https://doi.org/10.31820/f.34.2.13

Mirjana Borucinsky, Irena Bogunović

CRPLJENJE ENGLESKIH RIJEČI IZ KORPUSA HRVATSKOGA JEZIKA[1]

dr. sc. Mirjana Borucinsky, Sveučilište u Rijeci, Pomorski fakultet
mirjana.borucinsky@pfri.uniri.hr orcid.org/0000-0002-1132-9720
dr. sc. Irena Bogunović, Sveučilište u Rijeci, Pomorski fakultet
irena.bogunovic@pfri.uniri.hr orcid.org/0000-0002-2956-7014

izvorni znanstveni članak
UDK 811.163.42'322
811.111'322
rukopis primljen: 23. travnja 2021; prihvaćen za tisak: 15. rujna 2022.

Kao globalni jezik modernoga doba engleski je postao dominantan jezik davatelj. Danas se smatra da hrvatski jezik najviše posuđuje upravo iz engleskoga. Utjecaj engleskoga jezika na hrvatski vidljiv je u različitim funkcionalnim stilovima te na gotovo svim jezičnim razinama, no najizraženiji je na leksičkoj razini. U novije vrijeme, posebice u medijima i na društvenim mrežama, sve se češće javljaju neprilagođene engleske riječi, tj. riječi koje su zadržale izvorni oblik, a kojima se po potrebi dodaju hrvatski afiksi. Za sada još uvijek ne postoje konkretni podaci o takvim riječima u hrvatskome jeziku. U cilju pronalaženja engleskih riječi, u drugim su se jezicima koristile različite metode, od ručnih klasifikacija i korištenja postojećih jezičnih resursa do razvoja novih alata i/ili resursa. Međutim, jezične tehnologije za hrvatski jezik još uvijek su nedostatno razvijene. Stoga je cilj ovoga rada ispitati mogućnosti nekih od postojećih alata i resursa za crpljenje engleskih riječi i stvaranje baze engleskih riječi. U tu svrhu pretražívan je mrežni korpus

[1] Ovaj je rad financirala Hrvatska zaklada za znanost projektom *Engleske riječi u hrvatskome jeziku: identifikacija, afektivno-semantičko normiranje i ispitivanje kognitivne obrade bihevioralnim i neuroznanstvenim metodama* (UIP-2019-04-1576).

hrvatskog jezika hrWaC *pomoću platforme* Sketch Engine. *Ovom metodom dobiven je popis od 1217 engleskih riječi. Rezultati su pokazali da se pomoću dostupnih alata i resursa za hrvatski jezik može izraditi popis engleskih riječi i njihovih frekvencija, ali i da postoje brojni problemi zbog kojih se rezultati ne mogu smatrati u potpunosti pouzdanima. Isto tako, sam se postupak i dalje mora kombinirati s ručnim metodama i klasifikacijama. Zaključujemo da je za izradu cjelovite baze engleskih riječi u hrvatskome potrebno razviti nove alate i resurse koji bi omogućili automatsko crpljenje engleskih riječi iz korpusa hrvatskoga jezika.*

Ključne riječi: *engleske riječi; hrvatski jezik; računalno jezikoslovlje*

1. Uvod

Kao globalni jezik engleski je postao sveprisutan u glavnim domenama ljudskog djelovanja: obrazovanju (npr. Brannen, Piekkari i Tietze 2014; Marginson i van der Wende 2007), poslovanju (npr. Brannen, Piekkari i Tietze 2014; Gluszek i Hansen 2013; Graddol 2006) i slobodnim aktivnostima (npr. De Wilde, Brysbaert i Eyckmans 2019; Bogunović i Jelčić Čolakovac 2019; Lauricella, Cingel, Blackwell, Wartella i Conway 2014). Sveprisutnost engleskog jezika dovela je do naglog priljeva engleskih elemenata u mnoge jezike (npr. El-Dash i Busnardo 2002; Pulcini, Furiassi i Rodríguez González 2012), među kojima je i hrvatski (npr. Drljača Margić 2009; Hudeček i Mihaljević 2005). Utjecaj engleskog jezika na hrvatski vidljiv je u različitim funkcionalnim stilovima: znanstvenom (npr. Bogunović i Ćoso 2013; Matić 2017; Raos 2006), publicističkom (npr. Foro 2014; Brdar 2010; Mihaljević Djigunović, Cergol i Qingmin 2006), razgovornom (npr. Drljača Margić 2014, 2012; Mihaljević 2003) i administrativnom (npr. Jurič, Krampus i Račić 2013). Nadalje, utjecaj engleskog jezika vidljiv je i na mnogim jezičnim razinama (npr. Bogunović i Ćoso 2013; Drljača Margić 2009), pri čemu je najosjetljiviji leksik (npr. Hudeček i Mihaljević 2005).

U cilju očuvanja jezika, na leksičkoj se razini savjetuje uporaba hrvatskih istovrijednica, posebice u formalnom kontekstu koji zahtijeva standardni jezik (npr. Hudeček i Mihaljević 2005; Hudeček i Mihaljević 2015; Opačić 2007a; 2007b; Raos 2006). Ako u jeziku ne postoji odgovarajuća riječ, poseže se za različitim rješenjima poput višerječnih izraza, opisa ili pak uvođenja novih riječi. Pokušaji iznalaženja novih riječi nerijetko nailaze na otpore (npr. Patekar, 2019), dijelom i zbog toga što su

govornici dotad već prihvatili stranu riječ (npr. Muhvić-Dimanovski i Skelin Horvat 2008). Veliku ulogu u uvođenju novih riječi imaju i mediji (Drljača 2006). Primjerice, engleska riječ *selfie* u široku je uporabu ušla 2012. godine, a hrvatska riječ 'sebić' predložena je 2014. godine (Halonja i Hudeček 2014).

Dok se tradicionalni otpori prema stranim riječima, pa tako i onima posuđenima iz engleskoga, dijelom mogu okarakterizirati kao nastojanje da se spriječi strani utjecaj (Turk i Opašić 2008), svjedočimo i pristupima koji predlažu da se uporaba riječi iz engleskoga promatra u okviru globalizacije i globalnih trendova (Škifić i Mustapić 2012). U tom kontekstu, riječi i izrazi posuđeni iz engleskog jezika često preciznije opisuju pojavu te su uporabno praktičniji (npr. Drljača 2006). Stoga nešto fleksibilniji pristupi sugeriraju da bi hrvatski trebao biti otvoreniji za prihvaćanje engleskih riječi, ako se lako mogu prilagoditi pravilima hrvatskog jezika (npr. Peti-Stantić 2013).

S druge strane, uporaba riječi posuđenih iz engleskoga prihvatljivija je u neformalnom kontekstu koji ne podliježe standardu (npr. Hudeček i Mihaljević 2015). To je posebice vidljivo kod govornika adolescentske i mlađe odrasle dobi (npr. Nikolić-Hoyt 2005; Drljača Margić 2012; 2014; Skelin Horvat 2015). Istraživanja su pokazala da ove dobne skupine imaju pozitivne stavove prema uporabi takvih riječi u neformalnom kontekstu, dok domaće riječi smatraju prikladnijima u formalnom kontekstu (Drljača Margić 2014; 2012). Ipak, u domenama poput tehnologije, znanosti, industrije zabave i informacijskih tehnologija riječi posuđene iz engleskoga prosuđuju se pozitivno i u formalnom kontekstu (Matić 2017; Drljača Margić 2014; 2012). Primjerice, u domeni informacijskih tehnologija riječi posuđene iz engleskoga ponekad obilježava manja uporabna zahtjevnost u odnosu na domaće, često višerječne izraze (Škifić i Mustapić 2012). Stavovi govornika prema takvim riječima odgovaraju njihovoj uporabi s obzirom na kontekst (Drljača Margić 2014). Uporaba riječi iz engleskog jezika dijelom se može pripisati i prestižnom statusu engleskoga jezika (Drljača Margić 2009).

2. Riječi posuđene iz engleskoga jezika

Može se reći da hrvatski jezik danas najviše posuđuje upravo iz engleskoga (npr. Mihaljević Djigunović i Geld 2003). Engleski se jezik smatra prestižnim (npr. Crystal 2012), a znanje engleskog povezuje se s boljim

društvenim položajem i boljim životom (McKenzie 2010; Mihaljević Djigunović i Geld 2003). Upravo zbog tog prestižnog statusa umanjena je vjerojatnost da će se riječi posuđene iz engleskog prilagoditi jeziku primatelju (npr. McKenzie 2010; Nikolić-Hoyt 2005), te se nerijetko rabe u neprilagođenom obliku (npr. *freelancer, chat, e-mail*).

