STILOMETRIJA I JEZIČKI ALATI U FORENZIČKOJ LINGVISTICI

Jelena Redli
Filozofski fakultet, Univerzitet u Novom Sadu
redli@ff.uns.ac.rs

ŠTA JE FORENZIČKA LINGVISTIKA?

- ***Supspecijalistička oblast primene naučnih saznanja, teorija i metoda lingvistike prilikom analize uzoraka govornog i pisanog jezika u istražnim postupcima i sudskim sporovima u kojima je jezik deo dokaza, a ponekad i jedini dokaz***
- ***Primena lingvističke analize u pravnom i kriminalističkom kontekstu: ispitivanje jezika kako bi se otkrili ključi dokazi, odgovorilo na pitanja o autorstvu i osvetlile nijanse komunikacije unutar pravnog sistema.***
- ***Njena važnost leži u sposobnosti da raspetlja kompleksne slučajeve razlaganjem jezičkih nijansi.***

LINGVISTA U PRAVNOM KONTEKSTU

- ***Faza istrage***
- ***Faza suđenja***
- ***Žalbena faza***
- ***Privatni sporovi***

FAZA ISTRAGE

- ***Zahtevi za lingvističku analizu dolaze od institucija za sprovođenje zakona ili, u nekim zemljama, na poziv istražnog sudije (pružanje informacija o predmetu na osnovu lingvističke analize)***
- ***Uključivanje lingviste u istragu: analiza raznih vrsta beležaka, pisama, raznih tipova uzoraka snimljenog govora, tekstualnih i telefonskih poruka i specifičnih pretnji, validnosti oproštajnih pisama (uzrok i okolnosti smrti)***

FAZA SUĐENJA

- ***Pitanja na koja lingvista treba da odgovori:***
  - ***autorstvo (Ko je napisao tekst? Ko je govornik na ovom snimku?)***
- ***značenje i tumačenje (Da li ova reč znači x, y ili nešto treće? Na primer, šta znači kada neko kaže aha - da li aha uvek znači 'da'?)***
- ***analiza pretnje (Da li tekst sadrži pretnju?)***
- ***poreklo i oblikovanje teksta (Da li je tekst proizvod dvostrukog autorstva? Da li je pre pisan nego govoren? Da li je priznanje iznuđeno? Da li je osumnjičeni zaista priznao delo ili se tužiocu/policiji učinilo da jeste? itd.).***

FAZA ŽALBE

- ***Pokreće se tvrdnjom da su se pojavili novi dokazi ili da postojeće dokaze treba posmatrati na nov način.***
- ***Lingvistu angažuje odbrana ili tužilaštvo. Razlozi:***

➢ ***spor oko formulacije,***

➢ ***spor oko tumačenja,***

➢ ***spor oko autorstva izjave,***

➢ ***spor oko tumačenja izjave ili priznanja datog policiji,***

➢ ***može se tražiti novo tumačenje prethodnog forenzičkog izveštaja drugog forenzičara,***

➢ ***analiza uzoraka govora/teksta koji nisu dobro shvaćeni ili interpretirani, a uticali su na presudu i sl.***

POLJA PRIMENE

Razni tipovi kriminala:

- ***ubistvo***
- ***otmica***
- ***silovanje***
- ***seksualno zlostavljanje***
- ***ucena***
- ***korupcija***
- ***terorizam***
- ***trgovina ljudima/drogom…***

IZVORI DOKAZA

KRIVIČNA DELA

- ***razgovori s osumnjičenim i okrivljenim***
- ***razgovori u tajnim operacijama***
- ***govorne ili pisane ucene***
- ***govorni ili pisani zahtevi za otkup (otmice)***
- ***razgovori u slučajevima seksualnog zlostavljanja dece i odraslih***
- ***preteća komunikacija (pisma, imejlovi, SMS poruke, telefonski pozivi)***
- ***anonimne telefonske dojave interventnim službama…***

GRAĐANSKI SPOROVI

- ***ugovori***
- ***razna osiguranja***
- ***poslovne prevare***
- ***zaštitni znakovi***
- ***reklame***
- ***plagijati***
- ***testamenti…***

INSTRUMENTI FORENZIČKE ANALIZE I VEŠTAČENJA

- ***fonetika/fonologija***
- ***morfologija***
- ***leksikologija***
- ***sintaksa***
- ***semantika***
- ***pragmatika***
- ***analiza diskursa***
- ***teorija govornih činova***
- ***sociolingvistika***
- ***psiholingvistika***
- ***urbana i ruralna dijalektologija***

