DOI 10.5673/sip.60.2.8
UDK 070:004(497.5):81
Izvorni znanstveni rad

Germanizmi u medijskom prostoru*

Lidija Tepeš Golubić
Tehničko veleučilište u Zagrebu, Hrvatska
e-mail: ltepes2@tvz.hr
ORCID: 0000-0002-3297-047X

SAŽETAK Okosnicu istraživanja opisanom u članku čini popis germanizama formiran iz rječnika hrvatskoga jezika, rječnika stranih riječi i doktorskih disertacija koje se bave proučavanjem germanizama u hrvatskom jeziku. Pripremljeni popis germanizama omogućio je njihovu računalnu analizu u medijskom prostoru, odnosno hrvatskom mrežnom korpusu hrWaC-u, koji obuhvaća tekstove objavljene na *web*-u u razdoblju od 4 godine. Jezične tehnologije potpomognute naknadnom ručnom obradom podataka omogućile su utvrđivanje učestalosti korištenja germanizama u pisanim tekstovima.

Metodama automatske detekcije germanizama pronašli smo germanizme u svim njihovim oblicima koji su danas u uporabi u suvremenim tekstovima, čime je ujedno stvoren i čestotni rječnik germanizama.

Provedenim istraživanjem utvrđena je zastupljenost germanizama u tekstovima suvremenog hrvatskog jezika. Iako ukupni broj germanizama koji se danas pojavljuje u hrvatskim rječnicima i ostalim konzultiranim izvorima iz kojih je formiran osnovni popis iznosi 17 988 lema, u uporabi je u suvremenim tekstovima dokazano 8 400 lema. Potvrđenih 8 400 lema dokaz su tomu da se leksičko blago hrvatskog jezika zabilježeno u rječnicima nije izgubilo u suvremenom hrvatskom jeziku, nego je sustavno ušlo u korpus hrvatskog jezika te postoji u tekstovima koji nisu nužno standardnojezični.

Ključne riječi: medijski prostor, hrvatski *web*-korpus, računalna analiza, germanizam, lema.

* Dijelovi ovog rada prethodno su objavljeni 2016. godine u autoričinu doktorskom radu *Germanizmi u digitalnim novinskim korpusima hrvatskoga jezika* Sveučilišta u Zagrebu.

1. Uvod

Tragove utjecaja njemačkog jezika u hrvatskom jeziku nalazimo do današnjih dana. Iako se germanizmi češće koriste u govornom jeziku, dokaz koji se može naći u znanstvenoj literaturi pokazuje da se koriste i u pisanom jeziku, odnosno u medijskom prostoru ako internet i mrežne korpuse, koji se temelje na istom, promatramo kao medijski prostor.

Kao posljedica političkih konstelacija u europskoj povijesti hrvatski jezik bio je pod utjecajem njemačkog jezika do 20. stoljeća. Zbog intenzivnih političkih i gospodarskih kontakata germanizmi su ušli u hrvatski jezik u velikoj mjeri, a najveća akvizicija dogodila se između 1527. i 1868. godine kada je Hrvatska bila dio Habsburške Monarhije. S druge strane, od rane faze standardizacije hrvatskoga jezika do danas provodi se proces jezične purifikacije što je smanjilo broj germanizama u standardnom jeziku, tako da se danas rjeđe mogu naći u standardnom hrvatskom jeziku.

Germanizmi su u hrvatskim dijalektima i dalje snažno zastupljeni i vrlo dobro integrirani te prilagođeni fonološkoj, morfosintaktičkoj i semantičkoj razini jezika (Filipan-Žignić, 2007.). Na fonološkoj razini postoji nekoliko razlika u fonemskoj realizaciji riječi zbog nepostojanja hrvatskih ekvivalenata (npr. pf → f). Što se tiče morfosintaktičke razine, germanizmi su usvojili hrvatske deklinacijske i konjugacijske paradigme izuzevši grupu pridjeva koji pripadaju deklinacijskoj paradigmi koja je atipična za hrvatski jezik. Na semantičkoj razini mogu se čak naći i slučajevi gdje su germanizmi u hrvatskom jeziku postali pejorativi.

Medijski prostor za istraživanje germanizama čini hrvatski internetski prostor, odnosno mrežni korpus hrWaC prikupljen s *.hr* mrežne domene. Istraživanjem, računalnom obradom i naknadnom ručnom obradom podataka dobiven je razmjerno opsežan popis, odnosno rječnik germanizama koji može poslužiti za daljnja komparativna istraživanja kako u našem, tako i u drugim jezicima.

2. Pisani izvori za formiranje popisa germanizama

Pretpostavka istraživanju germanizama u hrvatskim korpusima pomoću jezičnih tehnologija formiranje je popisa germanizama koje nalazimo u relevantnim pisanim izvorima. Osnovni popis čine germanizmi prikupljeni iz rječnika hrvatskoga jezika, rječnika stranih riječi i regionalnih rječnika te germanizmi iz relevantnih znanstvenih i stručnih radova koji se bave istraživanjem germanizama u hrvatskom jeziku.

Popis germanizma, čija je pojavnost u nastavku istražena primijenjenim jezičnim tehnologijama, formiran je iz sljedećih rječnika i disertacija: „Rječnik hrvatskoga jezika“ urednika Jure Šonje (2000.); „Veliki rječnik hrvatskoga jezika“ autora Vladimira Anića

(2003.); „Novi rječnik stranih riječi“ Bratoljuba Klaića (2012.); „Rječnik stranih riječi“ autora Šime Anića, Nikole Klaića i Želimira Domovića (2002.); „Agramer. Rječnik njemačkih posuđenica u zagrebačkom govoru“ autorice Zrinjke Glovacki-Bernardi (2013.); „Esekerski rječnik“ Velimira Petrovića (2008.); disertacije Ive Medića, „Kulturno-historijsko značenje i lingvistička analiza njemačkih pozajmljenica kod zagrebačkih obrtnika (obrada metala, drva i kože): prilog pitanju utjecaja njemačkog jezika na hrvatske specijalne jezike.“ (1962.), disertacije Tea Bindera (1954.) te disertacije Velimira Piškorca, „Germanizmi u podravskom dijalektu“ (2001).

U svrhu ovog istraživanja određeno je da je germanizam svaka riječ preuzeta u hrvatski jezik bez obzira na to je li njemački pritom imao ulogu jezika posrednika ili je riječ izravno preuzeta iz njemačkog, odnosno austrijskog jezika u hrvatski jezik.

Etimološka odrednica koja je za potrebe našeg istraživanja *njem. – iz njemačkoga, njemački* oznaka je za rječničke natuknice preuzete iz njemačkog jezika. Preuzete natuknice prikazane su u nastavku na primjeru natuknice iz rječnika Vladimira Anića. Naime, „Rječnik hrvatskoga jezika“ Vladimira Anića sadrži 2 321 germanizam. Natuknica koja se odnosi na riječi preuzete iz njemačkog jezika označena je s *njem.*, a natuknice su formirane na sljedeći način:

> šminka ž 1. a. sredstvo za uljepšavanje lica b. sredstvo za uljepšavanje lika glumca prema ulozi 2. odjel u kazalištu, filmskim studijima i sl. koji brine o izgledu osobe koja nastupa 3. *žarg.* ono što je usmjereno na vanjski učinak, izazivanje površnih efekata; vanjština bez sadržaja
> ◊ *njem.* (Anić, 2003.: 1542)

Iz Anićeva rječnika, koji ovdje služi kao primjer, preuzete su sve riječi koje su okarakterizirane oznakom *njem.*, bez obzira na to uključuje li takva etimološka oznaka još koju, primjerice *lat.*, *tal.* i slično, što upućuje na to da je predmetna riječ u hrvatski jezik preuzeta iz nekog trećeg jezika, kroz njemački jezik kao posrednik:

> direktiva ž. 1. *pol. ideol.* izravna naredba o tome kako treba postupati, izdaje ju viši politički organ nižemu ili pojedinac nižima po hijerarhiji 2. općenito, smjernica, nalog, uputa
> ◊ *njem.* ← *fr.* (Anić, 2003.: 219)

Tako RHJ obuhvaća ukupno 921 germanizam, Klaićev „Novi rječnika stranih riječi“ 6 199 germanizama, a u „Rječniku stranih riječi“ autora Anića, Klaića i Domovića utvrđena su 1 193 germanizma. Regionalni rječnik „Agramer“ rječnik je njemačkih posuđenica u lokalnom zagrebačkom govoru koji obuhvaća 5 703 riječi njemačkog podrijetla, a Esekerski rječnik Velimira Petrovića tvori ukupno 7 731 natuknica.