Riječi posuđene iz engleskog jezika nazivaju se anglizmima (Filipović 1990). Posuđene se riječi kategoriziraju prema stupnju prilagođenosti i/ili uključenosti u jezik primatelj, u ovome slučaju hrvatski. Pritom treba razlikovati riječi koje su se potpuno ili dijelom prilagodile hrvatskom jeziku od onih koje su zadržale svoja izvorna obilježja. Na primjer, Kavgić (2013) anglizme dijeli na tri skupine, uzimajući u obzir njihov oblik i stupanj prilagodbe. U prvu skupinu svrstava očite anglizme, tj. elemente koji su se u većoj ili manjoj mjeri prilagodili jeziku primatelju (npr. *goal* – 'gol'). Drugu skupinu čine skriveni anglizmi odnosno leksički elementi čiji oblik podsjeća na onaj u jeziku primatelju, ali nosi značenje iz jezika davatelja (npr. *star* – 'zvijezda, popularna osoba'). Posljednjoj, trećoj skupini pripadaju tzv. sirovi anglizmi, odnosno leksički elementi preneseni u jezik primatelj bez ortografske prilagodbe, djelomične morfosintaktičke i fonološke prilagodbe te potpune semantičke prilagodbe (npr. *hat trick*). Görlach (2002b) razlikuje tri stupnja prilagodbe: 1. potpuna prilagođenost, pri čemu se riječ više ne prepoznaje kao strana, iako je zadržala fonološka, grafijska ili morfosintaktička obilježja jezika davatelja; 2. riječ u ograničenoj uporabi; te 3. riječ koja nije dijelom jezika primatelja, tj. kalk, posuđenica ili riječ koja je poznata isključivo dvojezičnim govornicima, a usko se veže uz britansku ili američku kulturu. Međeral (2016) riječi stranoga podrijetla dijeli prema stupnju uključenosti u jezik primatelj, pa tako razlikuje pet kategorija. Prvu kategoriju čine strane riječi u užemu smislu, odnosno riječi koje zadržavaju izvorna grafijska obilježja (npr. *gadget, brainstorming, makeover*). Takve su riječi često dijelom polusloženica (npr. *wellness*-centar, *offshore*-tvrtka), a po potrebi mogu primati hrvatske morfološke nastavke. Drugu skupinu čine tuđice, koje Međeral (2016) definira kao pravopisno prilagođene jeziku primatelju, no i dalje atipičnih fonoloških značajki (npr. 'lift', 'čips'). U treću skupinu svrstava prilagođenice, tj. riječi koje su se u potpunosti prilagodile jeziku primatelju (npr. 'ček', 'tim'), a iste ujedno čine i najbrojniju skupinu. U četvrtu skupinu ubrajaju se usvojenice, odnosno riječi koje su se toliko uklopile u hrvatski jezik da ih više ne doživljavamo stranima (npr. 'klub', 'tenk'), (usp. također Duvnjak Jardas 2019). Posljednju skupinu čine kalkovi ili

doslovne prevedenice (npr. *skyscraper* – 'neboder', *sustainable growth* – 'održivi razvoj'), (usp. također Duvnjak Jardas 2019).

Spomenute podjele pokazuju da u hrvatskome jeziku terminologija vezana uz jezično posuđivanje još uvijek nije ujednačena (npr. Muhvić-Dimanovski i Skelin Horvat 2006). Za riječi posuđene iz engleskog jezika koje su zadržale svoja izvorna obilježja rabe se različiti nazivi: anglizmi (npr. Hudeček i Mihaljević 2015), sirovi anglizmi (Kavgić 2013), posuđenice (Görlach 2002b), strane riječi (Međeral 2016), pseudoanglicizmi (Filipović 1990) ili engleske riječi (npr. Brdar 2010; Ćoso i Bogunović 2017). U ovome radu tematizirat će se isključivo riječi preuzete iz engleskog jezika s izvornim grafijskim obilježjima, a koje se mogu pojaviti s hrvatskim morfološkim nastavcima (npr. *freelanceri*, *chat*, *wellness, downloadati* itd.). Za njih će se rabiti naziv „engleske riječi" (usp. Ćoso i Bogunović 2017; Bogunović i Ćoso 2013; Brdar 2010).

Problem posuđivanja iz engleskog jezika prilično je detaljno istražen u kontekstu posuđivanja, prilagodbe i stavova. Za sada još uvijek ne postoje konkretni podaci o engleskim riječima u hrvatskome. Na primjer, još uvijek nije poznato koje se sve engleske riječi pojavljuju u hrvatskome i koliko su česte. Kako bi se odgovorilo na to pitanje, potrebno je poslužiti se računalno-jezikoslovnim resursima i alatima. Pritom najprije valja utvrditi omogućuju li postojeći resursi i alati optimalan način za pronalaženje engleskih riječi (npr. Núñez Nogueroles 2016; Balteiro 2011) ili je pak potrebno osmisliti nove (npr. Alex 2005; Andersen 2012; Alvarez-Mellado 2020). Stoga je cilj ovoga rada analizirati neke od postojećih alata i resursa te utvrditi omogućuju li pronalaženje engleskih riječi u hrvatskome jeziku.

3. Alati i resursi za crpljenje engleskih riječi

Jezične tehnologije obuhvaćaju jezične resurse, jezične alate te komercijalne proizvode (Tadić 2016). Jezični resursi predstavljaju jezičnu građu koja je digitalno usustavljena i služi za pretraživanje (*Mrežnik*), a uključuju korpuse i jezične zbirke te digitalne rječnike. Jezični su alati specijalizirani programi, razvijeni na temelju jezičnih resursa, koji omogućuju obradu postojećih resursa ili pak stvaranje novih. U komercijalne proizvode ubrajaju se rječnici, provjernici (pravopisa, gramatike, stila), te sustavi za diktiranje, strojno (potpomognuto) prevođenje i računalno potpomognuto učenje jezika (Tadić 2016).

U svrhu pronalaženja engleskih riječi u drugim su se jezicima koristile različite metode. Primjerice, Núñez Nogueroles (2016) koristi gotov popis engleskih riječi u španjolskome, koje potom pretražuje u nacionalnom korpusu. Sličnu metodu u domeni sporta primjenjuje i Balteiro (2011), pri čemu prilagođene i neprilagođene riječi iz engleskog jezika, prikupljene iz rječnika anglizama za španjolski jezik, pretražuje u nacionalnom korpusu kako bi dobila podatke o njihovoj učestalosti. Druge pak metode uključuju razvijanje novih alata i/ili resursa. Na primjer, s pretpostavkom da broj rezultata dobivenih Googleovim pretraživanjem može pokazati pripadnost nekom jeziku razvijen je nenadzirani sustav za prepoznavanje engleskih riječi u njemačkome koristeći postojeće leksičke baze i mrežno dostupne podatke (Alex 2005). Takav pristup može biti problematičan za jezike koji su nedovoljno zastupljeni na internetu, kao što je slučaj s hrvatskim. Mrežno pretraživanje može se izbjeći pretraživanjem leksikona u kombinaciji s n-gramima (npr. Furiassi i Hofland 2007), no ta metoda također neće odgovarati jezicima s nedovoljno razvijenim resursima, poput hrvatskoga. Ti se problemi mogu izbjeći pomoću metoda nadziranog učenja (npr. Alvarez-Mellado 2020; Serigos 2017; Castro i sur. 2016; Losnegaard i Lyse 2012). S druge strane, takve metode zahtijevaju označene podatke za treniranje klasifikatora, što znači da u slučaju da takvi podaci ne postoje, valja ih izraditi.

