# Ugotavljanje avtorstva besedil: primer »*Trenirkarjev*«

**Ana Zwitter Vitez**

Oddelek za prevajalstvo, Filozofska Fakulteta
Aškerčeva 2, 1000 Ljubljana
Trojina, zavod za uporabno slovenistiko
Dunajska 116, 1000 Ljubljana
ana.zwitter@guest.arnes.si

**Povzetek**

V prispevku predstavljamo analizo avtentičnega primera anonimnega besedila, ki je leta 2011 močno vznemirilo slovensko javnost. Avtorstvo besedila smo preverjali na korpusu 75 besedil 21 potencialnih avtorjev na podlagi vnaprej določenega nabora leksikalnih in berljivostnih značilk. Rezultati kažejo, da ima eden od potencialnih avtorjev zelo podobne vrednosti značilk, vendar v dani situaciji ni mogoče preveriti, ali je bil dejanski avtor besedila zajet v analizo ali ne.

**Authorship attribution: the 'Sportsuits' case**

In this paper we examine an authentic anonymous text which provoked intense reactions in Slovenian media in 2011. Within this authorship attribution task, a corpus of 75 texts written by 21 potential authors was analysed with a predefined set of lexical and readability features. The results show that one of the candidate authors resembles the anonymous text by most of the features although it is not possible to verify whether the actual author was included into the analysis or not.

## 1. Uvod

Ugotavljanje avtorstva besedil je v zadnjih desetletjih doživelo velik razmah zaradi hitrega razvoja postopkov analize velikih količin besedil, dodatno pa ga spodbuja veliko povpraševanje na področjih prava, kriminologije, literarnih ved in trženja. V zadnjih letih se namreč pogosto soočamo z naslednjimi pojavi:

- plagiati (doktorat nemškega obrambnega ministra, magisterij direktorja Lekarne Ljubljana),
- grozilna pisma (Janez Janša, Katarina Kresal),
- literarni psevdonimi (angleški blog *Belle de Jour*, slovenski roman *Čudoviti klon*),
- analiza profilov strank za potrebe trženja.

Ker je ugotavljanje avtorstva besedil izrazito interdisciplinarno področje, so med obstoječimi pristopi ogromne razlike. Na področju informatike prevladujejo študije z velikimi in dobro označenimi bazami podatkov, kontroliranim naborom značilk in natančno evalvacijo končnih modelov (Sebastiani 2002). Forenzični izvedenci pa se pogosto soočajo s kratkimi besedili brez možnosti analize primerljivega gradiva (Coulthard 2005).

V prispevku predstavljamo avtentičen primer ugotavljanja avtorstva besedila s pomočjo povprečne absolutne razlike med vrednostmi značilk.

## 2. Ugotavljanje avtorstva besedil

Če je tipična naloga ugotavljanja avtorstva besedil pripisati besedilo neznanega izvora enemu od potencialnih avtorjev, lahko z vidika strojnega učenja isto nalogo opišemo kot klasifikacijo besedil v več razredov (Sebastiani 2002, Stamatatos idr. 2001, Keselj idr. 2003).

Najprej je treba definirati značilke, torej lastnosti besedila, relevantne za klasifikacijo. S pomočjo izračunanih značilk je mogoče določeno besedilo predstaviti v obliki vektorja in tako kvantificirati določene lastnosti besedila. Pri tem različni raziskovalci upoštevajo različne kriterije: leksikalne (Sebastiani 2002, Argamon, Levitan 2005), grafemske (Keselj idr. 2003, Stamatatos 2006), skladenjske (Baayen idr. 1996, Hirst, Feiguina 2007) in semantične (McCarthy idr. 2006).

Na podlagi vektorjev značilk lahko izvedemo klasifikacijo besedil. Nekateri pristopi vsa besedila enega avtorja združijo v en dokument (angl. *compression-based approaches*), nato pa na podlagi te enote poskušajo kvantificirati avtorjev slog (Marton idr. 2005). Drugi pristopi vsako besedilo obravnavajo kot samostojno enoto, ki s svojimi lastnostmi prispeva h gradnji klasifikacijskega modela (Chaski 2005). Pri tem se je kot eden najbolj zanesljivih klasifikatorjev izkazala metoda podpornih vektorjev (SVM), ki ni občutljiva na šum in razpršenost podatkov (Li idr. 2006).

