravanim tekstovima biti uspješno pripisani Matošu korištenjem Jaccardovog koeficijenta, također je potvrđena na bigramima i trigramima teksta *Camao*, dok zajedničkih 4-grama i 5-grama nije bilo. Primjena Jaccardovog koeficijenta na n-grame u poznatim i provjeravanim tekstovima pokazala se kao osobito uspješna za bigrame, ali i za trigrame, dok s druge strane, 4-grami i 5-grami nisu bili toliko uspješni. Iako su ranija istraživanja (Wright, 2017) pokazala da su trigrami i 4-grami dobri pokazatelji autorstva, u

slučajevima manjih tekstova, kao što su ovi, ova se metoda pokazala manje uspješnom. Najveća je razlika između ranijih istraživanja primjene Jaccardovog koeficijenta to što su provjeravani tekstovi bili e-mailovi ili SMS poruke, koji su puno neformalniji i podložniji korištenju formulaičnih izraza (Wright, 2017). Wrightovo istraživanje iz 2017. godine zasnovano je na e-mailovima zaposlenika nekadašnje tvrtke Enron te je korpus korišten u tom istraživanju imao puno više pojavnica, ali i više formulaičnih izraza koje se često ponavljaju.

Istraživanje je pokazalo kako je u slučaju Antuna Gustava Matoša, autora koji u svojim djelima koristi vrlo širok vokabular, najuspješnija metoda za stilometrijsku provjeru autorstva Jaccardov koeficijent sličnosti bigrama, zatim trigrama i lema. Omjer različnica i pojavnica isto je tako uspješno utvrdio da se Matoš statistički značajno razlikuje od ostalih autora, međutim, činjenica da za ostale autore nije utvrđena statistički značajna razlika baca sumnju na ovu metodu, jer može voditi do statističke pogreške tipa dva i uzrokovati pogrešno zaključivanje. Dužina riječi i rečenica iz istog se razloga nije pokazala kao korisna metoda za utvrđivanje autorstva Antuna Gustava Matoša, međutim, ukazala je i na veliki problem varijacije unutar autora, koja može dovesti do pogrešnog zaključivanja.

7.1. Ograničenja i buduća istraživanja

Kao jedno od glavnih ograničenja ovog istraživanja pokazalo se to što tekstovi korišteni u istraživanju objavljeni kao književna djela. Samim time postoji mogućnost da su na tekstu radili ne samo autor, već i lektor, urednik, ali i ostali sudionici izdavačkog procesa. Iako je riječ o djelima koja su objavljena prije jednog stoljeća, ne može se izuzeti mogućnost vanjskog utjecaja na tekst autora. Za razliku od nespontanih tekstova kao što su pisma ili poruke, koja su često u fokusu forenzične analize teksta, potrebno je pristupiti utvrđivanju autorstva književnih tekstova uz određenu dozu opreza. Drugi nedostatak jest činjenica da ne postoje specijalizirani alati koji bi se mogli koristiti za utvrđivanje autorstva za hrvatski jezik, stoga preporučujemo nastavak istraživanja mogućnosti automatizacije dijela forenzične analize teksta u svrhu ubrzavanja procesa analize, ali i razvoja daljnjih znanja o izradi jezičnih alata za hrvatski jezik.

Isto tako, u istraživanju nisu obuhvaćene sve jezične značajke autora Antuna Gustava Matoša. Bilo bi zanimljivo dublje istražiti stilističke obrasce koji upućuju na idiolekt autora, ali i korištenje izraza iz stranih jezika, pogotovo utjecaj francuskog jezika, kako ortografski, tako i

leksički. Uz to, mogućnost daljnjeg istraživanja leži i u usporedbi s drugim autorima iz perioda moderne koji nisu bili uključeni u ovo istraživanje. Naposljetku, bilo bi vrlo korisno provesti analizu teksta koristeći i druge poznate metode u forenzičkoj lingvistici, kao što su drugi stilometrijski parametri, ali i parametri korišteni u stilističkoj analizi teksta, kako bi se utvrdile mogućnosti za daljnje razvijanje analize, ali i za usporedbu postojećih stilometrijski i stilističkih metoda.

8. Zaključak

Cilj istraživanja bio je utvrditi jezične značajke prema kojima se Antun Gustav Matoš razlikuje od ostalih autora uključenih u istraživanje, Milana Begovića, Joze Ivakića, Josipa Kozarca i Jure Turića, kao i provjeriti koje metode računalne i forenzičke lingvistike mogu pružiti najindikativnije rezultate. U tu svrhu korišteni su resursi *Sketch Engine* i *WordSmith Tools* kako bi se dobili podaci o tekstu i njegovim značajkama, a za analizu rezultata korištena je analiza varijance i Jaccardov koeficijent sličnosti.

