POVEJ MI KARKOLI IN POVEM TI, KDO SI: UGOTAVLJANJE AVTORSTVA BESEDIL

Ana Zwitter Vitez
Trojina, Zavod za uporabno slovenistiko, Ljubljana

UDK 81'42:81'322

V prispevku poudarjamo pomen interdisciplinarne raziskave jezikovnih parametrov, ki omogočajo ugotavljanje avtorstva ali profiliranje avtorja besedila v forenzičnem, literarnozgodovinskem ali gospodarskem kontekstu (anonimna grozilna pisma, literarna besedila neznanega izvora, profiliranje strank). Ker je tovrstne analize za slovenščino težko najti, predlagamo metodologijo luščenja skladenjskih, leksikalnih, semantičnih in znakovnih parametrov za potrebe kvantitativne obravnave avtorjevega osebnega sloga.

ugotavljanje avtorstva besedil, jezikovni korpusi, jezikovne tehnologije

The paper shows the importance of an interdisciplinary analysis of linguistic features that enable authorship detection or author profiling in forensic, literary, and in economic context (anonymous threats, literary texts of unknown authorship, client profiling). It also highlights the lack of realised analyses for Slovene, and outlines the methodology of extracting syntactic, lexical, semantic, and character features in order to quantify the author's personal style in Slovene.

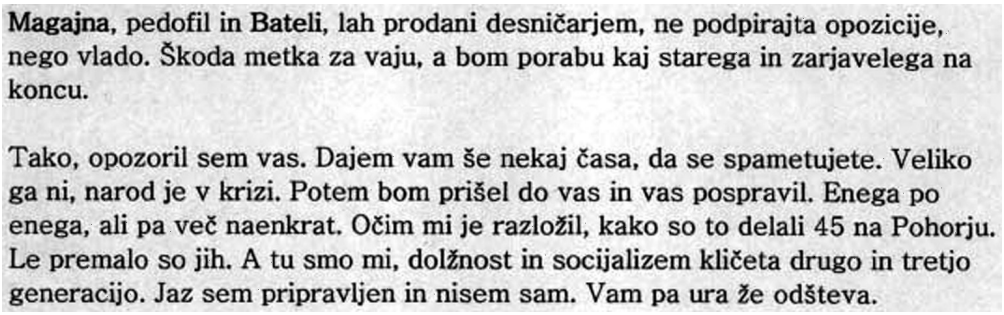
authorship attribution, author profiling, linguistic features, language technologies

1 Uvod

Ugotavljanje avtorstva je že od konca 19. stoletja vznemirljiva tematika na področju kriminalističnega preiskovanja, klasične filologije in literarne zgodovine. Pionirsko delo na področju ugotavljanja avtorstva glede na jezikovne lastnosti besedila je izvedel Mendenhall (1887), ko je na podlagi analize dolžine besed ugotavljal razlike med različnimi jeziki in različnimi avtorji. Ugotovil je, da kažeta histograma Shakespearja in Marlowa skoraj identične lastnosti, kar je presunljivo razkritje, saj je Marlowe v nepojasnjenih okoliščinah umrl dva tedna pred objavo Shakespearjevih prvih del.

Z razvojem jezikovnih korpusov in rudarjenja podatkov so analize s področja ugotavljanja avtorstva besedil še dodatno pridobile na prodornosti in zanesljivosti. Zato ne preseneča, da je ugotavljanje avtorstva preseglo osnovni izziv, da bi besedilo neznanega izvora pripisali enemu izmed potencialnih avtorjev, ampak razvija svoja podpodročja z določanjem osebnega profila avtorja besedila, detektiranjem plagiatorstva in ugotavljanjem stilnih neenotnosti pri večavtorskih besedilih. Danes je ugotavljanje avtorstva še posebej razvito na področjih prava in avtorskih pravic (Grant 2007), literarnih ved (Burrows 2002; Hoover 2004), pri kriminalističnih preiskavah (Coulthard 2007) in profiliranju kupcev v komercialne namene (Shaw idr. 2001).

Za boljšo predstavo o konkretnih nalogah ugotavljanja avtorstva navajamo primer besedila, primernega za kriminalistično preiskavo (Slika 1). Več o tem v razdelku 7.



Magajna, pedofil in Bateli, lah prodani desničarjem, ne podpirajta opozicije, nego vlado. Škoda metka za vaju, a bom porabu kaj starega in zarjavelega na koncu.

