

Oznaka poročila: ARRS-RPROJ-ZP-2015/199



ZAKLJUČNO POROČILO RAZISKOVALNEGA PROJEKTA

A. PODATKI O RAZISKOVALNEM PROJEKTU

1. Osnovni podatki o raziskovalnem projektu

Šifra projekta	Z7-4083
Naslov projekta	Jezikovnotehnoške analize za ugotavljanje osebnega profila avtorja besedila
Vodja projekta	24440 Ana Zwitter Vitez
Tip projekta	Zt Podoktorski projekt - temeljni
Obseg raziskovalnih ur	3400
Cenovni razred	A
Trajanje projekta	05.2013 - 07.2014
Nosilna raziskovalna organizacija	2923 Trojina, zavod za uporabno slovenistiko
Raziskovalne organizacije - soizvajalke	
Raziskovalno področje po šifrantu ARRS	7 INTERDISCIPLINARNE RAZISKAVE
Družbeno-ekonomski cilj	13.06 Humanistične vede - RiR financiran iz drugih virov (ne iz SUF)
Raziskovalno področje po šifrantu FOS	6 Humanistične vede 6.02 Jeziki in književnost

B. REZULTATI IN DOSEŽKI RAZISKOVALNEGA PROJEKTA

2. Povzetek raziskovalnega projekta¹

SLO

Področje ugotavljanja avtorstva besedil je v zadnjih dveh desetletjih doživelo silovit razmah na področjih prava in avtorskih pravic (Grant 2007), literarnih ved (Hoover 2004), kriminalističnih preiskav (Coulthard 2005) in profiliranja strank v komercialne namene (Shaw et al. 2001).

Cilj projekta Jezikovnotehnološke analize za ugotavljanje osebnega profila avtorja besedila je bil pridobiti znanje, prek katerega je mogoče ugotoviti najverjetnejšega avtorja neznanega besedila ali opredeliti njegov osebni profil (spol, starost, izobrazba, regionalna pripadnost). To znanje smo pridobili z naslednjo metodologijo:

- priprava besedil in izdelava referenčne baze spletnih besedil (55 avtorjev, 30.000 besedil, 1,2 milijona pojavnic)
- izračun leksikalnih in berljivostnih značilk za ugotavljanje avtorstva in določanje osebnega profila avtorja besedila,
- razvrščanje besedil (večinski klasifikator (Majority), podporni vektorji (SVM), najbližji centroid (NCC)),
- evalvacija (programa za strojno učenje Orange in Weka),
- izgradnja modela za ugotavljanje avtorstva in določanje osebnega profila avtorja besedila.

Analiza je pokazala, da pri ugotavljanju avtorstva besedil v slovenščini najbolje delujeta klasifikacija z metodo podpornih vektorjev (SVM) in klasifikacija z najbližjim centroidom (NCC) z značilkama, ki izračunata delež kompleksnih besed (rComplex) in stopnjo izobrazbe, potrebne za razumevanje besedila ob prvem branju (ARI). Rezultate analize smo uporabili na primeru avtentičnega anonimnega besedila in ugotovili najverjetnejšega avtorja. Rezultate smo posredovali pooblaščenim osebam pri Nacionalnem forenzičnem laboratoriju, ki se ukvarja z analizo anonimnih besedil.

ANG

Authorship attribution has developed immensely in the last two decades, especially in the fields of authorship law and copyright (Grant, 2007), literary studies (Hoover, 2004), criminology (Coulthard, 2005) and customer profiling for commercial purposes (Shaw et al., 2001).

The aim of the project Language technologies for determining the author's personal profile was to acquire knowledge that enables to detect an author of an anonymous text or to determine his profile (gender, age, education, region). This knowledge was gained using the following methodology:

- design and building a reference database of web texts (55 authors, 30,000 texts, 1,2 mio tokens),
- calculation of lexical and readability features,
- classification (Majority, Support Vector Machine (SVM), Nearest Cluster Centroid (NCC)),
- feature evaluation (programs for machine learning Orange and Weka),
- model for authorship attribution and author profiling.

The analysis showed that the best results for authorship attribution on the reference corpus are achieved when using classification with support vector machines (SVM) and the nearest centroids (NCC) with features relying on the proportion of complex words (rComplex) and on the readability formula ARI calculating the level of education necessary to understand a text at first reading. The analysis was also applied to detection of an authentic anonymous text and enabled to point out its most likely author. The results were submitted to the authorized person at the National Forensic Laboratory, which deals with anonymous text analysis.

3. Poročilo o realizaciji predloženega programa dela na raziskovalnem projektu²

1 Raziskovalna hipoteza

Izhodiščna hipoteza projekta Jezikovnotehnološke analize za ugotavljanje osebnega profila avtorja besedila je zasnovana na dejstvu, da je s pomočjo kakovostno zgrajenega in označenega korpusa besedil mogoče ugotoviti jezikovne parametre za slovenščino, s katerimi je mogoče kvantificirati avtorjeve jezikovne sledi v besedilu.

2 Potek dela

2.1 Gradnja baze

V prvem delovnem sklopu smo zajeli besedila, pridobili podatke o avtorjih in izgradili jezikoslovno označeno bazo besedil:

- zajem besedil (blogi po kriteriju branosti),
- čiščenje besedil,

- poenotenje metapodatkov o besedilu in avtorju (zvrst in leto zajema, spol, starost, regija, stopnja izobrazbe),
- pretvorba besedil v format xml,
- oblikoslovno označevanje in skladiščno razčlenjevanje.

