sciendo

Primjena NLP tehnologija na nisko resursna hrvatska narječja

Maja Polanec
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Marina Bagić Babac
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva

Sažetak

U sustavima za obradu prirodnog jezika (NLP) primjećuje se trend smanjenja učinkovitosti kada se primjenjuju na tekstove napisane nisko resursnim narječjem, umjesto standardnim jezikom. Ovisnosno parsiranje je važna komponenta u NLP sustavima, stoga bi njegovo unaprjeđenje moglo za posljedicu imati poboljšanje učinkovitosti tih sustava. Ovaj rad ima za cilj usporediti učinkovitost slovenskog i hrvatskog parsera za ovisnosno parsiranje kajkavskog narječja. Rezultati usporedbe će pružiti uvid u potencijal slovenskog parsera za parsiranje kajkavskog. Stvoren je skup podataka za ovisnosno parsiranje korištenjem paralelnih prijevoda knjige "Mali kraljević". Na temelju stvorenog skupa podataka je provedeno projiciranje oznaka iz isparsiranog hrvatskog standardnog jezika u kajkavsko narječje s ciljem dobivanja podataka za izračun UAS i LAS metrika za usporedbu hrvatskog i slovenskog parsera koji su implementirani pomoću Spacy knjižnice otvorenog koda. Hrvatski parser postigao je UAS rezultat od 0.47 i LAS rezultat od 0.30, što je manje uspješno od slovenskg parsera koji je ostvario rezultate od 0.52 za UAS i 0.34 za LAS. Dobiveni rezultati pokazuju da slovenski parser preciznije parsira kajkavsko narječje. Međutim, za donošenje općeg zaključka bilo bi potrebno proširiti skup podataka.

Ključne riječi: obrada prirodnog jezika; nisko resursno narječje; hrvatski jezik; ovisnosni parser

Vrsta članka: Preliminarno istraživanje

Primljeno: 14.6.2024.
Prihvaćeno: 4.1.2025.

DOI: 10.2478/crdj-2025-0008

Uvod

Sustavi za obradu prirodnog jezika (engl. Natural Language Processing) trebaju biti osjetljivi na razlike u narječjima kako bi osigurali pravednost i inkluzivnost prema govornicima različitih narječja. Ipak, prema Joshi i sur. (2024) i kod tekstova koji sadrže narječja jezika postoji trend pada učinkovitosti u sustavima za obradu prirodnog jezika u usporedbi s tekstovima na standardnom jeziku. Posljedično, korisničko iskustvo je lošije za govornike narječja te se sprječava ravnopravno sudjelovanje svih govornika u digitalnom svijetu. Qi i sur. (2019) navode da je ovisnosno parsiranje važna komponenta u raznim takvim sustavima obrade prirodnog jezika kao što su sustavi za označavanje semantičkih uloga, ekstrakciju odnosa i strojno prevođenje. Sukladno tome, postoji opravdana motivacija za istraživanje unaprjeđenja postojećih metoda ovisnosnog parsiranja, posebno u kontekstu narječja.

Banke stabala su osnova za oblikovanje ovisnosnih parsera. Za hrvatski jezik postoje dvije banke stabala. Prva takva banka stabala, koju je predstavio Tadić (2007), je HOBS (Hrvatska ovisnosna banka stabla). Ona je stvorena korištenjem dijela CW2000 podkorpusa koji predstavlja novinski korpus. Druga banka stabala koju su predstavili Agić i Ljubešić (2015) je „Croatian UD“, izrađena na temelju SETIMES.HR korpusa, koji je korpus novinskih članaka na balkanskim jezicima. Iako se, kako navode Farkaš i Filko (2022), dvije hrvatske banke stabala razlikuju u postupcima obilježavanja određenih sintaktičkih struktura korištenim prilikom njihove izrade i dalje im je zajedničko da su stvorene uporabom novinskog korpusa. Sva tri hrvatska narječja nisu podjednako zastupljena u novinama, zbog čega se čakavsko i kajkavsko narječje mogu smatrati niskoresursnim narječjima.

