Značaj digitalnog korpusa i jezičkih alata u jezičkoj analizi forenzičkih tekstova na srpskom jeziku

Naučni rad DOI: 10.18485/judig.2025.1.ch4

Jelena Redli[1] *0000-0003-2518-5286*

Apstrakt

Forenzička lingvistika (FL) je interdisciplinarna oblast koja kombinuje jezičke, pravne i digitalne tehnologije radi boljeg razumevanja jezičkih fenomena u pravnom i kriminalističkom kontekstu. U vidu pokušaja utvrđivanja autorstva FL se primenjuje od 18. veka još od prvih rasprava o autorstvu Biblije da bi se kasnije proširila na otkrivanje potencijalnih kriminalnih aktivnosti i sl. Za sve FL aktivnosti ključni su specijalizovani korpusi koji služe za obuku sistema tokom razvoja. Iako je upotreba ovih korpusa nezaobilazna u globalnom kontekstu, srpski jezik ostaje nedovoljno istražen zbog nedostatka specijalizovanih korpusa forenzičkih tekstova.

Ovaj rad istražuje potencijal kreiranja i primene korpusa forenzičkih tekstova na srpskom jeziku, fokusirajući se na njihovu ulogu u analizi autentičnosti teksta, identifikaciji autora i rešavanju pravnih sporova putem jezičkih dokaza. Metodologija koju rad predlaže obuhvata različite vrste forenzičkih tekstova kao što su policijski izveštaji, pravni dokumenti, preteće poruke i oproštajna pisma. Takođe se razmatraju napredni jezički alati za automatsku analizu teksta, kao i dubinsku analizu leksičke frekvencije, sintakse, stilskih markera i jezičkih obrazaca.

Pored teorijskog doprinosa, rad demonstrira praktičnu primenu korpusa u srpskom kriminalističko-pravnom kontekstu, ističući potencijalne izazove u razvoju kao što su tehnički, pravni i etički aspekti.

[1] Filozofski fakultet, Univerzitet u Novom Sadu, redli@ff.uns.ac.rs

Poseban naglasak stavlja se na poštovanje etičkih i pravnih smernica tokom prikupljanja i obrade osetljivih podataka.

Na kraju, rad pruža preporuke za buduća istraživanja i implementaciju, naglašavajući važnost interdisciplinarne saradnje između lingvista, pravnika, policije i IT stručnjaka u razvoju jezičkih resursa za forenzička istraživanja u Srbiji.

Ključne reči: forenzička lingvistika, srpski jezik, pravni jezik, digitalni korpus, jezički alati, forenzička analiza teksta, identifikacija autora.

1. Uvod

Forenzička lingvistika (FL) relativno je mlada disciplina koja se bavi analizom jezika u pravnim i kriminalističkim kontekstima. Njeni zvanični počeci datiraju iz 1968. godine, kada je švedski lingvista Jan Svartvik po prvi put primenio lingvističke metode u analizi izjava datih policiji u slučaju *Evans* (Svartvik, 1968). Forenzička lingvistika obuhvata širok spektar metoda i tehnika za analizu tekstova i govora u pravnim okvirima, s posebnim fokusom na jezičke dokaze koji se koriste u pravnim postupcima. Ona kombinuje lingvistička znanja s pravnim i kriminalističkim metodologijama kako bi se istražili složeni jezički fenomeni prisutni u forenzičkim tekstovima, kao što su izjave osumnjičenih, svedoka, policijski izveštaji, preteće poruke i drugi relevantni dokumenti.

Jedna od glavnih uloga FL jeste identifikacija autora spornih tekstova, analiza autentičnosti izjava, te evaluacija pravnih dokaza putem jezičke analize. Upotrebom metoda iz oblasti fonetike, leksikologije, sintakse, semantike, pragmatike, analize diskursa i lingvistike teksta forenzički lingvisti istražuju jezičke karakteristike tekstova u cilju utvrđivanja autorstva, verodostojnosti i drugih karakteristka datih izjava koje mogu poslužiti kao deo dokaza, a ponekad i kao jedini dokaz (Coulthard, 2005, p. 10; Olsson, 2010, p. 11).

Iako su istraživanja upotrebe digitalnih korpusa i alata u analizi forenzičkih tekstova na srpskom jeziku dugo bila ograničena, poslednjih godina zabeležen je značajan napredak (v. Vorkapić et al., 2017), pri čemu se u obradi jezika primenjuju kako tradicionalni, tako i savremeni alati, uključujući velike jezičke modele. Nedostatak specijalizovanih digitalnih korpusa forenzičkih tekstova na srpskom jeziku značajno ograničava mogućnosti za detaljnu analizu i praktičnu primenu jezičkih alata. U razvijenim pravnim sistemima, digitalni korpusi postaju važni alati za identifikaciju autora, analizu sintakse, leksike i stilskih markera u jeziku osumnjičenog ili nepoznatog autora nekog forenzičkog teksta.