Tako, opozoril sem vas. Dajem vam še nekaj časa, da se spametujete. Veliko ga ni, narod je v krizi. Potem bom prišel do vas in vas pospravil. Enega po enega, ali pa več naenkrat. Očim mi je razložil, kako so to delali 45 na Pohorju. Le premalo so jih. A tu smo mi, dolžnost in socijalizem kličeta drugo in tretjo generacijo. Jaz sem pripravljen in nisem sam. Vam pa ura že odšteva.

Slika 1[1]

2 Ugotavljanje avtorstva v mednarodnem merilu

Ena ključnih nalog na tem področju je odkriti tiste jezikovne parametre, ki kvantitativno določajo slog določenega avtorja. Pri večini jezikov se kot najboljši jezikovni parametri kažejo najmanj očitna jezikovna sredstva: funkcijske besede, skladenjski vzorci itn. Taka jezikovna sredstva uporabljamo nezavedno in jih zato najteže nadziramo.

V nadaljevanju predstavljamo raziskave glede na štiri tipe jezikovnih lastnosti, s katerimi so skušali raziskovalci kvantificirati osebne sledi avtorjev v (predvsem pisnih) besedilih: leksikalne, znakovne, oblikoskladenjske in semantične lastnosti.

2.1 Leksikalne lastnosti

Leksikalne lastnosti besedila ponavadi predstavljajo vektorji besednih frekvenc (Sebastiani 2002). Med najboljšimi leksikalnimi lastnostmi za ugotavljanje avtorstva so se pokazale t. i. funkcijske besede (predlogi, zaimki, pomožni glagoli, vezniki in členki), nad katerimi avtor nima nadzora (Burrows 1987; Argamon, Levitan 2005; Luyckx, Daelemans 2005). Vendar Stamatatos (2009) ugotavlja, da je ključna pomanjkljivost metode ugotavljanja leksikalnih lastnosti močna odvisnost od dolžine besedila.

2.2 Znakovne lastnosti

Pri metodi ugotavljanja znakovnih lastnosti besedil gre za ekstrahiranje znakovnih n-terčkov. Pri kvantificiranju avtorjevega sloga so se kot zelo uspešne izkazale študije Keselja idr. (2003), Stamatatosa (2006) in Diedericha idr. (2003). Primerjalna študija leksikalnih in znakovnih lastnosti istega korpusa (Grieve 2007) je pokazala, da so n-terčki učinkovitejše merilo za kvantificiranje avtorjevega sloga.

2.3 Skladenjske lastnosti

Metoda ugotavljanja skladenjskih lastnosti besedil sloni na ideji, da avtorji v besedilih nezavedno uporabljajo iste skladenjske vzorce. Zato velja, da je ta metoda bolj zanesljiva od metode leksikalnih lastnosti, vendar zahteva tudi bolj izdelana jezikovna orodja (npr. oblikoskladenjski označevalnik, razčlenjevalnik). Prvi so jo uporabili Baayen idr. (1996), za njimi pa še številne študije, npr. Stamatatos idr. (2000), Luyckx, Daelemans (2005), Hirst, Feiguina (2007).

2.4 Semantične lastnosti

Iskanje semantičnih lastnosti besedila je danes osnovano na semantični mreži Wordnet, ki omogoča iskanje sinonimov in hipernimov besed. Wordnet so uporabile različne študije; med najbolj znanimi je McCarthy idr. (2006), s katero je mogoče detektirati semantične podobnosti med besedami.

1 www.politikis.si/?p=15680

Korpusi, ki jih uporabljajo raziskave ugotavljanja avtorstva, so ponavadi omejeni na določen besedilni žanr in opremljeni z natančnimi podatki o avtorjih besedil, zato predstavljajo tudi dragocen narodnozgodovinski dokument.

3 Ugotavljanje avtorstva na Slovenskem

Na področju profiliranja avtorja besedila izstopa študija *Uporabnost spoznanj sociolingvistike in psiholingvistike za kriminalistično preiskovanje* (Umek, Brglez 2009), v kateri avtorja analizirata laično sposobnost prepoznavanja nekaterih potez jezikovnih sledi avtorja v besedilu: spol, starost in stopnjo izobrazbe. Raziskava kaže, da je intuitivno dokaj lahko zaznati spol in starostno skupino avtorja, manj pa ostale lastnosti.