Kako bi se unaprijedile metode ovisnosnog parsiranja niskoresursnih narječja, potrebno je razviti osnovni model (engl. baseline model) za ta narječja koji bi služio kao referentna točka za usporedbu performansi naprednijih metoda. Kao osnovni model mogu se koristiti modeli dobiveni učenjem bez primjera (engl. zero-shot learning), gdje se model treniran na visokoresursnom jeziku donoru primjenjuje na niskoresursni jezik primatelj bez dodatne prilagodbe. U radu Zampieri i sur. (2017) su tako razvijeni neki osnovni modeli za zadatak međujezičnog ovisnosnog parsiranja za različite jezične parove među kojima je bio i par hrvatski-slovenski. Učenje bez primjera je primijenjeno u radu Scherrer (2014), gdje je model španjolskog parsera korišten kao osnovni model za parsiranje katalonskog jezika. Učenje bez primjera je također istraživano uz korištenje više jezika donora. Primjerice, Scherrer i Rabus (2017, 2019) su trenirali označivač (engl. tagger) na podacima slovačkog, ukrajinskog, ruskog i poljskog jezika, a zatim su ga direktno primijenili na rusinski jezik. Zampieri i sur. (2017) u svom radu ističu da je ovisnosno parsiranje norveškog jezika ostvarilo bolje rezultate kada je model treniran na podacima danskog i švedskog jezika u usporedbi s modelom treniranim isključivo na danskom ili isključivo na švedskom jeziku.

Cilj je ovog rada usporediti učinkovitost parsera istreniranog na hrvatskom jeziku s parserom istreniranim na slovenskom jeziku za primjenu kod parsiranja kajkavskog

narječja te vidjeti postoji li razlika u njihovom potencijalu za parsiranje kajkavskog. Rezultati takve usporedbe bi pružili uvid u opravdanost implementacije osnovnog modela za kajkavski parser korištenjem slovenskog parsera te odgovorili na pitanje bi li parsiranje kajkavskog imalo koristi i od podataka na slovenskom jeziku.

Rad je organiziran na sljedeći način. Poglavlje Skup podataka za nisko resursna narječja pruža pregled literature u kojoj su opisani različiti pristupi za stvaranje i proširenje skupova podataka za NLP obradu nisko resursnih narječja. U poglavlju Metodologija je objašnjen postupak stvaranja korištenog skupa podataka, opisana je knjižnica otvorenog koda koja je korištena za ovisnosno parsiranje te su prikazani primjeri dobiveni poravnanjem riječi i projiciranjem oznaka. U poglavlju Rezultati su hrvatski i slovenski parser uspoređeni korištenjem metrika UAS (engl. Unlabeled Attachment Score) i LAS (engl, Labeled Attachment Score). U poglavlju Diskusija su komentirani dobiveni rezultati, pružena su moguća objašnjenja za takve rezultate te su istaknuta ograničenja ovog rada.

Skup podataka za nisko resursna narječja

Policentrični jezici, poput primjerice engleskog, su standardni jezici u više država te svaka država koristi svoju varijaciju tog jezika. To za posljedicu ima prisutnost te varijacije u brojnim pisanim izvorima. Posljedično tome pri stvaranju skupa podataka za daljnju NLP obradu dostupna je velika količina materijala poput novina, digitaliziranih knjiga itd. Tako je nastao, primjerice, korpus DSLCC koji je sastavljen od kratkih tekstova iz novinskih članaka. Između ostalih jezika u korpusu su prisutni i hrvatski, srpski i bosanski. Korištenje dijela ovog korpusa koji se odnosi na hrvatski jezik za obradu hrvatskih narječja bez daljnjeg dopunjavanja skupa podataka ne bi bilo djelotvorno jer sva tri hrvatska narječja nisu podjednako zastupljena u novinama, koje su bile jedini izvor za stvaranje tog skupa podataka. Štoviše, jedino je štokavski zastupljen u značajnim količinama. Iako kod stvaranja skupa podataka za narječja nekih jezika, uglavnom policentričnih, izostaje ova problematika nepodjednake zastupljenosti ili potpunog izostanka određenog narječja, ona je skoro uvijek prisutna pri obradi narječja ostalih jezika ili kolokvijalnog govora određenih skupina. Tako se u radovima na tu temu mogu pronaći brojne metode za stvaranje skupa podataka za podzastupljena narječja i govore.