Rezultat prve faze raziskave je označena referenčna baza blogovskih besedil s pripadajočimi podatki o osebnem profilu avtorja besedila (spol, starost, izobrazba, regionalna pripadnost). Baza zajema 55 avtorjev in približno 30.000 besedil oz. 1,2 mio pojavnic)

2.2 Izračun značilke

Za označeni besedilni korpus smo izračunali značilke, specifične za vsakega od obravnavanih avtorjev.

Pri izračunu značilk smo upoštevali naslednje kategorije:

Leksikalne značilke: raznolikost besedišča, Brunetova formula, relativna frekvenca hapaksov v besedilu, Honoréjeva formula.

Berljivostne značilke: razmerje med številom besed in številom povedi, razmerje med številom znakov in številom besed, formula (Automated Readability index), formula Gunning Fog.

2.3 Razvrščanje in evalvacija značilk

Na podlagi izračunanih vrednosti značilk smo izvedli razvrščanje besedil z različnimi klasifikatorji

- večinski klasifikator (Majority)
- podporni vektorji (SVM)
- najbližji centroid (NCC)

Nato smo klasifikatorje ocenili z različnimi metrikami, implementiranimi v programa za strojno učenje Orange in Weka. Rezultati analize so pokazali, da za slovenska besedila najbolje delujeta klasifikacija z metodo podpornih vektorjev (SVM) in klasifikacija z najbližjim centroidom (NCC) z značilkama, ki izračunata delež kompleksnih besed (rComplex) in stopnjo izobrazbe, potrebne za razumevanje besedila ob prvem branju (ARI).

3. Ključni rezultati

Ključne ugotovitve in znanstvena spoznanja: Rezultati analize so pokazali, da pri ugotavljanju avtorstva besedil v slovenščini najbolje delujeta klasifikacija z metodo podpornih vektorjev (SVM) in klasifikacija z najbližjim centroidom (NCC) z značilkama, ki izračunata delež kompleksnih besed (rComplex) in stopnjo izobrazbe, potrebne za razumevanje besedila ob prvem branju (ARI). Rezultati, vezani na osebni profil avtorja besedila, še niso dovolj zanesljivi za širšo uporabo.

4. Učinki in uporaba rezultatov

Rezultate smo posredovali pooblaščenim osebam pri Nacionalnem forenzičnem laboratoriju, ki se ukvarja z analizo anonimnih besedil. V sodelovanju z omenjeno institucijo smo izvedli kontrastivno analizo statističnih metod in metode analize diskurza za doseganje optimalnih rezultatov pri ugotavljanju avtorstva besedila.

Poleg predvidenih aktivnosti smo na podlagi zaostrenih družbenopolitičnih okoliščin izvedli dodatno raziskavo s področja sovražnega govora. Gre za ugotavljanje avtorja anonimnega besedila, proti kateremu je informacijska pooblaščenka leta 2011 podala kazensko ovadbo, javnost pa ni nikoli izvedela prave identitete avtorja spornega besedila. Analizirali smo 75 besedil 21 podpisanih avtorjev, ki so bila na isti spletni strani objavljena tri mesece pred in tri mesece po objavi anonimnega besedila. Rezultati so pokazali, da je glede na klasifikacijske kriterije anonimnemu avtorju najbližje avtor G glede na značilke z največjo razlikovalno močjo (raznolikost besedišča, relativna frekvenca hapaksov v besedilu, Brunetova formula).

4. Ocena stopnje realizacije programa dela na raziskovalnem projektu in zastavljenih raziskovalnih ciljev³

Izhodiščna hipoteza: s pomočjo kakovostno zgrajenega in označenega korpusa besedil je mogoče ugotoviti določujoče jezikovne parametre za slovenščino, s katerimi lahko kvantificiramo avtorjeve jezikovne sledi v besedilu.

Ocena stopnje realizacije raziskovalne hipoteze: rezultati analize so pokazali, da je slog posameznega avtorja dejansko mogoče kvantificirati s pomočjo izračuna in evalvacije

leksikalnih in berljivostnih značilk. Pri ugotavljanju avtorstva besedil najbolje delujeta klasifikacija z metodo podpornih vektorjev (SVM) in klasifikacija z najbližjim centroidom (NCC) z značilkama, ki izračunata delež kompleksnih besed (rComplex) in stopnjo izobrazbe, potrebne za razumevanje besedila ob prvem branju (ARI).

5. Utemeljitev morebitnih sprememb programa raziskovalnega projekta oziroma sprememb, povečanja ali zmanjšanja sestave projektne skupine⁴

Načrt raziskovalnega projekta smo uresničili.