Unatoč tome što je forenzička lingvistika relativno nova znanstvena disciplina, ranija istraživanja, kao i stvarni slučajevi, pokazali su da ima potencijal za široku primjenu, pogotovo u području utvrđivanja autorstva. Iako se danas govori o dva pristupa utvrđivanju autorstva, stilometrijskom i stilističkom pristupu, koji se razlikuju u načinu dolaženja do rezultata, ali imaju isti cilj, danas sve više istraživanja stavlja u fokus upravo ujedinjavanje te dvije metode kako bi se došlo do što boljih rezultata.

U tu svrhu provedena je analiza 26 tekstova pet hrvatskih autora, s fokusom na pronalaženje značajki kojima se Antun Gustav Matoš razlikuje od ostalih autora. Provjera rezultata autora unutar grupe pokazala je kako postoje razlike u dužinama riječi i rečenice unutar grupe djela svakog od autora, što upućuje na to da je kod mjerenja dužina riječi i rečenica potrebna doza opreza, budući da ovakva mjerenja mogu dovesti do pogrešnog tumačenja rezultata. Drugi dio analize bila je jednosmjerna analiza varijance, kojom je utvrđeno da se Matoš razlikuje od ostalih autora prema dužini riječi tako da koristi duže riječi od ostalih autora, ali i prema omjeru različnica i pojavnica. Rezultati analize usmjerene na usporedbu omjera različnica i pojavnica ukazuje na to da Matoš u tekstovima koristi raznovrsnije riječi od ostalih autora. Nije utvrđena razlika između autora prema dužini rečenica.

S druge strane, Jaccardov koeficijent sličnosti zajedničkih lema dobiven usporedbom lema pokazao je kako je 5% poznatih tekstova neuspješno pripisano Jozi Ivakiću, no nakon 10% teksta, to jest, nakon što je tekst prešao 1900 pojavnica, leme su uspješno pripisane Matošu. Otkriveno je da je razlog pogrešno pripisivanja Ivakiću nizak broj lema u tekstu, koji upućuje na manju leksičku raznovrsnost od Matoša. Daljnja usporedba lema u tekstu *Camao* s 25 poznatih tekstova dovela je do uspješnog pripisivanja zajedničkih lema Antunu Gustavu Matošu. Naposljetku, provjera zajedničkih bigrama i trigrama rezultirala je 80% uspješnosti na provjeri poznatih uzoraka teksta, a kao posebno uspješni istaknuli su se bigrami i trigrami. U nastavku analize, provjerom n-grama u tekstu *Camao*, bigrami i trigrami uspješno su pripisani Matošu, međutim, provjeravani uzorak i poznati uzorak nije sadržavao zajedničke 4-grame i 5-grame.

9. Reference

Begović, M. (n.d.). *Izabrane pripovijetke*. eLektire.skole.hr

Bohner, G. (2001). Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology*, *40*(4), 515-529.

Bosanac, S. i Štefanec, V. (2011). *N-gram Overlap in Automatic Detection of Document Derivation*. 3rd International Conference. The Future of Information Sciences: INFuture2011 – Information Sciences and e-Society.

Chaski, C. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics (The International Journal of Speech Language and the Law) 8*(1), 1–65.

Chaski, C. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence 4*(1), 1– 14.

Cotterill, J. (2010). How to use corpus linguistics in forensic linguistics. U A. O'Keefe i M. McCarthy (ur.), *The Routledge Handbook of Forensic Linguistics*, 578-590.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics 24*(4), 431–447.

Coulthard, M. (2020). „Experts and opinions: In my opinion". U M. Coulthard, A. May i R. Sousa-Silva (ur.), *The Roultedge Handbook of Forensic Linguistics*, (str. 523-538). Routledge.

Coulthard, M. i Johnson, A. (2007). *An Introduction to Forensic Linguistics*. Routledge.

Coulthard, M., Johnson, A. i Wright, D. (2017). *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge.

Daller, H., Van Hout, R., i Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197–222.

Fadden, L. i Disner, S.F. (2014). Forensic Linguistics. U G. Bruinsma i D. Weisburd (ur.), *Encyclopedia of Criminology and Criminal Justice*, (str. 1729-1741). Springer.

Grant, T. (2008). Approaching questions in forensic authorship analysis. U J. Gibbons i M. Teresa Turell (ur.), *Dimensions of Forensic Linguistics*, 215–229.