Na primjer, u radu Jørgensen i sur. (2016) za treniranje modela koji pridjeljuje POS oznake za tekstove povezane sa AAVE (African-American Vernacular English), koji je varijanta engleskog kojom se služi većina Afroamerikanaca radničke i srednje klase, skup podataka je sačinjen od objava na Twitteru, tekstova serije Žica (engl. The Wire) i stihova pjesama. Ista bi metodologija mogla biti primijenjena na hrvatska narječja. Repertoar pjesama na čakavskom i kajkavskom je dovoljno velik i raznolik, a postoje i serije i filmovi snimani na tim narječjima. Međutim, problem, pogotovo s pjesmama pisanim narječjem, je da su one često vrlo stilski obilježene i govore o konkretnim temama pa bi modeli naučeni na takvim podatcima slabije generalizirali.

Alternativni izvor za prikupljanje podataka s interneta koji je i korišten u prethodno navedenom radu su društvene mreže koje, kako navodi Bagić Babac (2023), bolje nego tekstovi pjesama opisuju svakodnevni govor. Alshutayri i Atwell (2017) u svom radu istražuju takav pristup pribavljanja podataka na primjeru arapskih dijalekata. Korištena društvena mreža bio je Twitter, a autori su se susreli sa problematikom koja je prisutna i kod hrvatskih narječja, a to je neravnomjerna zastupljenost različitih narječja. Pošto je cilj dobiti izbalansirani skup podataka zabilježili su da je za neka narječja bilo potrebno više vremena za pokretanje ekstraktora teksta.

Podaci za treniranje modela u kontekstu narječja često se dobivaju transkripcijom govora izvornih govornika. Takav način dobivanja podataka bi za čakavski i kajkavski imao potencijala dati najbolje i najkorisnije rezultate jer bi precizno opisivao jezik koji se koristi u svakodnevnom govoru. Transkripcija govora može se provoditi sustavima za automatsko prepoznavanje govora. Tako su primjerice Ali i sur. (2016) izradili korpus arapskog dijalekta. Transkripcija se također može provodoti i ručno kao što je primjerice u radu Scherrer i sur. (2019) izrađen korpus švicarskih i njemačkih dijalekata. Stvoreni skupovi podataka se mogu proširiti augmentacijom podataka. Primjeri takvih augmentacija su korišteni u radu Vania i sur. (2019). Oni su koristili preoblikovanje ovisnosnog stabla te proširivanje podataka generiranjem novih rečenica. Preoblikovali su ovisnosna stabla uklanjanjem dijelova rečenica, odnosno pojednostavljenjem rečenica te rotacijom riječi u rečenici oko korijenskog glagola. Rezultati takvih operacija su neprirodno oblikovane rečenice, no prema autorima rada to ne šteti učenju, a modelu pruža više varijacije u primjerima. Ističu da takav pristup može koristiti jezicima s fleksibilnim redoslijedom riječi u rečenici i jezicima s bogatom morfologijom, među koje spada i hrvatski jezik.

Metodologija

Stvaranje skupa podataka

Većina kajkavskog teksta dostupna online je iz stare literature koja nije odraz suvremenog kajkavskog narječja radi čega nije izabrana kao resurs za stvaranje skupa podataka za ovaj rad, već je skup podataka stvoren sa suvremenim kajkavskim tekstom. „Mali kralevič“ je prijevod novele „Mali kraljević“, francuskog autora Antoine de Saint-Exupérya, na suvremeni kajkavski. Knjiga je prevedena 2018. godine, a prijevod su odradili Akoš Anton Dončec i Đuro Blažeka. Za potrebe stvaranja skupa podataka razmatrana je još i uporaba titlova serije „Gruntovčani“ i „Nosila je rubac črleni“ po uzoru na rad Jørgensen i sur. (2016), ali pošto u vrijeme pisanja rada nisu bili javno dostupni gotovi titlovi, uzimanje kajkavskog iz knjige je predstavljalo efikasniji, a jednako vrijedan proces stvaranja skupa podataka.