Na srpskom jeziku, dosad nije kreiran specijalizovan korpus forenzičkih tekstova, što predstavlja velik izazov u kriminalističko-pravnom kontekstu. Stoga postoji potreba za razvojem korpusa koji bi uključivao različite vrste tekstova relevantnih za FL.

Zato je glavni cilj ovog rada da istraži značaj digitalnog korpusa i jezičkih alata u forenzičkoj analizi tekstova na srpskom jeziku. Kreiranje specijalizovanog korpusa omogućilo bi lakšu identifikaciju autora, analizu autentičnosti tekstova i unapređenje postupaka u pravnim sporovima. Ovaj rad pruža metodološki okvir za formiranje takvog korpusa i daje smernice za njegovu praktičnu primenu u pravnim i kriminalističkim istragama. Time bi srpski jezik postao deo globalne mreže forenzičkih istraživanja, uz poštovanje etičkih i pravnih smernica za obradu osetljivih podataka.

2. Pregled literature

2.1. Korpusna lingvistika i forenzička lingvistika

Korpusna lingvistika je postala temeljna metoda za analizu jezika u različitim oblastima, uključujući i FL. Razvoj velikih, digitalnih zbirki tekstova omogućio je lingvistima da proučavaju jezičke fenomene na način koji ranije nije bio moguć. S. Blackwell, jedno od prominentnih imena u ovom polju, ističe da, iako je korpusna lingvistika prisutna od šezdesetih godina XX veka, njena primena u FL relativno je nova (Blackwell, 2009, p. 5). Korpusni alati omogućavaju preciznu analizu sintakse, leksike i frazeologije, što pomaže u identifikaciji autora i analizi autentičnosti spornih tekstova poput pretećih pisama ili anonimnih poruka.

Jedan od glavnih problema u ranim radovima na polju FL bio je korišćenje opštih korpusa kao što je Sinklerov *Bank of English*[2] umesto specifičnih forenzičkih korpusa, što je često dovodilo do nejasnih rezultata (Ibid.). Korpusna lingvistika i forenzička lingvistika sada se sve više povezuju, kako Kultard naglašava, posebno u analizi pravnih i kriminalističkih tekstova (Coulthard, 2013, p. 202). Njegova istraživanja pokazala su značaj specijalizovanih korpusa za analizu stilskih markera u pravnim tekstovima, čime se povećava tačnost identifikacije autora i verodostojnosti dokumenata.

[2] Bank of English je danas dostupan preko CQPWeb platforme, na serveru Univerziteta u Birmingemu, što omogućava nastavnom osoblju i studentima ovog univerziteta da pristupe ovom i drugim korpusima kroz jedinstven interfejs. Pristup sistemu zahteva registraciju putem imejl adrese sa domenom bham.ac.uk. (v. University of Birmingham CQPWeb server: http://cqpweb.bham.ac.uk).

2.2. Primena korpusa u forenzičkim analizama

Korpusi su ključni u forenzičkim analizama zbog svoje sposobnosti da pruže uvid u frekvenciju pojavljivanja određenih jezičkih elemenata u tekstu. Na primer, istraživanja Dejvida Vrajta o n-gramima pokazala su kako specifični nizovi reči mogu biti jedinstveni za određenog autora, čime se značajno olakšava proces atribucije autorstva (Wright, 2017). Ova metoda je naročito korisna u slučajevima gde postoji sumnja na plagijat ili kada se analizira autentičnost tekstova kao što su anonimne poruke. Korpusni pristup omogućava lingvistima da sistematski analiziraju obrasce koji se često gube u tradicionalnim metodama analize.

Pored identifikacije autora, korpusi se koriste i za analizu leksičke frekvencije i stilskih varijacija unutar i između pravnih tekstova. Stanislav Gozdž-Roškovski je detaljno proučavao frazeologiju i varijacije u pravnim diskursima, ukazujući na to kako se određene fraze ponavljaju u specifičnim pravnim kontekstima, što je od suštinskog značaja za identifikaciju autora (Goźdź-Roszkowski, 2021, pp. 1521, 1526, 1529–1530). Korpusna analiza, takođe, omogućava otkrivanje evaluativnog jezika koji može biti ključan u pravnim postupcima.

2.3. Nedostatak korpusa na srpskom jeziku

Na srpskom jezičkom terenu, upotreba korpusnih alata u forenzičkoj lingvistici još uvek je u začetku. Iako su postignuti značajni pomaci u izgradnji srpskog korpusa, poput SrpKora (Utvić, 2014, str. 247 i dalje), specijalizovani forenzički korpusi još uvek ne postoje. Nedostatak ovakvih korpusa na srpskom jeziku prestavlja ozbiljan izazov i ograničava potencijal za istraživanja u oblasti FL, posebno u analizama autentičnosti tekstova i atribuciji autorstva.