Za ugotavljanje avtorstva besedil uporabljajo statistično analizo naslednje študije: analiza dolžine stavkov in besed za ugotavljanje plagiatorstva v besedilu (Dović 2002), analiza funkcijskih besed kot možnih jezikovnih parametrov za ugotavljanje avtorstva besedila (Limbek 2008) in razporeditev n-terčkov na koncih povedi (Jakopin 2003). Težava je v tem, da so omenjene raziskave omejene na korpuse v obsegu nekaj literarnih del, za ugotavljanje verodostojnih jezikovnih parametrov, na podlagi katerih bi lahko določili avtorstvo besedil za slovenščino, pa bi potrebovali jezikovni korpus, uravnotežen glede na spol, starost, izobrazbo in regionalno pripadnost avtorjev besedil.

Področje ugotavljanja avtorstva je tesno povezano z jezikovnimi viri in orodji za posamezni jezik. Dobro novico s tega področja predstavlja dejstvo, da je slovenščina danes opremljena z milijardnim pisnim korpusom Gigafida,[2] milijonskim govornim korpusom Gos[3] in naslednjimi jezikovnimi orodji:

- oblikoslovni označevalnik[4] in skladenjski razčlenjevalnik[5] ter
- orodje Text-Garden, namenjeno analizi podatkov v besedilih, razvito na Institutu Jožef Stefan.

Za slovenščino je torej tema ugotavljanja avtorstva še vedno praktično neraziskana, vendar kaže dobre možnosti za kakovostne raziskave zaradi dobro razvitih jezikovnih orodij in virov.

4 Kaj lahko naredimo

Za izhodiščno hipotezo vzemimo dejstvo, da je s pomočjo kakovostno zgrajenega[6] in označenega korpusa besedil mogoče ugotoviti jezikovne parametre za slovenščino, s katerimi je mogoče kvantificirati avtorjeve jezikovne sledi v besedilu. Na podlagi ugotovljenih jezikovnih parametrov in izbranih metod strojnega učenja je mogoče sklepati o tem, ali je besedilo neznanega izvora delo enega od potencialnih avtorjev, oziroma določiti profil neznanega avtorja, ki je tvoril to besedilo (spol, starost, izobrazbo, regionalno pripadnost in psihometrične lastnosti).

Če izhodiščna hipoteza drži, bo raziskava odgovorila na naslednji vprašanji:

- ko razpolagamo z nekaj potencialnimi avtorji: kdo je najverjetnejši avtor besedila neznanega izvora,
- ko razpolagamo z besedilom neznanega izvora: kakšen je osebni profil avtorja besedila (spol, starost, izobrazba, regionalna pripadnost, psihološki profil[7]).

2 www.gigafida.net
3 www.korpus-gos.net
4 http://oznacevalnik.slovenscina.eu
5 http://razclenjevalnik.slovenscina.eu
6 Zgradba korpusa je odvisna od končnih lastnosti besedila, ki jih želimo izluščiti na podlagi jezikovnih parametrov besedila. Največkrat ugotavljamo spol, starost, izobrazbo, regijsko pripadnost in psihološki profil avtorja, zato v korpus uvrstimo besedila avtorjev, ki pripadajo omenjenim besedilnim parametrom.
7 Za potrebe profiliranja avtorjev se uporablja Mednarodni vprašalnik za določanje osebnosti (International Personality Item Pool – IPIP).

5 Metodologija

Metodologija za ugotavljanje avtorstva in profiliranje avtorja besedila sledi osnovnim etapam strojnega učenja z elementi jezikoslovnih analiz:

a) izdelava referenčnega korpusa,
b) razvrščanje besedil in izdelava modelov,
c) luščenje jezikovnih parametrov za slovenščino,
d) evalvacija modelov za ugotavljanje avtorstva in profiliranje avtorja.

a) Izdelava referenčnega korpusa

Za izdelavo referenčnega korpusa za raziskave avtorstva je treba:

– pridobiti besedila iz različnih virov (obstoječih korpusov slovenščine, spletnih strani in posameznih avtorjev),
– določiti kategorije za vnos atributov o besedilu in avtorju (zvrst in leto zajema, spol, starost, stopnja izobrazbe),
– izvesti avtomatsko oblikoslovno označevanje in skladenjsko razčlenjevanje,
– označiti besedila s podatki o avtorjih.

b) Izdelava modelov za določanje avtorstva

Ta faza je namenjena določitvi optimalnih jezikovnih lastnosti in statičnih metod za ugotavljanje avtorstva besedil, zato predvideva:

– ugotavljanje učinkovitosti različnih jezikovnih parametrov za določanje avtorstva besedil: leksikalnih (npr. bogatost besedišča, frekvence besed), znakovnih (npr. znakovni n-terčki), skladenjskih (npr. večbesedna zaporedja besed) in semantičnih parametrov (npr. sinonimi),
– sprotno preverjanje z ročno jezikovno analizo.