Prije svega desetak godina smatralo se da hrvatski jezik pripada jezicima s vrlo slabo razvijenim jezičnim tehnologijama (Tadić, Brozović-Rončević i Kapetanović 2012; Tadić 2003). Ta tvrdnja danas je djelomično još uvijek točna, posebice kada jezične tehnologije za hrvatski jezik usporedimo sa slovenskim jezikom, ili pak engleskim, španjolskim i njemačkim. Međutim, u posljednjih desetak godina razvili su se mnogi resursi, prije svega pod vodstvom prof. Marka Tadića (npr. HR4EU) te brojne leksičke i terminološke baze (npr. Hrvatska psiholingvistička baza, Peti-Stanić i sur. 2018) kao i specijalizirani korpusi (npr. ParlaMint – HR 2.1). Pretragom dostupnih resursa za hrvatski jezik nije pronađen obuhvatan popis engleskih riječi s podacima o njihovoj učestalosti. U nedostatku takvog popisa autori s jedne strane pribjegavaju prikupljanju podataka iz različitih izvora, poput reklama, govora, različitih tekstova, natpisa s kojima se svakodnevno susreću (npr. Ćoso i Bogunović 2017; Mišić-Ilić i Lopičić 2011; Görlach 2001), a s druge strane donose djelomične popise temeljene na analizi prikupljenog korpusa ili rječničke građe, najčešće vezane uz određenu domenu (npr. Patekar 2019; Drljača 2016; Hudeček i Mihaljević 2015; Bogunović i Ćoso

2013; Brdar 2010; Runjić-Stoilova i Pandža 2010). Rječnici također predstavljaju važan izvor engleskih riječi (npr. Görlach 2001; Filipović 1990). Neke se engleske riječi mogu pronaći u *Rječniku neologizama* (Muhvić-Dimanovski, Skelin Horvat i Hriberski 2016) koji redovito bilježi ulaz novih riječi u hrvatski jezik. Portal *Bolje je hrvatski!* Instituta za hrvatski jezik i jezikoslovlje također bilježi ulaz stranih riječi u hrvatski te predlaže hrvatske zamjene za strane riječi. Pritom ne djeluje kao baza, niti nudi učestalost pojavnica, već se unosi pretražuju tako da se upiše traženi pojam ili pak abecednim redom. *Kontekst.io* je pak tražilica temeljena na komputacijskom jezičnom modelu koja omogućuje pretraživanje pojmova iz hrvatskog, slovenskog i srpskog mrežnog korpusa te pruža uvid u učestalost kao i slične unose. Primjerice, pretragom riječi *shopping*, izlistani su rezultati svih povezanih čestica prema učestalosti pojavljivanja i sličnosti (*šoping* 4.75, 91%; *shoping* 1.62, 86% i sl.)

Korpusi predstavljaju važan jezični resurs jer pružaju empirijske podatke za potvrđivanje ili opovrgavanje hipoteza i dobivanje odgovora na istraživačka pitanja (Borucinsky 2017). Za hrvatski jezik trenutno postoje tri korpusa općega jezika: *Hrvatski nacionalni korpus (HNK)*, *Hrvatska jezična riznica (Riznica)* te *Hrvatski mrežni korpus* (*hrWaC*). *Hrvatski mrežni korpus* (Ljubešić i Klubička 2016) trenutno je najopsežniji korpus hrvatskoga jezika, a dobiven je tzv. metodom puzanja (engl. *crawling*) top domene .hr, kojim je obuhvaćeno 8, 388 URL-a. Uz tekstove standardnog hrvatskog jezika, kao što su primjerice mrežne stranice službenih i javnih tijela, u korpusu su zastupljeni i tekstovi iz različitih blogova, reklama, korisničkih komentara, rasprava i sl. Stoga, između ostaloga, može dati uvid i u način preuzimanja riječi iz engleskoga jezika, a njegova najveća prednost očituje se u veličini i raznolikosti tekstova . S druge pak strane, jedna od najvećih mana mrežnog korpusa jesu izvornost i pouzdanost tekstova, te pitanje autorskih prava (usp. također Fletcher 2011). Nadalje, mrežni korpus ne zastupa sve funkcionalne stilove u jednakoj mjeri pa prema strogoj definiciji nije reprezentativan. No, ako pojam reprezentativnosti shvatimo u širem smislu i ako za polazište uzmemo činjenicu da takav korpus zrcali jezik u stvarnoj uporabi (npr. Núñez Nogueroles 2016; Furiassi 2008), onda ga možemo smatrati reprezentativnim za proučavanje dane jezične pojave.

Od dostupnih računalno-jezikoslovnih alata za pretragu korpusa može se koristiti *Sketch Engine* (Kilgarriff i sur. 2004) koji podržava više od 90 jezika, među njima i hrvatski. Njegovi algoritmi analiziraju autentične

tekstove koji sadrže nekoliko milijardi riječi kako bi se identificirale karakteristične pojave u jeziku, ali i one manje karakteristične. *Sketch Engine* sadrži više od 500 jezičnih resursa, tj. korpusa, među njima i *hrWaC* (Ljubešić i Klubička 2016), trenutno najveći mrežni korpus hrvatskoga jezika.

Kao što je razvidno iz navedenoga, postojeći jezični resursi za hrvatski jezik ne nude cjelovitu bazu engleskih riječi u hrvatskome niti podatke o njihovoj učestalosti. Problemu pronalaženja engleskih riječi u hrvatskome može se pristupiti na više načina. Najjednostavniji pristup je ručna pretraga koja, zbog dugotrajnosti procesa i ograničenog opsega, ne bi dala zadovoljavajuće rezultate. Sljedeći način uključuje korištenje postojećih alata i resursa, a mogućnosti takvog pristupa ispitat će se u ovom istraživanju.

4. Metoda

4.1. Pribor/Materijali

U ovome radu ispituju se mogućnosti crpljenja engleskih riječi iz hrvatskih tekstova pomoću nekih od postojećih računalno-jezikoslovnih resursa i alata za hrvatski jezik. Točnije, korišteni resurs jest *Hrvatski mrežni korpus* (Ljubešić i Klubička 2016), dok je od alata korišten *Sketch Engine* (Kilgarriff i sur. 2004).

HrWaC sadrži ukupno 1,405,794,913 pojavnica iz 3,611,090 dokumenata sastavljenih iz triju potkorpusa: 1. domena *.hr* – potkorpus koji obuhvaća 98,49 % pojavnica, 2. *Slobodna Dalmacija* koji čini 6,1 % korpusa, 3. *Večernji list* s udjelom od 3,52 %. Najnovija inačica resursa – *hrWaC* 2.2 RFTagger temelji se na Multext East ver. 5, s preciznošću lematizacije od oko 98%, morfosintaktičnog označavanja od 87 %, a označavanja vrsta riječi (POS-only) od 97% (usp. Agić, Ljubešić i Merkler 2013).

Uz morfosintaktičko označavanje, korpusi koji su dostupni putem platforme *Sketch Engine* sadrže i označavanje prema vrstama riječi, što olakšava njihovu pretragu. Za pronalaženje engleskih riječi u hrvatskome korištena je oznaka za strane riječi tj. [tag="Xf"].

4.2. Postupak

Prvi korak bio je usporediti popise domena prema količini teksta bez kriterija oznake Xf i s kriterijem oznake Xf kako bi se dobio uvid u najveće mrežne stranice u *hrWaC*-u, kao i one s najvećim brojem engleskih riječi.

Pritom su izdvojene mrežne stranice s najvećim brojem stranih riječi, odnosno onih označenih kao Xf.

Nakon toga uslijedio je postupak crpljenja engleskih riječi pomoću oznake Xf i čišćenje dobivenih rezultata.

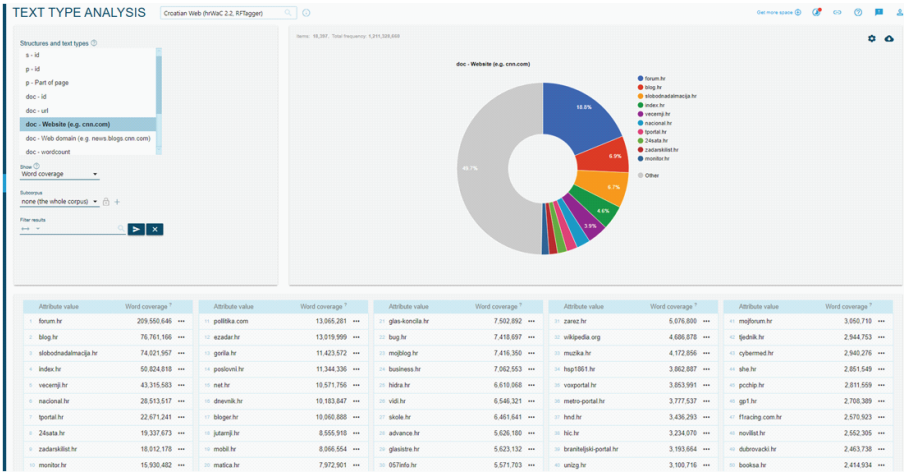
Pojavnice dobivene preko oznake stranih riječi potom su analizirane po učestalosti pojavljivanja i prema početnim postavkama pretrage. Točnije, uspoređeni su rezultati pretrage prema lemi, obliku riječi te velikim/malim slovima. Nakon usporedbe i dobivenih rezultata, uslijedila je ručna klasifikacija riječi. Određene riječi nisu uvrštene u konačni popis prema sljedećim kriterijima:

1. riječi iz stranog jezika koji nije engleski (npr. *Welt* iz njemačkoga, *pour* iz francuskoga);
2. vlastita imena ili toponimi (npr. *Wave Boat Sealver, Cambridge*);
3. lažni anglizmi (npr. *gastro show*);
4. pogrešno označene riječi (hrvatski prijedlog *van* označen kao engleska imenica *van* ‘kamion’);
5. gramatičke riječi (npr. *of, with*) koje nisu obuhvaćene ovim istraživanjem;
6. kratice i akronimi (npr. *btw, USA*);
7. pogrešno napisane riječi (npr. *coffe*);
8. idiosinkratične pojave (npr. *pussyhit*).