KLJUČNA PITANJA FLING

- ***forenzička fonetika i analiza govornih uzoraka: fonetska transkripcija, identifikacija govornika (komparacija govornika i određivanje lingvističkog profila nepoznatog govornika), forenzička analiza govora u kreiranju postupka predočavanja radi prepoznavanja…***
- ***jezik pravnih dokumenata***
- ***jezik policije i službi za sprovođenje zakona***
- ***obavljanje razgovora sa decom i osetljivim svedocima***
- ***interakcija u sudnici***
- ***lingvistički dokaz i svedočenje veštaka na sudu***
- ***pripisivanje autorstva i plagijati***

NASTANAK FLING

1966: Jan Svartvik je veštačio sporne iskaze Timotija Evansa, osuđenog na smrt vešanjem.

Identifikovao postojanje dva različita stila (obrazovanog pisanog stila i markiranog govornog stila) i kvantifikovao njihove razlike

- ***1968. objavio Evansove izjave: Slučaj za forenzičku lingvistiku.***
- ***U akademsku nomenklaturu uveo termin forenzička lingvistika.***

PRIMENE FLING

1. ***Identifikacija autora anonimnih ili kontroverznih tekstova (atribucija)***

- ***Anoniman tekst → tehnike: analiza stilskih karakteristika (izbor reči, sintaksičke strukture, greške u pisanju)***

2. ***Analiza pretnji, ucena i drugih oblika zloupotrebe jezika***

- ***Procena jezičkih obrazaca radi identifikacije potencijalne agresije, manipulacije ili prikrivene poruke***
- ***Procena stvarne opasnosti od pretnji i utvrđivanje da li određeni tekstovi mogu predstavljati krivična dela***

3. ***Razumevanje jezičkih dokaza u sudskim procesima***

- ***Pravna dokumenta → interpretacija testamenta, ugovora, zakonskih propisa i drugih pravnih materijala***
- ***Razjašnjavanje dvosmislenosti i nejasnoća u jeziku*** → cilj: ***jasnije i pravno ispravnije tumačenje***

RAČUNARSKA LINGVISTIKA U SLUŽBI FORENZIČKE LINGVISTIKE

Računarska lingvistika i automatska analiza jezičkih podataka

obrada i analiza ogromnih količina jezičkih podataka

algoritmi za obradu jezika mogu analizirati hiljade dokumenata u sekundi, identifikujući specifične jezičke obrasce

utvrđivanje potencijalnog autora (elektronska komunikacija: imejlovi, poruke na društvenim mrežama)

