Izvorni znanstveni rad

Goranka Blagus Bartolec
Institut za hrvatski jezik i jezikoslovlje, Zagreb
gblagus@ihjj.hr

Ivana Matas Ivanković
Institut za hrvatski jezik i jezikoslovlje, Zagreb
imatas@ihjj.hr

Korpus umom korisnika (na što treba pripaziti u korpusno utemeljenom istraživanju)

U radu će se iz korisničke perspektive uputiti na odstupanja i jezične pogreške na koje se može naići tijekom pretrage korpusa i analize dobivenih podataka. Istraživanje je rađeno na temelju primjene korpusa hrWaC (izrađen u alatu NoSketch Engine i Sketch Engine) za potrebe rada na projektima u kojima trenutačno sudjelujemo: e-GLAVA (*Baza hrvatskih glagolskih valencija*), *Leksikon hrvatskih prijedloga, Kolokacijska baza hrvatskoga jezika, Rječnik velikoga i maloga početnog slova,* MREŽNIK, COST akcija PARSEME). Uz navođenje tipskih primjera odstupanja jezične će pogreške biti grupirane prema tipu te će se odrediti njihova jezična razina. Namjera je pridonijeti poboljšanju i kvaliteti korpusa, čime će se omogućiti pouzdano istraživanje određenoga jezičnog fenomena.

Ključne riječi: hrWaC, hrvatski jezik, korisnička perspektiva, odstupanja i pogreške u korpusu

1. Uvod

U suvremenim lingvističkim istraživanjima korpus je postao nezaobilazan izvor prikupljanja podataka na svim jezičnim razinama. Izrada korpusnih alata za hrvatski jezik koji ubrzavaju i olakšavaju pretragu slijedi postojeće dosege europske korpusne i računalne lingvistike. Hrvatski jezik pretraživ je na trima računalnim korpusima: na *Hrvatskoj jezičnoj riznici* (http://riznica.ihjj.hr/index.hr.html) Instituta za hrvatski jezik i jezikoslovlje, na *Hrvatskome nacionalnom korpusu* (http://filip.ffzg.hr/cgi-bin/run.cgi/first_form) te na *Hrvatskome mrežnom korpusu* – hrWaC-u (http://nlp.ffzg.hr/resources/corpora/hrwac/). Sva tri korpusa imaju obilježja općega i referentnoga korpusa koji se koristi za utvrđivanje osnovnih karakteristika jezika (Nesselhauf 2005: 2–3).[1] hrWaC i *Hrvatski nacionalni*

1 Osim općih referentnih korpusa Nesselhauf (isto) s obzirom na vrste tekstova razlikuje i povijesne korpuse, regionalne korpuse, učeničke korpuse, višejezične ili paralelne korpuse te govorne korpuse.

korpus (HNK) kao javno dostupni korpusi lematizirani su i tagirani[2], što omogućuje identifikaciju vrste riječi i morfoloških oblika, pa i složenija pretraživanja (primjerice sintaktičkih obrazaca). O izradi korpusa hrvatskoga jezika napisano je više radova (Tadić 2003 i 2009, Ćavar i Brozović 2012, Ljubešić i Klubička 2014), dok su radovi o upotrebi korpusa s korisničkoga gledišta rjeđi (Blagus Bartolec i Matas Ivanković 2017 i 2018). U ovom se radu iznose iskustva rada na korpusu iz korisničke perspektive kad se rezultati dobiveni pretragom korpusa primjenjuju u jezikoslovnim istraživanjima. Primjeri koji se pritom donose dobiveni su kao rezultat različitih pretraga korpusa hrWaC za potrebe projekata na kojima surađujemo u Institutu za hrvatski jezik i jezikoslovlje. Iako su u uvodnom dijelu navedena sva tri hrvatska korpusa, u ovom radu u analizu je uključen samo hrWaC zbog lake dostupnosti te najvećega broja pojavnica. I hrWaC i HNK izrađeni su u platformi Sketch Engine, tako da u pozadinskim alatima između tih dvaju korpusa ne postoji razlika. S obzirom na vrstu tekstova *Hrvatska jezična riznica* i HNK sadržavaju „pročišćenije" tekstove, dakle tekstove iz područja književnosti i publicistike te su i jezično pravilniji, dok hrWaC obuhvaća velik broj tekstova (blogovi, forumi) u kojima postoji visok stupanj odstupanja od jezičnih pravila te može poslužiti i kao pokazatelj upotrebe jezika koja je bliža govorenomu jeziku i nije toliko pod utjecajem norme. Pristup hrWaC-u dostupan je preko javne, NoSketch Engine, inačice te preko Sketch Engine inačice, što omogućuje naprednije pretrage prema skicama riječi.