Grant, T. (2020). Text messaging forensics. U M. Coulthard, A. May i R. Sousa-Silva (ur.), *The Roultedge Handbook of Forensic Linguistics*, (str. 508-522). Routledge.

Grant, Tim. 2013. Txt 4N6: Method, consistency and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy 21*(2), 467–494.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, *22*(3), 251-270.

Heydon, G. (2005). *The language of police interviewing*. Palgrave Macmillan.

Ivakić, J. (n.d.). *Pripovijetke-Uspomene*. eLektire.skole.hr

Johnstone, B. (1996). *The Linguistic Individual: Self Expression in Language and Linguistics*. Oxford University Press.

Johnstone, B. (2000). The individual voice in language. *Annual review of anthropology*, *29*(1), 405-424.

Joula, P. (2008). *Authorship Attribution*. Now Publishers Inc.

Kaštelan, J. (1956). *Lirika A.G. Matoša: doktorska disertacija*. [Doktorska disertacija].

Kozarac, I. (n.d.). *Odabrane pripovijetke.* eLektire.skole.hr

Kruh, L. (1982). A basic probe of the Beale cipher as a bamboozlement. *Cryptologia*, *6*(4), 378-382.

Langacker, R. (1988). A usage-based model. U B. Rudzka-Ostyn (ur.), *Topics in Cognitive Linguistics*, 127-161.

Matoš, A.G. (1900). *Novo iverje.* eLektire.skole.hr

Matoš, A.G. (1909). *Umorne priče.* eLektire.skole.hr

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

Miškulin Saletović, L. i Kišiček, G. (2012). Contribution to the Analysis of Witness Statements in the Croatian Language. *Suvremena ligvistika*, *38*(73), 73-88.

Nini, A. i Grant, T. (2013). Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis. *The International Journal of Speech, Language and the Law 20*(2), 173–202.

Olsson, J. (2010). *Forenzička lingvistika*. Nakladni zavod Globus.

Perkins, R., & Grant, T. (2018). Native language influence detection for forensic authorship analysis: Identifying L1 Persian bloggers. *International Journal of Speech, Language & the Law*, *25*(1).

Prideaux, G. D. (2011). Linguistic Contributions to the Analysis of Hate. *International Journal of Law, Language & Discourse*, 27.

Saferstein, V. (2017). *Criminalistics: An Introduction to Forensic Science*. Pearson.

Schilling, N., & Marsters, A. (2015). Unmasking Identity: Speaker Profiling for Forensic Linguistic Purposes. *Annual Review of Applied Linguistics*, *35*, 195-214.

Shapero, J. J., & Blackwell, S. A. (2012). "There are letters for you all on the sideboard': What can linguists learn from multiple suicide-note writers". U S. Tomblin, N. MacLeod, R. Sousa-Silva i M. Coulthard (ur.) *Proceedings of The international association of forensic linguists' tenth biennial conference*, (str. 225-244). Aston University

Shi, Y., & Lei, L. (2022). Lexical Richness and Text Length: An Entropy-based Perspective. *Journal of Quantitative Linguistics*, *29*(1), 62-79.

Shuy, R. W. (2005). *Creating language crimes: How law enforcement uses (and misuses) language*. Oxford University Press on Demand.

Shuy, R. W. (2007). Language in the American courtroom. *Language and Linguistics Compass*, *1*(1-2), 100-114.

Stygall, G. (2010). Legal writing: Complexity. U M. Coulthard, A. May i R. Sousa-Silva (ur.), *The Roultedge Handbook of Forensic Linguistics*, (str. 32-47). Routledge.

Svartvik, J. (1968). *The Evans Statements: A case for Forensic Linguistics*. University of Gothenburg Press.

Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik*. Ex Libris.

The Innocence Project. (n.d.). *Overturning Wrongful Convictions Involving Misapplied Forensics*. Posjećeno 15. svibnja 2022. https://innocenceproject.org/overturning-wrongful-convictions-involving-flawed-forensics/

The National Registry of Exonerations. (n.d.). *Browse the National Registry of Exonerations*. Posjećeno 15. svibnja 2022. https://www.law.umich.edu/special/exoneration/Pages/browse.aspx?View={B8342AE7-6520-4A32-8A06-4B326208BAF8}&FilterField1=Contributing_x0020_Factors_x0020&FilterValue1=False%20or%20Misleading%20Forensic%20Evidence

Turić, J. (n.d.). *Pripovijesti*. eLektire.skole.hr

Wright, D. (2014). Stylistics versus Statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails. [Doktorska disertacija, University of Leeds]. White Rose eTheses Online. https://etheses.whiterose.ac.uk/8278/

Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, *22*(2), 212-241.