6. Najpomembnejši znanstveni rezultati projektne skupine⁵

Znanstveni dosežek			
1.	COBISS ID	51943522	Vir: COBISS.SI
	Naslov	SLO	Iskanje anonimnega avtorja
		ANG	Le décryptage de l'auteur anonyme
	Opis	SLO	V prispevku analiziramo anonimno besedilo in ga primerjamo s 75 besedili 21 znanih avtorjev. Analiza temelji na metodi podpornih vektorjev (SVM), ki omogoča določanje razlik in podobnosti med primerjanimi besedili na podlagi značilk besedišča in berljivosti. Rezultati kažejo, da so specifične enega izmed opazovanih avtorjev precej podobne besedilu neznanega izvora glede na raznolikost besedišča, Brunetovo formulo in relativno frekvenco hapaksov v besedilu.
		ANG	In this paper we compare an anonymous text to 75 texts written by 21 known authors. The analysis is based on support vector machines (SVM), which allows to detect differences and similarities between the compared texts on the basis of lexical and readability features. The results show that one of the authors properties resemble significantly to the anonymous text, especially according to lexical diversity, Brunet formula and hapax relative frequency.
	Objavljeno v	Znanstvena založba Filozofske fakultete; Linguistica; 2013; Letn. 53, št. 1; str. 91-101; Avtorji / Authors: Zwitter Vitez Ana	
Tipologija		1.01 Izvirni znanstveni članek	
2.	COBISS ID	56746594	Vir: COBISS.SI
	Naslov	SLO	Stateška sredstva za prevzem besede v francoščini in slovenščini
		ANG	Les moyens stratégiques pour prendre la parole en français et en slovène
	Opis	SLO	Cilj raziskave je opredeliti najučinkovitejša strateška sredstva v francoščini in slovenščini, ki jih uporabljajo govorci, ko se borijo za prevzem besede. Raziskava sloni na treh spontanah govornih izmenjavah, ki smo jih analizirali na prozodični, oblikoskladenjski in diskurzivni ravni. Rezultati kažejo, da so govorci, ki so bili bolj učinkoviti pri prevzemanju besede, tvorili številne premore, ponavljanja in samopopravke, pa tudi bistveno daljše diskurzivne preambule kot govorci, ki so besedo prepustili.
		ANG	The main challenge of the paper was to identify the most effective strategic structures when two speakers want to speak at the same time in French and in Slovene. The study is based on three spontaneous conversations which were analyzed on prosodic, morphosyntactic and discourse level. The results show that speakers who were efficient in taking their turn produced many pauses, repetitions and auto-corrections as well as longer discourse preambles than speakers who were less successful in taking (or keeping) their turns.
	Objavljeno v	Znanstvena založba Filozofske fakultete; Vestnik za tuje jezike; 2014; Letn. 6, št. 1; str. 63-77; Avtorji / Authors: Zwitter Vitez Ana	

	Tipologija	1.01 Izvirni znanstveni članek	
3.	COBISS ID	16771606	Vir: COBISS.SI
	Naslov	SLO	Gradnja, transkribiranje in uporaba referenčnega govornega korpusa
		ANG	Compilation, transcription and usage of a reference speech corpus
	Opis	SLO	V članku predstavimo postopek izgradnje referenčnega govornega korpusa in njegovo uporabo na primeru korpusa govorne slovenščine Gos. Osredotočimo se na zgradbo korpusa, postopek snemanja, označevanja posnetkov in transkribiranja na dveh ravneh (pogovorni in standardizirani zapis) ter predstavimo spletni vmesnik (konkordančnik) in dostopnost izvirnih korpusnih gradiv.
		ANG	This paper describes design and compilation of a reference speech corpus and its distribution to potential users, as it was done in the case of the Slovene corpus GOS. The corpus structure and experiences with recording, labelling system, and two levels of transcription (pronunciationbased and standardized) are described, as well as the main characteristics of the corpus interface (web concordancer) and the availability of the original corpus files.
	Objavljeno v	Springer; Language resources and evaluation; 2013; Vol. 47, iss. 4; str. 1031-1048; Impact Factor: 0.518; Srednja vrednost revije / Medium Category Impact Factor: 1.694; A'': 1; A': 1; WoS: EV; Avtorji / Authors: Verdonik Darinka, Kosem Iztok, Zwitter Vitez Ana, Krek Simon, Stabej Marko	
	Tipologija	1.01 Izvirni znanstveni članek	
4.	COBISS ID	55987554	Vir: COBISS.SI
	Naslov	SLO	Ugotavljanje avtorstva besedil: primer "Trenirkarjev"
		ANG	Authorship Attribution: the example of 'Voters in Sportsuits'
	Opis	SLO	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.
		ANG	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.
	Objavljeno v	Institut Jožef Stefan; Jezikovne tehnologije; 2014; Str. 131-134; Avtorji / Authors: Zwitter Vitez Ana	
	Tipologija	1.08 Objavljeni znanstveni prispevek na konferenci	
5.	COBISS ID	17960982	Vir: COBISS.SI
	Naslov	SLO	Baza Broadcast News in Korpus govorne slovenščine Gos: iskanje enotnih smernic za nadaljnje delo
		ANG	The Slovene BNSI broadcast news database and reference speech corpus GOS: Towards the uniform guidelines for future work
	Opis	SLO	Cilj prispevka je raziskati značilnosti obstoječih govornih baz v funkciji potreb jezikoslovnih raziskav in jezikovnih tehnologij, ki predstavljajo ključni področji njihove uporabe. V tem namen primerjamo bazo za

		avtomatsko razpoznavanje govora BNSI Broadcast News in referenčni Korpus govorne slovenščine Gos. Na koncu postavimo enotne smernice za nadaljnje delo.
	ANG	The aim of the paper is to search for common guidelines for the future development of speech databases in order to make them the most useful for both main fields of their use, linguistic research and speech technologies. We compare the Slovene speech database for automatic speech recognition – BNSI Broadcast News, and the Slovene reference speech corpus GOS, and outline possible common guidelines for future work.
Objavljeno v	ELRA; LREC 2014; 2014; Str. 2644-2647; Avtorji / Authors: Žgank Andrej, Zwitter Vitez Ana, Verdonik Darinka	
Tipologija	1.08 Objavljeni znanstveni prispevek na konferenci	