Stvorena su dva skupa podataka, jedan sa tekstom iz verzije knjige na hrvatskom standardnom jeziku, a drugi na kajkavskom. Oba skupa sadrže identičan broj rečenica, njih oko 500. Rečenice oba skupa su međusobno poravnate što je bilo ključno za

primjenu poravnanja riječi koja je prethodila projekciji anotacija. Primjer dijela skupa podataka iz kojeg je vidljivo da su hrvatske i kajkavske rečenice u skupu podataka međusobno poravnate prikazan je u sljedećoj tablici.

Tablica 1
Prikaz skupa podataka

STANDARDNI HRVATSKI JEZIK	**KAJKAVSKO NARJEČJE**
Dio 1	Del 1.
Kad mi je bilo šest godina vidio sam jednom veličanstvenu sliku, u knjizi o prašumi, koja se zvala Istinite priče.	Gda mi je bilo šest leti, videl sem jeno čudovito sliko v knigi o praega šumi koja se zove Istinite zgodbe.
Prikazivala je udava kako guta neku zvjerku.	Na sliki je bila velka kača koja je iti pôždrla neko zverino.
Evo kopije tog crteža.	Naslikal sem to sliko.
U knjizi je pisalo.	V knigi so pisali.
Udavi gutaju svoj plijen cijel cjelcat, bez žvakanja.	Velka kača pôždere svojega celoga plena, a ne vgrizne ga.
Nakon toga se više ne mogu maknuti, pa spavaju šest mjeseci, dok ga ne probave.	Onda se nemre gibati, zato prespi šest mesecov dok ga prebavla.
Zatim sam mnogo razmišljao o prašumskim pustolovinama, pa sam i sam uspio drvenom bojicom, nacrtati svoj prvi crtež.	V ovom cajto gliboko sem premišlaval o pustolovini v prašumaj i z farbanim klajbesom sem naslikal svojo prvo sliko.
Moj crtež broj 1, izgledao je ovako.	Glečte.
Pokazao sam svoje remek-djelo nekim odraslim osobama i upitao ih boje li se mojega crteža.	Pokazal sem svoje meštersko delo odraslim i pitam ak jih prestraši.

Izvor: Autorski rad

Spacy parser

Prema Borotić i sur. (2023), Spacy je knjižnica otvorenog koda namijenjena za prirodnu obradu jezika u Pythonu. U ovom radu je korištena za zadatke ovisnosnog parsiranja iako se može koristiti za niz NLP zadataka poput označivanja POS oznakama ili morfološku analizu. Kako navode Šandor i Bagić Babac (2024), Spacy podržava brojne jezike uključujući hrvatski i slovenski, čiji su parseri korišteni kasnije u radu. Oznake sa kojima rade modeli u knjižnici su kompatibilne sa Universal Dependencies shemom označavanja što omogućuje rad sa standardiziranim podacima neovisno o jeziku. U tablici 2 prikazane su oznake kojima Spacy parser označuje ovisnosne veze između riječi.

Tablica 2
UD oznake parsera

ROOT	acl	advcl	advmod	advmod:emph	amod
appos	aux	aux:pass	case	cc	ccomp
compound	conj	cop	csubj	csubj:pass	dep
det	discourse	expl:pv	fixed	flat	flat:foreign
goeswith	iobj	mark	nmod	nsubj	nsubj:pass
nummod	obj	obl	orphan	parataxis	punct

Izvor: Autorski rad

U praktičnom dijelu rada korišteni su hrvatski i slovenski ovisnosni parseri. Spacy nudi modele koji su trenirani na malim i velikim skupovima podataka. Za oba jezika su korišteni parseri trenirani na najvećem skupu podataka jer su ti parseri imali najbolje metrike.

Poravnanje riječi

Za poravnanje riječi između hrvatskih i kajkavskih rečenica korišten je alat SimAlign. Nedostatak ovog postupka je što je često nemoguće dobiti cijelu isparsiranu rečenicu, već je rezultat isparsirana rečenica kojoj nedostaju dijelovi. Primjer izlaza poravnanja riječi u jednoj duljoj rečenici prikazan je ispod. Iz primjera na slici je vidljivo da će rečenica biti nepotpuna kada se projiciraju oznake na kajkavski jer alat nije mogao svaku riječ kajkavske rečenice poravnati s nekom riječi hrvatske rečenice.

