KADA SE PORTALI „ZARAZE" KORONOM: RAZVOJ I USPOREDNA ANALIZA ČLANAKA PORTALA *INDEX.HR* 2019. I 2020. GODINE

Petra Bago

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

SAŽETAK *Svrha ovog rada jest predstaviti metodologiju, alate i rezultate usporedne računalne analize* online *članaka: od prikupljanja dokumenata i čišćenja jezičnih podataka za razvoj specijaliziranoga korpusa članaka do prikaza korištenih alata i usporedne statističke analize korpusa. Istraživanje je provedeno na dva specijalizirana korpusa razvijena upravo za potrebe istraživanja, a temelje se na 500 članaka u kategoriji „Vijesti" portala* Index.hr. *Jedan korpus temelji se na člancima objavljenima u predpandemijskoj 2019. godini, a drugi na temelju članaka objavljenih u pandemijskoj 2020. godini. Analizom podataka otkriveno je da je vokabular pandemijskoga korpusa značajno siromašniji od predpandemijskoga korpusa, da se u 2020. manje pisalo o susjednim državama RH nego 2019. godine te da se u predpandemijskom korpusu više spominju domaći gradovi nego inozemni, dok je suprotan slučaj u pandemijskome korpusu. Konačno, istražena je i primjerenost automatske ekstrakcije termina za identifikaciju specifičnih tema kojima se bave promatrani korpusi.*

KLJUČNE RIJEČI

STATISTIČKA ANALIZA KORPUSA, SPECIJALIZIRANI KORPUS, NOVINSKI ČLANCI, *SKETCH ENGINE*, *PYTHON*, INDEX.HR

Bilješka o autorici

Petra Bago :: Filozofski fakultet, Sveučilište u Zagrebu **::** pbago@ffzg.hr

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

UVOD

Dostupnost velike količine tekstova u digitalnom obliku omogućava njihovu analizu primjenom metoda i alata iz područja obrade prirodnoga jezika i računalne lingvistike. Navedene metode i alati imaju široku primjenu, pa se tako, između ostaloga, koriste i za potrebe analiza velike količine (digitalnih) tekstova, kakve čovjek nije u mogućnosti provesti ručnim metodama. Dosadašnja istraživanja o metodama i alatima za analizu digitalnih tekstova bave se njihovim razvojem, primjenom te prilagodbom za specifične potrebe istraživanja različitih društveno-humanističkih fenomena. U okviru ovoga rada nije moguće ponuditi njihov cjelovit pregled, stoga je nastavak usmjeren na prikaz istraživanja o tematskom modeliranju, analizi sentimenta i razvoju specijaliziranih korpusa u području novinarstva, koja predstavljaju teorijsko-metodološko polazište istraživanja predstavljenog u ovom radu. Općenito o području obrade prirodnoga jezika i računalne lingvistike v. npr. Jurafsky i Martin (2008), Manning i Schütze (1999) i Mitkov (2004).

Metode tematskoga modeliranja klase modela strojnoga učenja koriste se za automatsko otkrivanje tema u velikim skupovima dokumenata. Jacobi, Van Atteveldt i Welbers (2016) prikazuju primjenu alata za tematsko modeliranje na primjeru studije slučaja posvećene utvrđivanju načina na koje je *New York Times* pisao o temi nuklearne tehnologije u razdoblju od 1945. do 2013. godine. Autori su u tome istraživanju djelomično replicirali istraživanje Gamsona i Modiglianija iz 1989. godine. Podatkovni skup koji je analiziran za potrebe istraživanja sastoji se od 51 528 novinskih članaka prikupljenih iz *online* arhiva *New York Timesa*. Zanimljivo domaće istraživanje na temu tematskoga modeliranja objavio je Korenčić (2019) u doktorskoj disertaciji koja se bavi primjenom tematskih modela i metoda vrednovanja tematskih modela za potrebe analize medijske agende koja se provodi s ciljem stjecanja uvida u strukturu i zastupljenost medijskih tema. Istraživanje je provedeno na zbirkama političkih vijesti: 19 američkih portala i sedam hrvatskih portala. Istraživanje na temu pandemije bolesti COVID-19, a primjenjujući metode tematskoga modeliranja, proveli su Gozzi i sur. (2020) analizirajući novinske *online* članke, videa *mainstream* medija na *YouTubeu*, objave i komentare na društvenoj mreži *Reddit* te preglede tematskih stranica Wikipedije. Otkrili su da rast aktivnosti korisnika korelira s povećanjem medijskih objava, ali da česte medijske objave i visoke incidencije bolesti COVID-19 za posljedicu imaju nagli pad aktivnosti korisnika.

Analiza sentimenta područje je koje se bavi identifikacijom, klasifikacijom, kvantifikacijom i proučavanjem mišljenja, stavova, emocija i ostalih subjektivnih informacija u pisanim tekstovima. Analizirajući kolekciju od 1592 citata iz novinskih članaka na engleskome jeziku poznatoga izvora i predmeta, Balahur i sur. (2010) identificirali su tri podzadatka koje uključuje analiza sentimenta nad novinskim člancima: (1) definicija predmeta o kojem se prikuplja sentiment, (2) odvajanje dobrog i lošeg novinskog sadržaja od pozitivnog i negativnog sentimenta o predmetu izjašnjavanja, (3) analiza eksplicitno izraženih sentimenata bez dodatne interpretacije ili znanja. Ivanuš i Ivanuš (2016), koji su istraživanje proveli nad komentarima ispod članaka portala *Jutarnji.hr* i *Vecernji.hr*, prikazali su primjenu alata za analizu sentimenta na hrvatskome jeziku *Slavomjer*. Zaključili su da alat

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

ne prepoznaje kontekst i negaciju, te ne može prepoznati ironiju i sarkazam, što ga čini manje vjerodostojnim za izvršavanje toga zadatka. Ovdje valja napomenuti da je detekcija ironije i sarkazma još uvijek otvoreno pitanje u području obrade prirodnoga jezika za sve jezike – ne samo hrvatski jezik.

Istraživanja na temu pandemije bolesti COVID-19 primjenom kombinacije metoda tematskoga modeliranja i analize sentimenta proveli su Chandrasekaran i sur. (2020) analizirajući objave na društvenoj mreži *Twitter*. Klasificirali su objave u 26 užih i 10 širih tema te su pratili promjenu sentimenta prije i nakon proglašenja pandemije. Otkrili su zanimljiv podatak da se za neke teme (poput širenja i rasta broja slučajeva te simptoma) održao negativan sentiment, dok se za druge teme (poput prevencije te liječenja i oporavka) s vremenom promijenio sentiment iz negativnog u pozitivni. Primjenom navedenih metoda te analizirajući novinske članke portala *Tportal.hr* na temu COVID-19, Pandur i sur. (2021) kategorizirali su članke u 9 tema te zaključili da su sve teme povezane dominantno s negativnim sentimentom. Istu kombinaciju navedenih metoda primijenili su de Melo i Figueiredo (2021) proučavajući razvoj sentimenta u novinama i na društvenim mrežama na temu pandemije COVID-19 u Brazilu te otkrili da su česta spominjanja određenoga lijeka povezana s visokom političkom polarizacijom tijekom pandemije.

Razvoj specijaliziranih novinskih korpusa i njihova analiza predmet su mnogobrojnih istraživanja. Na primjer, Weir i Anagnostou (2007) u svojoj studiji istražuju primjenu korpusne analize škotskih dnevnih novina iz 2005. godine. U njoj autori opisuju razvoj specijaliziranoga novinskog korpusa s pomoću raznih dostupnih alata, prikazuju zastupljenost različitih fenomena u korpusu te uspoređuju obilježja pojedinih lingvističkih fenomena u novinskom i općem korpusu. Brindle (2015) je pak za analizu diskursa koristio korpus članaka iz dvoje tajvanskih novina na engleskome jeziku objavljenih u razdoblju od šest mjeseci nakon početka studentskoga Pokreta suncokreta 2014. godine, a istraživanje je proveo s ciljem utvrđivanja frekvencijske distribucije, kolokacija i konkordancija. Almazán-Ruiz i Orrequia-Barea (2020) analizirali su korpus glavnih naslova „ozbiljnih i senzacionalističkih novina u Velikoj Britaniji" objavljenih uslijed izbijanja pandemije COVID-19. Pearman i sur. (2021), prateći 102 novinska izvora iz 50 država na 11 jezika, primijetili su da se medijsko izvještavanje o bolesti COVID-19 smanjuje unatoč produbljivanju krize.