Duško Vitas i saradnici prepoznaju značaj razvoja resursa za obradu srpskog jezika (Vitas et al., 2003), ali trenutni resursi nisu prilagođeni potrebama forenzičkih analiza. Na primer, složena morfološka struktura srpskog jezika, slobodan redosled reči u rečenici i prisustvo dva pisma (ćiriličnog i latiničnog) predstavljaju specifične izazove za obradu jezika (Ibid, p. 2). Ovi tehnički izazovi otežavaju razvoj alata koji bi mogli omogućiti efikasnu analizu tekstova u forenzičkoj praksi.

Nikola Dobrić (2012) ukazuje na to da su zapadnobalkanski jezici napravili značajne korake u razvoju korpusa, ali ipak postoji potreba za specifičnijim alatima koji bi omogućili detaljniju analizu pravnih i kriminalističkih tekstova. Odsustvo specijalizovanih forenzičkih korpusa na srpskom jeziku predstavlja prepreku u napretku FL, posebno u oblastima

kao što su identifikacija autora i provera autentičnosti dokumenata u pravnim sporovima.

Razvoj forenzičkog korpusa za srpski jezik bio bi od ključnog značaja za unapređenje pravnih istraživanja i analize tekstova. Takav korpus omogućio bi ne samo precizniju analizu spornih tekstova već i unapredio razumevanje jezika u pravnom kontekstu, što bi znatno olakšalo rad lingvista i pravnih stručnjaka.

3. Predlog metodologije

3.1. Kreiranje korpusa forenzičkih tekstova na srpskom jeziku

Kreiranje specifičnog korpusa forenzičkih tekstova na srpskom jeziku zahteva pažljivo planiranje i odabir relevantnih tekstova koji će služiti kao reprezentativan uzorak za analizu. Ovaj korpus obuhvata različite vrste tekstova koji su od ključne važnosti u pravnim i kriminalističkim kontekstima. Posebnu pažnju treba posvetiti uključivanju tekstova kao što su policijski izveštaji, pravni dokumenti, izjave, preteće poruke, oproštajna pisma, zahtevi za otkupninu i apeli. Među njima se izdvajaju izveštaji nadležnih organa kao jedan od najvažnijih tipova, jer sadrže detalje o događajima koji se istražuju u sudskim postupcima, uključujući izjave svedoka i osumnjičenih. Oni pružaju uvid u poseban stil poznat kao policijski registar koji se koristi u policijskim dokumentima, čineći ih neophodnim delom korpusa.

Pravni dokumenti, uključujući tužbe, žalbe, presude i druge sudske spise, predstavljaju primere pravnog jezika i diskursa ključnog za analizu jezika u pravnim procesima. Takođe, korpus bi obuhvatio preteće poruke i oproštajna pisma – tekstove koji se često koriste u forenzičkim istragama za utvrđivanje autentičnosti i identifikaciju autora. Ovi tekstovi omogućavaju uvid u neformalni jezik i stil koji se koristi u kriminalnim aktivnostima, što ih čini važnim za forenzičku analizu.

Korpus treba da bude dinamičan, s mogućnošću redovnog ažuriranja i dodavanja novih tekstova, čime će se obezbediti njegova aktuelnost i relevantnost u vremenu.

3.2. Kriterijumi selekcije tekstova

Svi tekstovi koji bi bili uključeni u korpus pažljivo će se birati prema unapred definisanim kriterijumima kako bi se obezbedila tačnost i relevantnost podataka. Autentičnost tekstova je od suštinskog značaja – u korpus treba uključiti samo one tekstove koji su prošli kroz zvanične pravne ili policijske postupke i koji su sastavni deo zaključenih sudskih

predmeta. Reprezentativnost tekstova je takođe ključna, pri čemu treba osigurati da uzorci pokrivaju različite vrste pravnih i kriminalnih slučajeva, kao i raznolike stilove pisanja.

Osim toga, posebna pažnja posvetiće se obezbeđivanju varijabilnosti u tekstovima, uključujući tekstove različitih registara, od formalnog pravnog jezika do neformalnog jezika anonimnih poruka. Konačno, svi podaci će biti anonimizovani kako bi se zaštitila privatnost aktera, a proces izgradnje korpusa biće u potpunosti usklađen s etičkim smernicama.

3.3. Tehnički aspekti i jezički alati

Korpus će biti kreiran korišćenjem naprednih softverskih alata za analizu jezika, koji omogućavaju detaljnu analizu različitih jezičkih elemenata. Jedan od ključnih aspekata biće analiza leksičke frekvencije, koja omogućava prepoznavanje obrazaca u učestalosti reči, identifikaciju ključnih termina i uvid u leksičku strukturu tekstova. Alati poput SrpNet, AntConc, Unitex, SketchEngine i NoSketchEngine, WordSmith Tools, koji se često koriste u analizi jezika, pružiće mogućnost za pretragu i obradu tekstova na srpskom jeziku (Utvić, 2014, str. 208–243, Vitas et al., 2012, pp. 23, 34–35). Takođe, alati koje razvija NLP grupa na Matematičkom fakultetu Univerziteta u Beogradu (up. Vitas et al., 2003, para. 4), pružiće dodatne mogućnosti za analizu srpskog jezika u kontekstu forenzičkih tekstova.