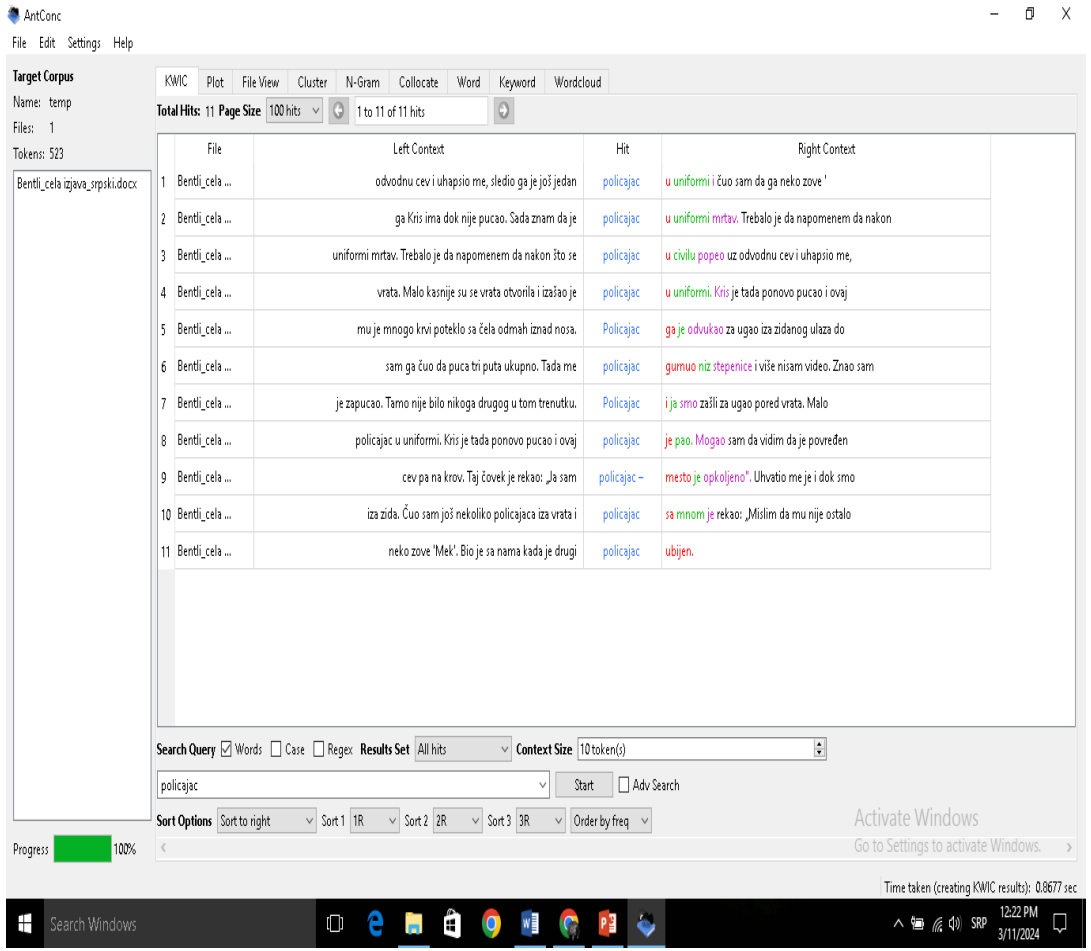
KVANTITATIVNA ANALIZA I SOFTVERI ZA ANALIZU TEKSTA

LINGUISTIC INQUIRY AND WORD COUNT (LIWC)

- ***Meri emocionalne, kognitivne i strukturne komponente jezika u tekstu.***
- ***Često se koristi za analizu emocionalnog sadržaja u pretnjama ili ucenama, pomažući stručnjacima da ocene ozbiljnost i autentičnost ovakvih poruka.***

https://www.liwc.app/demo

ANTCONC



- besplatni softver za analizu korpusa
- omogućava detaljnu analizu frekvencije reči, konkordansi i n-gram analizu.
- U slučaju identifikacije autora, može se koristiti za poređenje učestalosti određenih reči ili fraza između poznatih radova autora i teksta čiji autorstvo se istražuje.