Ukupan broj različitih germanizama koji smo dobili analizom rječnika i radova iznosi 17 988.

Usporedna analiza rječnika pokazala je da se određeni broj germanizama ne navodi u svim rječnicima već u jednom ili dva. Jednako tako, u nekim smo se situacijama susreli s različitim načinom bilježenja germanizama pa su (obje) inačice bilježenja unesene u naš popis.

Esekerski rječnik svojim korpusom uvelike odstupa od ostalih budući da bilježi njemačke riječi govornika esekerskog dijalekta čije su inačice uvjetovane materinskim dijalektom roditelja pojedinih govornika, njemačkim govornim jezikom dijela njemačkog stanovništva ili jezicima u dodiru, što sve svjedoči o jezičnoj razlici među Esekerima odraslim u različitim dijelovima Osijeka (Petrović 2008.: 6).

Iako ukupan popis sastavljen iz analiziranih rječničkih izvora i znanstvenih radova čini 17 988 germanizama, samo 8 400 germanizama pronađeno je pretraživanjem hrWaC *web*-korpusa.

Iz toga zaključujemo da je podskup germanizama pronađen u hrWaC-u (32% od svih prikupljenih germanizama, tj. 8 400 lema) relevantan podskup za istraživanje germanizama koji se još uvijek koriste u suvremenom hrvatskom jeziku. Budući da hrWaC sadrži tekstove s cijele *.hr web*-domene, daje i vjerodostojnu sliku ne samo hrvatskog standardnog jezika, nego i razgovornog jezika, dijalektalizama, regionalizama i žargonizama, dakle svega onoga što predstavlja cjeloviti sustav hrvatskoga suvremenog jezika. Stoga je činjenica da je presjek ukupnog broja germanizama pronađenih samo u rječničkim izvorima jednak presjeku ukupnog broja germanizama pronađenih u rječničkim i korpusnim izvorima dokaz tomu da se leksičko blago hrvatskog jezika zabilježeno u rječnicima nije izgubilo u suvremenom hrvatskom jeziku, nego je sustavno ušlo u korpus hrvatskog jezika te postoji u tekstovima koji nisu nužno standardnojezični, primjerice chatovi, forumski postovi, wiki-stranice, tweetovi, blogovi ili komentari na društvenim mrežama.

2. Metoda i formiranje korpusa

Prethodno su navedeni rječnički i ostali izvori koji su poslužili za formiranje popisa germanizama koji su se onda računalno priredili i pretraživali u računalnom korpusu hrvatskog jezika, *web*-korpusu hrWaC-u, koji je automatski prikupljen s internetske domene *.hr* i opsega je 1.9 milijardi pojavnica u verziji hrWaC 2.1.

Na temelju toga popisa formirala se baza podataka svih germanizama zabilježenih u spomenutim izvorima koja dodatno sadrži i izvornu njemačku riječ koja je posuđena u hrvatski jezik te njezin prijevod.

Naknadno su posebno označeni jednoznačni hrvatski ekvivalenti (ako postoje), a inače se u bazi uz svaki germanizam navodi opisni prijevod. Baza za svaki pojedini germanizam bilježi i izvore u kojima se isti pojavljuje te vrstu riječi, što se pokazalo važnim u detektiranju lažnih germanizama, tj. onih značenja u kojima je riječ podrijetlom iz hrvatskog, a ne iz njemačkog jezika (primjerice, *gol* – hrvatski pridjev u značenju *nag, neodjeven* i gol – germanizam u značenju *pogodak, zgoditak).*

Ova baza podataka predstavlja dodatni rezultat planiranog istraživanja te sama po sebi funkcionira kao rječnik germanizama te predstavlja stručni i znanstveni doprinos.

Prikupljeni germanizmi računalno su uspoređeni s javno dostupnim korpusom hrvatskoga jezika hrWaC-om (Ljubešić; Tomaž, 2011.) sastavljenim prema konceptu „Web as Corpus", koji sadrži tekstove prikupljene automatskom metodom s internetske domene *.hr* i opsega je 1.9 milijardi pojavnica u verziji 2.1. Korpus je morfosintaktički označen i lematiziran i omogućuje postavljanje upita od više riječi, tj. od sintagme, pretragu pomoću dodatnih lingvističkih obavijesti (npr. lema, vrsta riječi, gramatičke kategorije) te uporabu regularnih izraza.

Dosadašnja istraživanja germanizama provodila su se korištenjem književno-lingvističkog pristupa obradi germanizama uporabom računala samo za bilježenje i uređivanje podataka. Ovim se radom automatsko pronalaženje germanizama u svim njihovim morfološkim oblicima razradilo postojećim rječničkim bazama hrvatskoga standardnog jezika te automatskim generiranjem svih mogućih oblika pojedinih riječi hrvatskoga jezika čime se proširila baza germanizama hrvatskih rječnika i značajnih radova koji se bave jezičnom analizom germanizama.

3. HrWaC – hrvatski *web* korpus za istraživanje germanizama

Internet sadrži ogromne količine teksta na mnogo jezika i obuhvaća velik broj jezičnih varijeteta te ogroman skup tema. Jezikoslovci, računalni lingvisti i leksikografi sve ga više koriste kao izvor jezičnih podataka zbog njegove veličine i jer je jedini dostupan izvor za jezik i jezični varijetet koji je predmet njihova interesa te zato što je javno i trenutno dostupan. Kilgarriff i Grefenstette (2003.) u svom predgovoru specijalnom broju časopisa „Computational Linguistics" pišu o webu kao korpusu i definiraju korpus kao skup tekstova u kontekstu u kojem je korpus objekt jezične ili književne studije. *Web*-tekstovi proizvodi su raznovrsnih autora i, u usporedbi s recenziranim i lektoriranim papirnatim tekstovima, mogu nastati jeftino i brzo, bez mnogo promišljanja o pravopisnoj točnosti. *Web* je nečisti korpus, ali pravopisno ispravne konstrukcije imaju veću frekvenciju od neispravnih što je iznimno važno.

Korpusi se često koriste za ekstrakciju modela jezika: popisa ponderiranih riječi ili kombinacija riječi koje opisuju odnose među riječima te njihovu učestalost u određenoj domeni. U računalnoj obradi govora modeli jezika koriste se za predviđanje koje

su vjerojatne kombinacije riječi moguća interpretacija zvučnog vala. U pretraživanju informacija modeli se koriste za donošenje odluke o tome koje riječi su korisni pokazatelji teme, a u strojnom prevođenju za identificiranje dobrih kandidata za prijevod. Danas je zahvaljujući *web*-korpusima moguće parsati *web* i pretraživati ga prema lemama (različnicama), konstituentima (primjerice, pridjevska sintagma) i gramatičkim odnosima. Rječnici i leksikoni mogu se razviti izravno i lako korištenjem weba za razne, čak i vrlo rijetke jezike.

Hrvatski *web*-korpus hrWaC hrvatski je *web*-korpus veličine 1.2 milijardi pojavnica u verziji 1.0 (Ljubešić i Erjavec, 2011.) te 1.9 milijardi pojavnica u verziji 2.1 (http://nlp.ffzg.hr/resources/corpora/hrWaC/), sagrađen s ciljem dobivanja što čišćeg *web*-korpusa. Ljubešić i Erjavec u radu iz 2011. godine navode točnost od 97,9% i odziv od 70,7%. hrWaC je trenutno najveći korpus hrvatskog jezika, a verzija 1.0 nastala je 2011. godine prikupljanjem pojavnica pretraživanjem cijele *.hr* internetske domene što je rezultiralo korpusom od cca 1.9 milijardi pojavnica. Korpus je očišćen od HTML koda, lematiziran i automatski morfosintaktički označen pomoću CroTag sustava (Agić i sur., 2008.).