Nakon ručne klasifikacije pomoću konkordancija i svođenja na zajedničku lemu, bilo je potrebno završno ručno čišćenje zbog pojave određenog broja riječi označenih kao Xf u engleskom kontekstu, kao i pojave nekih riječi u identičnom kontekstu, ali drugom izvoru. Takve riječi nisu uvrštene u konačan popis.

5. Rezultati

Analiza vrsta tekstova pokazala je da su najveće mrežne stranice zastupljene u *hrWaC*-u: forum.hr, blog.hr, slobodnadalmacija.hr, index.hr i vecernji.hr te da čine 40,9 % svih mrežnih stranica zastupljenih u *hrWaC*-u (Slika 1).



Slika 1. *Najveće mrežne stranice zastupljene u hrWaC-u*

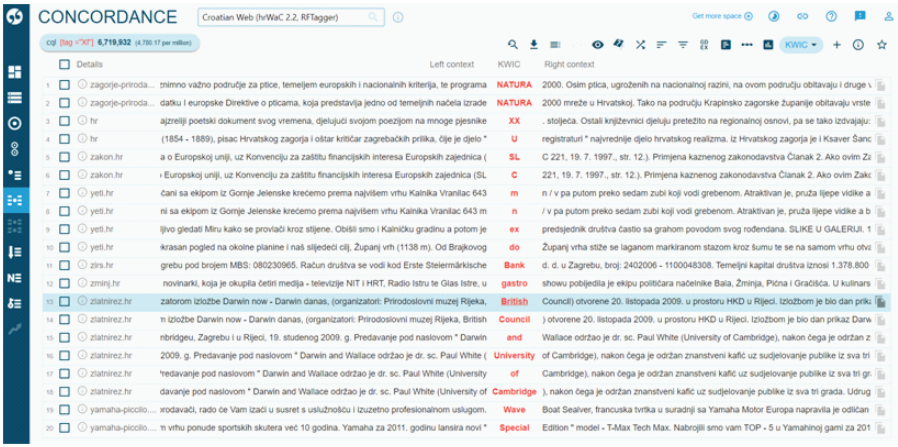
Najveći broj riječi označenih kao Xf nalazimo na mrežnim stranicama forum.hr, blog.hr, gorila. hr, index. hr itd. kao što je prikazano na Slici 2.

	Website (e.g. cnn.com)	Frequency
1	forum.hr	2,012,780
2	blog.hr	558,262
3	gorila.hr	499,607
4	tranexp.hr	158,391
5	index.hr	151,642
6	slobodnadalmacija.hr	150,040
7	muzika.hr	119,040
8	mobil.hr	99,016
9	vecernji.hr	83,772
10	tportal.hr	83,349

Rows per page: 10 1–10 of 12,924 1 / 1,293

Slika 2. *Mrežne stranice s najvećim brojem riječi označenih kao Xf u hrWaC-u*

U sljedećoj fazi pristupilo se crpljenju popisa neprilagođenih engleskih riječi iz korpusa. Pomoću oznake "Xf" od ukupnog je broja pojavnica u korpusu (1,405,794,913) izdvojeno 6,719,932 stranih riječi, odnosno 4,780.17 pojava na milijun riječi. Rezultat od 6,7 milijuna pojavnica sadrži veliku količinu šuma te ga je stoga valjalo pročistiti. Na Slici 3 prikazano je prvih 20 konkordancija dobivenih pretragom [tag="Xf"].



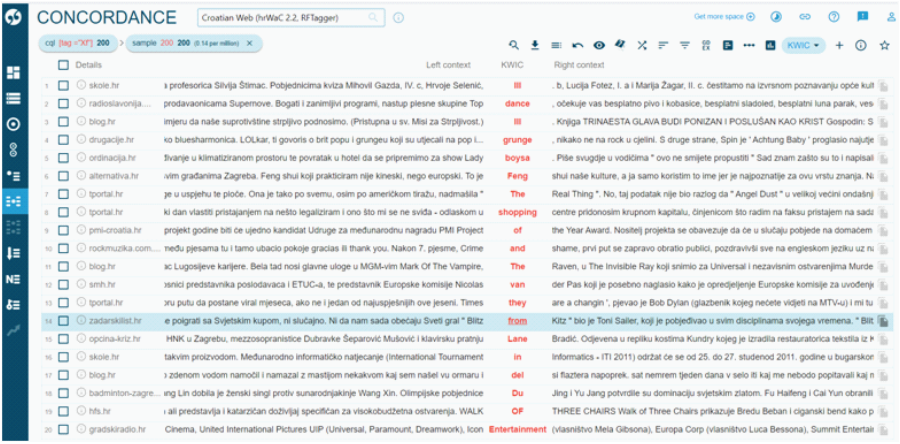
Slika 3. *Konkordancije dobivene pretragom [tag="Xf"]*

Pogledamo li popis prvih 20 konkordancija odnosno ključnih riječi u kontekstu (engl. *key word in context*, KWIC) prema obliku riječi za dobiveni rezultat, vidimo da su pojavnice:

1. iz stranog jezika koji nije engleski (*NATURA*);
2. vlastita imena ili toponimi (*Wave Boat Sealver, Cambridge*);
3. lažni anglizmi (*gastro show*);
4. pogrešno označene riječi (hrvatski prijedlog *do* označen je kao strana riječ) itd.

Od 20 prikazanih pojavnica, samo dvije (*ex* (*predsjednik*) i *special edition model*) odgovaraju kriterijima koji su definirani za neprilagođene engleske riječi u ovome istraživanju.

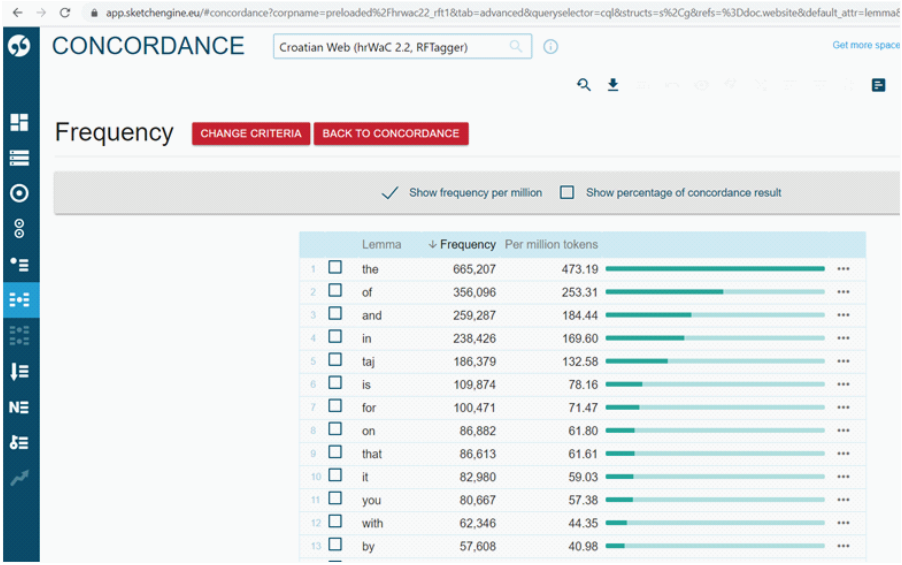
Rezultati dobiveni primjenom opcije nasumičnih rječničkih primjera pokazali su da nešto veći broj traženih pojavnica jesu engleske riječi (npr. *grunge*, *shopping* i *entertainment*), ali i ovaj popis sadrži previše šuma (Slika 4).



Slika 4. *Rezultati pretraga nasumičnih rječničkih primjera*

Analizom učestalosti za svaku su pojavnicu dobivene apsolutna frekvencija (engl. *absolute frequency, raw frequency*) i relativna frekvencija (engl. *relative frequency, normalized frequency*) te je moguće pregledavati i konkordancije za svaku riječ. Nešto drugačiji popisi pojavnica dobiju se ovisno o početnim postavkama pretrage, pa se tako primjerice za postavku pretrage „lema" dobije 2 209 unosa, a za postavku „oblik riječi" (engl. *word form*) 2 976 unosa. Razlika je u tome što se kod leme sve frekvencije zbrajaju, dok se za svaki oblik riječi računa zasebna frekvencija. Neznatna razlika dobivena je i promjenom postavki velika/mala slova.