7. Najpomembnejši družbeno-ekonomski rezultati projektne skupine⁶

	Družbeno-ekonomski dosežek		
1.	COBISS ID	53180258	Vir: COBISS.SI
	Naslov	SLO	Jezikoslovje in sovražni govor
		ANG	Linguists against the hostile discourse
	Opis	SLO	Če raziskava s področja ugotavljanja avtorstva besedil razkrije avtorja anonimnega besedila, ima lahko pomembne družbene posledice. Tak položaj odpira vprašanja, o katerih razpravljamo v prispevku: Ima raziskovalec pravico iskati avtorja anonimnega besedila? Bi moral glede na rezultate razkriti njegovo identiteto? Kje je meja med sovražnim govorom in pravico do svobode govora?
		ANG	When a research points out an anonymous author of an authentic text , these results can initiate important social consequences. This position opens several questions discussed in the paper: Does a researcher have the right to investigate authorship of an anonymous text? Is he obliged to reveal the author's identity? Where is the limit between hate speech and the liberty of speech?
	Šifra	B.03 Referat na mednarodni znanstveni konferenci	
	Objavljeno v	Institute of discourse and cultural studies; Center for contemporary chinese discourse studies; Zhejiang university; The fourth international conference on multicultural discourses. Hangzhou, China, october 24-26 2013; 2013; Str. 119; Avtorji / Authors: Zwitter Vitez Ana	
	Tipologija	1.12 Objavljeni povzetek znanstvenega prispevka na konferenci	
2.	COBISS ID	35723053	Vir: COBISS.SI
	Naslov	SLO	Osebnostne značilnosti v jezikovnih sledih
		ANG	Personal profiling through linguistic fingerprints
	Opis	SLO	Prenos metodologije ugotavljanja avtorstva besedil na primeru iz prakse: RTVSLO (http://ars.rtv slo.si/2013/02/osebnostne-znacilnosti-v-jezikovnih-sledih/)
		ANG	Transfer of the methodology of authorship attribution into practice: RTVSLO (http://ars.rtv slo.si/2013/02/osebnostne-znacilnosti-v-jezikovnih-sledih/)
	Šifra	F.17 Prenos obstoječih tehnologij, znanj, metod in postopkov v prakso	
	Objavljeno v	2013; Avtorji / Authors: Zwitter Vitez Ana	

	Tipologija	3.11 Radijski ali TV dogodek	
3.	COBISS ID	56964962	Vir: COBISS.SI
	Naslov	SLO	Jezikovne prakse "Novih govorcev": empirične raziskave
		ANG	Towards an empirical analyses of New Speakers' language practices
	Opis	SLO	Ana Zwitter Vitez je predstavnica Slovenije v odboru evropske Akcije Cost IS1306 New speakers in a Multilingual Europe. Osrednji cilj je preučiti izzive in priložnosti evropskega konteksta večjezičnosti, povezane s pritokom in percepcijo novih govorcev, ki se v svojem vsakodnevem družbenem delovanju soočajo z uporabo nematerne ali nenacionalnega jezika.
		ANG	Ana Zwitter Vitez is MC member in european Cost Action IS1306 New speakers in a Multilingual Europe. The aim of the Action is to investigate the challenges and the opportunities of multilingual speakers in the role of immigrants, transnational workers or within regional minorities.
	Šifra	D.03 Članstvo v tujih/mednarodnih odborih/komitejih	
	Objavljeno v	Universitat oberta de Catalunya; Second International Symposium on New Speakers in a Multilingual Europe; 2014; Str. 18-19; Avtorji / Authors: Arhar Holdt Špela, Zwitter Vitez Ana	
	Tipologija	1.12 Objavljeni povzetek znanstvenega prispevka na konferenci	
4.	COBISS ID	51943522	Vir: COBISS.SI
	Naslov	SLO	Ugotavljanje avtorstva besedil
		ANG	Le décryptage de l'auteur anonyme
	Opis	SLO	Korpus blogovskih besedil s pripadajočimi podatki o osebnem profilu avtorja besedila (spol, starost, izobrazba, regionalna pripadnost). Baza zajema 55 avtorjev in približno 30.000 besedil oz. 1,2 mio pojavnic.
		ANG	The Blog Corpus contains metadata about the authors' personal profile (gender, age, education, region). The database consists of 55 authors and 30.000 texts (1,2 mio words).
	Šifra	F.15 Razvoj novega informacijskega sistema/podatkovnih baz	
	Objavljeno v	Znanstvena založba Filozofske fakultete; Linguistica; 2013; Letn. 53, št. 1; str. 91-101; Avtorji / Authors: Zwitter Vitez Ana	
	Tipologija	1.01 Izvirni znanstveni članek	
5.	COBISS ID	55987554	Vir: COBISS.SI
	Naslov	SLO	Ugotavljanje avtorstva besedil: primer "Trenirkarjev"
		ANG	Authorship attribution: the example of 'Voters in Sportsuits'
	Opis	SLO	Korpus avtentičnih besedil potencialnih avtorjev anonimnega besedila, ki je leta 2011 vznemirilo slovensko javnost (21 avtorjev, 75 besedil, 55.000 pojavnic)
		ANG	Anonymous text corpus that provoked vivid public reactions in slovenan media in 2011. The database consists of 21 authors and 75 texts (55.000 words).
	Šifra	F.15 Razvoj novega informacijskega sistema/podatkovnih baz	
	Objavljeno v	Institut Jožef Stefan; Jezikovne tehnologije; 2014; Str. 131-134; Avtorji / Authors: Zwitter Vitez Ana	
	Tipologija	1.08 Objavljeni znanstveni prispevek na konferenci	

8. Drugi pomembni rezultati projektne skupine²

Poljudni nastopi:

ZWITTER VITEZ, Ana (intervjuvanec). Osebnostne značilnosti v jezikovnih sledih : Radio Slovenija : III. program : Ars program : oddaja Jezikovni pogovori. Ljubljana, 19. 2. 2013.