Šnajder i sur. (2013.) ukazuju na činjenicu da, osim nezaobilaznih pravopisnih i gramatičkih pogrešaka, hrWaC još uvijek sadrži netekstni sadržaj (primjerice, isječke kodova i strukturu oblikovanja), kodirane pogreške i sadržaj na stranim jezicima. Budući da to znatno utječe na jezičnu obradu, Šnajder i sur. 2013. su dodatno filtrirali korpus. Kao prvo, uklonjen je dio sadržaja hrWaC-a prikupljen s glavnih forumskih *web*-stranica i blogovskih *web*-stranica. Taj sadržaj uglavnom je negramatičan i rijetko sadrži dijakritike što je tipično za korisnički generirani sadržaj. U tom koraku uklonjena je trećina podataka. Preostali dio korpusa obrađen je tako da su rečenice međusobno segmentirane i rastavljene na pojavnice čime se dobilo 66 milijuna rečenica. Zatim je primijenjen niz heurističkih filtera na razini dokumenta i rečenice tako da su na razini dokumenta izbačeni svi dokumenti čija je duljina bila ispod unaprijed određenog praga, koji nisu sadržavali dijakritičke znakove, koji nisu sadržavali riječi s čestotnog popisa hrvatskih riječi, ili su sadržavali barem jednu riječ s popisa riječi stranog jezika (za srpski). Na rečeničnoj razini izbačene su rečenice čija je duljina bila ispod unaprijed određenog praga, koje su sadržavale nestandardne simbole, koje nisu sadržavale dijakritičke znakove ili su sadržavale previše stranih riječi s popisa riječi stranih jezika (za engleski i slovenski jezik). Konačna filtrirana verzija korpusa hrWaC sadrži 51 milijun rečenica i 1.2 milijarde pojavnica. HrWaC korpus javno je dostupan za preuzimanje, kao i detaljni opis koraka predobrade. Ovaj filtrirani korpus pogodan je za zadatke obrade prirodnog jezika u kojima je kvaliteta jezika važnija od pokrivenosti (npr. za parsanje).

Za morfosintaktičko označavanje, lematizaciju i ovisnosno parsanje hrWaC-a korišteni su javno dostupni alati s modelima razvijenima nad SETimes hrvatskim novinskim

korpusom (SETIMES.HR), koji je dio Southeast European Times (SETimes) paralelnog korpusa. SETimes je paralelni korpus engleskog jezika i jezika jugoistočne Europe, a obuhvaća sadržaj objavljen na news portalu SETimes.com koji objavljuje vijesti iz jugoistočne Europe na deset jezika: bugarski, bosanski, grčki, engleski, hrvatski, makedonski, rumunjski, albanski i srpski.

Ljubešić i Klubička u radu iz 2014. opisuju najnoviju verziju hrWaC korpusa (2.0) s 1.9 milijardi pojavnica. Podaci su dobiveni na webu pomoću „Brno pipeline" alata za obradu *web*-korpusa (Suchomel i Pomikalek, 2012.), kojim su prikupili sadržaj, detektirali kodiranje znakova hrvatskog jezika, ekstrahirali sadržaj i uklonili duplikate. Korpus je lematiziran pomoću CST lematizatora (Jongejan i Dalianis, 2009.), morfosintaktički označen HunPos označivačem (Halácsy et al, 2007.), a ovisnosna sintaksa kodirana je pomoću *mate-tools* alata za analizu prirodnog jezika (Bohnet, 2010.).[1] Svi modeli uvježbani su na hrvatskom označenom korpusu SETimes.HR od 90 000 pojavnica (Agić i Ljubešić, 2014.) koji su proširili s 50 000 dodatnih pojavnica iz raznih novinskih domena (nazivaju ga SETimes.HR +).

Za polazne URL adrese (eng. seed URLs) koristile su se početne stranice *web*-domena dobivene tijekom izgradnje prve verzija hrWaC korpusa. Broj tih polaznih URL-ova je 14 396.

4. Računalna obrada prikupljenih i za obradu pripremljenih germanizama

Hrvatski jezik pripada flektivnim jezicima i kao takav omogućava da se imenice, pridjevi i glagoli realiziraju u brojnim različitim oblicima koji označavaju padež, broj, glagolsko vrijeme, lice i druge gramatičke kategorije. Riječi koje pripadaju otvorenom vokabularu hrvatskog jezika mogu se u tekstovima pojaviti u različitim morfosintaktičkim oblicima. Da bi se u tekstu ti oblici povezali sa zajedničkim kanonskim oblikom (lemom), često se normaliziraju na zajednički korijen postupkom korjenovanja ili na osnovni kanonski oblik postupkom lematizacije.

Korjenovanje podrazumijeva uklanjanje afikasa, prefiksa i sufiksa iz oblika riječi da bi se dobio korijen zajednički svim oblicima. Korijen dobiven takvim postupkom ne mora nužno odgovarati pravom korijenu riječi u lingvističkom smislu. Pojednostavljeno, korijen riječi ono je što ostane nakon što istoj riječi odstranimo prefiks i sufiks. Morfološke varijante / oblici riječi imaju različite nastavke, ali u suštini opisuju isto značenje. Ove različite varijante, dakle, mogu biti spojene u zaseban reprezentativni oblik – korijen.

[1] https://code.google.com/p/mate-tools/

Lematizacija je proces sličan određivanju korijena riječi. Razlika je u tome što se prilikom lematizacije riječ, umjesto na korijen riječi, svodi na osnovni oblik, lemu.

Postoje različiti oblici u kojima se jedan germanizam može pojaviti i koji se razlikuju prema obliku, ali ne i prema značenju. Primjerice, riječi *bauštelom, bauštelama i bauštelі* su različiti oblici iste leme (*bauštela*) i te se riječi u korpusnoj lingvistici nazivaju različnicama (*eng. type),* dok se svako individualno pojavljivanje svake od tih riječi u korpusu naziva pojavnicom (*eng. token).* Dakle, niz svih oblika množine (nominativ, genitiv, dativ, akuzativ, vokativ, lokativ, instrumental) gore spomenutog primjera sastoji se od sljedećih 7 pojavnica: *bauštele, bauštela, bauštelama, bauštele, bauštele, bauštelama i bauštelama.* Budući da je pojavnica svako pojedinačno pojavljivanje riječi u korpusu, pod pojmom milijardni korpus (hrWaC ima 1.9 milijardi pojavnica) podrazumijevamo korpus od milijardu pojavnica. Različnica je pak jedinstveni oblik pojavnice iz korpusa. Dakle, u gornjem se nizu nalaze samo 3 različnice: *bauštele, bauštela, bauštelama*. S druge strane, germanizmi *bauštela* i *bauštelac* značenjski se razlikuju jer je riječ o dvjema različitim lemama.

Dakle, lema imenice njezin je oblik u nominativu jednine, primjerice *bauštela, bager*, a lema glagola infinitiv, primjerice *fušati*.

U nekim slučajevima, što je pokazalo i naše istraživanje, nije moguće jednoznačno odrediti lemu neke riječi. Primjerice riječ *lista* pripada sljedećim lemama: imenici *list* i glagolu *listati*.

Postupkom lematizacije moguće je riječi *lista* pridružiti ispravan natuknički oblik riječi.

Morfološki analizatori alati su koji riječ pridružuju njenoj lemi, a riječ se potom može pronaći u jezičnoj bazi ili leksikonu / rječniku. Morfološki generatori alati su koji lemi pridružuju sve odgovarajuće morfološke oblike (primjerice oblike jednine i množine za imenicu), a riječ se potom pronalazi u računalnom korpusu. Posljedično, morfološki analizator / generator važna je komponenta sustava obrade prirodnog jezika.

Flektivni sufiksi nosioci su sintaktičkih i semantičkih informacija potrebnih za sintaktičke i logičke analize rečenica. Za razliku od velikih indoeuropskih jezika, kao što su engleski i francuski, gdje je morfološka analiza često tako jednostavna da literatura o računalnoj obradi tih jezika obično izostavlja morfološku raspravu, analiza oblika riječi za slavenske ili ugrofinske jezične skupine često predstavlja problem.