10. Prilozi

10.1. Podaci o dužinama riječi

Tablica 9.

Dužine riječi u djelima Antuna Gustava Matoša.

Naslov	Broj riječi	Dužina riječi	Standardna devijacija
Balkon	3839	4,47	2,58
Cvijet sa raskršća	2538	4,23	2,48
Lijepa Jelena	3428	4,58	2,74
Put u ništa	5132	4,49	2,57
Ugasnulo svjetlo	2128	4,45	2,69

Tablica 10.

Dužine riječi u djelima Milana Begovića.

Naslov	Broj riječi	Dužina riječi	Standardna devijacija
Dva bijela hljeba	15344	3,99	2,33
Krzno od sibirske vjeverice	3608	4,27	2,53
Kvartet	7940	4,33	2,60
Nerotkinja	2555	3,74	2,23
Posljednji posjet	3009	3,94	2,33

Tablica 11.

Dužine riječi u djelima Ivana Kozarca.

Naslov	Broj riječi	Dužina riječi	Standardna devijacija
Didak i baka	1942	3,57	2,09
Gospodična Jelica	5068	4,05	2,43
Mlada žena	8206	3,83	2,28
Otrov	1260	3,50	2,11
Sestra	1933	3,52	2,12

Tablica 12.

Dužine riječi u djelima Joze Ivakića.

Naslov	Broj riječi	Dužina riječi	Standardna devijacija
Gnjili	1660	3,93	2,29
Moji ljudi	6109	4,16	2,48
Stara rana	2279	3,89	2,30
Sudoperka	4433	4,01	2,42
U nagonu	3548	4,08	2,48

Tablica 13.

Dužine riječi u djelima Jure Turića.

Naslov	Broj riječi	Dužina riječi	Standardna devijacija
Njihova ljubav	4875	3,73	2,16
Prosci	18387	3,95	2,37
Srce	2697	3,75	2,18
Tko je kriv	5159	3,79	2,26
U mraku	5043	3,94	2,31

Tablica 14.

Dužine riječi u Camau.

Naslov	Broj riječi	Dužina riječi	Standardna devijacija
Camao	6889	4,79	2,64

10.2. Podaci o dužinama rečenica

Tablica 15.

Dužine rečenica u djelima Antuna Gustava Matoša.

Naslov	Broj rečenica	Dužina rečenica	Standardna devijacija
Balkon	287	13,38	11,68
Cvijet sa raskršća	211	12,03	9,19
Lijepa Jelena	194	17,67	12,81
Put u ništa	391	13,13	9,43
Ugasnulo svjetlo	141	15,09	10,81

Tablica 16.

Dužine rečenica u djelima Milana Begovića.

Naslov	Broj rečenica	Dužina rečenica	Standardna devijacija
Dva bijela hljeba	1075	14,26	10,47
Krzno od sibirske vjeverice	227	15,89	11,28
Kvartet	555	14,30	10,06
Nerotkinja	225	11,36	7,22
Posljednji posjet	281	10,70	7,76

Tablica 17.

Dužine rečenica u djelima Ivana Kozarca.

Naslov	Broj rečenica	Dužina rečenica	Standardna devijacija
Didak i baka	112	14,82	12,96
Gospodična Jelica	476	12,83	10,57
Mlada žena	176	12,94	11,80
Otrov	378	11,73	9,07
Sestra	200	17,74	11,13

Tablica 18.

Dužine rečenica i u djelima Joze Ivakića.

Naslov	Broj rečenica	Dužina rečenica	Standardna devijacija
Gnjili	138	14,07	9,87
Moji ljudi	398	12,73	11,72
Stara rana	502	16,35	12,91
Sudoperka	134	9,40	6,03
U nagonu	145	13,33	7,11

Tablica 19.

Dužine rečenica u djelima Jure Turića.

Naslov	Broj rečenica	Dužina rečenica	Standardna devijacija
Njihova ljubav	278	17,54	10,59
Prosci	1,204	15,27	10,08
Srce	160	16,86	11,63
Tko je kriv	282	18,29	12,19
U mraku	369	13,66	10,61

Tablica 20.

Dužine rečenica u Camau.

Naslov	Broj rečenica	Dužina rečenica	Standardna devijacija
Camao	490	14,06	11,03

10.3. Podaci o omjerima različnica i pojavnica

Tablica 21.

Omjer različnica i pojavnica u djelima Antuna Gustava Matoša.

Naslov	Omjer različnica i pojavnica (%)
Balkon	49,52
Cvijet sa raskršća	50,39
Lijepa Jelena	50,96
Put u ništa	48,58
Ugasnulo svjetlo	55,03

Tablica 22.