Zadnja etapa ugotavljanja avtorstva besedil je evalvacija. Pri določanju stopnje natančnosti modelov igrajo pomembno vlogo dolžina besedil učnega korpusa (Marton idr. 2005, Hirst, Feiguina 2007), število potencialnih avtorjev (Koppel idr. 2006), in razporeditev besedil na posameznega avtorja (Stamatatos 2008).

## 3. Kontekst in hipoteza raziskave

Analizirali smo besedilo, ki je leta 2011 močno vznemirilo slovensko javnost. Besedilo je bilo objavljeno na uradni spletni strani ene od parlamentarnih strank in podpisano s psevdonimom Tomaž Majer. Nekaj dni po objavi je informacijska pooblaščenka anonimnega avtorja ovadila zaradi sovražnega govora, sodišče je ovadbo zavrglo, javnost pa se ni nehala spraševati o dejanskem avtorju besedila. Največkrat citirani elementi spornega besedila so se nanašali na interpretacijo zmage nasprotne stranke, ki naj bi ji botrovala udeležba "volivcev s tujim naglasom" in "volivcev v športnih oblačilih (trenirkah), ki so imeli na roki s kemičnim svinčnikom napisano številko, ki jo morajo obkrožiti na glasovnici" Zato se je besedila prijelo ime *Trenirkarji*.

Za potrebe raziskave smo formulirali naslednjo hipotezo: če je avtor besedilo anonimno objavil na uradni spletni strani stranke, obstaja velika verjetnost, da je na isti spletni strani objavil še kakšno svoje besedilo pod drugim ali pravim imenom.
Zato smo za potrebe raziskave analizirali besedila avtorjev, ki so na isti spletni strani objavljali tri mesece pred in tri mesece po objavi spornega besedila. Tako smo dobili korpus 75 besedil 21 podpisanih avtorjev s približno 55.000 pojavnicami, ki so precej neizenačeno razporejene po avtorjih (od 650 do 9000 pojavnic na avtorja).

## 4. Metodologija

Priprava besedil je zajemala:
- čiščenje besedil,
- pretvorbo iz html v format .txt,
- anonimizacijo besedil (avtorji so označeni z velikimi črkami, njihova besedila pa z zaporednim številom),
- tvorjenje glav dokumentov,[1]
- oblikoslovno označevanje (Grčar idr. 2012) in
- skladenjsko razčlenjevanje besedil (Dobrovoljc idr. 2012).

Analiza besedil je bila zasnovana na vnaprej pripravljenem naboru značilk besedišča in berljivosti[2], s katerim smo se želeli izogniti odvisnosti od tematike besedil.
Leksikalne značilke:
- raznolikost besedišča (*lexical density*),
- Brunetova formula (Brunet 1988): raznolikost besedišča neodvisno od dolžine besedila,
- hapax legomena (Holmes 1992): leme, ki se pojavijo samo enkrat v besedilu,
- Honoréjeva formula (Honoré 1979): razmerje med številom hapaksov in raznolikostjo besedišča.

Berljivostne značilke :
- formula Flesch-Kincaid: razmerje med številom besed in številom povedi
- formula Coleman-Liau: razmerje med številom znakov in številom besed,
- formula Automated Readability Index: izračun stopnje izobrazbe, potrebne za razumevanje besedila ob prvem branju,
- formula Gunning Fog (Gunning 1952): izračun števila let formalnega izobraževanja, potrebnih za razumevanje besedila po prvem branju.

Na podlagi naštetih formul smo izračunali povprečne absolutne razlike med vrednostmi značilk in postavili hipotezo o najverjetnejšem avtorju anonimnega besedila.