Svrha ovog rada jest predstaviti metodologiju, alate i rezultate usporedne računalne analize *online* članaka: od prikupljanja dokumenata i čišćenja jezičnih podataka za razvoj specijaliziranoga korpusa članaka, do prikaza korištenih alata i usporedne statističke analize korpusa. U radu su uspoređeni članci s jednoga portala objavljeni predpandemijske 2019. godine s člancima s istoga portala iz pandemijske 2020. godine. Istraživanje je provedeno s ciljem utvrđivanja zastupljenosti fenomena obrađenih u sadržaju članaka. S obzirom na to da je javnost 2020. godine iščekivala nove informacije o trenutnom stanju i saznanjima o bolesti COVID-19, pretpostavka je da su portali bili pod pritiskom brže objavljivati informacije tijekom cijele te godine. Pod tom pretpostavkom uspoređeni su osnovni statistički podatci o dužini članaka, očekujući da će članci iz 2020. godine biti kraći od članaka iz 2019. godine. Osim toga, s obzirom na to da je Svjetska zdravstvena organizacija

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

11. ožujka 2020. godine bolest COVID-19 proglasila pandemijom,[1] a bolest je uspjela zahvatiti sve kontinente,[2] cilj je bio utvrditi zastupljenost domaćih i međunarodnih gradova te susjednih i ostalih država u člancima iz 2019. i 2020. godine radi utvrđivanja mogućega pomaka fokusa s domaćih tema i tema iz susjedstva na globalne teme. Za analizu je odabran portal *Index.hr*,[3] koji je prema Reutersovu institutu za istraživanje novinarstva bio vodeći *online* brend u Hrvatskoj početkom 2020. godine (Newman i sur., 2020, str. 66). U istraživanju su korišteni alati *Sketch Engine* te razne *Python* skripte za obradu i vizualizaciju podataka koji automatizacijom olakšavaju i ubrzavaju analizu tekstova.

U nastavku rada slijedi cjelina posvećena osnovnim terminima iz područja obrade prirodnoga jezika koji su korišteni u ovom radu. Zatim su predstavljena istraživačka pitanja i hipoteze, metodologija te specijalizirani korpusi razvijeni za potrebe ovog istraživanja. U središnjem dijelu rada predstavljeni su rezultati analize i njihova interpretacija, nakon čega slijede zaključci i prijedlozi budućih istraživanja.

OSNOVNI TERMINI IZ PODRUČJA OBRADE PRIRODNOGA JEZIKA

U ovom su poglavlju ukratko predstavljeni pojmovi iz područja obrade prirodnoga jezika, računalne lingvistike i analitike teksta koji su korišteni u radu. Ta su tri područja interdisciplinarna, a njihove se metode, tehnike, pristupi i alati često međusobno preklapaju. Prije svega, zajednički im je predmet proučavanja, a to je tekst pisan prirodnim jezikom, tj. jezikom koji je nekome materinski (npr. hrvatski jezik).[4] Što se tiče zajedničkih tehnika, sva tri područja primjenjuju statističke tehnike te tehnike strojnog učenja. Jedan od njihovih zajedničkih ciljeva jest iz nestrukturiranih podataka oblikovati i strukturirati informacije i znanje. Među tim interdisciplinarnim područjima ne postoji nužno konsenzus o definicijama svih termina. U ovome se radu ne raspravlja o spornim definicijama već se polazi od definicija pojmova iz Pojmovnika[5] koji se nalazi na stranicama alata *Sketch Engine* (alata korištenog za potrebe ovog istraživanja) – jer autorica smatra da je u sklopu istraživanja, radi jasnoće i točnosti interpretacije rezultata, ključno prikazati termine onako kako su implementirani u alatima.

Pojavnica (engl. *token*) najmanja je jedinica koja sačinjava korpus. U pojavnice su uključeni svi oblici riječi koji se nalaze u tekstu, interpunkcijski znakovi, brojevi, kratice te sve ostalo što se nalazi među bjelinama. Postoje dvije vrste pojavnica: riječi (engl. *word*) i neriječi (engl. *non-word*). Bjeline nisu pojavnice. Prilikom brojanja koliko pojavnica sadržava korpus uključivat će se sve što se nalazi među bjelinama, neovisno o tome

[1] WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[2] Prvi potvrđeni slučaj zabilježen je u Kini 31. prosinca 2019. godine. Prvi potvrđeni slučaj zabilježen je u Hrvatskoj 20. veljače 2020. godine. Antarktika je posljednji kontinent na kojem su zabilježeni prvi slučajevi i to u prosincu 2020. godine, gotovo godinu dana nakon prvih otkrivenih slučajeva u Kini (Izvor: Reportan brote de coronavirus en base chilena en la Antártida, 21.12.2020., Infobae, https://www.infobae.com/america/agencias/2020/12/21/reportan-brote-de-coronavirus-en-base-chilena-en-la-antartida/)

[3] https://www.index.hr/

[4] Ključna je značajka prirodnoga jezika njegova višeznačnost. Nasuprot prirodnim jezicima formalni su jezici koje oblikuju ljudi kako bi jednoznačno komunicirali (npr. programski jezici omogućuju jednoznačnu komunikaciju s računalom, prometni znakovi omogućuju jednoznačnu komunikaciju ljudi u prometu i sl.).

[5] Glossary. https://www.sketchengine.eu/guide/glossary/

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

ponavlja li se neka pojava ili ne.[6] Tokenizacija ili opojavničenje proces je rastavljanja teksta na pojavnice,[7] dok se računalni alat koji automatski vrši taj proces zove tokenizator.[8]

Riječ (engl. *word*)[9] je vrsta pojavnice koja počinje nekim slovom abecede (npr. 'pandemiji', 'N1' itd.).

Neriječ (engl. *non-word*)[10] je pojavnica koja ne počinje nekim slovom abecede (npr. '2012.', '3D', ':' itd.).

U Pojmovniku alata *Sketch Engine* nedostaje definicija termina *različnica* (engl. *type*), ali iz dokumentacije je vidljivo kako autori alata koriste taj termin. Različnica je jedinstveni oblik pojavnice koji se prilikom izračunavanja broja različnica računa samo jednom iako se u korpusu pojavljuje više puta. Na primjer, rečenica 'I danas i sutra kiša će biti češća na Jadranu i uz njega' sadržava 14 pojavnica i 12 različnica.[11]

Lema je osnovni oblik riječi koji se najčešće nalazi u rječnicima.[12] Lematizacija je proces dodjeljivanja leme svakoj pojavnici u korpusu, a alat koji automatski vrši taj proces zove se lematizator. Lematizirani korpus omogućava pretraživanje prema lemi te vraćanje svih oblika te riječi.[13] Lematizacija je bitna obrada kod flektivno kompleksnih jezika kao što je hrvatski. Time se mogu grupirati svi oblici iste leme, npr. lema 'pandemija' povezana je sa svim ostalim oblicima te riječi: 'pandemije', 'pandemiji', 'pandemiju', 'pandemijo', 'pandemijom', 'pandemijama'. Osim za grupiranje, lematizacija je korisna za lakšu detekciju ključnih riječi i termina u korpusu.

Ključne riječi i termini pojavnice su i izrazi koji su tipični za promatrani korpus jer se u tom korpusu pojavljuju češće nego u općem jeziku ili nekom drugom korpusu koji je definiran kao referentni. U ovom smislu referentan korpus je svaki korpus koji služi za usporedbu ključnih riječi i termina s promatranim korpusom,[14] a promatrani korpus je korpus iz kojega su ekstrahirane ključne riječi i termini. Ključne riječi i termini mogu služiti za definiranje ili razumijevanje teme korpusa.[15] Ključne riječi sadržavaju samo jednu pojavnicu, dok termini zadrže dvije ili više pojavnica.

Automatska ekstrakcija (ključnih riječi i termina) proces je automatske identifikacije specifičnoga vokabulara određenog teksta.[16] Alat *Sketch Engine* kombinira statistiku i lingvističke kriterije za automatsku ekstrakciju usporedbom čestote pojavljivanja riječi i izraza u promatranom korpusu naspram referentnoga ili nekoga drugoga korpusa koji je definiran kao referentni. Osim češćega pojavljivanja u promatranome korpusu, termini

6 „token"u *Glossary*. https://www.sketchengine.eu/my_keywords/token/
7 „tokenization"u *Glossary*. https://www.sketchengine.eu/my_keywords/tokenization/
8 „tokenizer" u *Glossary*. https://www.sketchengine.eu/my_keywords/tokenizer/
9 „word" u *Glossary*. https://www.sketchengine.eu/my_keywords/word/
10 „non-word" u *Glossary*. https://www.sketchengine.eu/my_keywords/non-word/
11 Sve pojavnice se pojavljuju samo jednom, osim pojavnice „i" koja se pojavljuje tri puta.
12 „lemma" u *Glossary*. https://www.sketchengine.eu/my_keywords/lemma/
13 „Lemmatization" u *Glossary*. https://www.sketchengine.eu/my_keywords/Lemmatization/
14 „reference corpus" u *Glossary*. https://www.sketchengine.eu/my_keywords/reference-corpus-2/
15 „Keyword and term extraction" u priručniku *Quick start guide*. https://www.sketchengine.eu/quick-start-guide/keywords-and-terms-lesson/
16 „term extraction" u *Glossary*. https://www.sketchengine.eu/my_keywords/term-extraction/

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

moraju ispunjavati dodatni uvjet koji ovisi o jeziku teksta. U sklopu alata za određene su jezike raspisane gramatike koje definiraju dopuštenu strukturu termina za taj jezik.[17] Npr. u hrvatskome jeziku termini često imaju strukturu *pridjev* + *imenica* ili *imenica* + *imenica*. Alat nudi mogućnost automatske ekstrakcije termina za korpuse hrvatskoga jezika.