Rezultat te faze je kombinacija različnih jezikovnih parametrov, ki jih lahko uporabimo za izdelavo modela za pripisovanje avtorstva.

c) Luščenje optimalnih jezikovnih parametrov za profiliranje avtorja

Cilj te faze je določitev osebnega profila neznanega avtorja besedila. Zato je treba določiti jezikovne parametre, ki odločilno zaznamujejo avtorjev slog glede na avtorjev osebni profil (spol, starost, regijska pripadnost, izobrazba, psihometrične značilnosti).

d) Evalvacija modelov določanja avtorstva in profiliranja avtorja

V zadnji fazi se ugotavlja uspešnost avtomatskega ugotavljanja avtorstva in profiliranja avtorja z metodo uporabe korpusov, ki se od učnega korpusa razlikujejo v eni lastnosti (npr. v številu besedil na avtorja, številu potencialnih avtorjev in dolžini besedil).

6 Rezultati

Rezultate takšne raziskave lahko razvrstimo v tri skupine:

– referenčni korpus, ki predstavlja uravnoteženo bazo besedil različnih profilov avtorjev (glede na spol, starost, izobrazbo, regijo in psihometrične značilnosti),
– opis optimalnih kombinacij jezikovnih parametrov za ugotavljanje avtorstva besedil pri različnih besedilnih žanrih,
– opis optimalnih kombinacij jezikovnih parametrov za določitev osebnega profila avtorja besedila (spol, starost, izobrazba, regija, psihometrične značilnosti).

7 Uporabnost raziskav s področja določanja avtorstva besedil

Potreba po raziskavah s področja ugotavljanja avtorstva besedil je med drugim utemeljena z dejstvom, da se javne in nejavne osebnosti pogosto srečujejo s pojavom internetnih groženj in grozilnih pisem v tradicionalni obliki. V zadnjih nekaj letih so tovrstne grožnje doživeli Bush, Janša, K. Kresal, Žerjav, Jelinčič, Magajna, Batelli.[8]

8 www.dnevnik.si/novice/kronika/1042292860 in www.politikis.si/?p=15680

Zaradi izjemne dostopnosti besedil na spletu opažamo vse bolj pogost pojav plagiatorstva, o čemer pričajo doktorat nemškega obrambnega ministra Guttenberga, govor Janše ob dnevu državnosti 2008, monografija Cvikla, magistrska naloga Jakliča, izpit iz znanja nemškega jezika Mariniča itn.

Na področju literarnih študij lahko tovrstne raziskave rešijo (včasih pa tudi sprožijo) številne zagate, povezane z neznanimi ali nedokazanimi avtorstvi besedil, npr. pornografski roman *Čudoviti klon*, izdan pod psevdonimom Eva Pacher, roman *Sedem* z avtorjem Davidom Benjaminom ali predosamosvojitveni prispevki v tedniku *Mladina* z avtorjem (ali avtorji) Majdo Vrhovnik.

Profiliranje avtorja se uporablja tudi na področju kadrovanja oz. iskanja živih virov, kjer se pogosto govori o ljudeh kot kapitalu in potencialu podjetij (Schuler, Jackson 1999). Poznavanje jezikovnih parametrov, odločilnih za ugotavljanje osebnostnega profila avtorja, lahko namreč v podjetjih znatno izboljša izbor pravih kandidatov.

Profiliranje avtorja besedila je pomembno tudi v podjetništvu za potrebe poznavanja strank in njihovih kupnih navad. Zato podjetja gradijo baze kupnih navad in jezikovnih profilov strank ter se na podlagi njih odločajo o strategijah ponudbe in oglaševanja (Shaw idr. 2001).