## 5. Analiza

[1] Primer anonimizirane glave besedila: *A_1: Politična izprijenost in pošteni civilisti*
[2] Formule: **Lexical Density** = (different words / words) x 100, **Gunning Fog Index** = 0.4 x (ASL + ((SYW / words) x 100)), **ARI** = (0.5 x ASL) + (4.71 x ALW) - 21.43 **Coleman-Liau Grade** = 5.89 x ACW - 0.3 x sentences / (100 x words) - 15.8 **Flesch-Kincaid Grade Level** = (0.39 x ASL) + (11.8 x ASW) – 15.59

Najprej smo izračunali absolutne vrednosti značilk besedišča in berljivosti za vsako od analiziranih besedil. Tabeli 1 in 2 predstavljata rezultate za anonimno besedilo:

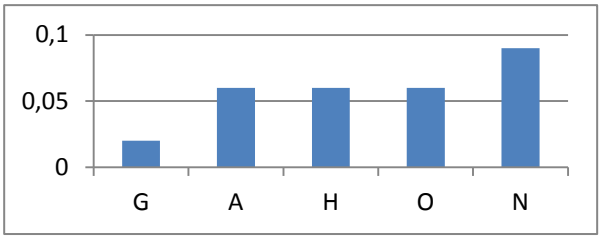
| **Značilka** | **Vrednost** |
|---|---|
| Raznolikost besedišča | 0,38 |
| Formula Brunet | 12,96 |
| Statistika Honoré | 1998,79 |
| Hapax legomena | 0,24 |

Tabela 1: Leksikalne značilke anonimnega besedila.

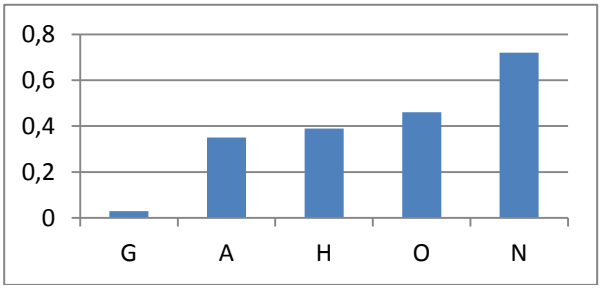
| **Značilka** | **Vrednost** |
|---|---|
| Razmerje št. besed/št. povedi | 21,24 |
| Razmerje št. znakov/št. besed | 5,14 |
| Indeks ARI | 13,38 |
| Formula Gunning Fog | 21,81 |

Tabela 2: Berljivostne značilke anonimnega besedila.

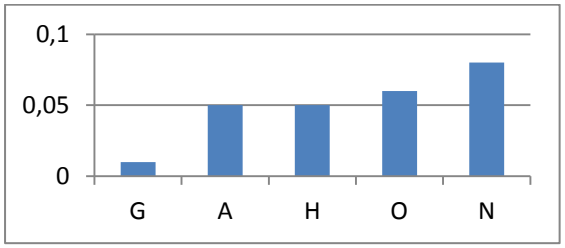
Na podlagi izračunanih formul smo opazovali povprečno absolutno razliko med vrednostmi značilk znanih avtorjev in anonimnega besedila.
V nadaljevanju (grafi 1 do 8) predstavljamo rezultate razvrščanj glede na upoštevane značilke. Pri vsaki značilki predstavljamo prvih pet avtorjev z najmanjšo povprečno absolutno razliko glede na anonimno besedilo (najmanjša povprečna absolutna razlika pomeni največjo podobnost z anonimnim besedilom).




Graf 1. Besedišče.
Graf 1 prikazuje razvrstitev avtorjev glede na razmerje med številom različnih lem in celokupnim številom lem v besedilu. Najbliže anonimnemu besedilu so vrednosti avtorja G.

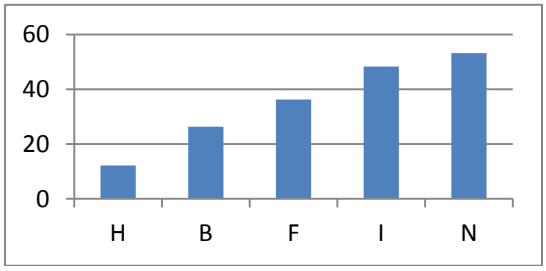



Graf 2. Brunet.
Graf 2 predstavlja avtorje z najmanjšo povprečno absolutno razliko glede na Brunetovo formulo, ki računa raznolikost besedišča glede na dolžino besedila.