https://www.laurenceanthony.net/software/antconc/

VOYANT TOOLS

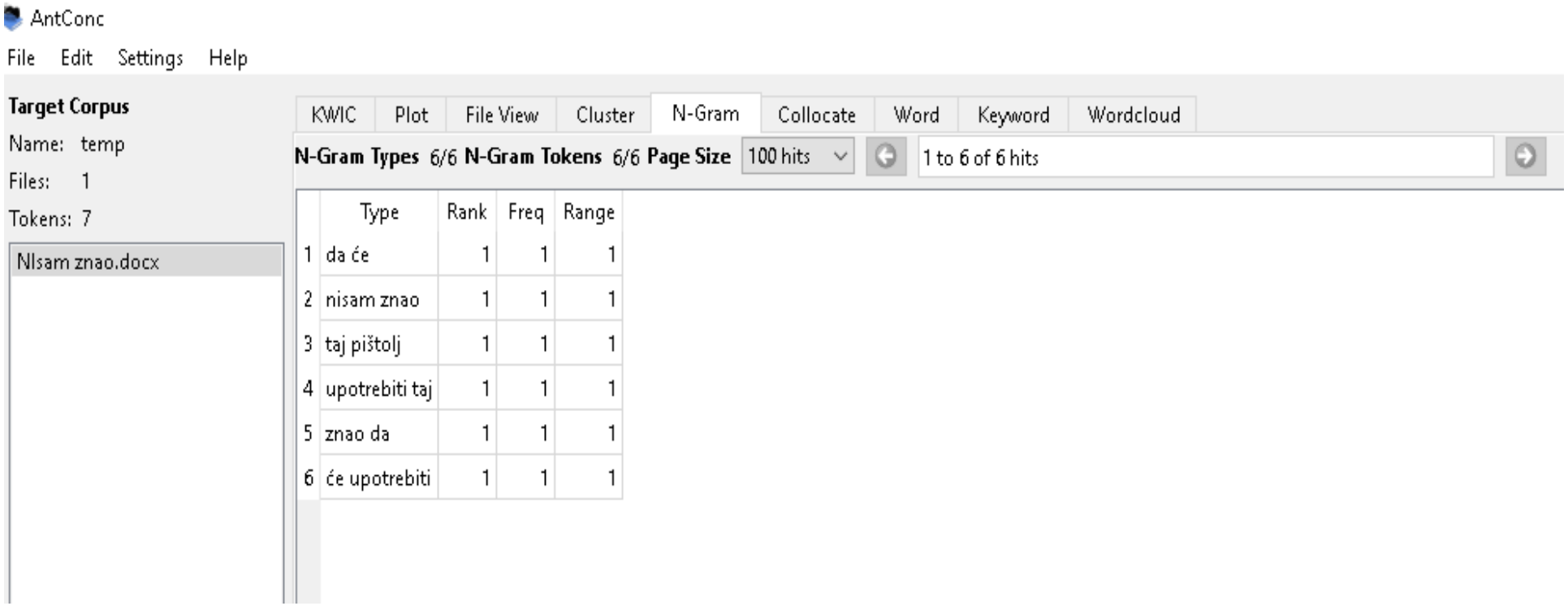
https://switchboard.clarin.eu/input

https://voyant-tools.org/

N-GRAMI

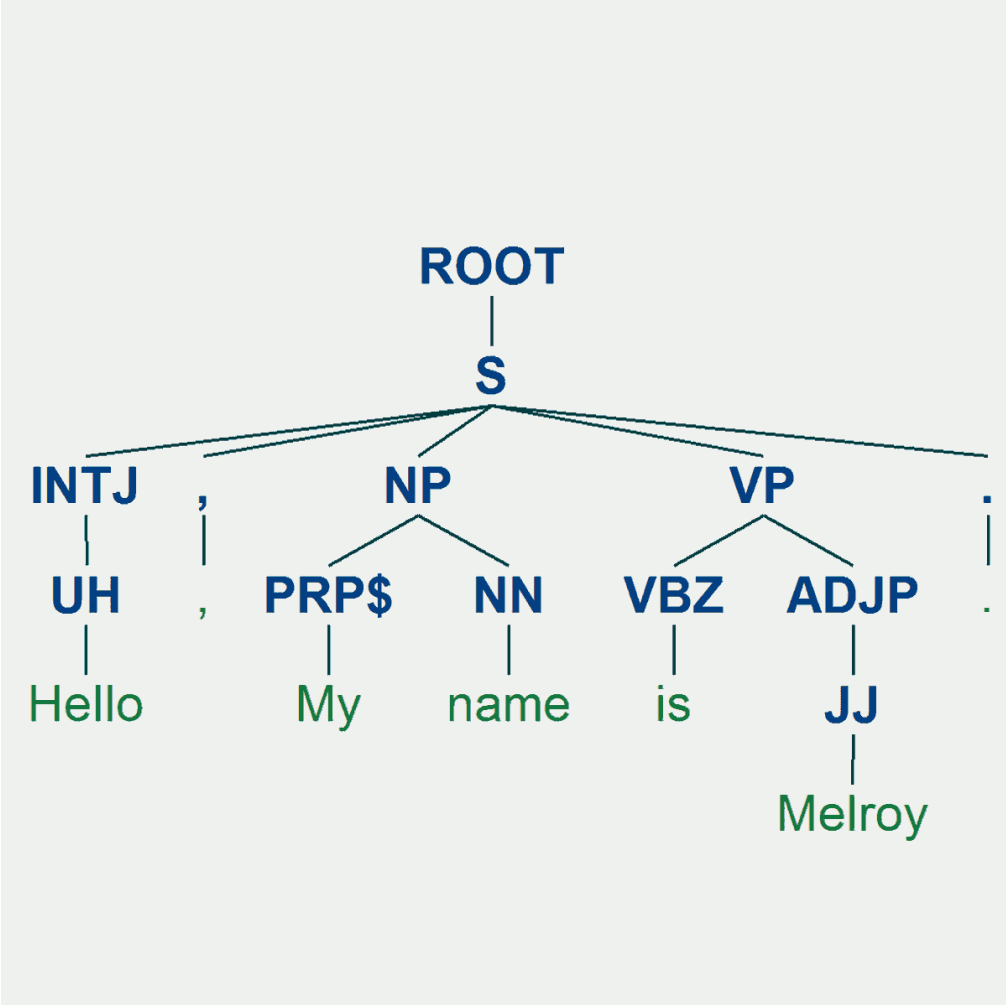
Jabuka je zelena

- ***1-grami (unigrami): "Jabuka", "je", "zelena“***
- ***2-grami (bigrami): "Jabuka je", "je zelena„***
- ***3-grami (trigrami): "Jabuka je zelena".***