Rezultati pretrage frekvencija pojavnica dobivenih preko oznake stranih riječi prikazani su na Slici 5.



Slika 5. *Frekvencije dobivene pretragom pojavnica za oznaku Xf po lemi*

U Tablici 1 prikazani su rezultati dvadeset najčešćih lema i oblika riječi u korpusu *hrWaC* s apsolutnim frekvencijama. Riječi koje se nalaze među prvih dvadeset, a dobivene su objema pretragama (lema i oblik riječi) otisnute su masnim slovima.

Tablica 1. *Prikaz 20 najčešćih lema i oblika u hrWaC-u*

Red. br.	Lema	Apsolutna frekvencija	Oblik riječi	Apsolutna frekvencija
1.	***top***	9637	***top***	38439
2.	***online***	7276	*all*	36373
3.	*daily*	6551	*not*	35975
4.	***times***	5837	***world***	28334
5.	*rock*	5826	*love*	26789
6.	***big***	5610	***time***	24122
7.	*info*	4854	***sex***	21261
8.	***world***	4843	***big***	20436
9.	***sex***	4764	***live***	20415

Red. br.	Lema	Apsolutna frekvencija	Oblik riječi	Apsolutna frekvencija
10.	***live***	4756	***times***	19535
11.	*wall*	4503	*man*	18802
12.	***full***	4272	*day*	18741
13.	*fashion*	3921	*open*	18630
14.	*gay*	3792	*life*	18458
15.	*reality*	3731	*grand*	17678
16.	*shopping*	3433	***online***	16852
17.	*bad*	3312	*like*	16453
18.	*story*	3310	*best*	16279
19.	*flash*	3272	***full***	16112
20.	*car*	3230	*do*	14866

Smatramo da ćemo dobiti najpovoljnije rezultate ako strane riječi tražimo kao oblike riječi, a ne leme jer ćemo time obuhvatiti i riječi poput *time* i *times*, a pretragu podesimo tako da obuhvati velika i mala slova. Nadalje, kako bi se optimizirala pretraga, podešen je kriterij pretrage na sve riječi koje su u korpusu označene kao Xf i sadrže više od tri znaka. tj. [word=". {3,}" & tag="Xf"], a ova pretraga rezultirala je je s 2289 riječi.

Nakon ručne klasifikacije pomoću konkordancije i svođenja na zajedničku lemu, od 2289 riječi preostalo je 1217 riječi, što znači da je 47% rezultata bio šum. Jedan od problema s kojima smo se susreli jest pojavljivanje riječi označenih kao Xf u engleskom kontekstu, što je trebalo ručno pregledati i ukloniti takve riječi s popisa. Drugi problem jest činjenica da se riječi pojavljuju u identičnom kontekstu, ali u drugom izvoru. Primjerice riječ 'shared' tri se puta pojavljuje u korpusu u identičnim rečenicama, ali iz drugih izvora, kao što je prikazano u primjeru (1).

> (1) *Također je odlučeno da se pokrene projekt objedinjavanja informacijskog sustava Hrvatskog zavoda za mirovinsko osiguranje (HZMO) u Shared Service Centru, koji se uspostavlja u Agenciji za podršku informacijskim sustavima i informacijskim tehnologijama* (APIS IT).

U Tablici 2 prikazano je prvih 50 riječi s frekvencijama pročišćenog popisa.

Tablica 2. *Prikaz 50 najfrekventnijih riječi u hrWaC-u označenih kao Xf nakon ručnog pročišćavanja*

Red. br.	Riječ (mala slova)	Apsolutna frekvencija	Relativna frekvencija
1.	*top*	38439	2.734.325
2.	*all*	36373	2.587.362
3.	*not*	35975	2.559.050
4.	*world*	28334	2.015.514
5.	*love*	26789	1.905.612
6.	*time*	24122	1.715.898
7.	*more*	23996	1.706.935
8.	*sex*	21261	1.512.383
9.	*big*	20436	1.453.697
10.	*live*	20415	1.452.203
11.	*times*	19535	1.094.043
12.	*man*	18802	1.337.464
13.	*day*	18741	1.333.125
14.	*open*	18630	1.325.229
15.	*life*	18458	1.312.994
16.	*grand*	17678	1.257.509
17.	*online*	16852	1.178.266
18.	*like*	16453	1.170.370
19.	*best*	16297	1.159.273
20.	*full*	16122	1.146.824
21.	*do*	14866	1.057.480
22.	*just*	14792	1.052.216
23.	*ad*	14649	1.042.044
24.	*go*	14118	1.004.272
25.	*music*	14052	955.189
26.	*only*	14013	996.803
27.	*make*	13771	979.588
28.	*art*	13747	977.881
29.	*house*	13412	902.906
30.	*home*	13383	951.988
31.	*high*	12860	914.785

Red. br.	Riječ (mala slova)	Apsolutna frekvencija	Relativna frekvencija
32.	*sorry*	12533	891.524
33.	*new*	12203	868.050
34.	*hard*	12174	865.987
35.	*gay*	11775	837.604
36.	*play*	11625	826.934
37.	*bad*	11608	804.528
38.	*daily*	11604	825.440
39.	*international*	11435	813.419
40.	*blue*	11187	795.778
41.	*jam*	11156	0.88064
42.	*night*	11125	791.367
43.	*non*	11104	789.873
44.	*ex*	10964	779.915
45.	*wall*	10920	776.785
46.	*flash*	10886	774.366
47.	*fashion*	10688	760.282
48.	*sun*	10247	728.911
49.	*national*	10241	728.485
50.	*info*	10178	724.003

Značajan problem predstavlja i činjenica da preko pretraživača stranih riječi nećemo dobiti očekivane riječi poput *host(ati), download(ati), link(ati)*, odnosno riječi koje primaju infinitivne nastavke i kao takve su automatski označene kao hrvatske riječi.

6. Rasprava

U prvom koraku istraživanja provedena je analiza vrsta tekstova koja je pokazala da su najveće mrežne stranice zastupljene u *hrWaC*-u: forum.hr, blog.hr, slobodnadalmacija.hr, index.hr i vecernji.hr, koji čine 40,9 posto svih mrežnih stranica zastupljenih u *hrWaC*-u. U drugom koraku utvrđeno je da su mrežne stranice s najvećim brojem riječi označenih kao Xf u *hrWaC*-u forum.hr, blog.hr, gorila. hr, index. hr, itd. Time je uvjetno potvrđena rasprostranjenost engleskih riječi u medijima. Naime, s obzirom na činjenicu da oznaka Xf ne daje pouzdane rezultate za crpljenje engleskih riječi iz

korpusa hrvatskoga jezika, ne možemo izvesti valjan zaključak o rasprostranjenosti engleskih riječi u medijima na temelju dobivenih podataka. Ipak, među deset izvora s najvećim brojem pojavnica našlo se čak šest informativnih i novinskih portala, čemu bi se mogao pripisati interes znanstvene zajednice za istraživanje upravo te domene (npr. Foro 2014; Brdar 2010; Mihaljević Djigunović i sur. 2006). Pritom valja istaknuti i važnu ulogu medija u prihvaćanju pojedinih riječi među govornicima (npr. Drljača Margić 2009; Muhvić-Dimanovski i Skelin Horvat 2008). Čini se da su engleske riječi najrasprostranjenije u razgovornom jeziku, na što upućuje podatak da su mrežne stranice s najvećim brojem stranih riječi forum.hr i blog.hr. Taj je podatak potvrđen istraživanjima koja navode veliku rasprostranjenost takvih riječi u razgovornom jeziku (npr. Ćoso i Bogunović 2017; Drljača Margić 2012; 2014; Mihaljević 2003). Mlađa je populacija u pravilu sklonija prihvaćanju trendova i novih tehnologija (npr. Ćoso i Bogunović 2017) pa se stoga može pretpostaviti da je ta populacija aktivnija na mrežnim forumima i blogovima. Također je poznato da mlađa populacija načelno ima pozitivan stav prema engleskim riječima, te ih u skladu s time i rabi (npr. Matić 2017; Drljača Margić 2014; 2012). Nadalje, uporaba engleskih riječi povezana je s boljom procjenom socijalne poželjnosti (npr. Ćoso i Bogunović 2017; El-Dash i Busnardo 2002), što bi također moglo igrati ulogu u učestaloj uporabi engleskih riječi na takvim mrežnim stranicama.

Crpljenje engleskih riječi iz korpusa hrWaC pomoću alata SketchEngine pokazalo je da takva pretraga rezultira s previše šuma. Razlog tomu jest činjenica da oznaka Xf nije primarno zamišljena za pronalaženje stranih riječi, već je to kategorija u koji se svrstavaju riječi koje označivač (engl. *tagger*) nije prepoznao, a to mogu biti riječi iz nekog stranog jezika koji nije nužno engleski, pogrešno napisane riječi, arhaične riječi, nestandardne riječi, *sleng* itd.