ZWITTER VITEZ, Ana (intervjuvanec) Jezik mladih: Radio Slovenija : Val 202. Ljubljana, 4.1.2015

ZWITTER VITEZ, Ana (intervjuvanec), VERDONIK, Darinka (intervjuvanec). Kakšno slovenščino govorimo? : o korpusu GOS : Radio Slovenija : III. program : Ars program : oddaja Jezikovni pogovori. Ljubljana, 23. 4. 2013.

ZWITTER VITEZ, Ana (predavatelj). Jezikoslovje in ugotavljanje avtorstva besedil. Lingvistični krožek, Filozofska fakulteta. Ljubljana, 10. 12. 2012.

ZWITTER VITEZ, Ana (predavatelj). Jezikovna forenzika: na sledi anonimnemu piscu besedila. Univerza na Primorskem, Fakulteta za humanistične študije. Koper, 1. 4. 2015

9.Pomen raziskovalnih rezultatov projektne skupine⁸

9.1.Pomen za razvoj znanosti⁹

SLO

Rezultati projekta so omogočili razvoj področja ugotavljanja avtorstva besedil na naslednjih ravneh:

- metodologija: raziskava je razvila interdisciplinarni dialog med področji jezikoslovja (gradnja in analiza korpusov), računalništva (metode strojnega učenja) in kriminologije (jezikovna forenzika),
- rezultati: za nadaljnje študije so na voljo korpus z označenimi lastnostmi avtorjev, izračunane vrednosti značilk in model za ugotavljanje avtorstva besedil,
- terminologija: objave v domačih in mednarodnih publikacijah so prispevale k poenotenju terminologije na področju ugotavljanja avtorstva besedil,
- prenos znanja: rezultati raziskave so bili vključeni v študijski proces, v poljudni obliki pa so bili predstavljeni tudi v medijih.

ANG

The results of the project have contributed to the field of authorship attribution at the following levels:

- methodology: the research has been conducted in an interdisciplinary dialogue between linguistics (corpus design and analysis), computer science (machine learning), and criminology (forensic linguistics),
- results: for further studies, corpus with authors' profile metadata, feature values and classification results are available,
- terminology: national and international publications have contributed to the unification of terminology in the field of authorship attribution,
- transfer of knowledge: results of the survey were incorporated into the study process and presented in the media.

9.2. Pomen za razvoj Slovenije¹⁰

SLO

Rezultati raziskave prispevajo k uresničevanju dveh strateških dokumentov:

- evropska iniciativa Digital Agenda for Europe, ki spodbuja izrabo digitalnih tehnologij,
- Resolucija o raziskovalni in inovacijski strategiji Slovenije 2011–2020 (3.2 Prenos znanja, 4.3 Razvoj raziskovalne infrastrukture, 4.5 Informacijska infrastruktura v podporo inovacijskemu sistemu).

Raziskava dopolnjuje kvantitativno analizo vsakodnevne jezikovne produkcije in izpostavlja uporabno vrednost tovrstnih analiz, zato so rezultati uporabni tako za znanstvene discipline (jezikoslovje, računalništvo, kriminologija) kot za naslednja področja:

- gospodarstvo (tržne analize): metodologijo raziskave lahko prilagodimo potrebam podjetij, ki na podlagi jezikovne produkcije strank prilagajajo strategije oglaševanja in razvoja produktov (Shaw et al., 2001),
- upravljanje s človeškimi viri: poznavanje jezikovnih parametrov, odločilnih za profiliranje avtorjev, bo v večjih podjetjih omogočilo izbor ustreznih kandidatov (Schuler et al. 1999),

- državni organi: na področju kriminalističnega preiskovanja lahko z izračunom in evalvacijo jezikovnih značilk določimo verjetnost, da je neznano besedilo napisal eden izmed znanih avtorjev,
- kulturna dediščina: nastala baza besedil z označenimi lastnostmi avtorjev omogoča nadaljnje analize avtorstva besedil (ob skrbnem varovanju osebnih podatkov in avtorskih pravic),
- medkulturni dialog: poznavanje razsežnosti analize avtentične jezikovne produkcije lahko izboljša razumevanje kompleksnosti družbenih odnosov in konfliktov v vsakdanjem življenju.

ANG

The research results contribute to the realization of two strategic documents:

- European initiative Digital Agenda for Europe, which promotes the use of digital technologies,
- Resolution on Research and Innovation Strategy of Slovenia 2011-2020 (3.2 Transfer of knowledge, 4.3 Research infrastructure development 4.5 Information infrastructure supporting innovation system).

The research complements the quantitative analysis of everyday language production and highlights the importance of applied linguistic analyses. The results can be exploited in different scientific disciplines (linguistics, informatics, criminology) and in the following fields:

- economy (market analysis): the methodology can be adapted to the needs of enterprises developing advertising strategies and product development on the basis of clients' language production (Shaw et al., 2001),
- human resource management: linguistic features for author profiling allow the selection of suitable candidates in large companies (Schuler et al. 1999),
- state authorities: in the field of criminal investigation, calculation and evaluation of lexical and readability features enables to detect whether one of the possible authors has written an anonymous text,
- cultural heritage: the resulting corpus with annotated metadata facilitates further analyses on authorship attribution and author profiling (with a careful protection of copyright and personal data)
- intercultural dialogue: awareness of the power of authentic language production analysis can improve the understanding of social relations and conflicts in everyday life.