Sintaksička analiza biće sprovedena korišćenjem alata kao što je Unitex/GramLab, koji omogućava prepoznavanje sintaksičkih obrazaca i dodavanje oznaka (Vitas et al. 2003; Krstev & Vitas, 2005; Utvić, 2014), ali ne sprovodi potpunu analizu rečeničnih struktura. Ova analiza omogućiće istraživačima da identifikuju određene sintaksičke konstrukcije i stilske karakteristike specifične za forenzičke tekstove. Iako se alat koristi za leksičku i sintaksičku obradu, uključujući segmentaciju, normalizaciju i disambiguaciju teksta, nije direktno namenjen za stilometrijske tehnike poput n-gram modela. Sistem se primarno oslanja na konačne automate i transduktore, ali n-gram modeli i drugi statistički stilometrijski alati nisu deo njegovih osnovnih mogućnosti. U kontekstu srpskog jezika, prema Vitasu i saradnicima (2012), razvijeni su različiti jezički resursi i alati, ali se dodatno naglašava potreba za naprednim tehnologijama i alatima koji omogućavaju sveobuhvatnu sintaksičku i semantičku analizu za bolje razumevanje složenih jezičkih struktura, što uključuje i forenzičku primenu.

Alati kao što su Unitex/GramLab i TXM nude dodatne opcije za analizu srpskog jezika u kontekstu leksičke i sintaksičke obrade.

Unitex/GramLab je alat za obradu prirodnog jezika koji koristi konačne automate za prepoznavanje reči i sintaksičkih obrazaca. Može se primeniti na srpski jezik uz odgovarajuće prilagođavanje leksikona i gramatičkih resursa, ali njegova ograničenja uključuju nedostatak duboke sintaksičke analize i odsustvo stilometrijskih tehnika kao što su n-gram modeli. S druge strane, TXM je softver za tekstometrijsku analizu koji omogućava frekvencijsku analizu, klasterizaciju i vizuelizaciju leksičkih odnosa u tekstualnim korpusima, te zahteva anotirane korpuse u XML formatu. TXM podržava Unicode i TEI standarde, što ga čini pogodnim za analizu srpskih korpusa, uz mogućnost primene različitih statističkih metoda za analizu tekstualnih struktura. Iako oba alata nude korisne funkcije za osnovnu analizu, za dublje stilometrijske analize i primenu statističkih modela preporučuje se korišćenje dodatnih alata, kao što su Python biblioteke (NLTK, scikit-learn) ili Stylo za naprednu obradu i poređenje tekstova.

Za stilometrijsku analizu, koja omogućava identifikaciju specifičnih jezičkih obrazaca karakterističnih za određenog autora ili tip teksta, obično se koriste drugačiji alati specifično razvijeni za statističku obradu teksta, kao što su R, Python biblioteke (npr. NLTK ili scikit-learn) ili namenski stilometrijski alati poput Stylo. Istraživanje Dejvida Vrajta (Wright, 2017) pokazuje da n-gram modeli, posebno nizovi reči između dve i šest reči (n-gramovi), mogu biti korisni kao stilski markeri za identifikaciju autora u forenzičkim analizama jer pružaju mogućnost poređenja specifičnih jezičkih obrazaca među tekstovima u različitim korpusima.

Srpski jezik se suočava sa specifičnim izazovima kada je reč o dostupnosti jezičkih alata za forenzičku jezičku analizu. Iako postoje nacionalni korpusi kao što je SrpKor, nedostatak specijalizovanih forenzičkih alata kao, na primer, onih koji bi omogućili duboko parsiranje a koji još uvek nisu razvijeni za srpski jezik (Vitas et al. 2012, p. 34), i dalje ograničava mogućnosti za dubinsku analizu tekstova u ovoj oblasti. Razvoj dodatnih softverskih alata prilagođenih potrebama FL na srpskom jeziku, kao i implementacija stilometrijskih tehnika u okviru postojećih sistema kao što je Unitex/GramLab i NooJ, biće ključni za unapređenje ove oblasti i poboljšanje efikasnosti forenzičkih analiza.

4. Analiza i diskusija

4.1. Mogućnosti analize forenzičkih tekstova

Korpus forenzičkih tekstova na srpskom jeziku nudi brojne mogućnosti za analizu autentičnosti, identifikaciju autora i pružanje jezičkih dokaza u pravnim postupcima. Analiza autentičnosti omogućava utvrđivanje verodostojnosti teksta, a u korpusu se mogu identifikovati stilski markeri specifični za određene vrste tekstova ili autorov jezički profil. Korišćenjem leksičke i stilometrijske analize, forenzički lingvisti mogu uporediti tekstove koji pripadaju različitim pravnim žanrovima i utvrditi da li tekst sadrži odstupanja koja bi ukazivala na manipulaciju ili falsifikovanje.