ISTRAŽIVAČKA PITANJA I HIPOTEZE

Cilj ovoga rada jest računalnim metodama analizirati osnovne značajke članaka na primjeru portala *Index.hr* te usporediti članke iz predpandemijske 2019. godine s člancima iz pandemijske 2020. godine. Pod pretpostavkom da su portali u pandemijskoj godini bili pod pritiskom što brže informirati javnost o novim saznanjima o bolesti COVID-19 – zbog čega je brzina objave informacija bila veća nego prethodne godine te je provjera informacija bila otežana ili nemoguća u novonastaloj situaciji – uspoređene su razne statistike koje opisuju dužinu članaka prateći je li došlo do njihova skraćivanja. Dodatni je cilj bio usporediti zastupljenost domaćih i stranih gradova te susjednih i ostalih država u analiziranim člancima radi praćenja pomaka u pokrivenosti različitih lokacija u novinarstvu. S obzirom na to da je pandemija bolesti COVID-19 zahvatila sve kontinente, očekivano je smanjenje zastupljenosti domaćih gradova nauštrb inozemnih te smanjenje zastupljenosti susjednih država nauštrb ostalih. Iznimno je očekivano da će se potresom pogođeni grad Zagreb jednako ili češće spominjati u pandemijskoj godini zbog izvještavanja javnosti o aktualnim događajima vezanim uz obnovu grada. Konačni cilj rada jest obrazložiti primjerenost metoda i alata iz obrade prirodnoga jezika za istraživanje pokrivenosti tema medijskihčlanaka automatskom ekstrakcijom ključnih riječi i termina upotrebom alata *Sketch Engine*.

Rad se zasniva na sljedeće tri hipoteze i istraživačkome pitanju:

H1: Članci portala Index.hr iz predpandemijske 2019. godine duži su od članaka istoga portala iz pandemijske 2020. godine.
H2: U člancima portala Index.hr iz predpandemijske 2019. godine zastupljenije su države u susjedstvu Republike Hrvatske od ostalih država, dok su u člancima istoga portala iz pandemijske 2020. godine zastupljenije ostale države od susjednih.
H3: U člancima portala Index.hr iz predpandemijske 2019. godine zastupljeniji su tuzemni gradovi od inozemnih, dok su u člancima istoga portala iz pandemijske 2020. godine zastupljeniji inozemni gradovi od tuzemnih.
IP1: Opcija automatske ekstrakcije ključnih riječi i termina u alatu *Sketch Engine* primjerena je za identifikaciju specifičnih tema kojima se bave medijski članci.

METODOLOGIJA

U ovoj cjelini rada slijedi prikaz tehnika i metoda koje su primijenjene pri prikupljanju i obradi podatkovnoga skupa koji je popraćen opisom razvoja specijaliziranih korpusa utemeljenih na uzorku članaka s portala Index.hr. U nastavku su ukratko prikazane dodatne

[17] „term" u *Glossary*. https://www.sketchengine.eu/my_keywords/term/

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

kvantitativne metode, tehnike i mjere korištene za ispitivanje i evaluaciju hipoteza i istraživačkoga pitanja.

Podatkovni skup i razvoj specijaliziranih korpusa

U istraživanju su korištene metode i tehnike iz područja obrade prirodnoga jezika za razvoj specijaliziranih korpusa polazeći od primjera dobre prakse (Sinclair, 2004). Za potrebe istraživanja odabran je portal *Index.hr*, koji je prema Reutersovu institutu za istraživanje novinarstva vodeći *online* brend u Hrvatskoj početkom 2020. godine (Newman i sur., 2020, str. 66). Razvijena su dva korpusa koja sadržavaju članke objavljene u kategoriji „Vijesti":[18] jedan korpus uključuje članke objavljene u predpandemijskoj 2019. godini (*Index-vijesti-2019*), a drugi u pandemijskoj 2020. godini (*Index-vijesti-2020*). Korpusi su izrađeni na sljedeći način.

Izračunata je veličina uzorka, odnosno broj članaka koji bi bio reprezentativan za odabrani portal u promatranoj godini. Budući da je nepoznat sveukupni broj članaka u kategoriji „Vijesti" koji su objavljeni promatranih godina, radi konzervativnosti određena je veličina populacije beskonačna. Potom je izračunata veličina reprezentativnog uzorka uzimajući u obzir veličinu populacije, granicu pogreške od 5 % te razinu pouzdanosti od 95 %. Dobiveni broj od 385 zaokružen je, ponovno radi konzervativnosti, na 500 te predstavlja broj potrebnih nasumičnih članaka u svakom od specijaliziranih korpusa kao dovoljnu količinu članaka koji čine reprezentativan uzorak dotičnoga portala za promatrane godine.

Za izgradnju korpusa korišteno je nekoliko alata, a primarno *Sketch Engine*[19] – alat za razvoj i upravljanje korpusima te (računalnu) analizu teksta (Kilgrarriff i sur., 2004; Kilgarriff i sur., 2014). U sklopu alata moguće je izgraditi jednojezični ili višejezični korpus na dva načina: pronalaskom tekstova na internetskim stranicama i učitavanjem vlastitih dokumenata. Alat također nudi dodatne mogućnosti obrade korpusa koje ovise o jeziku, pa tako za hrvatski jezik nudi npr. označavanje vrsta riječi i morfosintaktičko označavanje (engl. *morphosyntactic description*, MSD), lematizaciju, skice riječi (engl. *word sketch*), konkordancije, ekstrakciju nazivlja i dr.[20] Opcija izgradnje mrežnih korpusa temelji se na *WebBootCaT* tehnologiji (Baroni, 2006), koja ima mogućnost definiranja mrežne domene ili poddomene s koje zatim alat automatski preuzima internetske stranice na toj domeni ili poddomeni. Definirana je poddomena https://www.index.hr/vijesti. Osim definiranja poddomene nije moguće definirati neka druga ograničenja (npr. godina objave), što rezultira time da korpus sadržava uglavnom nedavno objavljene članke. To je svakako dobar pristup algoritma koji prednost daje nedavno objavljenim člancima kako bi se mogle istraživati suvremene i aktualne teme. Međutim, kako je sastavljanje korpusa provedeno u veljači 2021. godine, taj pristup nije optimalan za ovo istraživanje. Osim toga, u korpusu su se nalazili i članci koji nisu iz kategorije „Vijesti" jer algoritam radi po principu da prikuplja internetske stranice na koje upućuju već prikupljene stranice. To se događa jer stranice na koje neka druga stranica upućuje vjerojatno sadržavaju istu ili sličnu temu te program

[18] https://www.index.hr/vijesti
[19] https://www.sketchengine.eu/)
[20] https://www.sketchengine.eu/corpora-and-languages/croatian-text-corpora/

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

time širi sastav korpusa, uz ograničenje da ne izlazi izvan glavne domene (u ovom slučaju domene https://www.index.hr/). Zbog toga sastav korpusa nije u potpunosti odgovarao potrebama ovog istraživanja. Tablica 1. prikazuje sastav prve inačice korpusa *Index-vijesti*. Prikupljena su 434 članka iz 2020. godine iz kategorije „Vijesti", a samo 69 iz iste kategorije objavljenih 2019. godine, od sveukupno prikupljenih 1180 članaka. Drugim riječima, tek je oko 6 % prikupljenih članaka objavljeno u predpandemijskoj 2019. godini, a oko 38% u pandemijskoj 2020. godini. Veliku većinu ostalih prikupljenih članaka čine članci objavljeni 2021. godine, a mali dio iz godine 2018. ili prije. Zbog toga je odlučeno da će se za izgradnju korpusa *Index-vijesti-2020* pokrenuti mogućnost povećanja korpusa istim alatom radi prikupljanja članaka koji nedostaju. Prikupljeno je dodatnih 1808 članaka od kojih je trebalo ekstrahirati članke objavljene 2020. godine. Drugom iteracijom prikupljanja internetskih stranica uspješno je prikupljeno 500 članaka objavljenih 2020. godine za razvoj korpusa *Index-vijesti-2020*.

Tablica 1. **Sastav prve inačice korpusa *Index-vijesti***

Godina objave članka	Broj članaka iz kategorije „Vijesti"
2019.	69 (5,85 %)
2020.	434 (37,78 %)
Ostalo	677 (57,37 %))
Ukupno	1180 (100 %)

Budući da za izgradnju korpusa iz 2019. godine nije bilo moguće prikupiti dovoljan broj članaka isključivo putem alata *Sketch Engine*, primijenjene su tehnike automatskoga prikupljanja potrebnih URL adresa na članke objavljene predpandemijske godine. Odlučeno je prikupiti 750 URL adresa, a da korpus sačinjava prvih 500 nasumično odabranih članaka koji ispunjavaju dva uvjeta: da su objavljeni 2019. godine i da su objavljeni u kategoriji „Vijesti". Za prikupljanje URL adresa članaka upotrijebljen je modul *googlesearch*[21] iz *Python* biblioteke koji omogućuje automatsko pretraživanje Google tražilice te vraća definiran broj prvih rezultata upita. Definirano je da se prikupi prvih pet rezultata upita. Dodatno je generirano 150 nasumičnih datuma u 2019. godini koji će se koristiti u upitu. Upit za pretraživanje postavljen je tako da je definiran datum te su ograničeni rezultati na definiranu poddomenu (primjer upita: „03.06.2019. site:www.index.hr/vijesti"). Napisana je skripta u *Pythonu* tako da se vrši iteracija prema datumu, postavljaju upiti na Google tražilicu te pohranjuje u datoteku prvih pet rezultata, čime je prikupljeno 750 URL adresa s definirane poddomene. Nakon toga je uslijedila izgradnja mrežnog korpusa *Index-vijesti-2019* primjenom opcije u Sketch Engine alatu koja omogućava učitavanje popisa URL adresa poželjnih za sastav korpusa. Time je prikupljeno 750 članaka za korpus iz predpandemijske godine. Ako se u korpusu nije nalazio datum objave članka, takav članak je uklonjen iz korpusa. Također su uklonjeni članci koji nisu objavljeni 2019. godine. Iz preostalih članaka nasumičnim odabirom odabrano je prvih 500 članaka za predpandemijski korpus.