Prilikom određivanja korijena i sufiksa germanizama (tj. leme i sufiksa) koristili smo se Hrvatskim morfološkim leksikonom (HML: http://hml.ffzg.hr/) u sklopu predistraživanja koje je obuhvatilo podskup od 1 360 germanizama ekstrahiranih iz skupa pri-

kupljenih 17 988 germanizama. HML je leksikon koji se sastoji od oko 11 000 lema te više od 4 000 000 oblika riječi. Leksikon je namijenjen profesionalcima i sustavima za pretraživanje podataka, ali i za dubinsku analizu teksta i različite zadatke obrade prirodnojezičnih podataka, tj. tekstova na hrvatskom jeziku. Leksikon je distribuiran pod CC-BY-NC-SA licencom.

Nedostatak ovog leksikona za naše istraživanje vrlo je mali broj germanizama koji su pronađeni među tih 11 000 lema. U HML-u smo otkrili 734 germanizma (lema sa svim oblicima riječi) od 1 360 germanizama (lema) s popisa.

Lematizacija svih 17 988 germanizama izvršena je pomoću odostražnog rječnika, kako bi se kanonskom obliku pridružili odgovarajući nastavci. Ova metoda osigurala je točne rezultate u daljnjem procesu pronalaženja germanizama.

Prvi korak u procesu pronalaženje germanizama je stoga bio utvrditi sve moguće oblike za svaku lemu-germanizam koji su rezultat dekliniranja (imenica i pridjeva) i konjugacije (za glagole). Nakon toga, pripremljen je skup koji se sastoji od oblika nastalih dodavanjem sufiksa svakom korijenu riječi (osnovi koja je zajednička svim oblicima iste leme), što je čak i za izvorne hrvatske riječi često proces koji uključuje određeni postotak pogrešaka ako se radi automatski (Tepeš Golubić et al., 2013.).

Za pretragu tekstova u dnevnim novinama (što je bio prvi korak u istraživanju, prije pretraživanja *web*-korpusa) RegExp metodom (Regular Expression metoda) za svaku lemu stvoren je regularni izraz kojim se tekst može upariti s uzorkom. Nizovi za pretraživanje stvoreni su prema predlošku: ručno određenim korijenom riječi (najkraći zajednički niz znakova prije fonetske / morfološke promjene) kojem je dodan sufiks (ručno određen prema postojećim gramatičkim pravilima).

Primjer:

- korijen riječi & (sufiks1 | sufiks2 | sufiks3 | sufiksN...)
- grup & (a | u | om...)

Metoda je omogućila generiranje gramatičkih pravila na relativno ispravan način koristeći samo RegExp matricu.

Korijenski završeci čine veliku zbirku odsječaka koji povezuju korijen riječi sa sufiksima, primjerice za broj, padež, vrijeme ili lice. Na primjer, hrvatski jezik ima različite paradigme za ženski, muški i srednji rod, za živo i neživo, označena flektivnim sufiksima i različitim korijenskim završetcima u sedam padeža jednine i množine.

Primjer:

- lema → grupa

Korijen je određen kao: *grup* i dodani su sufiksi deklinacijske paradigme za imenice ženskog roda (a | e | i| | o | u | om | ama). Primjena RegExp matrice kao rezultat dala je konstrukciju koja izgleda ovako:

- \bgrup (a | e | i| | o | u | om | ama) \b

Navedeni regularni izraz omogućio je pronalaženje sljedećih flektivnih oblika leme *grupa* u novinskom tekstu: *grupa*, *grupe, grupi, grupu, grupo, grupom, grupama*.

RegExp metoda primijenjena je na manji uzorak germanizama u ovom predistraživanju (1 360 lema). No, određivanje korijena riječi nije uvijek dalo zadovoljavajuće rezultate pa se dodatno koristilo „pravilo posljednjeg znaka u lemi" za automatsko ispravljanje pogrešaka dobivenih RegExp metodom.

Primjer:

- lema → cement

Korijen je određen kao: *cement* i dodani su sljedeći sufiksi deklinacijske paradigme za imenice muškog roda (a | u | om...). Kao rezultat, primjena RegExp matrice dala je konstrukciju koja izgleda ovako:

- \ bcement (a | u | om...) \ b

Navedeni regularni izraz omogućio je pronalaženje sljedećih flektivnih oblika leme *cement* u novinskom tekstu: *cementa, cementu, cementom*, ali ne i *cement*. Stoga, kako bi dobiveni uzorak bio potpun, izdvojio se zadnji znak iz ove i ostalih lema, leme su potom uzlaznim poretkom razvrstane prema tom znaku, a znak je pretvoren u jedan od mogućih sufikasa.

- Rezultat: cement - t = cemen + (t | ta | tu | tam...)

Posljedično, metoda se pokazala uspješnom ne samo za cement, nego i za mnoge druge germanizme, kao što su *bankrot, bizmut, balast, bruderšaft, recept* itd.

Stvorili smo odostražni rječnik 17 988 germanizama prikupljenih iz opisanih izvora i sortiranih odostražno. Odostražni rječnik rječnik je u kojem riječi nisu poredane abecednim poretkom kao u tradicionalnom rječniku. Organizacija rječnika temelji se na razvrstavanju svakog leksičkog unosa prema njegovu posljednjem slovu / znaku i potom znakovima koji slijede od kraja prema početku riječi (primjerice, riječ *cement* je u odostražnom rječniku *tnemec*). Posljedično, u takvom se rječniku sve riječi koje imaju isti sufiks se pojavljuju slijedno. Za razliku od odostražnog rječnika, u standardnom rječniku natuknice su organizirane tako da se riječi s istim prefiksom pojavljuju slijedno budući da redoslijed sortiranja započinje prvim slovu leksičkog unosa / natuknice koje potom slijede sva ostala slova prema kraju riječi.

Takav odostražni rječnik koristan je za računalne lingviste, ali i za pjesnike koji traže riječi koje završavaju posebnim sufiksom ili antropologe te forenzičke stručnjake koji analiziraju oštećeni tekst (npr. kameni natpis ili spaljeni dokument) koji posjeduje samo završni dio određene riječi. Koristi se za pronalaženje riječi s točno zadanim sufiksom (tj. sufiksom koji je nosilac značenja, kao *-ina*, *-ost* itd.) i riječi koje posjeduju isti završetak kao i određenu riječ zadanu upitom.

Svakoj riječi, germanizmu, u našem odostražnom rječniku u bazi je pridružena oznaka morfosintaktičke kategorije i broj paradigme. Iz tih podataka moguće je izdvojiti heurističke informacije o ekvivalentnim klasama korijena.

Za svih 17 988 lema germanizama prikupljenih iz svih dostupnih izvora (rječnika i ostale opisane građe) generirani su svi oblici (različiti padežni oblici za imenice i pridjeve te jednina i množina glagola u prezentu, perfektu, imperfektu za nesvršene, aoristu za svršene te imperfektu i aoristu za dvovidne glagole, kao i krnji infinitiv te glagolski prilog sadašnji i prošli).

U Tablici 1 u nastavku je za primjer prikazano nekoliko glagola kojima su pridruženi nastavci glagolske paradigme *-ati*, potom je prikazan dio imenica muškog roda na *-nt* i nastavci za njihovu kraću / dugu množinu i, konačno, generirani oblici germanizama muškog roda na *-nt* (kraća / duža množina).

Tablica 1.
Germanizmi na -ati i pridruženi nastavci za paradigme

Germanizmi: glagolske leme na -ati i pridruženi nastavci za paradigme		
Germanizam	**Vrsta glagola**	**Paradigma**
ablendati	svr. / nesvr.	ti t m š 0 mo te ju o la lo li le j jmo jte jući vši h smo ste še hu
ablenduvati	nesvršeni	ti t m š 0 mo te ju o la lo li le j jmo jte jući vši h smo ste še hu
abluftati	nesvršeni	ti t m š 0 mo te ju o la lo li le j jmo jte jući vši h smo ste še hu
abmarkirati	svr. / nesvr.	ti t m š 0 mo te ju o la lo li le j jmo jte jući vši h smo ste še hu
abšmalcati	svršeni	ti t m š 0 mo te ju o la lo li le j jmo jte jući vši h smo ste še hu
abšminkati	nesvršeni	ti t m š 0 mo te ju o la lo li le j jmo jte jući vši h smo ste še hu
abštehati	nesvršeni	ti t m š 0 mo te ju o la lo li le j jmo jte jući vši h smo ste še hu

Svi generirani oblici upareni su s oblicima u hrWaC *web*-korpusu, a kao rezultat iz korpusa su ekstrahirane rečenice koje osim generiranog oblika sadrže i njegovu lijevu i desnu okolinu te izvor, tj. URL adresu. Naravno, budući da je korpus lematiziran pomoću CST lematizatora (Jongejan i Dalianis, 2009.), svim generiranim oblicima u hrWaC-u automatski su pridružene odgovarajuće leme.