U domenu identifikacije autora, korpus bi omogućio analizu idiolekatskih elemenata kroz upotrebu n-gramova i ponavljajućih fraza, čime se može identifikovati specifičan jezički obrazac karakterističan za pojedinca ili grupu. Na primer, istraživanja su pokazala da nizovi reči (n-gramovi) omogućavaju identifikaciju specifičnih jezičkih obrazaca koji se koriste za razlikovanje autentičnih tekstova od onih s nepoznatim autorom (Wright, 2017). Upotrebom ovih analiza, moguće je povezati sporne tekstove s potencijalnim autorima, čime se forenzički korpus pozicionira kao ključan alat u krivičnim istragama.

4.2. Praktični primeri i mogućnosti identifikacije autorstva

Praktična primena korpusa u identifikaciji autorstva može se prikazati kroz nekoliko konkretnih primera. Kada se ispituje autentičnost ovakvih tekstova, analiza stilskih markera — kao što su specifični nizovi reči, ton i struktura — može otkriti podudarnosti s jezičkim obrascima poznatih autora. Na primer, u slučaju poruke u kojoj je bila zahtevana uplata u zamenu za kompromitujući dokument s dokazima, korišćena je konstrukcija s leksičkim izborom koji bi mogao ukazivati na prisustvo idiolekta: *Stavite novac u kovertu, 1.000 € u dunavskom parku iza izviđačkog kampa, cd vam je zaboden između zida tu osavite kovertu a uzmite cd mozete doci odmah cd je vec tamo*. U ovakvom slučaju važno je utvrditi da li je reč o idiolektu, odnosno kolika je frekvencija ovakvog leksičko-gramatičkog spoja. Drugim rečima, pretragom korpusa s ciljem poređenja jezika analiziranih tekstova sa referentnim primerima, mogu se identifikovati obrasci koji nisu uobičajeni ili odstupaju od norme što bi pomoglo u donošenju preciznih zaključaka o potencijalnom autoru.

U ovakvim primerima, važno je utvrditi i da li pretnja ima realnu osnovu, odnosno srž, koja ukazuje na stvarnu opasnost ili je pretnja

usmerena na zastrašivanje bez konkretne namere za izvršenje. Praktična primena korpusa u identifikaciji autorstva i proceni ozbiljnosti pretnji može biti ključna za utvrđivanje da li poruka sadrži realnu pretnju koja zahteva dalju istragu. Na osnovu klasifikacije pretnji prema stepenu rizika, FBI je razvio smernice koje omogućavaju procenu ozbiljnosti pretnji putem analize specifičnih lingvističkih elemenata. Na primer, pretnje s visokim stepenom rizika obuhvataju direktne, uverljive izjave u vezi sa žrtvom i često uključuju detalje o oružju, mestu i vremenu napada, što signalizira da je preduzeta priprema za ostvarivanje pretnje (Nikolić Novaković, 2017, str. 104–105), dok pretnje kao što je *J***m ti mater, budalo matora, j****u ti majku, ubicu te, sve cu vas pobiti, za sve si ti kriva ljubice j****u ti majku, ubicu te, sredicu Vas Da li bi zelela da zaradis 20 eura, dobro si parce, gde ti je majka, jel otisla da se prodaje u belo roblje i tebe odvukla, sve cu vas pobiti*, koje uključuju uvredljive i vulgarne izraze, ali bez jasnih detalja o planiranom napadu, najčešće ukazuju na manji rizik jer nemaju srž. Takve pretnje se uglavnom karakterišu afektivnim izrazima besa i namerom da izazovu strah bez čvrste osnove za realizaciju. Upotrebom stilometrijskih metoda i procene rizika u analizi pretećih tekstova, korpus bi mogao omogućiti istražiteljima da razlikuju pretnje koje imaju visok stepen verodostojnosti od onih koje služe samo za zastrašivanje, čime se doprinosi preciznosti i efikasnosti u kriminalističkim istragama.

U sličnom kontekstu, analiza oproštajnih pisama može otkriti jedinstvene stilske elemente koji bi mogli ukazivati na lažno autorstvo. Prava oproštajna pisma obično imaju karakteristike kao što su izrazi ljubavi prema porodici, direktivna uputstva najbližima i odsustvo populističkih stavova poput iskazivanja sopstvene slabosti ili kukavičluka. Ove osobine razlikuju autentične poruke od fingiranih oproštajnih pisama, koja mogu sadržavati te elemente da bi izazvala emocionalni odgovor kod čitaoca ili lažno predstavila autorov stav i osećanja. Korisnost postojanja korpusa ogleda se u mogućnosti poređenja jezika analiziranih tekstova s referentnim korpusom kako bi se identifikovali obrasci koji odstupaju od norme, što doprinosi preciznosti analize i donošenju mišljenja o autentičnosti teksta. U jednom primeru iz stvarne prakse, samoubica je ostavio sledeće pismo: *Neno moja*, *ne mogu više da izdržim*. *Plačem za svim što sam izgubio. Naše anđele i tebe*. *Za sve sam ja kriv*. *Čuvaj naše devojčice*. *Sve vas volim puno*. Ovakva frazeologija, koja sadrži izraze sopstvene krivice, tuge i brige za porodicu, oslikava autentičan emotivni ton i privrženost najbližima, što je indikator autentičnosti oproštajnog pisma.