[21] https://pypi.org/project/googlesearch-python/

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

Nakon prikupljanja članaka za specijalizirane korpuse uslijedila je faza čišćenja. Svaki članak nalazi se u zasebnom XML elementu s atributima koji sadrže informaciju o URL adresi članka te neke dodatne atribute o identifikaciji i nazivu datoteke, koji pak ovisi o organizaciji poslužitelja na kojem se nalazi portal. Zadržan je samo atribut s URL adresom članka kao jedinstveni identifikator članka, a poluautomatskom metodom dodan je atribut s informacijom o datumu objave članka (vidi Primjer 1).

```
<index-2020-v6 url="https://www.index.hr/vijesti/clanak/sto-se-to-dogadja-sa-zrakom-u-
zagrebu/2148207.aspx" date='2020-01-15'>
```

Primjer 1.
Primjer korijenskoga elementa članka iz korpusa *Index-vijesti-2020*

Svaki odlomak[22] članka nalazi se u zasebnom XML elementu. Izbrisani su oni odlomci koji su sadržavali informaciju samo o autoru i/ili izvoru članka, datumu objave,[23] broju dijeljenja članaka te koji su sadržavali standardne tekstove koji ne pridonose statističkoj jezičnoj analizi članaka.[24] Također su izbrisani odlomci koji su sadržavali tekst na nekom drugom jeziku, pretežito engleskom. To su uglavnom tekstovi s Twittera koji su uključeni u tekst članka. Zapaženo je da je tekstova na drugim jezicima bilo malo u korpusu iz 2019. godine, dok ih je u korpusu iz 2020. godine mnogo više. Točni podatci o zastupljenosti dijelova tekstova na drugim jezicima nisu dostupni, ali bi se mogli naknadno rekonstruirati za potrebe budućeg istraživanja. Nakon što su oba korpusa očišćena, u svakom je korpusu nasumično odabrano po 500 članaka iz kategorije „Vijesti" koji su objavljeni promatrane godine te su spremni za statističku jezičnu analizu.

Alat *Sketch Engine* nudi dodatne mogućnosti obrade korpusa koje ovise o tehnologiji implementiranoj u pozadini alata koja može ovisiti o jeziku. Tako alat za hrvatski jezik nudi označavanje vrsta riječi i morfosintaktičko označavanje, lematizaciju, skice riječi, konkordancije, ekstrakciju ključnih riječi i termina i dr. Korpusi su automatski vertikalizirani, tokenizirani, lematizirani, označeni morfosintaktičkim oznakama[25] te rastavljeni na rečenice.

Za potrebe istraživanja provjerene su oznake vrsta riječi, dok se u detaljnije morfosintaktičke oznake nije ulazilo jer bi provjera MSD oznaka oduzela previše vremena (npr. samo imenica nosi dodatnih pet podataka o vrsti imenice, rodu, broju, padežu i živosti). Na primjer, riječ 'covid-19' često je imala oznaku pridjeva ili priloga iako je bila u funkciji imenice. Uz provjeru i ispravak vrsta riječi izvršena je provjera i ispravak lematizacije jer je za točno izvršavanje automatske ekstrakcije ključnih riječi i termina potrebno da su pojavnice povezane s ispravnim lemama. Navedeno je samo nekoliko pogrešaka kod

[22] Kraj odlomka znak je za prelazak u novi red. Tako se u zasebnim odlomcima, osim odlomaka u užem smislu, nalaze i glavni naslovi (šlageri A) te mali naslovi (šlageri B i mali šlageri).
[23] Ovaj je korak napravljen nakon što je datum objave članka stavljen kao atribut koji opisuje članak.
[24] Na primjer, „Opširnije", „Tekst se nastavlja ispod oglasa.", „Znate li nešto više o temi ili želite prijaviti grešku u tekstu?", „Stavovi izneseni u kolumnama i komentarima su osobni stavovi autora i ne odražavaju nužno stav redakcije Index.hr portala", „Želite li momentalno primiti obavijest o svakom objavljenom članku vezanom uz koronavirus instalirajte Index.me aplikaciju i pretplatite se besplatno na tag: koronavirus" i sl.
[25] Korpusi su označeni MULTEXT-East oznakama V5, koje su *de facto* standard za označavanje morfosintaktičkih oznaka za hrvatski jezik. Sadrže informaciju o vrsti riječi i ostale morfosintaktičke informacije. Više o oznakama na http://nl.ijs.si/ME/Vault/V5/msd/html/msd-hr.html

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

lematizacije budući da analiza pogrešaka nadmašuje okvire ovoga rada. Neke od primijećenih pogrešaka su sljedeće: 'Alemka' koja je lematizirana kao 'alemko'[26], 'Antifa' kao 'antif', 'covid-19' kao 'covida-19', 'Donald' kao 'donaldo', 'Trump' kao 'trumpe' te 'Vili' kao 'vila'.

Nakon opisanih obrada dobivene su konačne inačice korpusa *Index-vijesti-2019* i *Index-vijesti-2020* koji su neznatno različite veličine, što omogućava njihovu lakšu usporedbu. Korpus iz predpandemijske 2019. godine sadrži 465 788 pojavnica, a korpus iz 2020. 444 546 pojavnica. Svaka pojavnica uključuje sljedeće oznake: MSD oznaka, lema s oznakom za vrstu riječi, lema (vidi Primjer 2). Za detaljniji statistički opis korpusa vidi Tablicu 2. u šestoj cjelini rada.

```
<p>
<s>
Što         Pi3n-n      što-p       što
se          Px—sa       sebe-p      sebe
to          Pd-nsn      taj-p       taj
događa      Vmr3s       događati-v  događati
sa          Si          sa-s        sa
zrakom      Ncmsi       zrak-n      zrak
u           Sl          u-s         u
Zagrebu     Npmsl       zagreb-n    zagreb
<g/>
?           Z           ?-z         ?
</s>
</p>
```

Primjer 2.
Primjer odlomka (naslova članka) iz korpusa *Index-vijesti-2020*

METODE I TEHNIKE KVANTITATIVNOG ISTRAŽIVANJA

U radu je primijenjena deskriptivna statistika za testiranje hipoteza usporedbom zastupljenosti različitih fenomena u specijaliziranim korpusima članaka portala *Index.hr* razvijenih za potrebe istraživanja. Deskriptivna statistika je „[o]naj dio statističkih metoda koji se bavi opisivanjem činjenica, dobivenih opažanjem ili mjerenjem neke pojave" (Kolesarić i Petz, 2003, str. 35). Ovim se metodama ne dokazuju uzročno-posljedične veze, a interpretacije rezultata istraživanja ne uzimaju u obzir širi kontekst.

Frekvencijske distribucije koriste se za sažimanje glavnih značajki podatkovnoga skupa te ih se može opisati i kao empirijski ekvivalent vjerojatnosnoj distribuciji (Everitt i Skrondal, 2002, str. 174). U sklopu ovog istraživanja mjerene su i uspoređivane frekvencijske distribucije sljedećih fenomena: pojavnica, riječi, rečenica, odlomaka, različnica, lema, gradova i država.

[26] Sve leme su pisane malim slovom.

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

Omjer različnica i pojavnica (engl. *type-token ratio*, TTR) često se koristi kao jedna od kvantitativnih mjera rječničke raznolikosti. Iako ta mjera ima svoje nedostatke te je nestabilna pri usporedbi tekstova različitih dužina (vidi npr. Torruella i Capsada, 2013), u slučaju ovog istraživanja primjerena je jer se uspoređuju korpusi zanemarivo različitih veličina. Kod TTR mjere visok omjer upućuje na to da tekst sadrži mnogo različitih leksičkih jedinica, što znači da velik udio pojavnica u tekstu ima specifično značenje. Drugim riječima, visok omjer znači bogati vokabular. S druge pak strane, nizak omjer upućuje na to da se u tekstu nalazi malen broj specifičnih pojavnica te da su česte opće pojavnice (Westin, 2002, str. 77). Tako su Chaffe i Danielewicz (1987, str. 4 – 5) otkrili da je omjer različnica i pojavnica veći kod pisanoga jezika (0,22 u pismima, 0,24 u znanstvenim radovima) nego kod govorenoga (0,18 pri razgovoru i 0,19 pri predavanju). Također je primijećeno da se omjer smanjuje povećanjem teksta jer se povećanjem teksta mora više toga ponavljati, a poznavanje vokabulara je ograničeno i konačno (Torruella i Capsada, 2013).