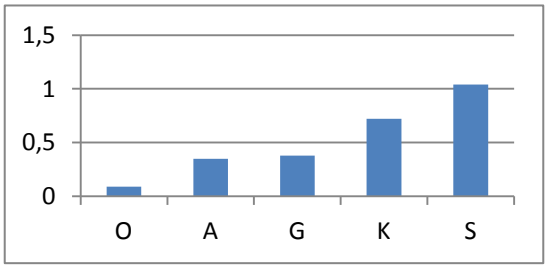

Graf 3. Hapax legomena

Graf 3 predstavlja avtorje, razvrščene glede na število lem, ki se pojavijo samo enkrat v besedilu. Tudi po tem kriteriju se avtor G najmanj razlikuje od anonimnega besedila.



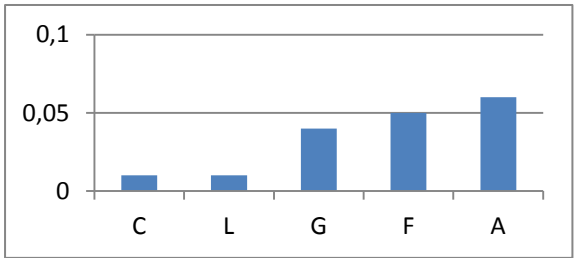

Graf 4. Honoré.

Graf 4 razvršča po formuli Honoré (Honoré 1979), ki računa razmerje med številom hapaksov in raznolikostjo besedišča. Najmanjšo povprečno absolutno razliko ima avtor H.



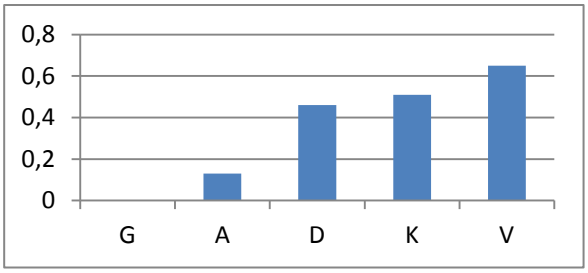

Graf 5. Število besed/število povedi.

Razmerje med številom besed in številom povedi, ki ga prikazuje graf 5, je pogosto uporabljan kriterij za določanje stopnje berljivosti besedila.



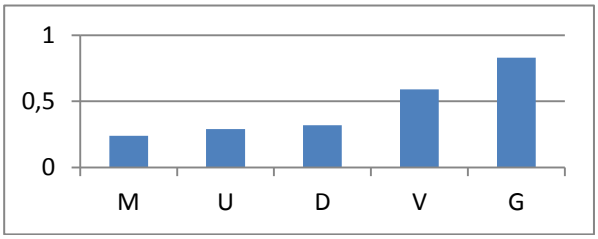

Graf 6. Število znakov/število besed.

Graf 6 izpostavlja razmerje med številom znakov in številom besed v analiziranih besedilih. Po tem kriteriju ima najmanjšo povprečno absolutno razliko od anonimnega besedila avtor C.




Graf 7. ARI.

Formula ARI (Automated Readability Index) izračuna okvirno stopnjo izobrazbe, ki jo zahteva neko besedilo za razumevanje ob prvem branju. Ta značilka izpostavlja avtorja G, ki po tem kriteriju dosega identične vrednosti kot anonimno besedilo.

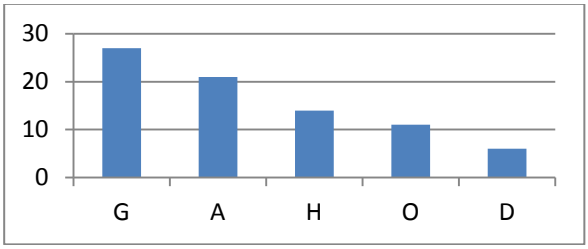



Graf 8. Gunning Fog.

Formula Gunning Fog predstavlja nekoliko drugačen izračun stopnje izobrazbe, ki jo zahteva neko besedilo, da ga bralec razume po prvem branju. Glede na ta kriterij kaže največjo podobnost z anonimnim besedilom avtor M.

## 6. Sinteza in omejitve raziskave

Če vsakemu od prvih petih avtorjev z najmanjšo povprečno absolutno razliko glede na anonimno besedilo pripišemo od 1 do 5 točk, dobimo naslednjo razvrstitev:




Graf 9: Stopnja podobnosti avtorjev z anonimnim besedilom.