8 Zaključek

V prispevku smo obravnavali pomen kakovostne analize jezikovnih parametrov, ki bi omogočila ugotavljanje avtorstva in določanje osebnega profila avtorja besedila v forenzičnem, literarnozgodovinskem ali gospodarskem kontekstu (anonimna grozilna pisma, literarna besedila neznanega izvora, profiliranje strank). Ker so tovrstne analize za slovenščino maloštevilne, predlagamo metodologijo luščenja skladenjskih, leksikalnih, semantičnih in znakovnih parametrov za potrebe kvantitativne obravnave avtorjevega osebnega sloga.

Glede na visoko uporabno vrednost tovrstnih raziskav se zdita verjetna dva scenarija: takšne jezikovne raziskave bodo finančno podprli državni organi ali pa jih bodo gmotno omogočili gospodarski subjekti. V prvem primeru bodo analize usmerjene v razkrivanje avtorstva anonimnih grozilnih pisem, plagiatorstva in literarnih besedil. Če bo raziskava financirana s strani gospodarskih subjektov, pa bo najverjetneje prispevala predvsem k razvoju določanja profila kandidatov in strank. Etična presoja bi dilemo prav gotovo nagnila k prvi možnosti.

Viri

ARGAMON, Shlomo, LEVITAN, Shlomo, 2005: Measuring the usefulness of function words for authorship attribution. *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*.

BAAYEN, Harald, van HALTEREN, Hans, TWEEDIE, Fiona, 1996: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11/3. 121–131.

BRGLEZ, Lucija, UMEK, Peter, 2009: Uporabnost spoznanj sociolingvistike in psiholingvistike za kriminalistično preiskovanje. *10. slovenski dnevi varstvoslovja: Varstvoslovje med teorijo in prakso*. Maribor: Fakulteta za varnostne vede.

BURROWS, John F., 2002: Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17/3. 267–287.

COULTHARD, Malcom, 2005: The linguist as expert witness. *Linguistics and the Human Sciences* 1/1.

DIEDERICH, Joachim idr., 2003: Authorship attribution with support vector machines. *Applied Intelligence* 19/1–2. 109–123.

DOVIĆ, Marijan, 2002: Podbevšek in Cvelbar: Poskus empirične preverbe namigov o plagiatorstvu. *Slavistična revija* 50. 233–249.

GRANT, Tim D., 2007: Quantifying evidence for forensic authorship analysis. *International Journal of Speech Language and the Law* 14/1. 1–25.

GRIEVE, Jack, 2007: Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22/3. 251–270.

HIRST, Graeme, FEIGUINA, Ol'ga, 2007: Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22/4. 405–417.

HOOVER, David, 2004: Testing Burrows' Delta. *Literary and Linguistic Computing* 19/4. 453–475.

JAKOPIN, Primož, 2003: Nizkoentropijski jezikovni model na besedilih Cirila Kosmača in Ivana Cankarja. Miran Hladnik, Gregor Kocijan (ur.): *Slovenski roman Obdobja 21*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. 421–428.

KEŠELJ, Vlado idr., 2003: N-gram-based author profiles for authorship attribution. *Proceedings of the Pacific Association for Computational Linguistics*. 255–264.

LIMBEK, Marko, 2008: Usage of Multivariate Analysis in Authorship Attribution: Did Janez Mencinger Write the Story »Poštena Bohinčeka«? *Metodološki zvezki* 5/1. 81–93.

LUYCKX, Kim, DAELEMANS, Walter, 2005: Shallow text analysis and machine learning for authorship attribution. *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands*.

McCARTHY, Philip. M. idr., 2006: Analyzing writing styles with coh-metrix. *Proceedings of the Florida Artificial Intelligence Research Society International Conference*. 764–769.

MENDENHALL, Thomas Corwin, 1887: The characteristic curves of composition. *Science* IX. 237–249.

SEBASTIANI, Fabrizio, 2002: Machine learning in automated text categorization. *ACM Computing Surveys* 34/1.

SHAW, Michael idr., 2001: Knowledge management and data mining for marketing. *Decision Support Systems* 31/1. 127–137.

STAMATATOS, Efstathios, 2009: A Survey of Modern Authorship Attribution Methods. 60/3. 538–556.

STAMATATOS, Efstathios, FAKOTAKIS, Nikos, KOKKINAKIS, George, 2000: Automatic text categorization in terms of genre and author. *Computational Linguistics* 26/4. 471–495.

ZHENG, Rong idr., 2006: A framework for authorship identification of online messages: Writing style features and classification techniques. *Journal of the American Society of Information Science and Technology* 57/3. 378–393.