Primjer jedne od datoteka koja sadrži rezultat za glagol *koštati* u jednom od oblika i isječak frekvencije leme *koštati* u korpusu s indeksom datoteka nalazi se u Tablici 2:

http://www.arsenal-croatia.hr/highbury/

Novi stadion koštao je 125 000 funti. .....Zapadna tribina koštala je 45 000 funti, dok je istočna probila budžet, na kraju koštajući 130 000, prije svega zbog skupe fasade.

Tablica 2.
Lema koštati, njena frekvencija u korpusu i indeks datoteka

Lema	Frekvencija	Indeks (naziv datoteke + pozicija riječi u tekstnoj datoteci)
Koštati	100740	00000111.txt:1581|00000263.txt:225|00000294.txt:470|00000294. txt:697|00000294.txt:914|00000294.txt:1469|00000294.txt:1653|00000294. txt:3290|00000299.txt:289|00000374.txt:1729|00000374.txt:1790|00000643. txt:50956|00000674.txt:3634|00001274.txt:368|00001296.txt:2792|00001421. txt:2348|00001523.txt:2272|00001691.txt:2859|00001833. txt:40635|00001915.txt:652|00002006.txt:833|00002006.txt:1647|00002050. txt:12049|00002095.txt:3447|00002144.txt:823|00002144.txt:1487|00002195. txt:1957|00002305.txt:8395|00002327.txt:108|00002353.txt:884|00002542. txt:2448|00002542.txt:3820|00002590.txt:25174|00002730. txt:34|00002733.txt:3323|00002859.txt:918|00002870.txt:11858|00002987. txt:2194|00003119.txt:982|00003144.txt:5187|00003162.txt:70|00003225. txt:690|00003373.txt:400|00003425.txt:2632|00003516.txt:4605|00003713. txt:1493|00003820.txt:4062|00003893.txt:6840|00004118.txt:1803|00004118. txt:2328|00004192.txt:3589|00004203.txt:8168|00004203. txt:51992|00004483.txt:1431|00004646.txt:1285|00004703. txt:2692|00004714.txt:1052|00004735.txt:1339|00004802.txt:1664|00004802. txt:1713|00004802.txt:2178|00004826.txt:7201|00004883.txt:2074|00004990. txt:21579|00004995.txt:2446|00005049.txt:2745|00005136.txt:459|00005136. txt:10043|00005136.txt:10445|00005136.txt:11064|00005161. txt:292|00005182.txt:14713|00005335.txt:3491|00005335.txt:3611|00005335. txt:3739|00005406.txt:8513|00005423.txt:35310|00005571. txt:6947|00005618.txt:1448|00005618.txt:2483|00005704.txt:528|00005874. txt:2051|00005932.txt:1154|00006012.txt:2914|00006059.txt:1768|00006059. txt:2044|00006110.txt:11696|00006413.txt:1578|00006492. txt:6081|00006553.txt:174971|00006555.txt:1585|00006566. txt:4005|00006614.txt:1134|00006690.txt:481|00006754.txt:1244|00006775. txt:279670|00006970.txt:20323|00007001.txt:15761|00007001.txt:15800

Kao konačni rezultat pretraživanja hrWaC *web*-korpusa, pronađeno je 8 400 lema germanizama (od 17 988 polaznih lema germanizama za pretraživanje). Tih 8 400 lema pronađeno je u 786 356 tekstnih datoteka koje sadrže jedan ili više germanizama, tj. lema u jednom ili više različitih oblika. Pretražena je cijela *.hr* domena, tj. korpus od 1.9 milijardi pojavnica.

5. Analiza germanizama ekstrahiranih iz rječnika i hrWac *web*-korpusa

Germanizmi se sukladno hrvatskoj gramatici dekliniranju, konjugiraju i, u slučaju pridjeva, sklanjaju, pri čemu ponekad dolazi do odstupanja u značenju riječi. U takvim slučajevima može doći do preklapanja osnovnog oblika germanizma s nekom hrvatskom riječi, ali i do preklapanja nekog od dekliniranih, konjugiranih ili sklanjanih oblika germanizama s nekom hrvatskom riječi.

Računalno pripremljeni i obrađeni tekstovi iz hrWaC-a dali su cijeli niz rečenica u kojima se nalaze zadani germanizmi. U tim se rečenicama germanizmi pojavljuju u osnovnom obliku, kao i u nekim od drugih flektivnih oblika. Analiza rečenica u kojima su detektirane riječi, koje je program prepoznao kao germanizme, pokazala je da dolazi do odstupanja. Naime, postoje germanizmi čiji se oblik riječi podudara s nekom hrvatskom ili, u pojedinim primjerima, čak s nekoliko hrvatskih riječi. Takvi su se „lažni" germanizmi u nekim primjerima pokazali znatno frekventnijim od „pravih" germanizama.

Analiza germanizama u primjerima, odnosno rečenicama vršila se u dvije faze. U prvoj fazi računalno su se obradili zadani germanizmi i njihova pojavnost u rečenicama pri čemu je u određenim primjerima došlo do preklapanja s hrvatskim riječima. U drugoj fazi ručno su se označile, odnosno isključile one riječi koje su „lažni" germanizmi odnosno hrvatske riječi.

Tijekom prve faze istraživanja germanizama u tekstovima s cijele *.hr* domene utvrđeno je da se pojavljuje ukupno 8 400 lema koje su bile označene kao germanizam. Njihova je učestalost pojavljivanja u tekstovima različita. Kao što smo ranije istaknuli, druga faza analize pokazala je da nisu sve detektirane riječi germanizmi, već da su neki germanizmi „lažni".

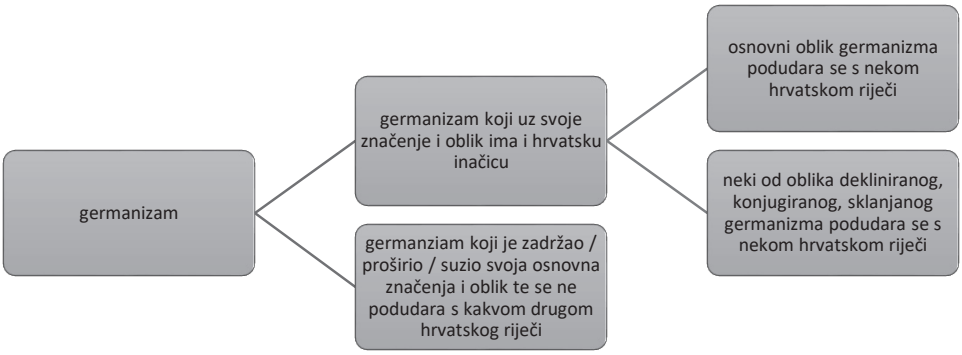
Od 8 400 lema ručnom analizom su 883 leme označene kao višeznačne, tj. te leme uz značenje germanizma posjeduju i značenje riječi iz standardnoga hrvatskog jezika (primjerice *grad* u značenju *stupanj* vs. *grad* u značenju *naseljenog mjesta).* Te 883 leme analizirane su u kontekstu, tj. u konkretnim rečenicama u kojima se pojavljuju u korpusu hrWaC. Pritom su se neke leme u korpusu javile i više od milijun puta (primjerice *grad* s frekvencijom od 1 476 107). Za leme koje su imale frekvenciju veću od 1 000, ručno je analizirano 999 rečeničnih primjera (75 germanizama s frekvencijom većom od 1000), a za one germanizme koji su imali frekvenciju manju od 1 000, analizirani su svi korpusni primjeri (807 germanizama s frekvencijom manjom od 1 000). U Tablici 3 u nastavku nalazi se prikaz isječka iz ukupnog popisa od 999 primjera leme *grad.*

Tablica 3.
Isječak iz ukupnog popisa s dijelovima rečenice ispred, lijevo te iza, desno, od germanizma