Raziskava je nastala v okviru avtentične situacije objave anonimnega besedila, ki je vznemirila slovensko javnost. Za potrebe raziskave smo postavili hipotezo, da je avtor na isti spletni strani verjetno objavil še kakšno drugo besedilo in ga podpisal s svojim pravim imenom. Hipoteza je verjetno utemeljena, vendar raziskava:

- ponuja premajhen nabor besedil, da bi lahko izvedli evalvacijo modela (Argamon, Levitan 2005),

- ne ponudi odgovora na vprašanje, ali je bil dejanski avtor besedila sploh vključen v analizo.

## 7. Zaključek

V prispevku poskušamo identificirati avtorja anonimnega besedila, ki je v slovenskih medijih sprožil številne odzive. Analiza zajema 75 besedil 21 znanih avtorjev in vključuje razvrščanje na podlagi leksikalnih in berljivostnih značilk.

Rezultati kažejo, da je med 21 potencialnimi avtorji glede na upoštevane kriterije razvrščanja najverjetnejši avtor anonimnega besedila avtor G. Vendar zaradi majhnega števila analiziranih besedil ne moremo izvesti evalvacije razvrščanja in preverjanja hipoteze.

Analiza bi pridobila na kredibilnosti, če bi lahko analizirali večje število besedil, potrdili razlikovalno moč upoštevanih značilk na večji bazi besedil ali dopolnili metodo s kvalitativno analizo diskurza.

**Primarna vira**

Besedilo *Volivci v trenirkah*:
http://www.delo.si/assets/media/other/20111211//Prispevek%20Toma%C5%BEa%20Majerja.pdf
Spletni arhiv stranke:
http://www.sds.si/arhiv?id=12

**Bibliografija**

S. Argamon, S. Levitan. 2005. Measuring the usefulness of function words for authorship attribution. P*roceedings of ACH/ALLC 2005*.

attribution. In Proceedings of the Pacific

R. Baayen, H. van Halteren, F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11/3, str. 121–131.

E. Brunet. 1988. Une mesure de la distance intertextuelle : la connexion lexicale. *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines*, 24/1, str. 81–116.

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

M. Coulthard. 2005. The linguist as expert witness. *Linguistics & the Human Sciences*, 1/1, str. 39–58.

K. Dobrovoljc, S. Krek, J. Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. *Zbornik Osme konference Jezikovne tehnologije*, str. 42-47.

M. Grčar S. Krek, K. Dobrovoljc. 2012. Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene. *Zbornik Osme konference Jezikovne tehnologije*, str. 42–47.

R. Gunning. 1952. *The Technique of Clear Writing*. New York: McGraw-Hill.

G. Hirst, O. Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22/4, str. 405– 417.

D. I. Holmes. 1992. A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society*, 155/1. str. 91-120.

A. Honoré. 1979. Some Simple Measures of Richness of Vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 7, str. 172–177.

V. Keselj, F. Peng, N. Cercone, C. Thomas. 2003.. N-gram-based author profiles for authorship attribution. *Proceedings of the Pacific Association for Computational Linguistics*, str. 255-264.

M. Koppel, J. Schler, S. Argamon, E. Messeri. 2006. Authorship attribution with thousands of candidate authors. *Proceedings of the 29th ACM SIGIR*, str. 659-660.

K. Luyckx, W. Daelemans. 2005. Shallow text analysis and machine learning for authorship attribution. *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands*. str. 149–160.

Y. Marton, N. Wu, L. Hellerstein. 2005. On compression-based text classification. *Proceedings of the European Conference on Information Retrieval*, str. 300–314.

P. M. McCarthy, G. A. Lewis, D. F. Dufty, D. S. McNamara. 2006. Analyzing writing styles with coh-metrix. *Proceedings of the Florida Artificial Intelligence Research Society International Conference*, 764-769.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*,34/1, str. 1-47

E. Stamatatos. 2006. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15/5, str. 823-838.

E. Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management, 44/2, str. 790-799.

E. Stamatatos, N. Fakotakis, G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35/2, str. 193-214.

E. Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*. 60/3, str. 538–556.

R. Zheng, J. Li, H. Chen, Z. Huang. 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, str. 378-393.

A. Zwitter Vitez. 2012. Authorship Attribution: Specifics for Slovene. *Slavia Centralis* 5/1, str. 75–85.