4.3. Uloga sintaksičke analize u identifikaciji autora

Prema Kerol Časki, sintaksička analiza može značajno doprineti identifikaciji autora u forenzičkim istraživanjima, posebno kada se radi o kratkim, fragmentiranim izjavama. Ona predlaže stilometrijske metode zasnovane na kvantitativnoj analizi sintaksičkih struktura kao što su dužina reči, frekvencija rečeničnih struktura i raspored sintaksičkih markera, pri čemu se postiže visoka tačnost u razlikovanju autora, čak i kada je količina podataka ograničena (Chaski, 2001, 2005). U okviru ovih metoda nekada se koristio softver ALIAS, koji je omogućavao lematizaciju, proračunavanje frekvencije reči, kategorizaciju interpunkcije i sintaksičku analizu. Program je bio zasnovan na sintaksičkom pristupu autorskoj atribuciji i obuhvatao je automatsku segmentaciju teksta na rečenice, označavanje gramatičkih kategorija (Part-Of-Speech tagging), kategorizaciju interpunkcijskih znakova i identifikaciju sintaksičkih struktura. Međutim, ALIAS se danas ne koristi budući da su u međuvremenu razvijeni savremeniji alati zasnovani na generativnoj veštačkoj inteligenciji, posebno veliki jezički modeli, koji se pokazali kao izuzetno efikasni u stilometrijskim i forenzičkim analizama različitih tekstualnih zadataka (Michelet & Breitinger, 2024). Ovi modeli omogućavaju dublju obradu teksta i preciznije razlikovanje autora, naročito u tekstovima sa nepoznatim autorstvom.

Istraživanja K. Časki ukazuju na značaj prepoznavanja specifičnih jezičkih struktura kao stilometrijskih markera koji otkrivaju jedinstvene jezičke obrasce karakteristične za pojedince, što je od izuzetne važnosti za formiranje i primenu forenzičkog korpusa na srpskom jeziku. Takvi markeri mogu biti presudni u preciznoj identifikaciji autora i u proceni legitimnosti spornih tekstova. Uvođenje ovih modela u analizu tekstova na srpskom predstavljalo bi značajan iskorak u razvoju forenzičke lingvistike.

Metodološki pristup Časki zasniva se na sintaksičkoj klasifikaciji interpunkcijskih znakova kao markera autorstva i naglašava značaj sofisticiranih jezičkih alata koji mogu da razlikuju suptilne varijacije unutar tekstova, uključujući raspored sintaksičkih granica sintagmi i rečenica (Chaski, 2005, pp. 5–7). Kombinacijom tradicionalnih stilometrijskih metoda i naprednih sintaksičkih analiza postiže se veća tačnost i pouzdanost u identifikaciji autora u digitalnim dokazima (Chaski, 2001, pp. 2–3, 7), što bi činilo srpski forenzički korpus vrednim resursom u istragama u kojima su identitet autora i autentičnost teksta ključni za pravne postupke.

U tom kontekstu, posebno je važan pojam markiranosti, koji u lingvistici označava asimetričnu binarnu opoziciju — jedan element (nemarkirani) predstavlja neutralnu ili češću formu, dok je drugi

(markirani) redi, formalno obeležen ili stilski naglašen. U sintaksi, nemarkirane strukture su uobičajene i lakše za obradu, dok markirane mogu biti složenije i ređe. Upravo zahvaljujući tim razlikama, sintaksička markiranost može omogućiti identifikaciju autorstva unutar istog autora, kao i razlikovanje između različitih autora (Chaski, 1997).

5. Potencijalni izazovi, bezbednosni i etički aspekti

U radu s tekstovima kao što su policijski izveštaji, pretnje, izjave svedoka i drugi pravno relevantni dokumenti, nezaobilazna su pitanja zaštite podataka. Iako su pravni aspekti privatnosti regulisani važećim zakonodavstvom, uključujući i Opštu uredbu o zaštiti podataka (GDPR),[3] u savremenom digitalnom okruženju sve veću pretnju predstavljaju bezbednosni rizici, posebno hakerski napadi. Izveštaji vladinih institucija i pravni dokumenti često su meta sajber napada, čime se ugrožava integritet i poverljivost forenzičkog materijala (Mayer, 2017). Stoga je pri kreiranju i korišćenju digitalnog korpusa neophodna primena visokih standarda zaštite, uključujući enkripciju, kontrolu pristupa, pseudonimizaciju osetljivih podataka i redovno ažuriranje bezbednosnih protokola. Pored tehničkih mera, važno je istaći etičku odgovornost istraživača u sprečavanju moguće zloupotrebe podataka. Forenzičkolingvistička analiza, naročito u pravnim postupcima, može imati direktne posledice po pojedince. Zbog toga je neophodno osigurati transparentnost metodologije, informisanost relevantnih strana i minimizovanje rizika od pogrešne interpretacije ili instrumentalizacije jezičkih dokaza.