Automatska ekstrakcija termina ima primjenu u prevoditeljstvu i terminologiji, ali i analitici teksta, gdje se koristi za dubinsku analizu nestrukturiranih tekstova – za što se upravo primjenjuje u ovom radu. Struktura termina ovisi o jeziku, pa se tako u hrvatskom termini često sastoje od pridjeva i imenice u nominativu (npr. Europska unija) ili imenice u nominativu i imenice u genitivu (npr. priziv savjesti) i sl. Stoga je bitno da je korpus označen na razini vrsta riječi ili čak da sadrži morfosintaktičke oznake. Za flektivno kompleksne jezike kao što je hrvatski kod ekstrakcije termina dodatno je bitna lematizacija koja grupira sve oblike iste leme (npr. termini „Europskoj uniji" i „Europskom unijom" povezuju se s kanonskim oblikom termina „Europska unija"). Alat *Sketch Engine* nudi mogućnost automatske ekstrakcije ključnih riječi i termina za hrvatski jezik tako da se promatrani korpus usporedi s nekim drugim referentnim (često općim) korpusom. Kao referentni korpus korišten je hrvatski mrežni korpus *hrWaC* (Ljubešić i Klubička, 2014). Ključne riječi poredane su silazno prema posebnoj mjeri (tzv. engl. *keyness score*) kojom se mjeri svojstvo pojavnice ili niza pojavnica da budu kandidati za ključnu riječ ili termin (više o mjeri u Kilgarriff, 2009).

REZULTATI ISTRAŽIVANJA

Za potrebe istraživanja razvijena su dva specijalizirana korpusa, svaki sastavljen od 500 članaka objavljenih na portalu *Index.hr*: jedan korpus sadrži članke iz predpandemijske 2019., a drugi članke iz pandemijske 2020. godine. Iz Tablice 2. vidljivo je da su korpusi neznatno različite veličine, što čini korpuse jednostavno usporedivima. U daljnjem prikazu temeljnoga statističkog opisa prvo će biti predstavljen broj koji se odnosi na korpus iz 2019. godine, a zatim broj koji se odnosi na korpus iz 2020. godine. Oba korpusa sadrže otprilike jednak broj pojavnica (465 788 naspram 444 546), riječi (400 358 naspram 384 109), rečenica (21 376 naspram 20 737) i odlomaka (10 888 naspram 11 552). Kao što je prethodno navedeno, u fazi čišćenja korpusa primijećeno je da se u korpusu iz 2020. godine nalazilo značajno više teksta pisanog nekim drugim jezikom osim hrvatskim. U većini je slučajeva to bio engleski jezik i uglavnom se radilo o objavama s *Twittera*. Razlika u 21 242 pojavnice među korpusima je neznačajna i iznosi manje od 5 % ukupne veličine korpusa, dok je razlika u 16 249 riječi također neznačajna i iznosi oko 4 % ukupne veličine kor-

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

pusa. Zanimljiv podatak je da pandemijski korpus jedino po broju odlomaka nadmašuje predpandemijski (11 522 naspram 10 888). Korpusi se značajno razlikuju u broju različnica i lema. Korpus *Index-vijesti-2019* sadržava preko 11 500 različnica više od korpusa iz 2020. godine (57 765 naspram 46 160) te gotovo 5000 lema više (23 852 naspram 18 866), što je vidljivo u Tablici 2. S obzirom na to da je razlika u broju različnica značajna, za očekivati je i rezultat TTR mjere koji je za korpus iz 2019. godine veći (0,12) od korpusa iz 2020. godine (0,10).

Tablica 2. **Statistički opis korpusa *Index-vijesti-2019* i *Index-vijesti-2020***

	Index-vijesti-2019	**Index-vijesti-2020**
Broj članaka	500	500
Broj pojavnica	465.788	444.546
Broj riječi	400.358	384.109
Broj rečenica	21.376	20.737
Broj odlomaka	10.888	11.552
Broj različnica	57.765	46.160
Broj lema	23.852	18.866
TTR	0,12	0,10

U Tablici 3. predstavljeni su podatci o prosječnoj dužini članka iz kojih je vidljivo da *Index-vijesti-2020* sadrži jedan do dva odlomaka više po članku od *Index-vijesti-2019* (23,10 naspram 21,78), međutim sadrži u prosjeku jednu rečenicu manje po članku (41,47 naspram 42,75). U tablici su također vidljivi i podatci o prosječnom broju pojavnica, riječi, različnica i lema na razini članka, gdje je primijećeno da su vrijednosti iz predpandemijske godine veći za svaki promatrani fenomen od vrijednosti iz pandemijske godine.

Tablica 3. **Prosječna dužina članka**

	Index-vijesti-2019	**Index-vijesti-2020**
Broj odlomaka	21,78	23,10
Broj rečenica	42,75	41,47
Broj pojavnica	931,58	889,09
Broj riječi	800,72	768,22
Broj različnica	115,53	92,32
Broj lema	47,70	37,73

Dodatno je izračunata prosječna dužina odlomaka, a vrijednosti su predstavljene u Tablici 4. Zamijećeno je da su odlomci vrlo kratki u oba korpusa, a sadrže otprilike jednu do dvije rečenice (1,96 za *Index-vijesti-2019* i 1,80 za *Index-vijesti-2020*). Zanimljiv je i podatak

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

da korpus iz 2019. godine sadrži otprilike tri do četiri riječi više po odlomku nego korpus iz 2020. godine (36,77 naspram 33,25).

Tablica 4. **Prosječna dužina odlomka**

	Index-vijesti-2019	**Index-vijesti-2020**
Broj rečenica	1,96	1,80
Broj pojavnica	42,78	38,48
Broj riječi	36,77	33,25
Broj različnica	5,31	4,00
Broj lema	2,19	1,63

Nadalje izračunata je prosječna dužina rečenice u oba korpusa te je primijećeno da je približno jednaka, a sadrži između 18 i 19 riječi (vidi Tablica 5.).

Tablica 5. **Prosječna dužina rečenice**

	Index-vijesti-2019	**Index-vijesti-2020**
Broj pojavnica	21,79	21,44
Broj riječi	18,73	18,52
Broj različnica	2,70	2,23
Broj lema	1,12	0,91

Budući da je bilo jednostavno izvući sve naslove preko URL-ova u zaglavlju, dodatno je izračunat i broj pojavnica u naslovima članaka. Uočeno je da oba korpusa imaju jednaku prosječnu duljinu naslova, koja iznosi između 10 i 11 pojavnica, s tim da korpus iz 2020. godine ima neznačajno duže naslove (10,60 za *Index-vijesti-2019* i 10,73 za *Index-vijesti-2020*).

Dodatno su prikazana oba korpusa u obliku oblaka riječi radi lakše vizualizacije oba korpusa (vidi Sliku 1 i 2).

Slika 1.
Oblak riječi korpusa *Index-vijesti-2019*

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

Slika 2.
Oblak riječi korpusa
Index-vijesti-2020

U sljedećoj su fazi iz oba korpusa izvučene informacije o deset najčešće spominjanih država i gradova. Za podatke o državama, uključena je i Europska unija iako je to međunarodna i nadnacionalna organizacija više europskih država.

Tablica 6. **Deset najčešće spominjanih država i učestalost njihova spominjanja**

Rang	Index-vijesti-2019	Index-vijesti-2020
1.	Hrvatska (1129)[27]	Hrvatska (1033)
2.	Europska unija (466)[28]	Europska unija (396)
3.	BiH (141)[29]	SAD (395)
4.	SAD (132)[30]	Kina (279)
5.	Njemačka (124)	Austrija, Njemačka (182)
6.	Srbija (85)	Italija (161)
7.	Finska (73)	Francuska (152)
8.	Irska (67)[31]	Velika Britanija (115)
9.	Kina (66)	Turska (88)
10.	Velika Britanija (62)	Švedska (62)

Iz Tablice 6. vidljivo je da se u oba korpusa najviše spominju Hrvatska i Europska unija, dok se ostatak popisa prilično razlikuje po učestalosti spominjanja država koje se nalaze na obje liste. Zanimljivo je da se u pandemijskom korpusu Kina spominje 279 puta, dok se u predpandemijskom spominje 66 puta. SAD se 2020. godine spominje 395 puta, a 2019. godine 132 puta. U predpandemijskom korpusu primjećuje se često spominjanje zemalja s kojima Republika Hrvatska graniči, i to BiH i Srbije, dok se te dvije države ne nalaze na

[27] Zbrojeni su brojevi pojavljivanja lema „hrvatska" (1052) (označena kao imenica) i „rh" (77).
[28] Zbrojeni su brojevi pojavljivanja lema „eu" (195), „europa" (162) i „unija" (109). Budući da je automatskom metodom izvučen podatak o državama i gradovima, a leme „europa" i „unija" mogu se odnositi i na druge termine od Europske unije, u budućem istraživanju potrebno je razriješiti tu višeznačnost.
[29] Zbrojeni su brojevi pojavljivanja lema „bih" (109) i „bosna" (32), ali ne i „hercegovina" (34). Obrazloženje tome je da ako se već u članku nalazio puni naziv Bosne i Hercegovine, uzet je onaj manji broj pojavljivanja jednog od dva entiteta. Postoji mogućnost da se u člancima pisalo samo o jednom entitetu bez spominjanja država BiH. Stoga je potrebno u budućem istraživanju razriješiti tu višeznačnost.
[30] Prikazan je broj samo leme „sad", ali ne i „amerika" jer je uvidom u korpus uočeno da se lema uglavnom ne odnosi na SAD.
[31] Budući da je automatskom metodom izvučen podatak o državama i gradovima, ovdje su uključena Republika Irska i Sjeverna Irska, odnosno cijeli otok Irske. Stoga je potrebno u budućem istraživanju razriješiti tu višeznačnost.

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

popisu iz 2020. godine. U korpusu *Index-vijesti-2020* od zemalja s kojima graniči RH nalazi se Italija, koja se ne nalazi na popisu iz 2019. godine. Ostale države koje se nalaze na popisu iz 2019. godine, a ne nalaze se na popisu iz 2020., jesu Finska i Irska, dok se isključivo na popisu iz 2020. nalaze još Austrija, Francuska, Turska i Švedska.