Slika 1
Projekcija oznaka na kajkavski

```
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (7, 7), (8, 8), (9, 9), (10, 10), (11, 11),
(12, 12), (13, 13), (14, 15), (15, 16), (16, 17), (17, 18), (18, 19), (19, 20)]
```

Izvor: Autorski rad

Projiciranje oznaka iz hrvatskog u kajkavski

Primijenjen je parser za hrvatski na tekst „Malog kraljevića“ te je korištenjem izlaza parsera i rezultata poravnanja riječi stvorena banka stabala za tekst „Malog kraleviča“. Kod usporedbe rezultata projiciranja oznaka sa rezultatima direktne primjene hrvatskog parsera na kajkavski uočljivo je da projiciranje oznaka daje bolje rezultate. Na primjerima ispod uspoređeni su rezultati za projiciranje oznaka i direktno parsiranje kajkavskog teksta parserom predtreniranom na hrvatskom tekstu, u navedenom poretku.

Rečenice u primjerima su “Mel sem pitko vodo za komaj osem dni.“ („Imao sam pitke vode jedva za osam dana.“) i „Samo so mi rekli.“ („One su mi odgovorile.“).

```
0  Mel  _  _  _  _  0  ROOT  _  _
1  sem  _  _  _  _  0  aux   _  _
```

```
   2  pitko _  _  _  _  3  amod  _  _
    3  vodo_  _  _  _  0  obj   _  _
4  komaj  _  _  _  _  3  advmod  _  _
     5  za _  _  _  _  7  case _  _
6  osem   _  _  _  _  7  nummod  _  _
    7  dni   _  _  _  _  0  obl   _  _

    0  Mel  _  _  _  _  6  nsubj_  _
    1  sem  _  _  _  _  0  flat   _  _
    2  pitko _  _  _  _  0  flat   _  _
    3  vodo_  _  _  _  2  flat   _  _
     4  za _  _  _  _  5  case _  _
  5  komaj  _  _  _  _  0  nmod   _  _
  6  osem   _  _  _  _  6  ROOT   _  _
   7  dni   _  _  _  _  7  ROOT   _  _

    0  Samo   _  _  _  _  3  nsubj_  _
     1  so _  _  _  _  3  aux  _  _
     2  mi_  _  _  _  3  obj  _  _
   3  rekli  _  _  _  _  3  ROOT   _  _

0  Samo   _  _  _  _  1  advmod  _  _
     1  so _  _  _  _  3  nsubj_  _
     2  mi_  _  _  _  3  obj  _  _
   3  rekli  _  _  _  _  3  ROOT   _  _
```

Primjeri su prikazani u CONLL-u formatu koji se sastoji od tri tipa redaka:

- Retci s riječima koji sadrže anotaciju riječi u 10 polja odvojenih znakovima tabulatora. Polja su redom: indeks riječi, sama riječ, lemma, UPOS, XPOS, FEATS, HEAD, DEPREL, DEPS, MISC. Podebljana polja su ona koja su popunjena u korištenim CONLL-u datotekama. Ostala polja su prazna jer se ne tiču ovisnosnog parsiranja.
- Prazni retci koji označavaju granice rečenica. Posljednji redak svake rečenice je prazan redak.
- Komentari na razini rečenice koji počinju znakom hash (#).

Rezultati

Metrike koje su korištene za evaluaciju rezultata su UAS i LAS. To su metrike koje se i inače koriste za evaluaciju točnosti parsiranja modela parsera. UAS se računa kao broj riječi koje su točno povezane sa svojom glavom kroz ukupan broj riječi u rečenici, dok je LAS nešto napredniji i uzima u obzir i točnost oznake ovisnosne veze, odnosno računa se kao broj točno povezanih riječi s točno označenom vrstom veze kroz ukupan broj riječi u rečenici. Usporedbe radi, hrvatski parser iz Spacy knjižnice postiže UAS ocjenu 0.87 i LAS ocjenu 0.8, dok za slovenski parser te ocjene iznose 0.87 i 0.84.

Projicirane oznake opisane u prošlom poglavlju ovdje su uspoređene sa izlazom prvo hrvatskog pa potom slovenskog parsera. Kako se projiciranje oznaka temelji na poravnanju riječi, posljedično je nemoguće tim postupkom dobiti cijelu isparsiranu rečenicu, već je rezultat isparsirana rečenica kojoj nedostaju dijelovi. Posljedično tome su se se izlazi hrvatskog i slovenskog parsera uspoređivali samo sa riječima koje su „preživjele" projiciranje.