6. Zaključak

Ovaj rad doprinosi razvoju FL u Srbiji, posebno u kontekstu kreiranja i primene digitalnog specijalizovanog korpusa forenzičkih tekstova na srpskom jeziku. Istaknute su mogućnosti koje ovakav korpus pruža u analizama autentičnosti teksta, identifikaciji autora i rešavanju pravnih sporova upotrebom jezičkih dokaza. Korpus bi obuhvatio raznovrsne forenzičke tekstove, uključujući policijske izveštaje, pretnje i oproštajna pisma, čime bi se omogućilo dublje razumevanje različitih jezičkih obrazaca unutar kriminalističko-pravnih diskursa.

Dodatna vrednost rada ogleda se u metodologiji koja bi se razvila za formiranje korpusa, kao i u uspostavljanju kriterijuma za selekciju

[3] General Data Protection Regulation (GDPR) je pravni okvir Evropske unije koji reguliše prikupljanje, obradu i čuvanje ličnih podataka (European Union, 2016). U kontekstu forenzičke lingvistike, GDPR zahteva strogu kontrolu pristupa i obrade tekstova koji sadrže podatke o ličnosti, posebno u slučajevima koji uključuju osetljive pravne ili policijske informacije.

relevantnih i autentičnih tekstova. Takođe, u radu je ukazano na važnost prilagođavanja softverskih alata specifičnostima srpskog jezika koji se suočava sa složenom morfologijom i slobodnim redosledom reči. Ovaj pristup ne samo da omogućava analizu specifičnih diskursa već doprinosi i razvoju alata za obradu srpskog jezika, što ima široko primenljiv potencijal.

Na osnovu dosadašnjih nalaza, preporučuje se nastavak rada na stvaranju sveobuhvatnog forenzičkog korpusa na srpskom jeziku, uz proširenje na dodatne vrste tekstova poput zahteva za otkupninu i krivičnih prijava. Dalji razvoj ovakvog korpusa zahteva interdisciplinarnu saradnju lingvista, pravnika i IT stručnjaka kako bi se osigurala precizna analiza i poštovanje zakonskih regulativa o zaštiti podataka.

Preporučuje se da buduća istraživanja budu usmerena na razvoj i primenu naprednih softverskih rešenja koja bi mogla automatizovati procese računanja leksičke frekvencije, sintaksičke analize i stilometrijskih modela specifičnih za srpski jezik. Korišćenje savremenih NLP tehnika, kao što su n-gram modeli i prepoznavanje jedinstvenih jezičkih obrazaca, dodatno bi unapredilo tačnost identifikacije autora i verodostojnosti teksta u pravnim kontekstima.

Na kraju, neophodna je kontinuirana edukacija stručnjaka u vezi s etičkim smernicama i zaštitom privatnosti prilikom obrade forenzičkih tekstova. Uspostavljanje korpusa predstavlja osnovu za dalji razvoj forenzičke lingvistike u Srbiji, a njegova implementacija bi značajno unapredila kvalitet i pouzdanost pravnih analiza, doprinoseći globalnoj mreži forenzičkih istraživanja i pravnih resursa.

Literatura

[1] Blackwell, S. (2009). Why forensic linguistics needs corpus linguistics. *Comparative Legilinguistics*, *1*, 5–19. https://doi.org/10.14746/cl.2009.01.01

[2] Chaski, C. (1997). Who wrote it? Steps toward a science of authorship identification. *National Institute of Justice Journal*, *233*(233), 15-22.

[3] Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic linguistics*, *8*, 1–65.

[4] Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International journal of digital evidence*, *4*(1), 1–13.

[5] Coulthard, Malcolm (2005). Some forensic applications of forensic linguistics. *Revista Veredas de Estudos Linguisticos*, *9*, 9–28.

[6] Coulthard, M. (2013). On the use of corpora in the analysis of forensic texts. *International Journal of Speech, Language and the Law*, *1*(1), 27–43.

[7] Dobrić, N. (2012). Savremeni jezički korpusi na zapadnom Balkanu–istorijat, trenutno stanje i budućnost. *Slavistična revija*, *60*(4), 677–692. https://srl.si/ojs/srl/article/view/COBISS_ID-51417186

[8] European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union*, *L 119*, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[9] Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, *34*(5), 1515–1540. https://doi.org/10.1007/s11196-021-09860-8

[10] Krstev, C. & Vitas, D. [2005]. Corpus and Lexicon – Mutual Incompletness. In P. Danielsson & M. Wagenmakers (eds), *Proceedings of the Corpus Linguistics Conference*, Vol. 4, Birmingham. Retrieved May 6, 2025 from https://www.researchgate.net/profile/Cvetana-Krstev/publication/266882123_Corpus_and_Lexicon_-Mutual_Incompleteness/links/544792880cf22b3c14e0ed6b/Corpus-and-Lexicon-Mutual-Incompleteness.pdf

[11] Michelet, G., & Breitinger, F. (2024). ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models. *Forensic Science International: Digital Investigation*, *48*, 301683. Retrieved May 6, 2025 from https://www.sciencedirect.com/science/article/pii/S2666281723002020

[12] Mayer, J. (2017). Government hacking. *Yale LJ*, *127*, 570.