U Tablici 7. predstavljeno je deset najčešće spominjanih gradova u korpusima, iz koje je vidljivo da je najčešće spominjani grad Zagreb (425 puta u predpandemijskom i 151 puta u pandemijskom korpusu). Na popisu iz 2019. godine nalazi se sedam hrvatskih gradova (Zagreb, Split, Rijeka, Dubrovnik, Zadar, Osijek i Vukovar), dok se na popisu iz 2020. nalaze samo četiri (Zagreb, Vukovar, Split i Zadar). Od stranih gradova na popisu iz 2019. godine nalaze se Sarajevo, Bruxelles, London, New York i Beograd, dok se na pandemijskom popisu nalaze New York, Wuhan, Washington, Minneapolis, Pariz i London.

Tablica 7. **Deset najčešće spominjanih gradova i čestota njihova pojavljivanja**

Rang	Index-vijesti-2019	Index-vijesti-2020
1.	Zagreb (425)	Zagreb (151)
2.	Split (177)	New York (136)
3.	Rijeka (129)[32]	Wuhan (122)
4.	Dubrovnik (92)	Washington (87)
5.	Zadar (73)	Vukovar (58)
6.	Osijek (64)	Minneapolis (55)
7.	Sarajevo (47)	Pariz (53)
8.	Vukovar (46)	Split (40)
9.	Bruxelles, London, New York (35)	Zadar (39)
10.	Beograd (34)	London (36)

U fazi istraživanja korisnosti automatske ekstrakcije terminologije ekstrahirano je prvih 50 ključnih riječi te termina iz oba korpusa (vidi Tablicu 8.). Iz popisa je vidljivo da u predpandemijskom korpusu 32 od 50 ključnih riječi sačinjavaju uglavnom prezimena i imena ili posvojni pridjevi nastali od imena ili prezimena (npr. „kolinda", „plenković", „plenkovićev" i sl.) dok se u pandemijskom korpusu pojavljuju 20 od 50 puta. Primijećeno je da se u pandemijskom korpusu ključnih riječi isto 20 puta pojavljuje i nazivlje vezano za pandemiju bolesti COVID-19 (npr. koronavirus, samoizolacija, *lockdown*, epidemiološki i sl.).

Popis termina daje korisniji uvid u sintagme koje se češće koriste u proučavanim korpusima, iz čega se bolje mogu iščitati teme koje su se obrađivale u promatranim godinama. Može se primijetiti da se 2019. godine pisalo o raznim temama, npr. kurikularnoj reformi, koeficijentima složenosti poslova, Aspergerovu sindromu, seriji *Igra prijestolja* i sl. S druge pak strane, sintagme vezane za pandemiju bolesti COVID-19 sačinjavaju 39 od 50 termina.

[32] Budući da se lema „rijeka" može odnositi na grad Rijeku i na vodotok, u budućem istraživanju potrebno je razriješiti tu višeznačnost.

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

Tablica 8. **Prvih 50 ključnih riječi i termina oba korpusa**

	Ključne riječi		Termini	
	Index-vijesti-2019	Index-vijesti-2020	Index-vijesti-2019	Index-vijesti-2020
1	najviše	Koronavirus	lažna vijest	nošenje maski
2	kolinda	covid 19	vatikanski ugovor	pandemija koronavirusa
3	plenković	Trump	kurikularna reforma	nacionalan stožer
4	kuščević	Trumpov	stajati u odgovoru	stožer civilne zaštite
5	trump	Capak	smjer mora	kolektivan imunitet
6	plenkovićev	Samoizolacija	aspergerov sindrom	epidemiološka mjera
7	kbr	Pandemija	negativan bod	nositi maske
8	grabar-kitarović	Plenković	istanbulska konvencija	nov koronavirus
9	bunjac	Biden	istražni zatvor	epidemiološka situacija
10	epstein	Wuhan	povećanje koeficijenata	epidemija koronavirusa
11	uhljeb	Korona	koeficijent složenosti poslova	slučaj zaraze
12	pernar	zavadlav	plinarsko naselje	širenje koronavirusa
13	palfi	lockdown	islamska država	širenje virusa
14	bagdadi	markotić	imovinska kartica	nositi masku
15	thunberg	sars-cov-2	zavod za hitnu medicinu	nov soj
16	trumpov	macron	priziv savjesti	širenje zaraze
17	brexit	donald	igra prijestolja	pozitivan na covid 19
18	steubner	karantena	koeficijent složenosti	početak pandemije
19	duhaček	beroš	živ zid	nov slučaj
20	meteoalarm	najviše	visoka energija	civilna zaštita
21	mađarević	epidemiološki	odlučivanje o sukobu interesa	nacionalan stožer civilne zaštite
22	kalifat	pennsylvanija	složenost poslova	slučaj koronavirusa
23	škoro	paty	odlučivanje o sukobu	lažna vijest
24	gingivalis	alemka	hitna medicina	pozitivan na koronavirus
25	d8	novozaražen	indexov novinar	trumpov administracija
26	dron	epidemiolog	bijel patuljak	teoretičar zavjera
27	pecolaj	bidenov	unitedov let	borba protiv koronavirusa
28	đakić	epidemija	osoba s aspergerovim sindromom	stroža mjera
29	epsteinov	antifa	seks bez pristanka	brojenje glasova
30	amazonija	distanciranje	u smjeru mora	policijsko ubojstvo
31	bedić	lauc	vrlo visoka energija	trumpov stožer
32	vuličević	pence	kvaliteta zraka	bijela kuća
33	makjanić	minneapolis	broj ministarstava	preboljeti covid 19
34	šuica	qanon	plenkovićev vlada	znanstven savjet
35	kurikularan	hzjz	bruto plaća	kirurška maska

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

36	rora	respirator	detencijski centar	policijski sat
37	grško	migrant	kolona u smjeru	smrt na milijun
38	škaro	covid	električan bicikl	stopa smrtnosti
39	prstac	cdc	koncentracija ozona	smrt na milijun stanovnika
40	asperger	melania	korist računa	test na koronavirus
41	žuvić	floyd	umjetna inteligencija	širenje covida 19
42	aspergerov	r0	prekid trudnoće	pandemija covida 19
43	ćimić	penrose	pedofilski otok	crn život
44	yammat	hebdo	snimanje službene osobe	infektivna bolest
45	škorin	zaražen	prosvjed za klimu	policijsko nasilje
46	divjak	tvit	stan u zvonimirovoj ulici	teorija zavjere
47	greta	zaraza	naknada za uplate	broj smrti
48	index	kurz	naplaćivati naknadu	razvoj cjepiva
49	fazlagić	božinović	razlog iseljavanja	izvanredno stanje
50	ozon	elektorski	vozan park	kolona sjećanja

DISKUSIJA

H1: Članci portala Index.hr iz predpandemijske 2019. godine duži su od članaka istoga portala iz pandemijske 2020. godine.

Oba korpusa sadržavaju otprilike jednak broj pojavnica, riječi i rečenica (vidi Tablicu 2.), no predpandemijski korpus je ipak neznatno veći. Korpusi se u broju pojavnica razlikuju u manje od 5 % ukupne veličine korpusa, u broju riječi oko 4 %, a u broju rečenica oko 3 %. Jedan od mogućih razloga zašto je korpus iz 2020. godine neznatno manji jest veća prisutnost objava na drugim jezicima, uglavnom na engleskom jeziku s *Twittera*, koje su brisane u fazi čišćenja korpusa. Kao zanimljiv podatak zamijećeno je da pandemijski korpus jedino po broju odlomaka nadmašuje predpandemijski za otprilike 6 %. Međutim, korpusi se značajno razlikuju u broju različnica i lema. Predpandemijski korpus sadrži 25 % više različnica (57 765 naspram 46 160) i 26 % više lema (23 852 naspram 18 866) od pandemijskog. S obzirom na te rezultate, očekivana je razlika u rezultatu TTR mjere, koja je za predpandemijski korpus veća (0,12) od korpusa iz 2020. godine (0,10). Iako predpandemijski korpus sadrži više pojavnica od pandemijskog korpusa, razlika od 5 % premala je da se H1 smatra dokazanom. Također nije dokazana pretpostavka da je 2020. godine pritisak na portale da se što brže objave nove informacija o bolesti COVID-19 utjecao na objavljivanje kraćih članaka. Međutim, znatna razlika u broju različnica i lema te razlika u rezultatu TTR mjere upućuje na to da je vokabular u pandemijskom korpusu siromašniji što može biti posljedica toga da su članci pokrivali manje različitih tema nego u predpandemijskom korpusu. Navedenu interpretaciju manjeg rezultata TTR mjere u pandemijskom korpusu nadopunit će analiza provedena u svrhu odgovora na istraživačko pitanje.

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

H2: U člancima portala Index.hr iz pandemijske 2019. godine zastupljenije su države u susjedstvu Republike Hrvatske od ostalih država, dok su u člancima istoga portala iz pandemijske 2020. godine zastupljenije ostale države od susjednih.