Nakon ručnog pregledavanja pojavnica i izdvajanja onih riječi koje ne spadaju pod engleske riječi (npr. riječi iz stranog jezika koji nije engleski, lažni anglizmi, vlastita imena, toponimi, internacionalizmi, pogrešno označene riječi, pogrešno napisane riječi itd.) dobiven je popis najučestalijih engleskih riječi. Pritom valja napomenuti da postoje razlike u rezultatima ovisno o izboru leme, oblika riječi, velikih, malih slova, što također predstavlja problem u takvoj vrsti pretrage. Nakon što su kriteriji pretrage podešeni na način da daju optimalne rezultate, dobiveno je 2289 riječi, a nakon ručne klasifikacije popis je sadržavao 1217 riječi. Problematične su bile one pojavnice koje su dijelom tekstova na engleskome jeziku, primjerice prijevodi hrvatskih mrežnih

stranica (npr. unizg.hr). Takve pojavnice nisu uzete u obzir, već samo one koje su uklopljene u hrvatski tekst. Da bi se to provjerilo, potrebno je za svaku pojavnicu detaljno pregledati kontekst pomoću opcije KWIC i odrediti odgovara li kriterijima. Za utvrditi i pročistiti navedeno, također je bilo potrebno ručno pročišćavanje rezultata. Značajan problem predstavlja i činjenica da preko pretraživača stranih riječi nećemo dobiti očekivane riječi poput *host(ati), download(ati), link(ati)*, odnosno riječi koje primaju infinitivne nastavke te su kao takve automatski označene kao hrvatske riječi. S obzirom na to da i te riječi pripadaju kategoriji neprilagođenih engleskih riječi (npr. Međeral 2016; Bogunović i Ćoso 2013; Ćoso i Bogunović 2017; Filipović 1990) te su temom ovog rada, rezultati takve pretrage mogu se smatrati nepotpunima. Kako je razvidno iz navedenih primjera, takve su riječi vrlo česte u domenama tehnologije i računarstva gdje se engleske riječi često rabe (npr. Škifić i Mustapić 2012), a mlađa populacija upravo u tim domenama iskazuje sklonost prema engleskim riječima (npr. Matić 2017; Drljača Margić 2012; 2014). U tu kategoriju može se svrstati i glagol *googlati*, koji je problematičan ne samo zbog hrvatskog infinitivnog nastavka, već i zbog činjenice da je izveden iz naziva *Google* koji se može svrstati pod vlastita imena čime ne bi bio uvršten u popis, a zapravo je značajan.

U konačnici, rezultati dobiveni uporabom računalno-jezikoslovnih alata nisu zadovoljavajući jer nismo pronašli riječi koje smo očekivali, kao npr. *selfie, influencer, celebrity*, tj. 'novije riječi'. Jedan od mogućih razloga jest činjenica da je *hrWaC* prikupljan u razdoblju između 2011. i 2013. godine. Kao primjer možemo uzeti već spomenutu englesku riječ *selfie,* koja je u širu uporabu u svijetu ušla 2012. godine (Halonja i Hudeček 2014). Drugi mogući razlog jest da ova vrsta pretrage ne daje pouzdane rezultate.

Cilj ovoga rada bio je utvrditi omogućuju li neki od postojećih jezičnih resursa i alata za hrvatski jezik pronalaženje engleskih riječi u hrvatskome jeziku. Nakon provedene analize utvrđeno je nekoliko problema. Prije svega, u *hrWaC*-u, kao najopsežnijem hrvatskom mrežnom korpusu, zastupljene su razne vrste tekstova, no funkcionalni stilovi nisu jednako zastupljeni. Isto tako, podaci dobiveni analizom ovog korpusa ne zrcale trenutno jezično stanje, jer korpus nije posuvremenjen niti mu se redovito dodaju novi tekstovi. Stoga se javlja potreba za stvaranjem korpusa koji će ponuditi novije podatke, a koji bi se mogao koristiti u kombinaciji s podacima dobivenim iz *hrWaC*-a. Nadalje, oznaka Xf nije primarno namijenjena za crpljenje stranih riječi već je to kategorija u koju se svrstavaju riječi koje označivač nije prepoznao, a to mogu biti kratice, akronimi, tipfeleri, strane

riječi itd. Stoga pretraga preko te oznake zahtijeva ručno pročišćavanje u nekoliko koraka. Pored toga, neke riječi zbog hrvatskih morfoloških nastavaka nisu prepoznate kao strane riječi, što ovaj način crpljenja stranih riječi čini nepouzdanim.

Liste riječi i frekvencije najčešće su polazište za daljnja istraživanja (no ne moraju uvijek dati optimalne rezultate), a ovo istraživanje pokazalo je da rezultati znatno variraju. Frekvencije ne ovise samo o tekstovima koji čine korpus, već i o sustavima za obradu korpusa, stoga je potrebno poznavati mogućnosti i ograničenja alata, otkriti razloge za nezadovoljavajuće rezultate i pokušati ih riješiti. Rezultati ovog istraživanja pokazali su da se ovim alatima i resursima za hrvatski jezik mogu pronaći neke engleske riječi u hrvatskome, no isto tako da se popis dobiven ovom metodom ne može smatrati cjelovitim, pouzdanim i reprezentativnim. Stoga se postavlja pitanje vrijedi li se baviti korpusima s obzirom na navedene poteškoće. Odgovor na to pitanje je potvrdan jer takva analiza ne zahtijeva veliku financijsku potporu, a puno je brža od ručnog crpljenja riječi. Nije optimalna, no može se kombinirati s drugim metodama, tj. ručnom klasifikacijom. Ipak, za stvaranje baze engleskih riječi čini se potrebnim dopuniti postojeće resurse (korpuse) novijom građom te razviti nove alate koji će preciznije i učinkovitije klasificirati engleske riječi u hrvatskome jeziku. Pored značajnog doprinosa u istraživanju pojave engleskih riječi u hrvatskome, ovaj istraživački smjer uvelike bi pridonio i razvoju računalno-jezikoslovnih resursa i alata za hrvatski jezik, ali i eksperimentalnih istraživanja koja zahtijevaju jasne, precizne i pouzdane podatke koji će omogućiti zadovoljavajuću eksperimentalnu kontrolu.

7. Zaključak

Ovim istraživanjem željeli smo ispitati omogućuju li neki od postojećih jezikoslovnih resursa i alata za hrvatski jezik stvaranje baze engleskih riječi i njihovih učestalosti ili pak postoji potreba za razvijanjem novih. Točnije, koristila se platforma SketchEngine za pretraživanje najopsežnijeg mrežnog korpusa hrvatskog jezika *hrWaC*. Rezultati su pokazali da postoji potreba za novijim podacima koji bi se mogli kombinirati s rezultatima dobivenima pretraživanjem *hrWaC*-a. Rezultati su također uvjetno pokazali da su engleske riječi najrasprostranjenije u razgovornom i publicističkom stilu, što potvrđuje dosadašnji znanstveni interes za te domene. Nadalje, pretraga riječi polučila je različite rezultate ovisno o vrsti pretrage, a najbo-

ljom opcijom pokazala se pretraga po obliku riječi s obuhvaćenim velikim i malim slovima. Postupak je rezultirao velikom količinom šuma, te su u radu opisani kriteriji ručne klasifikacije. Naposljetku, dobiven je popis od 1217 engleskih riječi. Zaključno, u radu su opisane mogućnosti opisanih resursa i alata za hrvatski jezik, koji se mogu koristiti za izradu popisa engleskih riječi s frekvencijama, no i njihova ograničenja u vidu različitih rezultata ovisno o vrsti pretrage, potrebe za ručnim pročišćavanjem i klasifikacijom. Rad također donosi i opisane rezultate, tj. dobiven popis engleskih riječi koji, iako nedovoljno pouzdan za neka istraživanja, daje uvid u broj, vrstu i pojavnost engleskih riječi u hrvatskome jeziku.

Naposljetku, rad predstavlja primjer kvantitativnog lingvističkog istraživanja, a rezultati mogu doprinijeti novim spoznajama o ograničenjima korpusnih istraživanja nad postojećim korpusima hrvatskoga jezika.

Popis literature

Agić, Željko, Nikola Ljubešić, Danijela Merkler (2013) „Lemmatization and Morphosyntactic Tagging of Croatian and Serbian", *The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing Workshop Proceedings*, ur. Jakub Piskorski, Lidia, Pivovarova, Hristo Tanev i Roman Yangarber, The Asscociation for Computational Linguistics, Stroudsburg, USA, 48–57.