Tablica 3 prikazuje performanse hrvatskog i slovenskog parsera evaluiranih na 500-tinjak rečenica iz knjige „Mali kralevič".

Tablica 3
Rezultati hrvatskog i slovenskog parsera

	Hrvatski parser	Slovenski parser
UAS	0.47	0.52
LAS	0.30	0.34

Izvor: Autorski rad

Rezultati pokazuju da slovenski parser ostvaruje bolje performanse od hrvatskog pri primjeni na tekstove na kajkavskom narječju, postižući više vrijednosti i za UAS i za LAS metriku. Kao što je uobičajeno, vrijednosti UAS-a nadmašuju one LAS-a, no u našem radu ta je razlika izraženija nego kod hrvatskog i slovenskog parsera iz Spacy knjižnice. Primjerice, za hrvatski parser u Spacy-ju razlika između UAS-a i LAS-a iznosi 0.07, dok u našem radu doseže 0.17. Za slovenski parser razlika u Spacy-ju je 0.03, a u našem radu 0.18. To ukazuje da model u našem radu znatno teže određuje ispravnu oznaku ovisnosne veze nego što pravilno povezuje riječi s odgovarajućom glavom.

Diskusija

Slovenski parser preciznije parsira kajkavsko narječje, što se može objasniti sličnostima između kajkavskog narječja i slovenskog jezika. Kako navodi Celinić (2020), kajkavština je, kroz brojne izoglose, povezana sa slovenskim jezikom, a kajkavsko narječje dijeli mnoge osobine sa slovenskim jezikom. Na primjer, u oba govora prisutan je protetski v- ispred refleksa u, a dvojina, koja je očuvana u slovenskom jeziku, djelomično je prisutna i u kajkavskom narječju. Celinić (2020) također tvrdi da su mnoge leksičke značajke zajedničke slovenskom jeziku.

U ovom radu je korišten skup podataka koji se sastoji od oko 500 poravnatih rečenica hrvatskog standardnog jezika i kajkavskog narječja. Zbog primjene metode projiciranja oznaka koja se temelji na poravnanju riječi, dio riječi u rečenicama je „izgubljen". Takav skup podataka je premalen za generalizaciju zaključka da je slovenski parser superiorniji od hrvatskog parsera u parsiranju kajkavskog narječja. Također, korištenom skupu podataka nedostaje raznolikosti pošto se temelji isključivo na jednom književnom dijelu.

Zaključak

U ovom radu su opisani različiti pristupi stvaranju skupa podataka za nisko resursna narječja. Stvoren je skup podataka za ovisnosno parsiranje korištenjem paralelnih prijevoda knjige „Mali kraljević". Provedeno je projiciranje oznaka iz isparsiranog hrvatskog standardnog jezika na kajkavsko narječje kako bi se dobili podaci za izračun UAS i LAS metrika. Konačno, korištenjem tih metrika, uspoređeni su hrvatski i slovenski parser za parsiranje kajkavskog narječja. Slovenski parser je pružio bolje rezultate od hrvatskog parsera, no zbog ograničenosti korištenog skupa podataka ne može se zaključiti da slovenski parser općenito daje bolje rezultate.

Praktični značaj ovog istraživanja leži u zaključku da prilikom parsiranja kajkavskog narječja uvijek treba koristiti i slovenske parsere, s obzirom na to da slovenski jezik dijeli mnoge sličnosti s kajkavskim, čime se omogućava preciznija analiza sintaktičke strukture kajkavskih tekstova**.** Svakako bi imalo smisla kod kreiranja osnovnih modela za ispitivanje naprednijih tehnika parsiranja kajkavskog implementirati osnovni model slovenskim parserom.

U budućim istraživanjima potrebno je proširiti skup podataka kako bi se omogućilo donošenje općeg zaključka, s obzirom na to da je ovaj rad ograničen veličinom korištenog skupa podataka. Skup podataka može se proširiti izradom novog skupa podataka, a također se može koristiti augmentacija podataka kao u radu Vania i sur. (2019).