Analiza podataka pokazala je da se na popisu deset najčešće spominjanih država (vidi Tablicu 6.) u predpandemijskom korpusu nalaze susjedne države BiH i Srbija, koje zajedno čine oko 10 % pojava s popisa. Međutim, te dvije države ne nalaze se na popisu deset najčešće spominjanih država u pandemijskom korpusu iz 2020. godine. Od susjednih se država nalazi Italija koja čini oko 5 % pojava s pandemijskoga popisa. Pretpostavka je da se pojavila na pandemijskom popisu zbog teške situacije vezane za pandemiju COVID-19 koja ju je zahvatila početkom 2020. godine. Zanimljiv je podatak da se u pandemijskom korpusu Kina spominje 4,22 puta više, SAD gotovo trostruko više, a Velika Britanija 1,89 puta više nego u predpandemijskoj 2019. godini. Dodatno je zanimljiv podatak da se Hrvatska spominje manje u pandemijskom popisu te čini sveukupno oko 34 % pojavljivanja, dok u predpandemijskom popisu čini oko 48 % pojava. Svakako je potrebno naglasiti da predpandemijski popis sadrži 2345 pojava, a pandemijski 3045. Konačno je zaključeno da je H2 dokazana te da su se Hrvatska i susjedne zemlje u predpandemijskom korpusu češće spominjale nego u pandemijskom korpusu. To je u skladu s pretpostavkom da su se članci u pandemijskom korpusu više bavili globalnim temama jer je sama pandemija globalna pojava.

H3: U člancima portala Index.hr iz predpandemijske 2019. godine zastupljeniji su tuzemni gradovi od inozemnih, dok su u člancima istoga portala iz pandemijske 2020. godine zastupljeniji inozemni gradovi od tuzemnih.

Analiza podataka deset najčešće spominjanih gradova (vidi Tablicu 7.) pokazuje da se hrvatski gradovi spominju u 84 % slučajeva u predpandemijskom korpusu, a u svega 37 % slučajeva u pandemijskoj 2020. godini. Drugim riječima, inozemni gradovi se u predpandemijskom korpusu spominju u samo 16 % slučajeva, dok se u panedmijskom spominju u 63 % slučajeva. Tri najčešće spominjana grada u predpandemijskom popisu tri su najveća grada prema Popisu najvećih gradova u RH s popisa stanovništva iz 2011. godine.[33] Zanimljivo je primijetiti i da se Zagreb na oba popisa nalazi kao najčešće spominjani grad. Međutim, u pandemijskom korpusu iz 2020. godine spominje se čak 2,8 puta rjeđe nego u 2019. godini. Ako se uzme u obzir činjenica da je početkom godine, 22. ožujka 2020., Zagreb pogodio razoran potres od 5,5 stupnjeva po Richteru, postavlja se pitanje zašto se potresom pogođeni grad spominje manje nego inače. Uz pretpostavku da vodeći *online* brend izvještava o svim važnim aktualnim događajima, može se zaključiti da nisu imali o čemu informirati javnost na temu obnove grada Zagreba. Alternativna interpretacija navedenih podataka može svjedočiti i o tome da je slaba zastupljenost vijesti o obnovi vezana uz uredničku politiku portala, no autorica je ipak sklonija prvom tumačenju. Konačno, H3 je dokazana te su se češće spominjali inozemni gradovi u pandemijskom korpusu od tuzemnih te da se obrnuto može tvrditi za predpandemijski korpus.

[33] Popis najvećih gradova u Republici Hrvatskoj preuzet je s popisa stanovništva iz 2011. godine i dostupan je na poveznici: https://www.dzs.hr/Hrv/censuses/census2011/results/htm/H01_06_01/H01_06_01.html

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

IP1: Opcija automatske ekstrakcije ključnih riječi i termina u alatu *Sketch Engine* primjerena je za identifikaciju specifičnih tema kojima se bave medijski članci.

Analiza podataka pokazala je da popis ključnih riječi (vidi Tablicu 8.) iz predpandemijskog korpusa sadrži u 64 % slučajeva prezimena i imena ili posvojne pridjeve nastale od prezimena ili imena, dok ta skupina čini samo 40 % pandemijskog korpusa. Također 40 % ključnih riječi iz pandemijskog popisa čini nazivlje vezano za pandemiju bolesti COVID-19. Informativniji su popisi termina iz kojih se bolje mogu iščitati specifične teme kojima se bave korpusi promatranih godina. Na primjer, primijećeno je da se 2019. godina bavila kurikulnom reformom, koeficijentima složenosti poslova, Aspergerovim sindromom, serijom *Igra prijestolja* i sl., dok se 78 % sintagmi iz pandemijske godine odnosi na pandemiju. Proučavajući broj različnica i lema za potrebe ispitivanja H1, primijećena je manja TTR mjera za pandemijsku godinu što je indiciralo da su članci u pandemijskoj godini pokrivali manje tema nego u predpandemijskoj godini. Međutim, automatska ekstrakcija termina nam omogućuje dubinsku analizu toga zaključka. Svakako treba imati na umu jedno bitno ograničenje automatske ekstrakcije termina, a to je činjenica da popis ključnih riječi i termina ne nudi uvid u sve sintagme koje se pojavljuju u korpusu, već samo u one sintagme koje su specifične za promatrani korpus. Time se ne dobiva uvid u sve teme koje se pojavljuju u korpusu, već samo u one teme koje se u promatranom korpusu pojavljuju češće nego u referentnom korpusu. Stoga je zaključeno da je opcija primjerena za identifikaciju specifičnih tema za promatrani korpus, ali ne i svih tema koje se pojavljuju u korpusu.

ZAKLJUČCI I BUDUĆA ISTRAŽIVANJA

Istraživanje je provedeno na dva specijalizirana korpusa razvijena upravo za potrebe istraživanja, a temelje se na 500 članaka u kategoriji „Vijesti" portala *Index.hr* koji je prema Reutersovu institutu za istraživanje novinarstva vodeći *online* brend u Hrvatskoj početkom 2020. godine. Jedan korpus temelji se na člancima objavljenima u predpandemijskoj 2019. godini, a drugi na temelju članaka objavljenima u pandemijskoj 2020. godini. Cilj rada jest kvantitativnim metodama, tehnikama i mjerama usporediti strukturu i sadržaj članaka iz dvaju korpusa. Za potrebe rada korišten je alat *Sketch Engine* koji služi za razvoj i analizu korpusa te razne Python skripte za obradu i vizualizaciju podataka.

Iz analize podataka nije dokazano da su članci portala *Index.hr* iz predpandemijske 2019. godine duži od članaka istoga portala iz pandemijske 2020. godine jer su korpusi tek neznatno različitih dužina. Međutim, prilikom te analize otkriveno je da je vokabular pandemijskog korpusa značajno siromašniji od predpandemijskog korpusa. Do sličnog su zaključka došli Beliga i sur. (2021) koji su proveli longitudinalnu analizu sadržaja povezanog uz bolest COVID-19 hrvatskih *online* medija primjenjujući razne metode obrade prirodnoga jezika. Zanimljivo je da su autori zaključili da je moguća posljedica velikog preklapanja često korištenih termina u prvih 13 mjeseci pandemije upravo uzak fokus izvješćivanja u određenim razdobljima. Dodatno je dokazano da su u člancima iz 2019. godine zastupljenije države u susjedstvu RH od ostalih država, dok su u člancima iz 2020. godine

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

zastupljenije ostale države od susjednih država RH. Također je dokazano da su u predpandemijskom korpusu zastupljeniji tuzemni gradovi od inozemnih, dok se suprotno može tvrditi za korpus iz 2020. godine. Prilikom ove analize primijećeno je da se Zagreb čak 2,8 puta rjeđe spominje u 2020. godini nego u prethodnoj, iako ga je početkom pandemijske godine pogodio razoran potres. Konačno, istražena je i primjerenost automatske ekstrakcije ključnih riječi i termina koja se nudi u alatu *Sketch Engine* za identifikaciju specifičnih tema kojima se bave promatrani korpusi. Zaključeno je da je alat primjeren za predloženu uporabu, uz upozorenje da se ne može dobiti pregled svih tema, već samo onih koje su karakteristične za promatrani korpus, odnosno u ovom slučaju za otkrivanje specifičnih tema koje je pokrivao portal *Index.hr* 2019. i 2020. godine. Istraživanja poput ovog mogu dati uvid u uredničku politiku medijskih portala, ali i u (ne)aktivnosti različitih razina vlasti i ostalih aktera o kojima mediji izvještavaju javnost te služiti kao kronika važnih aktualnih i zanimljivih događaja promatranog razdoblja.

Glavno ograničenje istraživanja u tome je što se zaključci odnose samo na jedan portal. Istraživanje bi bilo korisno proširiti na druge portale, na više promatranih godina te uključiti više članaka u korpusu. Time bi se mogle raditi usporedbe prema godinama, prema portalima ili njihovom kombinacijom. Metodologija i tehnike razvijene u sklopu ovog istraživanja omogućuju provedbu istraživanja ciljanih godina, pa čak i kraćeg razdoblja (npr. dan objave članaka). U planu je razvoj korpusa za 2021. godinu. Budući da je potkraj 2020. godine Banovinu, nažalost, pogodio razorniji potres od onog u Zagrebu iste godine, želja nam je utvrditi zastupljenost gradova, terminologije vezane za potres te proučiti je li i u 2021. godini pandemija dominantna tema. Iako osiromašenje vokabulara može zvučati alarmantno za budućnost hrvatskoga jezika, svakako je potrebno dulje vrijeme pratiti taj fenomen te utvrditi je li pojava ograničena na pandemijsku godinu ili će se protezati kroz duže razdoblje. Osim za provođenje ovakvih kvantitativnih istraživanja, korpusi (uz dodatnu obradu) se mogu primijeniti za provođenje analize diskursa, koja kombinira lingvističku i društvenu analizu teksta, za istraživanje fenomena infodemije[34] detaljnom kvalitativnom analizom informacija i svih njezinih izvedenica u medijskim člancima, kao i za druge vrste kvantitativnih i kvalitativnih istraživanja.