Omjer različnica i pojavnica u djelima Milana Begovića.

Naslov	Omjer različnica i pojavnica (%)
Dva bijela hljeba	26,85
Krzno od sibirske vjeverice	41,02
Kvartet	36,31
Nerotkinja	40,70
Posljednji posjet	38,70

Tablica 23.

Omjer različnica i pojavnica u djelima Ivana Kozarca.

Naslov	Omjer različnica i pojavnica (%)
Didak i baka	48,13
Gospodična Jelica	37,37
Mlada žena	44,56
Otrov	37,38
Sestra	38,42

Tablica 24.

Omjer različnica i pojavnica u djelima Joze Ivakića.

Naslov	Omjer različnica i pojavnica (%)
Gnjili	37,59
Moji ljudi	33,17
Stara rana	29,02
Sudoperka	40,48
U nagonu	35,08

Tablica 25.

Omjer različnica i pojavnica u djelima Jure Turića.

Naslov	Omjer različnica i pojavnica (%)
Njihova ljubav	29,25
Prosci	24,42
Srce	36,67
Tko je kriv	30,08
U mraku	35,40

Tablica 26.

Omjer različnica i pojavnica u Camau.

Naslov	Omjer različnica i pojavnica (%)
Camao	45,68

10.4. Rezultati usporedbe lema

Tablica 27.

Rezultati usporedbe lema poznatih tekstova.

Veličina uzorka 1	Veličina uzorka 2	Antun Gustav Matoš (%)	Milan Begović (%)	Joza Ivakić (%)	Josip Kozarac (%)	Jure Turić (%)
5%	**95%**	7.22%	5.07%	7.39%	5.77%	5.59%
10%	**90%**	11.40%	8.22%	10.64%	8.93%	8.59%
15%	**85%**	14.07%	9.97%	12.49%	10.85%	10.64%
20%	**80%**	15.43%	11.49%	13.40%	11.90%	11.99%
25%	**75%**	16.76%	12.65%	13.97%	12.92%	12.87%

10.5. Rezultati usporedbe n-grama

Tablica 28.

Rezultati usporedbe bigrama poznatih tekstova.

Veličina uzorka 1	Veličina uzorka 2	Antun Gustav Matoš (%)	Milan Begović (%)	Joza Ivakić (%)	Josip Kozarac (%)	Jure Turić (%)
5%	**95%**	0.73%	0.35%	0.53%	0.47%	0.36%
10%	**90%**	1.24%	0.64%	1.00%	0.87%	0.66%
15%	**85%**	1.69%	0.91%	1.42%	1.19%	0.93%
20%	**80%**	2.11%	1.17%	1.77%	1.47%	1.17%
25%	**75%**	2.09%	1.19%	1.76%	1.50%	1.20%

Tablica 29.

Rezultati usporedbe trigrama poznatih tekstova.

Veličina uzorka 1	Veličina uzorka 2	Antun Gustav Matoš (%)	Milan Begović (%)	Joza Ivakić (%)	Josip Kozarac (%)	Jure Turić (%)
5%	**95%**	0.08%	0.01%	0.05%	0.02%	0.03%
10%	**90%**	0.11%	0.05%	0.10%	0.04%	0.05%
15%	**85%**	0.17%	0.07%	0.12%	0.05%	0.06%
20%	**80%**	0.24%	0.10%	0.17%	0.09%	0.08%
25%	**75%**	0.18%	0.09%	0.17%	0.08%	0.08%

Tablica 30.

Rezultati usporedbe 4-grama poznatih tekstova.

Veličina uzorka 1	Veličina uzorka 2	Antun Gustav Matoš (%)	Milan Begović (%)	Joza Ivakić (%)	Josip Kozarac (%)	Jure Turić (%)
5%	**95%**	0.02%	0.00%	0.00%	0.00%	0.00%
10%	**90%**	0.01%	0.001%	0.00%	0.00%	0.00%
15%	**85%**	0.01%	0.001%	0.00%	0.00%	0.001%
20%	**80%**	0.06%	0.001%	0.00%	0.00%	0.001%
25%	**75%**	0.03%	0.001%	0.00%	0.00%	0.001%

Tablica 31.

Rezultati usporedbe n-grama u Camau i poznatim tekstovima.

Veličina n-grama	Antun Gustav Matoš (%)	Milan Begović (%)	Joza Ivakić (%)	Josip Kozarac (%)	Jure Turić (%)
2	2.75%	1.68%	2.36%	2.11%	1.45%
3	0.17%	0.11%	0.16%	0.08%	0.06%
4	0.00%	0.00%	0.00%	0.00%	0.00%