[13] Nikolić-Novaković, L. (2017). *Forenzička lingvistika*. *Primena metoda forenzičke lingvistike u kriminalističkim istragama pretnje*. Banjaluka: Evropski defendologija centar za naučna, politička, ekonomska, socijalna, bezbjednosna, sociološka i kriminološka istraživanja.

[14] Olsson, J. (2010). *Forenzička lingvistika*. Zagreb: Globus.

[15] Svartvik, J. (1968). *The Evans statements: A case for forensic linguistics*. Göteborg: University of Göteborg. Retrieved December 20, 2020 from https://www.thetext.co.uk/Evans%20Statements%20Part%202.pdf

[16] Utvić, M. V. (2014). *Izgradnja referentnog korpusa savremenog srpskog jezika* (Publication No. 30969143) [Doktorska disertacija, University of Belgrade]. ProQuest Dissertations & Theses.

[17] Vitas, D., Krstev, C., Obradovic, I., Popovic, L., & Pavlovic-Lazetic, G. (2003). An overview of resources and basic tools for processing of Serbian written texts. In *Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics*.

[18] Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., & Stanojević, M. (2012). *The Serbian language in the digital age*. META-NET

White Paper Series, G. Rehm, H. Uszkoreit (eds.).
http://dr.rgf.bg.ac.rs/s/repo/item/0000764

[19] Vorkapić, D., Tomašević, A., Mladenović, M., Stanković, R., & Vulović, N. (2017). Digital Library From A Domain Of Criminalistics As A Foundation For A Forensic Text Analysis. In *International Scientific Conference "Archibald Reiss Days" Thematic Conference Proceedings Of International Significance*, Belgrade, 7–9 November 2017.

[20] 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.

The Importance of a Digital Corpus and Linguistic Tools in the Linguistic Analysis of Forensic Texts in the Serbian Language

Jelena Redli

Summary

Forensic linguistics (FL) is an interdisciplinary field that combines linguistic, legal, and digital technologies to analyze language use in legal and criminal contexts. Its primary applications include author identification, plagiarism detection, and criminal investigations, relying heavily on specialized linguistic corpora. However, forensic linguistic research on the Serbian language remains underdeveloped due to the absence of dedicated forensic text corpora, limiting the applicability of advanced analytical methods.

This paper investigates the potential for developing and applying a forensic text corpus for Serbian, highlighting its role in text authenticity verification, authorship attribution, and the resolution of legal disputes through linguistic evidence. The proposed methodology involves compiling and analyzing various types of forensic texts, including police reports, legal documents, threatening messages and suicide notes. The study also explores the integration of advanced computational linguistic tools for automated text analysis, including lexical frequency profiling, syntactic parsing, stylistic marker detection, and forensic stylometry techniques.

A key challenge in constructing a Serbian forensic corpus is addressing the linguistic complexity of the language, including its rich morphological structure, free word order, and dual script usage (Cyrillic

and Latin). Additionally, the research identifies major technical, legal, and ethical barriers in corpus development, particularly regarding data privacy, anonymization, and the ethical use of sensitive legal texts. The paper underscores the necessity of interdisciplinary collaboration between linguists, legal experts, law enforcement agencies, and IT professionals to ensure the corpus is both methodologically rigorous and legally compliant.

Beyond theoretical contributions, the paper presents potential real-world applications of a Serbian forensic corpus, such as its use in criminal investigations, fraud detection, and threat assessment. By analyzing forensic texts with computational methods, law enforcement and judicial authorities could improve the accuracy of authorship identification and linguistic profiling. The study also raises critical questions for future research, including:

How can computational forensic linguistic tools be adapted to Serbian's linguistic structure?

What legal frameworks have to be established to protect privacy while enabling forensic text analysis?

How can forensic corpora be continuously updated to reflect emerging linguistic trends in digital communication?

The findings suggest that establishing a specialized forensic corpus for Serbian would significantly enhance forensic linguistic research and its application in legal practice. Furthermore, the creation of such a resource would align Serbian forensic linguistics with global trends, facilitating cross-linguistic comparisons and the advancement of digital forensic methodologies. The paper concludes with a call for interdisciplinary efforts to bridge the gap between linguistic research, law enforcement needs, and technological advancements in forensic text analysis.

Keywords: forensic linguistics, Serbian language, legal language, digital corpus, linguistic tools, forensic text analysis, authorship identification.