Literatura

>Almazán-Ruiz, E. i Orrequia-Barea, A. (2020). The British Press' Coverage of Coronavirus Threat: A Comparative Analysis Based on Corpus Linguistics. *Çankaya University Journal of Humanities and Social Sciences, 14*(1), 1-22.

>Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B. i Belyaeva, J. (2010). Sentiment Analysis in the News. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2216-2220.

>Baroni, M., Adam, K., Pomikálek, J. i Rychlý, P. (2006). WebBootCaT: a Web Tool for Instant Corpora. *Proceedings of the 12th EURALEX International Congress*, 123-131.

[34] O novosti termina „infodemija" govori da je u obliku samostalne rječničke natuknice uključen u Oxfordski engleski rječnik u travnju 2020. godine (https://public.oed.com/updates/new-words-list-april-2020/), a objavljen *online* dva mjeseca kasnije (https://www.oed.com/view/Entry/88407009). Termin je definiran kao „proliferacija raznolikih, često neutemeljenih informacija koje se odnose na krizu, kontroverze ili događaje, koji se brzo i nekontrolirano šire putem vijesti, *online* i na društvenim mrežama, a smatra se da pojačavaju nagađanja i zabrinutost javnosti". Iako se termin u tom značenju pojavio još 2003. godine, tek pojavom pandemije bolesti COVID-19 2020. godine počeo se često upotrebljavati u javnom diskursu, a fenomen je postao predmet raznih znanstvenih istraživanja (v. npr. Eysenbach, 2020; Medford i sur., 2020; Orso i sur., 2020; Zarocostas, 2020,).

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 6.6.2021.

>Beliga, S., Martinčić-Ipšić, S., Matešić, M., Vuksanović, I. P. i Meštrović, A. (2021). Infoveillance of the Croatian Online Media During the COVID-19 Pandemic: One-Year Longitudinal Study Using Natural Language Processing. *JMIR public health and surveillance*, *7*(12), e31540. https://doi.org/10.2196/31540
>Brindle, A. (2016). A corpus analysis of discursive constructions of the Sunflower Student Movement in the English-language Taiwanese press. *Discourse & Society, 27*(1), 3-19. https://doi.org/10.1177/0957926515605957
>Chafe, W. i Danielewicz, J. (1987). Properties of spoken and written language. *Technical Report No. 5*. University of California Carnegie Mellon University.
>Chandrasekaran, R., Mehta, V., Valkunde, T. i Moustakas, E. (2020). Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study. *Journal of Medical Internet Research*, *22*(10), e22624. https://doi.org/10.2196/22624
>Eysenbach, G. (2020). How to fight an infodemic: the four pillars of infodemic management. *Journal of Medical Internet Research*, *22*(6), e21820. https://doi.org/10.2196/21820
>Gamson, W. A. i Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology, 95*(1), 1-37. https://doi.org/10.1086/229213
>Everitt, B. i Skrondal, A. (2002). *The Cambridge dictionary of statistics* (Vol. 106). Cambridge University Press.
>Gozzi, N., Tizzani, M., Starnini, M., Ciulla, F., Paolotti, D., Panisson, A. i Perra, N. (2020). Collective response to media coverage of the COVID-19 pandemic on Reddit and Wikipedia: mixed-methods analysis. *Journal of Medical Internet Research*, *22*(10), e21597. https://doi.org/10.2196/21597
>Ivanuš, R. i Ivanuš, Ž. (2016). Učinkovitost Slavomjera, alata za analizu sentimenta, na primjeru Jutarnjeg i Večernjeg lista. *Zbornik radova Život u digitalnom dobu: Društveni aspekti, druga međunarodna znanstveno-stručna konferencija iz marketinga i komunikacija „Fedor Rocco"*: 354-366.
>Jacobi, C., Van Atteveldt, W. i Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89-106. https://doi.org/10.1080/21670811.2015.1093271
>Jurafsky, D. i Martin, J. H. (2008). *Speech and Language Processing.* Prentice Hall.
>Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of the Corpus Linguistics Conference.* https://doi.org/10.1007/s40607-014-0009-9
>Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. i Suchomel, V. (2014). The Sketch Engine: Ten Years on. *Lexicography (1)*, 7-36. https://doi.org/10.1007/s40607-014-0009-9
>Kilgarriff, A., Rychlý, P., Smrz, P. i Tugwell, D. (2004). The Sketch Engine.*Proceedings of the 11th EURALEX International Congress,* 105-115.
>Kolesarić, V. i Petz, B. (2003). *Statistički rječnik : Tumač statističkih pojmova*. Naklada Slap.
>Korenčić, D. (2019). *Računalni postupci za modeliranje i analizu medijske agende temeljeni na strojnome učenju* [Doktorska disertacija. Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva].
>Ljubešić, N. i Klubička, F. (2014). {bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9),* 29-35.
>Manning, C. i Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
>Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M. i Lehmann, C. U. (2020, srpanj). An "infodemic": leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. *Open forum infectious diseases,* 7(7): ofaa258). https://doi.org/10.1093/ofid/ofaa258
>de Melo, T. i Figueiredo, C. M. (2021). Comparing news articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach. *JMIR Public Health and Surveillance*, *7*(2), e24585. https://doi.org/10.2196/24585
>Mitkov, R. (ur.). (2004). *The Oxford Handbook of Computational Linguistics*. Oxford University Press.
>Newman, N., Fletcher, R., Schulz, A., Andı, S. i Nielsen, R. K. (2020). *Reuters Institute Digital News Report 2020.* Oxford: Reuters Institute for the Study of Journalism.

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

>Orso, D., Federici, N., Copetti, R., Vetrugno, L. i Bove, T. (2020). Infodemic and the spread of fake news in the COVID-19-era. *European Journal of Emergency Medicine*, *27*(5), 327-328. https://doi.org/10.1097%2FMEJ.0000000000000713
>Pandur, M. B., Dobša, J., Beliga, S., i Meštrović, A. (2021). Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal. *Proceedings of the 24th International Multiconference Information Society*, 155-158.
>Pearman, O., Boykoff, M., Osborne-Gowey, J., Aoyagi, M., Ballantyne, A. G., Chandler, P., Daly, M., Doi, K., Fernández-Reyes, R., Jiménez-Gómez, I., Nacu-Schmidt, A., McAllister, L., McNatt, M., Mocatta, G., Petersen, L. K., Simonsen, A. H. i Ytterstad, A. (2021). COVID-19 media coverage decreasing despite deepening crisis. *The Lancet Planetary Health*, *5*(1), e6-e7. https://doi.org/10.1016/S2542-5196(20)30303-X
>Sinclair, J. (2004). Corpus and Text - Basic Principles. U Martin Wynne (ur.) *Developing Linguistic Corpora: a Guide to Good Practice*. https://users.ox.ac.uk/~martinw/dlc/.
>Sketch Engine Guide, https://www.sketchengine.eu/guide/
>Torruella, J. i Capsada, R. (2013). Lexical statistics and tipological structures: a measure of lexical richness. *Procedia-Social and Behavioral Sciences,* 95, 447-454. https://doi.org/10.1016/j.sbspro.2013.10.668
>Weir, G. R. S. i Anagnostou, N. K. (2007). Exploring newspapers: a case study in corpus analysis. *Proceedings of ICTATLL 2007,* 12-19.
>Westin, I. (2002). *Language change in English newspaper editorials*. Brill | Rodopi. https://doi.org/10.1163/9789004334007
>Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

IZVORNI ZNANSTVENI RAD / DOI: 10.20901/ms.13.25.2 / PRIMLJENO: 16.6.2021.

WHEN NEWS SITES "CATCH" THE CORONAVIRUS: DEVELOPMENT AND COMPARATIVE ANALYSIS OF THE 2019 AND 2020 ARTICLES PUBLISHED ON THE *INDEX.HR* NEWS PORTAL

Petra Bago

ABSTRACT *The goal of this paper is to present the methodology, tools and results of comparative computational analysis of newspaper online articles: from the collection of documents and the cleaning of language data for the development of specialized corpora of newspaper articles, to the presentation of the tools used and the comparative statistical analysis of the corpora. The research was conducted on two specialized corpora developed precisely for the purpose of this research, based on 500 newspaper articles in the category "News" of the* Index.hr *news portal. One corpus is based on articles published in the pre-pandemic year 2019, and the other is based on articles published in the pandemic year 2020. By analyzing the data, we found that the vocabulary of the pandemic corpus is significantly poorer than the pre-pandemic corpus, that in 2020 less was written about the neighboring states of the Republic of Croatia than in 2019, and that the pre-pandemic corpus mentioned domestic cities more than the foreign ones, while the opposite can be argued for the pandemic corpus. Finally, we also investigated the adequacy of automatic term extraction to identify specific topics covered in the observed corpora.*

KEY WORDS

STATISTICAL CORPUS ANALYSIS, SPECIALIZED CORPUS, JOURNAL ARTICLES, SKETCH ENGINE, PYTHON, INDEX.HR

Author's note

Petra Bago :: Faculty of Humanities and Social Sciences, University of Zagreb **::** pbago@ffzg.hr