2. O korpusu općenito

Jezična su istraživanja u 21. stoljeću postala nezamisliva bez pomoći računala. Primarna građa kojom se pritom služimo jest računalni korpus – skup strojno čitljivih tekstova nekoga jezika sastavljen po određenom kriteriju (Klobučar Srbić 2008). Korpus pomaže u pretraživanju, prikupljanju i obradi podataka na kojima se temelji zaključak o pojavi ili fenomenu koji se istražuje ili se potvrđuje hipoteza utemeljena na drugim izvorima te se analizom dobivenih podataka unapređuju gramatička, leksikografska i druga jezikoslovna istraživanja. Korpusna lingvistika koja „u najširem smislu obuhvaća istraživanje jezika zasnovanog na korpusu" (Utvić 2013: 1) danas je veoma razvijano područje i postala je predmetom zanimanja jezikoslovaca i računalnih stručnjaka, a korisnicima su dostupni različiti korpusi. Kako tumači Tadić (2003: 156), korpus je „pisani ili govorni jezični resurs, koji je prikupljen i obilježen u cilju: analize jezika kojom se utvrđuju njegova svojstva, analize ljudskog ponašanja (u domeni jezične upotrebe) u određenim situacijama, obuke sustava kako bi se njegovo ponašanje prilagodilo specifičnim jezičnim okolnostima, empirijske provjere neke jezične teorije, izrade testa za neku jezično

2 *Riznica* je također lematizirana i tagirana, no dostupna je samo preko Sketch Enginea (Riznica v0.1).

inženjersku tehniku ili pak primjene kojom se utvrđuje njeno funkcioniranje u praksi." Kao takvi, računalni su korpusi „kodirani na standardan i dosljedan način s nakanom da budu otvoreni za računalno pretraživanje" (Tadić 1998: 337).

3. Nedostatci korpusa

Osim pozitivnih karakteristika računalnih korpusa opisanih u uvodnom dijelu, koje olakšavaju i ubrzavaju znanstvenoistraživački rad, korisnici uočavaju i neke nedostatke tijekom oblikovanja same pretrage ili tijekom analize dobivenih rezultata pretrage. Dobiveni rezultati sadržavaju i određene pogreške na svim jezičnim razinama te konačna statistička interpretacija na kojoj korisnik temelji svoje istraživanje zbog toga katkad može biti upitna. Izvor odstupanja i pogrešaka koji nastaju pretragom korpusa mogu biti: 1. korisnik korpusa, 2. tekstovi koje određeni korpus obuhvaća, 3. sam korpus, odnosno način na koji je tagiran, što katkad dovodi do pogrešaka u pronalaženju morfoloških oblika ili, za hrvatski jezik, pojave neuobičajenih zapisa (primjerice, u pretrazi riječi *sova* korpus prikazuje primjer *McDonald'sova omiljena meta su djeca*). Hunston (2002), istražujući neke engleske korpuse (*Bank of English*, *Timeov* korpus), navodi četiri nedostatka korpusnih izvora, a koji se mogu primijeniti i u analizi hrvatskih korpusa: 1. prije primjene korpusa korisnik mora biti svjestan i eventualnih nedostataka kako bi se korpusima mogao koristiti na pravilan način; na prvome mjestu tvrdi kako korpus korisnicima ne daje informacije o tome što je u jeziku moguće, a što nije, već što je frekventno, a što nije, 2. korpus prikazuje isključivo svoj sadržaj, a ne jezik u cijelosti zbog čega se zaključci o jeziku ne trebaju tretirati kao čvrste činjenice, 3. korpus ne može pružiti informacije, već samo dokaze, što znači da korpus može pružiti velik broj primjera korištenja jezika, a interpretirati ih se može intuicijom i dodatnim jezičnim znanjem i 4. korpus prezentira jezik izvan svojega konteksta, što znači da ne može u potpunosti prenijeti neverbalni dio jezične komunikacije kao što je govor tijela te vokalizacija, naglasci, ton glasa i ostale paralingvističke informacije. Iznimka su govorni korpusi u kojima je korisnicima omogućeno preslušavanje govora određenog jezika. S tih se aspekata u nastavku pristupa analizi odstupanja i jezičnih pogrešaka u analiziranome hrvatskom korpusu.

4. Što je pogreška/odstupanje

U postojećim jezikoslovnim izvorima pogreška i odstupanje različito se definiraju (od općih do usko terminoloških opisa), ovisno o vrsti izvora i korisnicima kojima se te definicije namijenjene. U *Školskom rječniku* (2012) *pogreška* (*pogrješka*) se definira s ljudskoga aspekta kao 'neispravan postupak, ono što je napravljeno loše i netočno', a *odstupanje* je 'postupanje suprotno od onoga što je zadano ili očekivano'. U bazi *Struna pogreška* se definira kao 'odstupanje pojedinoga mjerenja fizičke veličine od srednje vrijednosti izvedene na temelju statističke obradbe svih

mjerenja', a *odstupanje* je 'razlika između opažene i referentne vrijednosti / razlika između izmjerene i standardne vrijednosti'. Pogreške i odstupanja najčešće su nenamjerni, no mogu biti i namjerni, kao rezultat jezične igre. Budući da je riječ o korpusnoj analizi jezika, u ovom radu pogreškom ili odstupanjem smatramo svaki otklon od standarda, tj. kodificirane norme. Pri opisu pogrešaka i odstupanja koja proistječu iz rada na korpusu moguće je primijeniti podjelu pogrešaka prema Jelaska i Bjedov (2015) te Pala, Rychlý i Smrž (2003), a prema izvoru pogrešaka mogu se izdvojiti tri skupine:

- pogreške korisnika
- pogreške u tekstu
- pogreške u korpusu.