SOFTVERI ZA ANALIZU SINTAKSE TEKSTA

STANFORD PARSER



spaCy

SyntaxNet

https://devpost.com/software/how-to-use-syntaxnet

TreeTagger

http://corpora.lancs.ac.uk/tree-tagger/

- ***JGAAP (Java Graphical Authorship Attribution Program) dizajnirao je Patrick Juola (Duquesne University).***
- ***Besplatan program za tekstualnu analizu, kategorizaciju teksta i atribuciju autorstva.***
- ***Prednosti JGAAP-a:***

☑***Pruža opsežne mogućnosti analitička prilagođavanja kao što su kanonizatori (normalizacija tekstova), selekcija (šta se uklanja iz podataka), analitički događaji (karakteristike poput n-grama, dužine reči, itd.), i metode analize (Burrows Delta, Hi-kvadrat, itd.).***

☑***Može obrađivati više tekstova i izvoditi razne vrste analitika odjednom.***

☑***Lako funkcioniše sa Javom, bez potrebe za dodatnim softverom ili znanjem programiranja.***

☑***Grafički korisnički interfejs (GUI) pruža uputstva i smernice u vezi sa različitim statističkim opcijama kao što su selekcija, metode analize i opcije frekvencija.***

- ***Nedostaci JGAAP-a:***

☒***Ne generiše vizualizacije podataka, već samo proizvodi sirove statističke rezultate.***

☒***Iako je korisnički vodič veoma obuhvatan, nije ažuriran od verzije 5.1 (2013).***

- ***Zanimljiva činjenica: Juola i JGAAP su najpoznatiji po otkriću da je J.K. Rowling autorka knjige "The Cuckoo's Calling„ (Robert Galbright).***

STILOMETRIJA I NJEN DOPRINOS FLING

1. ***Šta je stilometrija i kako se koristi za identifikaciju autora?***

- ***Metod analize teksta koji se fokusira na kvantitativno ispitivanje jezičkih karakteristika teksta kako bi se identifikovali autori nepoznatih ili spornih tekstova.***
- ***Koristi različite statističke i kompjuterske tehnike → posebno korisna kada su tekstovi anonimni ili kada postoje sumnje oko autorstva.***

2. ***Metode stilometrije***

- ***Analiza frekvencije reči – ispitivanje koliko često se određene reči ili fraze pojavljuju u tekstu, npr. međutim ili stoga***
- ***Analiza sintaktičke strukture – analizira rečenične strukture autora, npr. česta upotreba pasiva ili kompleksnost rečeničnih oblika.***
- ***Upotreba specifičnih fraza i idioma***

STILOMETRIJSKI ALATI U FORENZIČKIM ISTRAGAMA

UNABOMBAŠ

- ***Analiza njegovih tekstova pomogla je da se utvrdi autorstvo „Manifesta“ koji je bio ključan u identifikaciji Teda Kazinskog kao osumnjičenog.***
- ***Lingvisti su identifikovali značajne jezičke sličnosti koje su ukazivale na to da je Kazinski autor Manifesta.***
- ***Ova analiza je bila ključna u njegovom daljem procesuiranju.***

KO JE ZA TASTATUROM?

- ***Stilometrija se pokazala kao ključni alat u identifikaciji autora kontroverznih elektronskih dokumenata.***
- ***Bivši zaposleni u vladi optužen je za slanje uvredljivih imejlova svom nadređenom. Nakon što je otpušten, tužio je vladu tvrdeći da je bilo koji njegov kolega mogao pristupiti njegovom računaru i poslati imejlove.***
- ***Stilometrijska analiza, posebno analiza sintaktičkih obrazaca, pokazala je da su sporni imejlovi bili u skladu sa njegovim stilom pisanja. Rezultati analize su bili presudni u povlačenju njegove tužbe protiv vlade.***

Svaki tekst je obrađen korišćenjem programa ALIAS, koji je razvila Chaski (1997, 2001) u svrhu stvaranja baze podataka tekstova, lematizacije, računanja frekvencije leksičkih jedinica, izračunavanja dužine reči, rečenica i teksta, brojanja interpunkcijskih granica, označavanja rečeničnih elemenata, sortiranja n-grafa i n-grama, kao i potkategorizacije markiranosti. ALIAS je sposoban da pruži veliki broj lingvističkih promenljivih. Međutim, u ovom istraživanju koristile su se samo tri vrste promenljivih: interpunkcijske, sintaktičke i leksičke.