Alex, Beatrice (2005) „An unsupervised system for identifying English inclusions in German text", 43. *Annual Meeting of the Association for Computational Linguistics* (*ACL 2005),* ur. Chris, Callison-Burch i Stephen Wan, Michigan, SAD, Ann Arbor, 133–138.

Alvarez-Mellado, Elena (2020) „An Annotated Corpus of Emerging Anglicisms in Spanish Newspaper Headlines", *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching, LREC 2020,* ur. Solorio Thamar, Choudhury Monojit, Bali Kalika, Sitaram Sunayana, Das Amitava, i Diab Mona, Marseille: European Language Resources Association, 1–8.

Andersen, Gisle (2012) „Semi-automatic approaches to Anglicism detection in Norwegian corpus data", *The Anglicization of European Lexis*, ur. Cristiano, Furiassi, Virginia, Pulcini i Félix Rodríguez González, Amsterdam-Philadephia: John Benjamins Publishing Company, 111–130, https://doi.org/10.1075/z.174.09.

Balteiro, Isabel (2011) „A reassessment of traditional lexicographical tools in the light of new corpora: sports Anglicisms in Spanish", International Journal of English Studies, 11, 2, 23–52, https://doi.org/10.6018/ijes/2011/2/149631.

Bogunović, Irena, Bojana Ćoso (2013) „Engleski u hrvatskome: znanstveni izričaj biomedicine i zdravstva", Fluminensia 25, 2, 177–191.

Bogunović, Irena, Jasmina Jelčić Čolakovac (2019) „Uloga neformalnih aktivnosti u nenamjernom usvajanju jezika: Povezanost uporabe jezika i jezičnog znanja", Fluminensia 31, 2, 181–199, https://doi.org/10.31820/f.31.2.15.

Bolje je hrvatski!, https://bolje.hr/, posjet 25. listopada 2020.

Borucinsky, Mirjana (2017) „Korpusansätze in der Sprachforschung: Mit besonderer Rücksicht auf korpusgebundene Untersuchungen der kroatischen Sprache", *Applied Linguistics Research and Methodology – Proceedings from the 2015 CALS conference*, ur. Kristina Cergol Kovačević i Sanda Lucija Udier, Frankfurt am Main: Peter Lang, 255–269, https://doi.org/10.3726/b10830.

Brannen, Mary Yoko, Piekkari, Rebecca, Susanne Tietze (2014) „The multifaceted role of language in international business: Unpacking the forms, functions and features of a critical challenge to MNC theory and performance", Journal of International Business Studies, 45, 495–507, http://dx.doi.org/10.1057/jibs.2014.24.

Brdar, Irena (2010) „Engleske riječi u jeziku hrvatskih medija", Lahor, 10, 217–232.

Castro, Dayvi, Ellen Polliana Souza, Lima De Oliveira (2016) „Discriminating Between Brazilian and European Portuguese National Varieties On Twitter Texts", *Proceedings of the 5th Brazilian Conference on Intelligent Systems*, Recife, 265–270.

Crystal, David (2012) *English as a global language*, 2. izdanje, New York, Cambridge University Press.

Ćoso, Bojana, Irena Bogunović (2017) „Person perception and language: A case of English words in Croatian", Language & Communication 53, 25–34, https://doi.org/10.1016/j.langcom.2016.11.001.

De Wilde, Vanessa, Brysbaert, Mark, June Eyckmans (2019) „Learning English through out-of-school exposure: Which levels of language proficiency are attained and which types of input are important?",

Bilingualism: Language and Cognition 23, 1, 171–185, http://dx.doi.org/10.1017/S1366728918001062.

Drljača, Branka (2006) „Anglizmi u ekonomskome nazivlju hrvatskoga jezika i standardnojezična norma", Fluminensia 18, 1, 65–85.

Drljača Margić, Branka (2009) „Latentno posuđivanje u hrvatskome i drugim jezicima – posljedice i otpori", Rasprave Instituta za hrvatski jezik i jezikoslovlje, 35, 1, 53–71.

Drljača Margić, Branka (2012) „Croatian university students' perception of stylistic and domain-based differences between Anglicisms and their native equivalents", *Languages, Literatures and Cultures in Contact: English and American Studies in the Age of Global Communication*, Vol. 2, *Language and Culture*, ur. Marta, Dąbrowska, Justyna, Lesniewska i Beata Piąte, Krakow, Tertium, 109–126.

Drljača Margić, Branka (2014) „Contemporary English influence on Croatian: A university students' perspective", *Language Contact Around the Globe. Proceedings of the LCTG3 Conference*, ur. Amel, Koll-Stobbe i Sebastian Knospe, Frankfurt am Main/Berlin/Bern/Bruxelles/New York/Oxford/Wien, Peter Lang, 73–92.

Duvnjak Jardas, Ivana (2019) „Anglicizmi u sportskoj terminologiji u hrvatskom jeziku", Zbornik radova Veleučilišta u Šibeniku 1,2, 185–194.

El-Dash, Linda Gentry, JoAnne Busnardo (2002) „Brazilian attitudes toward English: dimensions of status and solidarity", International Journal of Applied Linguistics 11, 57–74, http://doi.org/10.1111/1473-4192.00004.

Filipović, Rudolf (1990) *Anglicizmi u hrvatskom ili srpskom jeziku: porijeklo-razvoj-značenje,* Zagreb, Školska knjiga.

Fletcher, William H. (2011) „Corpus analysis of the world wide web", *The encyclopedia of applied linguistics*, ur. Carol A., Chapelle, Hoboken, New Jersey: Wiley-Blackwell, http://dx.doi.org/10.1002/9781405198431.wbeal0254.

Foro, Mirjana (2014) „Leksička razina publicističkog stila", Hrvatistika 7: 151–164.

Furiassi, Cristiano, Knut Hofland (2007) „The retrieval of false anglicisms in newspaper texts", *Corpus Linguistics 25 Years on. Language and Computers* 62, *Studies in Practical Linguistics*, ur. Roberta Facchinetti, Amsterdam/New York: Rodopi, 347–363.

Furiassi, Cristiano (2008) „What dictionaries leave out: new non-adapted Anglicisms in Italian", *Investigationg English with Corpora. Studies in Honor of Maria Teresa Prat*, ur. Aurelia, Martelli i Virginia Pulcini, Monza, Polimetrica International Scientific Publisher, 153–169.

Gluszek, Aagata, Karolina Hansen (2013) „Language attitudes in the Americas", *The Social Meanings of Languages, Dialects, and Accents: An International Perspective*, ur. Giles, Howard i Bernadette Watson, New York, Peter Lang, 26–44.

Görlach, Manfred (ur.) (2001) *A Dictionary of European Anglicisms: A Usage Dictionary of Anglicisms in Sixteen European languages*, New York, Oxford University Press.

Görlach, Manfred (ur.) (2002a) *An Annotated Bibliography of European Anglicisms*, Oxford, Oxford University Press.

Görlach, Manfred (ur.) (2002b) *English in Europe*, Oxford, Oxford University Press.

Graddol, David (2006) *English next: Why Global English may mean the end of „English as a foreign language*", London, British Council.

Halonja, Antun, Lana Hudeček (2014) „Pokloni mi svoj *selfie*", Hrvatski jezik 2: 26–27.

Hrvatska jezična riznica, Institut za hrvatski jezik i jezikoslovlje, http://riznica.ihjj.hr/dokumentacija/index.hr.html, posjet 10. studenog 2020.

Hrvatski nacionalni korpus, https://web.archive.org/web/20160606073223/http://www.hnk.ffzg.hr/, posjet 10. studenog 2020.

HR4EU, https://www.hr4eu.hr, posjet 18. srpnja 2022.

Hudeček, Lana, Milica Mihaljević (2005) „Nacrt za višerazinsku kontrastivnu englesko-hrvatsku analizu", Rasprave Instituta za hrvatski jezik i jezikoslovlje, 31, 107–151.

Hudeček, Lana, Milica Mihaljević (2015) „Anglizmi na stand-byu", Hrvatski jezik 2, 1–10.

Jurič, Boris, Vicko Krampus, Marta Račić (2013) „Anglizmi u hrvatskome poslovnom jeziku – tržništvo ili marketing", *Napredak* 154, 4, 567–579.

Kavgić, Aleksandar (2013) „Intended communicative effects of using borrowed English vocabulary from the point of view of the addressor: Corpus-based pragmatic analysis of a magazine column", Jezikoslovlje 14, 2,3, 487–499.

Kontekst.io, https://www.kontekst.io/hrvatski, posjet 30. listopada 2020.