Literatura

1. Agić, Ž., & Ljubešić, N. (2015, September). Universal dependencies for Croatian (that work for Serbian, too). In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing* (pp. 1–8).
2. Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., Bell, P., & Renals, S. (2016). Automatic dialect detection in Arabic broadcast speech. In *Proceedings of INTERSPEECH 2016* (pp. 2934–2938). San Francisco, CA, USA.
3. Alshutayri, A., & Atwell, E. (2017). Exploring Twitter as a source of an Arabic dialect corpus. *International Journal of Computational Linguistics (IJCL), 8*(2).
4. Bagić Babac, M. (2023). Emotion analysis of user reactions to online news. *Information Discovery and Delivery, 51*(2), 179–193. https://doi.org/10.1108/IDD-04-2022-0027
5. Borotić, G., Granoša, L., Kovačević, J., & Bagić Babac, M. (2023). Effective spam detection with machine learning. *Croatian Regional Development Journal, 3*(2), 43–64. https://doi.org/10.2478/crdj-2023-0007
6. Celinić, A. (2020). Kajkavian. *Hrvatski dijalektološki zbornik, 24*, 1–37.

7. Farkaš, D., & Filko, M. (2022). Obilježavanje koordinacije u ovisnosnim bankama stabala. *Jezikoslovlje, 23*(2), 193–214.

8. Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., & Dippold, D. (2024). Natural language processing for dialects of a language: A survey. *arXiv preprint* arXiv:2401.05632. https://arxiv.org/abs/2401.05632

9. Jørgensen, A., Hovy, D., & Søgaard, A. (2016). Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1115–1120). San Diego, CA, USA.

10. Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2019). Universal dependency parsing from scratch. *arXiv preprint* arXiv:1901.10457. https://arxiv.org/abs/1901.10457

11. Scherrer, Y. (2014, August). Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects* (pp. 30–38).

12. Scherrer, Y., & Rabus, A. (2017, April). Multi-source morphosyntactic tagging for spoken Rusyn. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (pp. 84–92).

13. Scherrer, Y., & Rabus, A. (2019). Neural morphosyntactic tagging for Rusyn. *Natural Language Engineering, 25*(5), 633–650. https://doi.org/10.1017/S1351324919000202

14. Scherrer, Y., Samardžič, T., & Glaser, E. (2019). Digitising Swiss German – How to process and study a polycentric spoken language. *Language Resources and Evaluation, 53*(4), 735–769. https://doi.org/10.1007/s10579-019-09459-5

15. Šandor, D., & Bagić Babac, M. (2024). Sarcasm detection in online comments using machine learning. *Information Discovery and Delivery, 52*(2), 213–226. https://doi.org/10.1108/IDD-01-2023-0002

16. Tadić, M. (2007). Building the Croatian dependency treebank: The initial stages. *Suvremena lingvistika, 33*(63), 85–92.

17. Vania, C., Kementchedjhieva, Y., Søgaard, A., & Lopez, A. (2019). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1105–1116).

18. Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., & Aepli, N. (2017, April). Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (pp. 1–15).

O autorima

Maja Polanec je studentica završne godine diplomskog studija računarstva na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu. Prije dvije godine završila je preddiplomski studij računarstva na istoj instituciji, sa završnim radom pod naslovom "Modeli strojnog učenja za klasifikaciju glazbenih žanrova na skupu podataka AudioSet". Njezini profesionalni interesi uključuju strojno učenje i obradu prirodnog jezika. Uz studij radi honorarno kao programerka. Autoricu možete kontaktirati na **maja.polanec@fer.hr**.

Marina Bagić Babac je izvanredna profesorica na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu, gdje je stekla titule diplomiranog inženjera, te magistra i doktora znanosti. Također je stekla titutlu diplomiranog novinara na Fakultetu političkih znanosti Sveučilišta u Zagrebu. Aktivno sudjeluje na nekoliko međunarodnih projekata u području umjetne inteligencije. Članica je programskih odbora nekoliko međunarodnih znanstvenih konferencija i časopisa, te recenzentica u brojnim međunarodnim časopisima. Njezini istraživački interesi uključuju umjetnu inteligenciju, strojno učenje, obradu prirodnog jezika i analizu društvenih mreža. Autoricu možete kontaktirati na **marina.bagic@fer.hr**.