Chaski (2001) je pokazala je da sintaktički klasifikovana interpunkcija ima nešto bolje performanse od jednostavnih interpunkcijskih obeležja u razlikovanju autora, dok se istovremeno čuva intraautorska klasifikacija. Autori mogu deliti isti niz obeležja, ali je važna njihova pozicija. Ovaj pristup korišćenju interpunkcije kao autorskog identifikatora - sintaksički klasifikovana interpunkcija – veoma je različit od pristupa – jednostavna interpunkcijska obeležja – koji zagovara ispitivanje spornih dokumenata, forenzičku stilistiku, kao i druge računske stilometrijske studije. U pristupima sa jednostavnom interpunkcijom, broje se sami znaci interpunkcije, poput zareza, dvotačaka, uzvičnika, itd. U pristupu sintaksički klasifikovane interpunkcije, znaci (bez obzira na to šta tačno predstavljaju) broje se prema vrsti granice ili ruba koju interpunkcija obeležava.

Nakon što se svaki tekst automatski podeli na rečenice, korisnik sarađuje sa ALIAS-om kako bi kategorizovao interpunkciju unutar svake rečenice prema sintaksičkoj granici koju obeležava. Ove sintaksičke granice su klauza, fraza i morfema. Obeležja kraja klauze mogu biti zarezi, tačka-zarez, crtice; posebna obeležja se ne broje odvojeno, već se broji svako obeležje kraja klauze. Granica fraze može biti obeležena crticama ili zarezima; broji se obeležena granica kraja klauze. Unutrašnje leksičke granice obično su morfemske granice, pri čemu je morfema minimalna jedinica značenja. Na primer, engl. reč re-invent uključuje dve morfeme, [re] i [invent], a crtica obeležava granicu morfeme [re]. Broje se morfemske granice koje su markirane nekom interpunkcijom (npr. crticom). ALIAS zatim izvozi ove brojeve sintaksički klasifikovane interpunkcije u tabelu.

EKSPERIMENT: AI-GENERISAN TEKST NA TVITERU

- ***Stilometrijska analiza je korišćena za detekciju AI-generisanih tvitova unutar vremenskih linija korisnika na Tviteru.***
- ***Fokus: razlike između tvitova napisanih od strane ljudi i onih generisanih pomoću AI.***
- ***Analizirane su frazeologija, interpunkcija i lingvistička raznolikost → detektovane promene u stilu pisanja koje ukazuju na prelazak sa ljudskog autora na AI.***
- ***Predložena nova arhitektura za detekciju promene tačke autorstva, odnosno trenutka kada AI počinje da generiše tvitove.***
- ***Ključno za identifikaciju pretnji na društvenim mrežama i potencijalne dezinformacije.***
- ***Rezultati: stilometrijske karakteristike značajno poboljšavaju performanse detektora teksta generisanog pomoću AI, posebno kada su u pitanju kratki tvitovi koji sadrže ograničene količine semantičkih informacija.***

„WE SHALL BE WATCHING YOU, YOU'RE GOING TO DIE"

- ***Analizirane pretnje koje dolaze i od terorista i od neterorista.***
- ***Fokus:*** ***upotreba zamenica i vrsta rečenica.***
- ***Korišćen softver Wordsmith Tools za analizu učestalosti reči → detaljna analiza leksike i gramatičkih struktura unutar korpusa pretnji.https://www.lexically.net/wordsmith/***
- ***Nalaz:*** ***različita upotreba zamenica u pretnjama koje dolaze od terorista i onih koje dolaze od neterorista.***

Na primer, u pretnjama koje nisu povezane s terorizmom, češće se koristi 1. l. jd., dok teroristi preferiraju 1. l. mn.