Kilgarriff, Adam, Rychlý, Pavel, Smrž, Pavel, David Tugwell (2004) „Itri-04-08 the sketch engine", Information Technology 105–116.

Lauricella, Alexis R., Drew P. Cingel, Courtney K. Blackwell, Ellen Wartella, Annie Conway (2014) „The mobile generation: Youth and adolescent ownership and use of new media", Communication Research Reports, 31, 4, 357–364, http://doi.org/10.1080/08824096.2014.963221.

Losnegaard, Gyri Smørdal, Gunn Inger Lyse (2012) „A data-driven approach to anglicism identification in Norwegian, *Exploring newspaper language: using the web to create and investigate a large corpus of modern Norwegian*, ur. Gisele Andersen, John Benjamins, 131–154.

Ljubešić, Nikola, Filip Klubička. 2016. {bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian. In Felix Bildhauer & Roland Schäfer (eds.), Proceedings of the 9th web as corpus workshop (WaC-9), 29–35. Gothenburg: Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/W14-0405.

Marginson, Simon, Marijk van der Wende (2007) „Globalisation and Higher Education", *OECD Education Working Papers*, OECD Publishing, https://doi.org/10.1787/19939019.

Matić, Daniela. (2017) „Perception of the English element in the scientific register of Croatian ICT university educational material with graduate ICT students", Jezikoslovlje, 18, 2, 319–345.

McKenzie, Robert M. (2010) *The Social Psychology of English as a Global Language: Attitudes, Awareness and Identity in the Japanese Context*, New York, Springer.

Međeral, Krešimir (2016) „Jezične bakterije – pomagači ili štetočine u jezičnome organizmu?", Hrvatski jezik 3, 1–10.

Mihaljević, Milica (2003) *Kako se na hrvatskome kaze WWW? Kroatistički pogled na svijet računala*, Zagreb, Hrvatska Sveučilišna naklada.

Mihaljević Djigunović, Jelena, Kristina Cergol, Li Qingmin (2006) „Utjecaj medija na nenamjerno usvajanje engleskog vokabulara", *Jezik i mediji – Jedan jezik: više svjetova*, ur. Jagoda Granić, Zagreb, Split: Hrvatsko društvo za primijenjenu lingvistiku, 445–452.

Mihaljević Djigunović, Jelena, Renata Geld (2003) „English in Croatia today: Opportunities for incidental vocabulary acquisition", Studia Romanica et Anglica Zagrabiensia, 43: 335–352.

Mišić-Ilić, Biljana, Vesna Lopičić (2011) „Pragmatički anglicizmi u srpskom jeziku", Zbornik Matice srpske za filologiju i lingvistiku 54, 1, 261–273.

Mrežnik, http://ihjj.hr/mreznik/, posjet 27. svibnja 2021.

Muhvić-Dimanovski, Vesna (2005) *Neologizmi: problemi teorije i primjene*, Zagreb, Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Muhvić-Dimanovski, Vesna, Anita Skelin Horvat (2008) „Contests and nominations for new words – why are they interesting and what do they show", Suvremena lingvistika, 65, 1–26.

Muhvić-Dimanovski, Vesna, Anita Skelin Horvat, Diana Hriberski (2016). R*ječnik neologizama u hrvatskome jeziku*, www.rjecnik.neologizam.ffzg.unizg.hr, posjet 12. studenog 2020.

Núñez Nogueroles, Eugenia E. (2016) „Anglicisms in CREA: A Quantitative Analysis in Spanish Newspapers", Language Design, 18, 215–242.

Nikolić-Hoyt, Anja. (2005) „Englesko-hrvatski jezično-kulturni dodiri", *Jezik u društvenoj interakciji, Zbornik radova sa savjetovanja održanoga 16. i 17. svibnja u Opatiji,* ur. Diana Stolac, Nada, Ivanetić i Boris Pritchard, Zagreb, Rijeka, Hrvatsko društvo za primijenjenu lingvistiku, 353–358.

Núñez Nogueroles, Eugenia Esperanza (2016) „Anglicisms in CREA: A Quantitative Analysis in Spanish Newspapers", Language Design, 18, 215–242.

Opačić, Nives (2007a) „Prodor engleskih riječi u hrvatski jezik", Jezik, 54, 22–27.

Opačić, Nives (2007b) „Odakle stižu ozbiljnije prijetnje hrvatskom jeziku: izvana ili iznutra?", Lahor, 4, 279–291.

ParlaMint-HR 2.1, https://www.clarin.si/noske/run.cgi/corp_info?corpname=parlamint21_hr&struct_attr_stats=1, posjet 18. srpnja 2022.

Patekar, Jakob (2019) „Prihvatljivost prevedenica kao zamjena za anglizme", Fluminensia 31,2, 143–179, https://doi.org/10.31820/f.31.2.17.

Peti-Stantić, Anita (2013) „Domaće je (naj)bolje", *Javni jezik kao poligon jezičnih eksperimenata*, ur. Barbara Kryžan-Stanojević, Zagreb, Srednja Europa, 39–51.

Peti-Stantić, Anita i sur. (2018). Hrvatska psiholingvistička baza (HPB). Projekt HRZZ-IP-2016-06-1210. Modeliranje mentalne gramatike hrvatskoga: ograničenja informacijske strukture.

Pulcini, Virginia, Cristiano Furiassi, Felix Rodríguez González (2012) „The lexical influence of English on European languages: From words to phraseology", *Anglicization of European Lexis*, ur. Virginia, Pulcini, Cristiano, Furiassi i Felix Rodríguez González, Amsterdam, Philadelphia, John Benjamins Publishing Company, 1–27.

Raos, Nenad (2006) „O potrebi razlikovanja hrvatskoga i engleskog jezika", Arh Hig Rada Toksikol 57, 405–412.

Runjić-Stoilova, Anita, Anamarija Pandža (2010) „Prilagodba anglizama u govoru na hrvatskim televizijama", Croatian Studies Review 6, 1, 229–240.

Serigos, Jacqueline Ray Larsen (2017) *Applying Corpus and Computational Methods to Loanword Research:New Approaches to Anglicisms in Spanish*, doktorska disertacija, Austin: University of Texas.

Skelin Horvat, Anita (2015) „Jezik adolescenata – od slenga preko kolokvijalnoga do standarda", *Nestandardni hrvatski jezik prema standardnom hrvatskom jeziku: zbornik radova*, ur. Anđa Suvala, i Jasna Pandžić, Zagreb: Institut za hrvatski jezik i jezikoslovlje; Agencija za odgoj i obrazovanje, 67–72.

Škifić, Sanja, Emilija Mustapić (2012) „Anglizmi i hrvatsko računalno nazivlje kroz prizmu jezičnog konflikta i jezične ideologije", Jezikoslovlje 13, 3, 809–839.

Tadić, Marko (2003) *Jezične tehnologije i hrvatski jezik*, Zagreb, Exlibris.

Tadić, Marko, Dunja Brozović-Rončević, Amir Kapetanović (2012) *Hrvatski jezik u digitalnom dobu,* Heidelberg, Springer, http://doi.org/10.1007/978-3-642-30882-6.

Tadić, Marko (2016) Jezici i jezične tehnologije u Hrvatskoj. Radionica ELRC-a u Hrvatskoj, Zagreb, http://www.lr-coordination.eu/sites/default/files/Croatia/04_S5_Zagreb_Languages_and_LT_in_Croatia_HR.pdf, posjet 10. srpnja 2020.

Turk, Marija, Maja Opašić, (2008) „Linguistic borrowing and purism in the Croatian language", Suvremena lingvistika, 65, 73–88.

SUMMARY

Mirjana Borucinsky, Irena Bogunović

EXTRACTING ENGLISH WORDS FROM A CORPUS OF CROATIAN

As the *lingua franca* of the modern age, English has become the dominant donor language for many languages, including Croatian. The influence of English on Croatian is evident across different registers and linguistic levels, especially the lexical one. Recently, more and more English words have started to appear in their unadapted form (e.g., *freelancer, chat, e-mail*) in Croatian, especially in the news and social media. English words can be extracted from corpora either manually, by using existing corpus linguistics tools or by developing new tools. The aim of this paper is to analyse whether the existing tools for Croatian can yield a list of unadapted English words. For that purpose, the web corpus (*hrWaC*) was analysed using the *Sketch Engine* platform. A list of 1217 English words was composed using this method. The results showed that it is possible to compile a list of English words and their frequencies with the help of the available tools and resources for the Croatian language, but also that there are many problems due to which the results cannot be considered completely reliable. Moreover, the procedure itself still has to be combined with other manual methods and classifications, and there is a need for the development of new tools for automatic extraction of English words from a corpus of Croatian.

Key words: *English words; Croatian language; corpus linguistics*