10. Samo za aplikativne projekte in podoktorske projekte iz gospodarstva!

Označite, katerega od navedenih ciljev ste si zastavili pri projektu, katere konkretne rezultate ste dosegli in v kakšni meri so doseženi rezultati uporabljeni

Cilj		
F.01	Pridobitev novih praktičnih znanj, informacij in veščin	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.02	Pridobitev novih znanstvenih spoznanj	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.03	Večja usposobljenost raziskovalno-razvojnega osebja	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.04	Dvig tehnološke ravni	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	

	Uporaba rezultatov	
F.05	Sposobnost za začetek novega tehnološkega razvoja	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.06	Razvoj novega izdelka	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.07	Izboljšanje obstoječega izdelka	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.08	Razvoj in izdelava prototipa	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.09	Razvoj novega tehnološkega procesa oz. tehnologije	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.10	Izboljšanje obstoječega tehnološkega procesa oz. tehnologije	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.11	Razvoj nove storitve	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.12	Izboljšanje obstoječe storitve	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.13	Razvoj novih proizvodnih metod in instrumentov oz. proizvodnih procesov	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.14	Izboljšanje obstoječih proizvodnih metod in instrumentov oz. proizvodnih	

	procesov	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.15	Razvoj novega informacijskega sistema/podatkovnih baz	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.16	Izboljšanje obstoječega informacijskega sistema/podatkovnih baz	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.17	Prenos obstoječih tehnologij, znanj, metod in postopkov v prakso	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.18	Posredovanje novih znanj neposrednim uporabnikom (seminarji, forumi, konference)	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.19	Znanje, ki vodi k ustanovitvi novega podjetja ("spin off")	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.20	Ustanovitev novega podjetja ("spin off")	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.21	Razvoj novih zdravstvenih/diagnostičnih metod/postopkov	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.22	Izboljšanje obstoječih zdravstvenih/diagnostičnih metod/postopkov	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.23	Razvoj novih sistemskih, normativnih, programskih in metodoloških rešitev	

	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.24	Izboljšanje obstoječih sistemskih, normativnih, programskih in metodoloških rešitev	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.25	Razvoj novih organizacijskih in upravljavskih rešitev	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.26	Izboljšanje obstoječih organizacijskih in upravljavskih rešitev	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.27	Prispevek k ohranjanju/varovanje naravne in kulturne dediščine	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.28	Priprava/organizacija razstave	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.29	Prispevek k razvoju nacionalne kulturne identitete	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.30	Strokovna ocena stanja	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.31	Razvoj standardov	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.32	Mednarodni patent	
	Zastavljen cilj	☐ DA ☐ NE

	Rezultat	
	Uporaba rezultatov	
F.33	Patent v Sloveniji	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.34	Svetovalna dejavnost	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	
F.35	Drugo	
	Zastavljen cilj	☐ DA ☐ NE
	Rezultat	
	Uporaba rezultatov	

Komentar

--

11.Samo za aplikativne projekte in podoktorske projekte iz gospodarstva!**Označite potencialne vplive oziroma učinke vaših rezultatov na navedena področja**

	Vpliv	Ni vpliva	Majhen vpliv	Srednji vpliv	Velik vpliv	
G.01	Razvoj visokošolskega izobraževanja					
G.01.01.	Razvoj dodiplomskega izobraževanja	☐	☐	☐	☐	
G.01.02.	Razvoj podiplomskega izobraževanja	☐	☐	☐	☐	
G.01.03.	Drugo:	☐	☐	☐	☐	
G.02	Gospodarski razvoj					
G.02.01	Razširitev ponudbe novih izdelkov/storitev na trgu	☐	☐	☐	☐	
G.02.02.	Širitev obstoječih trgov	☐	☐	☐	☐	
G.02.03.	Znižanje stroškov proizvodnje	☐	☐	☐	☐	
G.02.04.	Zmanjšanje porabe materialov in energije	☐	☐	☐	☐	
G.02.05.	Razširitev področja dejavnosti	☐	☐	☐	☐	
G.02.06.	Večja konkurenčna sposobnost	☐	☐	☐	☐	
G.02.07.	Večji delež izvoza	☐	☐	☐	☐	
G.02.08.	Povečanje dobička	☐	☐	☐	☐	
G.02.09.	Nova delovna mesta	☐	☐	☐	☐	
G.02.10.	Dvig izobrazbene strukture zaposlenih	☐	☐	☐	☐	
G.02.11.	Nov investicijski zagon	☐	☐	☐	☐	
G.02.12.	Drugo:	☐	☐	☐	☐	

G.03	Tehnološki razvoj					
G.03.01.	Tehnološka razširitev/posodobitev dejavnosti	☐	☐	☐	☐	
G.03.02.	Tehnološko prestrukturiranje dejavnosti	☐	☐	☐	☐	
G.03.03.	Uvajanje novih tehnologij	☐	☐	☐	☐	
G.03.04.	Drugo:	☐	☐	☐	☐	
G.04	Družbeni razvoj					
G.04.01	Dvig kvalitete življenja	☐	☐	☐	☐	
G.04.02.	Izboljšanje vodenja in upravljanja	☐	☐	☐	☐	
G.04.03.	Izboljšanje delovanja administracije in javne uprave	☐	☐	☐	☐	
G.04.04.	Razvoj socialnih dejavnosti	☐	☐	☐	☐	
G.04.05.	Razvoj civilne družbe	☐	☐	☐	☐	
G.04.06.	Drugo:	☐	☐	☐	☐	
G.05.	Ohranjanje in razvoj nacionalne naravne in kulturne dediščine in identitete	☐	☐	☐	☐	
G.06.	Varovanje okolja in trajnostni razvoj	☐	☐	☐	☐	
G.07	Razvoj družbene infrastrukture					
G.07.01.	Informacijsko-komunikacijska infrastruktura	☐	☐	☐	☐	
G.07.02.	Prometna infrastruktura	☐	☐	☐	☐	
G.07.03.	Energetska infrastruktura	☐	☐	☐	☐	
G.07.04.	Drugo:	☐	☐	☐	☐	
G.08.	Varovanje zdravja in razvoj zdravstvenega varstva	☐	☐	☐	☐	
G.09.	Drugo:	☐	☐	☐	☐	