STILOMETRIJSKA DETEKCIJA ONLINE PREVARA U UGOVORIMA

- ***Fokus: efikasnost različitih stilometrijskih softvera u detekciji studenstdkih prevara u slučajevima unajmljujivanja drugih osoba za pisanje zadataka.***
- ***Šta je analizirano:***

1. ***Test.dokumenti koji simuliraju prevare, poredeći tekstove za koje se veruje da su napisani od strane poznatog autora sa onima napisanim od drugih. Tekstovi su bili iz akademskih izvora i bili su konzistentni u temi i stilu kako bi se tačno testirale sposobnosti softvera.***

- ***Korišćeni stilometrijski softveri:***

1. ***Signature Stylometry System (SSS): pružio osnovnu analizu i poređenje na osnovu tekstualnih karakteristika kao što su dužina reči i rečenica, ali je pokazao nisku tačnost u testovima.***
2. ***JGAAP: ponudio sveobuhvatniju analizu, uključujući opcije za kanonizatore, pokretače događaja i različite metode analize → umereno dobri rezultati, sa varirajućim rezultatima zavisno od korišćene metode analize.***
3. ***JStylo Authorship Attribution Framework: proizveo najkonzistentnije i najtačnije rezultate u različitim testovima, ispitivajući karakteristike kao što su dužina rečenice i ocene čitljivosti.***

- ***Rezultati:***

1. ***Efikasnost alata je varirala, pri čemu je JStylo pokazao najvišu tačnost i konzistentnost u identifikaciji tačnog autorstva tekstova. JGAAP i SSS su imali mešovite rezultate, sa tačnošću koja je varirala u zavisnosti od složenosti analize i specifičnih podešavanja koji su korišćeni.***

- ***Studija je zaključila da stilometrijski softver ima značajan potencijal za detekciju prevara u akademskim tekstovima. Međutim, tačnost zavisi u velikoj meri od izbora softvera, korišćenih podešavanja i prirode tekstova koji se analiziraju. Predlaže se dalje istraživanje i unapređenje ovih alata kako bi se poboljšala pouzdanost i upotrebljivost u akademskim okruženjima.***

- ***Tehnike stilometrije. Upotrebljene su različite stilometrijske tehnike tekstualne analize.***
- ***U leksičkoj analizi analizirane su karakteristike zasnovanih na rečima i karakterima, što ima prednost da je otporna na „šum" u tekstu (npr. greške u pravopisu i gramatici).***
- ***Leksička analiza je obuhvatila n-grame reči (tj. ponavljanje reči "n" puta u dokumentu, gde je "n" broj pojavljivanja), frekvencije reči, broj reči po rečenici, broj rečenica i bogatstvo rečnika.***
- ***Razmatrane su i strukturne karakteristike kao što su uvlačenja, pravopisne greške, gramatika i reči specifične za određene društvene ili kulturne pozadine.***
- ***Sintaksičke karakteristike, koje uključuju interpunkciju i delove govora, takođe su pojačale analizu.***

PITANJA O IMPLIKACIJAMA UPOTREBE JEZIČKIH ALATA U PRAVNOM KONTEKSTU

1. ***Privatnost i poverljivost:*** ***Kako se osigurava da upotreba jezičkih alata ne krši privatnost pojedinaca čiji se tekst analizira? Postoje li adekvatne zaštite za osjetljive podatke koji se mogu pojaviti u analiziranom materijalu?***
2. ***Tačnost i pouzdanost:*** ***Koliko su pouzdani jezički alati u identifikaciji autora, naročito u kontekstu sličnih stilova pisanja ili kada autori namerno menjaju svoj stil? Kako se procenjuje i verifikuje tačnost tih alata?***
3. ***Pravednost u primeni:*** ***Da li se jezički alati primjenjuju jednako na sve slučajeve i pojedince? Postoji li rizik od pristrasnosti, bilo na osnovu jezika, dijalekta ili stilskih razlika koje bi mogle neproporcionalno uticati na određene grupacije?***
4. ***Transparentnost metodologije:*** ***Da li su metode koje se koriste u lingvističkoj analizi dovoljno transparentne tako da mogu biti podložne nezavisnoj reviziji i proveri?***
5. ***Konzistentnost i standardizacija:*** ***Postoje li standardizovani protokoli za upotrebu lingvističkih alata u pravnom kontekstu? Kako se osigurava da se analiza sprovodi na konzistentan i nepristrasan način?***
6. ***Odgovornost i nadzor:*** ***Ko je odgovoran za nadzor upotrebe lingvističkih alata u pravnom sistemu? Kako se reguliše i kontroliše upotreba tih alata da bi se izbegle zloupotrebe?***