Germanizam *grad* s još dva hrvatska značenja i kontekstom							
Lema	**Vr**	**Frekv.**	**Opis značenja**	**Broj značenja**	**Odabir**	**Lijevo**	**Desno**
grad	m	1476107	(1) stupanj; (2) naseljeno mjesto; (3) ledena kiša, tuča	3	2	će uskoro biti pušteni u pogon	. Grad Zagreb osigurao je sr
grad	m	1476107	(1) stupanj; (2) naseljeno mjesto; (3) ledena kiša, tuča	3	2	i hrvatskom jeziku . Rijeka je	grad sa zanimljivom prošlosti
grad	m	1476107	(1) stupanj; (2) naseljeno mjesto; (3) ledena kiša, tuča	3	2	borbe za državu i za slobodu ,	grad u kojemu je utemeljena H
grad	m	1476107	(1) stupanj; (2) naseljeno mjesto; (3) ledena kiša, tuča	3	2	m u lice naš simbol stradanja	- grad Vukovar . Nije li dost
grad	m	1476107	(1) stupanj; (2) naseljeno mjesto; (3) ledena kiša, tuča	3	2	tromjesečni pilot-projekt koji	Grad Zagreb razvija kao moguć
grad	m	1476107	(1) stupanj; (2) naseljeno mjesto; (3) ledena kiša, tuča	3	2	g razgledavanja najužeg centra	grada , odnosno barokne jezgr
grad	m	1476107	(1) stupanj; (2) naseljeno mjesto; (3) ledena kiša, tuča	3	2	kend na šest lokacija u centru	grada : na Glavnom kolodvoru

Grafička podjela germanizama koje smo dobili računalnom detekcijom u korpusu i potom analizirali prikazana je na Slici 1.

Slika 1.
Podjela germanizama



Analiza germanizama u kontekstu tako je dala podjelu germanizama u dvije osnovne skupine pri čemu je kriterij za podjelu bila podudarnost germanizma s riječima hrvatskog jezika u nekom ili u svim flektivnim oblicima, odnosno izostanak takve podudarnosti.

Prvu skupinu čine oni germanizmi koji imaju značenje, odnosno značenja koja se vežu za izvornu njemačku riječ, koja su preuzeta zajedno s tom riječju ili su uporabom u praksi prilagođena i modificirana, no i dalje čine germanizam.

Drugu skupinu čine oni germanizmi kod kojih dolazi do preklapanja s nekom hrvatskom riječi. To znači da s jedne strane imamo germanizam, a s druge strane neku drugu riječ iz hrvatskih rječnika čiji se oblici podudaraju, no značenja su različita. Ova druga skupina podijeljena je na dvije podskupine s obzirom na to pripadaju li germanizam i hrvatska riječ istoj vrsti riječi (primjerice *grad* u značenju *stupanj* vs. *grad* u značenju *naseljenog mjesta)* ili različitim vrstama riječi (primjerice *lak* u značenju *otopina* vs. *lak* u značenju *lagan,* tj. pridjev vs. imenica).

U nastavku ćemo prikazati svega nekoliko primjera najfrekventnijih germanizama koji pripadaju ovoj drugoj skupini, one koji su u korpusu hrWaC detektirani u više od 999 primjera. Također ćemo prikazati nekoliko germanizama koji se javljaju u vrlo malom broju primjera, u rečenici, dvije ili tri, što je neznatno u odnosu na veliki broj računalno pretraženih datoteka.

5.1. Isječak iz primjera najfrekventnijih germanizama:

Primjer germanizma „faks“ (frekvencija u hrWaC korpusu: 45819)

Dvije natuknice nalaze se na Hrvatskom jezičnom portalu pod tumačenjem riječi „faks“. Natuknica koja se veže za englesku riječ *facsimile* donosi sljedeće: 1. *admin. žarg.* telefaks i 2. *meton.* tekst, dokument koji se šalje telefaksom [*u tom faksu razrađeni su uvjeti prodaje*]; faksimil.

Germanizam *faks* veže se za natuknicu *fakultet*, od njemačke riječi *Fakultät*: 1. visokoškolska obrazovna ustanova određenog profila [*pravni fakultet*; *šumarski fakultet*], 2. *meton.* a. zgrada visokoškolske ustanove; faks b. visokoškolsko obrazovanje [*završiti fakultet*].

U analiziranom materijalu riječ se pojavljuje u značenju stroja za prijenos podataka u 127 od 999 analiziranih primjera:

(…) pismeni zahtjev korisnika putem faksa, e-maila te na Internet (…)
(…) primjerice, čak ako vam ne radi faks uređaj, možete zamoliti (…)

Germanizam *faks* javlja se u čak 850 od 999 analiziranih rečenica

Vidi mene, šta govorim, a nisam uspio ni faks završiti.
Drugi su oni koji su faks upisali prije 2007.
Isto kao što na faks idete na praktikum, iako (…)

Primjer germanizma „šalica“ (frekvencija u hrWaC korpusu: 19529)

Na Hrvatskom jezičnom portalu zadanom uvjetu šalica odgovaraju dvije natuknice.

U prvoj natuknici riječ je o *dem.* od šala; mala šala

Druga natuknica donosi sljedeća tumačenja: 1. manja posuda s ručkom sa strane iz koje se pije crna ili bijela kava, čaj itd. [*porculanska šalica*; *keramička šalica*; *kavena šalica*; *šalica za crnu kavu*] te 2. *meton.* količina tekućine ili sipke tvari koja stane u šalicu [*šalica brašna*]. Natuknica se povezuje s natuknicom šäla, riječi preuzete od njemačke *Schalle*. Na Portalu je natuknica šäla opisana kao regionalizam: 1. školjka (zahodska) te 2. zdjela, ob. kostšala.

Od ukupno 999 rečenica u kojima je detektiran oblik riječi šalica, predmetna je riječ germanizam u čak 980 primjera:

(…) kineskoj trgovini opazio prekrasnu šalicu za čaj i (…)
(…) da se ostaci kave ohlade u šalici.
U ruci i sad držim šalicu vrućeg čaja.

5.2. Isječak iz primjera najrjeđih germanizama:

Primjer germanizma „hausfrau“ (frekvencija u hrWaC korpusu: 3)

Tumačenje riječi *hausfrau* pronađeno na Hrvatskom jezičnom portalu (HJP) odnosi se isključivo na značenja preuzeta iz njemačkog jezika i označena je kao regionalizam: 1. *ona koja iznajmljuje stan ili sobu u vlastitoj kući; gazdarica, kućevlasnica*; 2. *ona koja vodi brigu o kući, čistoći hodnika i stubišta; bedinerica, hauserica, pospremačica*; njem. *Hausfrau*.

Primjer: *(…) brižna i štedljiva Hausfrau, a ne uobičajena veličanstvena (…)*

Primjer germanizma „špancerati“ (frekvencija u hrWaC korpusu: 2)

Pretraga pojmom „špancerati“ nije dala rezultate, već je pretraga prilagođena. Riječ špancírati se, opisana na Hrvatskom jezičnom portalu kao zastarjeli regionalizam, ima sljedeće značenje: 1. kretati se, pješačiti umjerenom brzinom radi odmora i razonode; šetati; 2. pren. besposličariti, bespotrebno traćiti vrijeme [on se špancira okolo] te 3. pejor. javno se pokazivati s kim kad se očekuje skrivanje ili diskrecija (npr. ljubavni par) od njemačke riječi *spazieren*.

Primjer: *(…) i ubojice mednami slobodno španceraju, i ni jem niti las zglave , radi (…)*
Koli glave s španceraju, kak kolo od vil, lepi beli (…)

5.3. Popisani germanizmi i rezultati dobiveni pretraživanjem hrWaC-a

Iako ukupan popis sastavljen iz analiziranih rječničkih izvora i znanstvenih radova čini 17 988 germanizama, 8 400 germanizama pronađeno je pretraživanjem hrWaC-a.

Najveći broj germanizama pronađenih u hrWaC-u, njih 2 678 dao je regionalni rječnik „Agramer. Rječnik njemačkih posuđenica u zagrebačkom govoru“ autorice Glovacki-Bernardi (2013.). Također, iz Petrovićeva „Esekerskog rječnika“ (2008.), 2 104 germanizama pronađeno je u hrWaC-u. Od ukupno 8 400 germanizama, izvor za 1 638 germanizma bio je „Veliki rječnik hrvatskoga jezika“ autora Vladimira Anića (2003.).