Komentar

--

12.Pomen raziskovanja za sofinancerje¹¹

	Sofinancer		
1.	Naziv		
	Naslov		
	Vrednost sofinanciranja za celotno obdobje trajanja projekta je znašala:		EUR
	Odstotek od utemeljenih stroškov projekta:		%
	Najpomembnejši rezultati raziskovanja za sofinancerja		Šifra
	1.		
	2.		
	3.		

		4.		
		5.		
	Komentar			
	Ocena			

13. Izjemni dosežek v letu 2014¹²

13.1. Izjemni znanstveni dosežek

--

13.2. Izjemni družbeno-ekonomski dosežek

ZWITTER VITEZ, Ana. Ugotavljanje avtorstva besedil: primer "Trenirkarjev". V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.). Language technologies : proceedings of the 17th International Multiconference Information Society - IS 2014, October 9th - 10th, 2014, Ljubljana, Slovenia : volume G. Ljubljana: Institut Jožef Stefan, 2014, str. 131-134.

Opis dosežka oziroma učinka:

V prispevku predstavimo analizo anonimnega besedila, ki vsebuje elemente sovražnega/žaljivega govora. Rezultati izračuna leksikalnih in berljivostnih značilk kažejo, da je eden od 21 potencialnih avtorjev zelo verjetno dejanski avtor spornega besedila.

Vprašanja, ki za politično nemotiviranega, a družbeno odgovornega raziskovalca ostajajo odprta, so:

- ali rezultati raziskave zadoščajo za sklepanje o dejanskem avtorju spornega besedila,
- ali smo dolžni o rezultatih raziskave obvestiti javnost,
- kje je meja med žaljivim govorom in svobodo govora.

C. IZJAVE

Podpisani izjavljam/o, da:

- so vsi podatki, ki jih navajamo v poročilu, resnični in točni
- se strinjamo z obdelavo podatkov v skladu z zakonodajo o varstvu osebnih podatkov za potrebe ocenjevanja ter obdelavo teh podatkov za evidence ARRS
- so vsi podatki v obrazcu v elektronski obliki identični podatkom v obrazcu v pisni obliki
- so z vsebino zaključnega poročila seznanjeni in se strinjajo vsi soizvajalci projekta

Podpisi:

*zastopnik oz. pooblaščen oseba
raziskovalne organizacije:*

in

vodja raziskovalnega projekta:

Trojina, zavod za uporabno
slovenistiko

Ana Zwitter Vitez

ŽIG

Kraj in datum:

Ljubljana

13.3.2015

Oznaka poročila: ARRS-RPROJ-ZP-2015/199

¹ Napišite povzetek raziskovalnega projekta (največ 3.000 znakov v slovenskem in angleškem jeziku) [Nazaj](#)

² Napišite kratko vsebinsko poročilo, kjer boste predstavili raziskovalno hipotezo in opis raziskovanja. Navedite ključne ugotovitve, znanstvena spoznanja, rezultate in učinke raziskovalnega projekta in njihovo uporabo ter sodelovanje s tujimi partnerji. Največ 12.000 znakov vključno s presledki (približno dve strani, velikost pisave 11). [Nazaj](#)

³ Realizacija raziskovalne hipoteze. Največ 3.000 znakov vključno s presledki (približno pol strani, velikost pisave 11) [Nazaj](#)

⁴ V primeru bistvenih odstopanj in sprememb od predvidenega programa raziskovalnega projekta, kot je bil zapisan v predlogu raziskovalnega projekta oziroma v primeru sprememb, povečanja ali zmanjšanja sestave projektne skupine v zadnjem letu izvajanja projekta, napišite obrazložitev. V primeru, da sprememb ni bilo, to navedite. Največ 6.000 znakov vključno s presledki (približno ena stran, velikost pisave 11). [Nazaj](#)

⁵ Navedite znanstvene dosežke, ki so nastali v okviru tega projekta. Raziskovalni dosežek iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A" ali A'. [Nazaj](#)

⁶ Navedite družbeno-ekonomske dosežke, ki so nastali v okviru tega projekta. Družbeno-ekonomski rezultat iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A" ali A'.

Družbeno-ekonomski dosežek je po svoji strukturi drugačen kot znanstveni dosežek. Povzetek znanstvenega dosežka je praviloma povzetek bibliografske enote (članka, knjige), v kateri je dosežek objavljen.