Ukupno 2 972 germanizma su iz „Novog rječnika stranih riječi“ autora Bratoljuba Klaića (2012.), a 909 iz „Rječnika stranih riječi“ autora Anića, Klaića i Domovića (2001.). Od rječničkih izvora najmanji broj germanizama potječe iz „Rječnika hrvatskoga jezika» urednika Šonje i sur. (2000.), svega 667. Rječnik Šonje i suradnika tako je ukupno dao najmanji broj germanizama pronađenih u hrWaC-u.

Broj germanizama popisanih iz doktorskih disertacija pronađenih u hrWaC-u je kako slijedi u nastavku: od 8 400 germanizama, 1 398 je germanizama koje je Velimir Piškorec popisao i opisao u disertaciji koja se bavi istraživanjem germanizama u podravskom dijalektu (2001.), 1 297 je onih germanizama koji su popisani iz disertacije Tea Bindera te 839 iz doktorske disertacije Ive Medića. Broj svih različitih germanizama kojima su izvori bile tri disertacije (Binder, Piškorec i Medić) i koji su pronađeni u korpusu hrWaC je 95. Ako taj popis usporedimo s popisom 160 različitih germanizama prikupljenih iz te 3 disertacije, možemo zaključiti da je 60% germanizama iz zajedničkog rječničkog blaga popisanog u disertacijama pronađeno u korpusu.

Slika 2.
Isječak iz popisa germanizama pronađenih u HrWaC-u

Lema	Frekv u hrWaC-u
neprešan	6
nesprešan	6
pariran	6
foringa	6
mašingever a	6
florentiner	6
renumeira n	6
zokna	6
muzirati	6
pošmirglan	6
zglihati	6
promenadn	6

Lema	Frekv u hrWaC-u
kistihant	5
hofrat	5
koncentraci oni	5
istupljen	5
rosselsprun g	5
kipan	5
regrutni	5
njurgan	5
grenadirma rš	5
narba	5
nažveglati	5

Lema	Frekv u hrWaC-u
šrajbmašin a	5
liniran	5
cvikcange	5
špancirat	5
šiltkapa	5
cvikan	5
zafarban	5
licitirat	5
kuglov	5
štucer	5
štrihan	5
šmekerica	5
[illegible]	5

Lema	Frekv u hrWaC-u
nereprezent iran	4
nekondenzi ran	4
šrajtoflin	4
neofucan	4
štekan	4
štamtiš	4
nerehabiliti ran	4
štrapaciran	4
nestratificir an	4
ciferblat	4

6. Zaključak

Jezične tehnologije omogućile su nam da utvrdimo koliko često se germanizmi iz pisanih izvora pojavljuju u hrvatskom medijskom prostoru.

Za potrebe istraživanja utvrđeno je koji se germanizmi pojavljuju u suvremenom hrvatskom jeziku, a kao izvor za sastavljanje takvog popisa poslužili su suvremeni standardni hrvatski rječnici (rječnici hrvatskoga jezika i rječnici stranih riječi), rječnici (germanizama) regionalnog tipa i doktorske disertacije. Analiza tih izvora dala je popis od ukupno 17 988 različitih germanizama. Računalna usporedna analiza izvora pokazala je da se germanizmi ne navode u svim izvorima, već se pojedini germanizmi javljaju samo u nekim izvorima gdje se, primjerice ističe Petrovićev „Esekerski rječnik" koji bilježi njemačke riječi govornika esekerskog dijalekta. Tijekom analize u nekim smo situacijama naišli i na različite načine bilježenja germanizama pa su (obje) inačice bilježenja uvrštene u naš popis.

Popis germanizama jezičnim se tehnologijama analizirao u računalnom korpusu hrvatskoga jezika hrWaC tijekom 4 godine.

Pretraživanjem hrWaC-a pronađeno je 8 400 germanizama od ukupno 17 988 popisanih iz različitih izvora. Tih 8 400 lema pronađeno je u 786 356 tekstnih datoteka u kanonskom ili nekom drugom obliku. Nakon računalno provedene ručne analize pokazalo se da postoje germanizmi čiji se oblik podudara s nekom hrvatskom riječi. U nekim su se slučajevima takvi „lažni" germanizmi pokazali bitno frekventniji od „pravih" germanizama".

Istraživanjem je utvrđeno da se u korpusu najčešće pojavljuju germanizmi koji potječu isključivo iz rječnika stranih riječi. Analiza frekvencije germanizama iz presjeka „Novog rječnika stranih riječi" autora Bratoljuba Klaića (2012.) i „Rječnika stranih riječi" autora Anića, Klaića i Domovića (2001.) pokazala je da njih 965 u hrWaC-u ima frekvenciju veću od 1 000.

Analizom je utvrđeno da se od 17 988 germanizama popisanih iz svih izvora u korpusu hrWaC pojavljuje njih 47%. No, od tog postotka samo se 1 285 germanizama u hrWaC-u pojavljuje s frekvencijom većom od 1 000. Također, u razdoblju od godine dana u dnevnim novinama (njih 4) pronađena je svega 191 lema – germanizam (18 284 pojavnica). Zaključujemo da se germanizmi u hrvatskom jeziku pojavljuju sustavno, ali rjeđe no što bismo očekivali s obzirom na stoljetni intenzivni utjecaj njemačkog jezika u našim krajevima.

Germanizmi bez svojeg ekvivalenta u hrvatskom jeziku pojavljuju se češće od germanizama koji imaju hrvatski ekvivalent. Potvrđenih 8 400 germanizama ručno je provjeravano na Hrvatskom jezičnom portalu, a primjeri koji se nisu mogli utvrditi

na portalu HJP provjereni su u njihovim polazišnim izvorima i sukladno tome određen je njihov status, odnosno imaju li hrvatski ekvivalent ili ne. Pokazalo se da 52% germanizama analiziranih ovom metodom ima hrvatski ekvivalent.

Istraživanje provedeno nad opsežnim računalno i ručno analiziranim tekstovima pokazalo je da se u suvremenim tekstovima upotrebljava dio popisanih germanizama. Ujedno je analiza rečeničnih primjera dala naslutiti da se germanizmi ne pojavljuju u svim svojim oblicima, već su neki od, primjerice, za imenicu, dekliniranih oblika, češći (nominativ, akuzativ), dok se neki uopće ne pojavljuju ili se pak pojavljuju samo u osnovnom obliku. Tako primjerice imamo germanizam „bofl", imenicu muškog roda, koja se na Hrvatskom jezičnom portalu tumači kao 1. v*rlo loša roba, roba s greškom [bofl roba]; otpaci, škart* te 2. pren. duhovni, umjetnički itd. proizvod takve vrijednosti. U primjerima iz analiziranog korpusa potvrđen je germanizam i to samo u kanonskom obliku „bofl": *Bofl je doista jeftin. (…) Dovode bofl igrače koji ne mogu igrati ni u drugoj ligi*. Bilo bi zanimljivo istražiti u kojem se obliku germanizmi, neovisno je li riječ o imenicama, pridjevima, glagolima ili prilozima, upotrebljavaju u hrvatskom jeziku, odnosno koji flektivni oblici nisu svojstveni germanizmima.

Važan doprinos provedene računalne obrade germanizama oformljena je baza podataka koja, sama po sebi, funkcionira kao rječnik germanizama te predstavlja stručni i znanstveni doprinos. Razradilo se automatsko pronalaženje germanizama u svim njihovim morfološkim oblicima uz pomoć postojećih rječničkih baza hrvatskoga standardnoga jezika te automatskim generiranjem svih mogućih oblika pojedinih riječi hrvatskoga jezika, čime se proširila baza germanizama hrvatskih rječnika i značajnih radova koji se bave jezičnom analizom germanizama.

Rezultat istraživanja germanizama u digitalnom korpusu mehanizam je analize germanizama u hrvatskom jeziku koji se ujedno može primijeniti i na neke druge jezike.

7. Smjernice za buduća istraživanja

S obzirom na situaciju s germanizmima i njihovim ekvivalentima u hrvatskom jeziku, a uzevši u obzir naš korpus, ovo se istraživanje može dodatno razraditi. Istraživanje germanizama provedeno je nad opsežnim računalno i ručno analiziranim tekstovima te je dokazano da se u suvremenim tekstovima upotrebljava dio popisanih germanizama. Poluautomatska analiza rečeničnih primjera pokazala je da se germanizmi ne pojavljuju u svim svojim oblicima, već su neki, primjerice za imenicu, deklinirani oblici češći (nominativ, akuzativ), dok se neki uopće ne pojavljuju ili se pak pojavljuju samo u osnovnom obliku.

Za potrebe ovog istraživanja germanizme smo označili samo u odnosu na činjenicu imaju li ekvivalent u hrvatskom jeziku ili ne. No, formirani popis pokazao je da posto-

je oni germanizmi koji imaju, uvjetno rečeno, „jednostavan“ ekvivalent u hrvatskom jeziku, kao što su to primjerice *frištik*, *reg. prvi dnevni obrok; zajutrak, doručak, fruštik* ili pak *cušpajz, reg. kulin. ukuhano povrće, ob. kao prilog glavnom jelu; varivo*.

Postoje i oni germanizmi koji imaju nekoliko tumačenja u hrvatskom jeziku, poput germanizma *druker, kibic, navijač (na sportskim utakmicama)* ili tumačenja *kopča od dva dijela (jedan dio ulazi u drugi); driker, šuština, učvrsnica*. Glagol *heklati* može značiti *izrađivati čipke ili slične radove posebnom iglom (kukicom); kačkati, kukičati*, no i a*) praviti stalno iste nesvjesne pokrete rukom (zbog tika, bolesti i sl.) b) motati usnama i jezikom u govoru (zbog pijanstva, umora i sl.) c.) u pijanom stanju bauljati cestom; posrtati, teturati*.

Za primjere germanizama koji uz „jednostavan“ ekvivalent (za *heklanje* bi to bio hrvatski ekvivalent *kukičanje*) imaju dodatna tumačenja, moglo bi se istražiti u kojem se tumačenju u suvremenim tekstovima pojedini germanizmi učestalije koriste.

Zanimljivo bi bilo dodatno istražiti frekvenciju pojavljivanja germanizama u korpusima hrvatske narječne građe te korpusima različitih razvojnih faza hrvatskoga jezika, posebice od druge polovice XIX. stoljeća naovamo što bi činilo dobru podlogu jezičnim i jezičnopovijesnim istraživanjima hrvatskoga jezika.

Literatura

1. Agić, Ž. and Ljubešić, N. (2014). *The SETimes.HR linguistically annotated corpus of Croatian*. U Proceedings of LREC 2014.
2. Agić, Ž.; Tadić, M. and Dovedan, Z. (2008). Combining Part-of-Speech Tagger and Inflectional Lexicon for Croatian, in: Erjavec, T. and Žganec Gros, J. (Eds.). *Proceedings of the 6th Language Technologies Conference*.
3. Binder, T. (1954). *Die deutschen Lehnwörter in der kroatischen Essegger Mundart*. Dissertation zur Erlangung des Doktorgrades an der philosophischen Fakultät der Universität Wien.
4. Björkelund, A.; Bohnet, B.; Hafdell, L.; Nugues, P. (2010). *A high-performance syntactic and semantic dependency parser*. In Coling 2010: Demonstration Volume, 33-36. Beijing, August 23. - 27. 2010.
5. Filipan-Žignić, B. (2007). Germanismen in der kroatischen Sprache (Germanisms in Croatian language), u: Stein, B (Ur.). *Wege zu anderen Sprachen und Kulturen*. Verlag Dr. Kovač. Hamburg, 23-43.
6. Halacsy, P.; Kornai, A. and Oravecz, C. (2007). *HunPos: an open source trigram tagger.* U Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL ’07, 209-212. Stroudsburg, PA, USA: Association for Computational Linguistics.

7. Jongejan, B. and Dalianis, H. (2009). *Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike*. U Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 145-153.
8. Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics,* 29 (3): 333-348.
9. Ljubešić, N. and Erjavec, T. (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene, in: Habernal, I. and Matousek, V. (Eds.). *Text, Speech and Dialogue, Lecture Notes in Computer Science*. Berlin / Heidelberg: Springer, 95-402.
10. Ljubešic, N. and Klubicka, F. (2014). *{bs, hr, sr} WaC – web corpora od Bosnian, Croatian and Serbian*. In Proceedings of the 9th Web as Corpus Workshop (Wac-9), 29-35.
11. Medić, I. (1962): *Kulturno-historijsko značenje i lingvistička analiza njemačkih pozajmljenica kod zagrebačkih obrtnika (obrada metala, drva i kože)*. Unveröffentl. Diss. Zagreb.
12. Piškorec, V. (2001). *Germanizmi u podravskom dijalektu*. Disertacija. Filozofski fakultet Sveučilišta u Zagrebu.
13. Suchomel, V. and Pomikálek, J. (2012). Efficient Web crawling for large text corpora. In *Proc Seventh Web as Corpus Workshop (WAC7)*, 39-43. Lyon, France.
14. Šnajder, J.; Padó, S. and Agić, Ž. (2013). *Building and Evaluating a Distributional Memory for Croatian.* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Sofia: Association for Computational Linguistics, 784-789.
15. Tepeš Golubić, L. (2016). *Germanizmi u digitalnim novinskim korpusima hrvatskoga jezika*. Doktorski rad. Zagreb.
16. Tepeš Golubić, L.; Mikelić Preradović, N. and Boras, D. (2013). Semi-automatic detection of germanisms in Croatian newspaper texts, in: Vetulani, Z. and Uszkoreit, H. (Eds.).*Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznan, Poland: Fundacja Uniwersytetu im. A. Mickiewicza, 173-177.

Rječnici

1. Anić, V. (2003). *Veliki rječnik hrvatskog jezika*. Zagreb: Novi liber.
2. Anić, V. i Goldstein, I. (1999). *Rječnik stranih riječi*. Zagreb: Novi liber.
3. Anić Š.; Klaić N. i Domović Ž. (2002). *Rječnik stranih riječi*. Zagreb: SANI-PLUS.
4. Glovacki-Bernardi, Z. (2013). *Agramer. Rječnik posuđenica u zagrebačkom govoru*. Zagreb: Novi liber.
5. Klaić, B. (2012). *Novi rječnik stranih riječi*. Zagreb: Školska knjiga.

6. Petrović, V. (2008). *Esekerski rječnik*. Zagreb: Filozofski fakultet.
7. Šonje, J. (glavni urednik) (2000). *Rječnik hrvatskoga jezika*. Zagreb: Leksikografski zavod Miroslav Krleža i Školska knjiga.
8. *Veliki rječnik hrvatskoga standardnog jezika*. Za izdavača: Ante Žužul (2015). Zagreb: Školska knjiga.

Izvorni znanstveni rad

German Loanwords in Digital Environment

Lidija Tepeš Golubić
Zagreb University of Applied Sciences, Croatia
e-mail: ltepes2@tvz.hr

Abstract

The framework of this research make German loan words compiled from the dictionaries of the Croatian language, dictionaries of foreign words, and PhD thesis studying German loanwords in the Croatian language. The list of German loanwords found in the contemporary Croatian texts was analised using linguistic technologies supported by subsequent manual data processing. The compiled list of gathered loan words allowed for their computational analysis in the hrWaC web corpus comprising the texts from the whole .hr domain collected in a period of four years.

This research has established the presence of the German loan words in contemporary Croatian language texts. Although the total number of German loan words that appeared in contemporary dictionaries of the Croatian language and other sources consulted during the research was 17 988 lemmas, we were able to confirm the usage for only 8 400 lemmas in contemporary texts.

Applying the automatic detection method, we have found all German loan words used in contemporary texts in all of their forms, simultaneously creating the German loan words frequency dictionary. The confirmed 8 400 lemmas serve as proof that the lexical treasure of the Croatian language recorded in dictionaries has not been lost in the contemporary Croatian language. On the contrary, it has systematically entered into the web corpus of the Croatian language and is found in the texts that do not necessarily belong to the standard language.

Key words: digital environment, Croatian web corpus hrWaC, computational analysis, German loanword, lemma.