Povzetek družbeno-ekonomskega dosežka praviloma ni povzetek bibliografske enote, ki ta dosežek dokumentira, ker je dosežek sklop več rezultatov raziskovanja, ki je lahko dokumentiran v različnih bibliografskih enotah. COBISS ID zato ni enoznačen, izjemoma pa ga lahko tudi ni (npr. prehod mlajših sodelavcev v gospodarstvo na pomembnih raziskovalnih nalogah, ali ustanovitev podjetja kot rezultat projekta ... - v obeh primerih ni COBISS ID). [Nazaj](#)

⁷ Navedite rezultate raziskovalnega projekta iz obdobja izvajanja projekta (do oddaje zaključnega poročila) v primeru, da katerega od rezultatov ni mogoče navesti v točkah 6 in 7 (npr. ni voden v sistemu COBISS). Največ 2.000 znakov, vključno s presledki. [Nazaj](#)

⁸ Pomen raziskovalnih rezultatov za razvoj znanosti in za razvoj Slovenije bo objavljen na spletni strani: <http://sicris.izum.si/> za posamezen projekt, ki je predmet poročanja [Nazaj](#)

⁹ Največ 4.000 znakov, vključno s presledki [Nazaj](#)

¹⁰ Največ 4.000 znakov, vključno s presledki [Nazaj](#)

¹¹ Rubrike izpolnite / prepišite skladno z obrazcem "izjava sofinancerja" <http://www.arrs.gov.si/sl/progproj/rproj/gradivo/>, ki ga mora izpolniti sofinancer. Podpisan obrazec "Izjava sofinancerja" pridobi in hrani nosilna raziskovalna organizacija – izvajalka projekta. [Nazaj](#)

¹² Navedite en izjemni znanstveni dosežek in/ali en izjemni družbeno-ekonomski dosežek raziskovalnega projekta v letu 2014 (največ 1000 znakov, vključno s presledki). Za dosežek pripravite diapozitiv, ki vsebuje sliko ali drugo slikovno gradivo v zvezi z izjemnim dosežkom (velikost pisave najmanj 16, približno pol strani) in opis izjemnega dosežka (velikost pisave 12, približno pol strani). Diapozitiv/-a priložite kot priponko/-i k temu poročilu. Vzorec diapozitiva je objavljen na spletni strani ARRS <http://www.arrs.gov.si/sl/gradivo/>, predstavitev dosežkov za pretekla leta pa so objavljena na spletni strani <http://www.arrs.gov.si/sl/analize/dosez/>. [Nazaj](#)

Obrazec: ARRS-RPROJ-ZP/2015 v1.00a

74-46-44-96-1A-C2-36-35-18-80-5E-CB-1F-BC-F5-16-F6-DE-2A-81

Priloga 1

VEDA : 7 INTERDISCIPLINARNE RAZISKAVE

Področje: 6.02 Jeziki in književnost

Dosežek 1:

ZWITTER VITEZ, Ana. Ugotavljanje avtorstva besedil: primer "Trenirkarjev". V: ERJAVEC, T. (ur.), ŽGANEC GROS, J. (ur.). *Jezikovne tehnologije : zbornik 17. mednarodne multikonference Informacijska družba - IS 2014* Ljubljana: Institut Jožef Stefan, 2014, str. 131-134.

Vir: [COBISS.SI-ID [55987554](#)]

Author : Anonymous

Vocabulary features: standard deviation

Author	DRB	BI	HS	HL	DRL	BI-L	HS-L	HL-L
G	0.02 ± 0.11	0.14 ± 0.61	70.30 ± 372.30	0.02 ± 0.11	0.02 ± 0.12	0.03 ± 0.95	77.01 ± 193.59	0.01 ± 0.11
A	0.05 ± 0.02	0.11 ± 0.04	56.17 ± 230.80	0.05 ± 0.01	0.06 ± 0.02	0.35 ± 0.02	95.76 ± 254.19	0.05 ± 0.00
H	0.06 ± 0.06	0.22 ± 0.27	15.38 ± 74.07	0.06 ± 0.06	0.06 ± 0.06	0.39 ± 0.44	12.16 ± 83.69	0.05 ± 0.05
O	0.06 ± 0.02	0.26 ± 0.18	420.85 ± 315.92	0.07 ± 0.03	0.06 ± 0.03	0.46 ± 0.33	197.93 ± 154.95	0.06 ± 0.03
D	0.07 ± 0.04	0.34 ± 0.18	196.48 ± 175.76	0.07 ± 0.04	0.10 ± 0.04	0.84 ± 0.30	316.35 ± 83.34	0.09 ± 0.04
M	0.10 ± 0.02	0.47 ± 0.05	242.78 ± 160.22	0.10 ± 0.02	0.12 ± 0.03	1.01 ± 0.13	378.17 ± 135.76	0.12 ± 0.02
N	0.10 ± 0.04	0.48 ± 0.13	1.36 ± 58.86	0.10 ± 0.04	0.09 ± 0.05	0.72 ± 0.29	53.15 ± 134.35	0.08 ± 0.04
P	0.10 ± 0.06	0.44 ± 0.30	183.11 ± 43.80	0.10 ± 0.05	0.10 ± 0.04	0.72 ± 0.32	167.55 ± 109.35	0.09 ± 0.03
J	0.12 ± 0.00	0.48 ± 0.00	38.12 ± 0.00	0.12 ± 0.00	0.15 ± 0.00	1.10 ± 0.00	91.45 ± 0.00	0.13 ± 0.00

Opis dosežka oziroma učinka:

V prispevku predstavimo analizo anonimnega besedila, ki vsebuje elemente sovražnega/žaljivega govora. Rezultati izračuna leksikalnih in berljivostnih jezikovnih značilk kažejo, da je eden od 21 potencialnih avtorjev zelo verjetno dejanski avtor spornega besedila.

Vprašanja, ki za politično nemotiviranega, a družbeno odgovornega raziskovalca ostajajo odprta, so:

- ali rezultati raziskave zadoščajo za sklepanje o dejanskem avtorju spornega besedila,
- ali smo dolžni o rezultatih raziskave obvestiti javnost,
- kje je meja med žaljivim govorom in svobodo govora.