4.1. Pogreške korisnika

U jednostavnoj pretrazi, u kojoj se unosi samo određena riječ (lema), oblik riječi ili određena sveza riječi, korisnik zbog nepažnje može pogriješiti i krivo upisati riječ pa ne dobiti potvrdu. U složenijim pretragama s pomoću korpusnoga upitnog jezika (engl. *Corpus Query Language / CQL*) korisnik može nepotpuno ili netočno postaviti pretragu: primjerice, neprecizno postaviti regularni izraz ili netočno upotrijebiti oznake koje se primjenjuju u korpusu. Korpusni alati, dakle, omogućuju sofisticiranu pretragu koja u konačnici može ponuditi kvalitetne podatke, međutim, pri takvoj složenoj pretrazi od korisnika se očekuje specifično računalno znanje kako bi došao do podataka koje traži (primjena regularnih izraza ili filtara), što određuje konačan rezultat pretrage. Rezultat takve pretrage ovisi, dakle, o korisnikovu poznavanju mogućnosti složenije pretrage, odnosno od korisnika se očekuje primjena računalnih tehnologija, a to je znanje koje često jezikoslovci i drugi istraživači primarno ne posjeduju te im je potrebna dodatna edukacija.

4.2. Pogreške u tekstu i stilska ograničenost korpusnih tekstova

Pogreške u tekstu proizlaze iz tekstova koji su preuzeti i obrađeni u korpusu. Pritom treba istaknuti da uzrok tekstnih pogrešaka nije korpusna platforma preko koje korisnik pretražuje tekstove, nego su izvor pogreške ili odstupanja sami tekstovi. Tekstovi u hrWaC-u najvećim dijelom pripadaju trima stilovima hrvatskoga standardnog jezika: publicističkomu stilu (tekstovi s različitih novinskih portala), razgovornomu stilu (tekstovi različitih foruma i blogova, što uključuje mnogo razgovornih oblika i nekonvencionalnih načina zapisa koji su problematični u tagiranju, a samim time i u pretrazi korpusa) te administrativnomu stilu (tekstovi zakona, propisa, službenih stranica različitih tijela i udruga). Također treba uzeti u obzir i ograničenost izvora, tj. često se navodi isti izvor ili isto sadržajno područje, pa se primjeri ponavljaju ili su slični.

4.2.1. *Tekstne pogreške prema stupnju ovladavanja kodificiranom jezičnom normom*

„Budući da se jezik otprilike s dvanaest godina automatizira, na kraju obveznoga školovanja, koje učenici u Hrvatskoj završavaju s četrnaest-petnaest godina, očekuje se ovladanost kodificiranom normom hrvatskoga jezika na fonološkoj, morfološkoj i sintaktičkoj razini." (Jelaska i Bjedov 2015: 228). Iako su izvori tekstova u korpusima (ponajprije u hrWaC-u) tematski i stilski različiti, što utječe i na stupanj otklona od kodificirane norme, u skladu s navedenim tumačenjem, može se očekivati da je većina autora tekstova tijekom obrazovnoga procesa ovladala standardnojezičnim pravilima hrvatskoga jezika te da će ih primjenjivati u tekstovima, no pokazalo se da postoje odstupanja koja mogu biti potvrda neusvojenosti tih pravila ili njihova neprimjenjivanja kao rezultat prilagodbe razgovornomu stilu (na blogovima i forumima) u kojima su ta odstupanja izraženija. Jelaska i Bjedov s obzirom na stupanj ovladavanja kodificiranom normom upućuju na proširenost netočne jezične proizvodnje te razlikuju: *propuste* kao „slučajne jezične pogrješke koje imaju uglavnom nejezične čimbenike, poput umora, rastresenosti, napetosti i slično" (isto: 229), *pogrješke* kao „ono što se tijekom proizvodnje može izbjeći (...), primjerice u pisanju svakodnevnih, najčešćih riječi poput *neću* umj. *neću, bjel, svjet* umj. *bijel, svijet*" (isto: 230), *odstupanja* kao „međujezične osebujnosti, jedinice koje se razlikuju od onih u ciljanomu idiomu"[3] i *dvojnosti* kao „više od jedne prihvatljive jezične jedinice koje se usporedno rabe, a obje ili sve pripadaju normi. Mogu biti dvojnosti (dublete), trojnosti (triplete), koje su govornici slobodni birati." (isto: 231).

Sve četiri kategorije otklona od kodificirane norme uočljive su i pri analizi podataka dobivenih pretragom korpusa. Propusti su česti u tekstovima koji su preuzeti s foruma:[4] *... s otprilike istom prednošću* ***isred*** *trećeplasiranih vršnjaka iz VK "Jarun " / U svakom slučaju, nešto se važno trebalo* ***dogodi t i*** *kada su tri* ***mudraca-svećenika zaratustre-astrologa*** *krenula prema Palestini čak iz Perzije. /...* ***zamračilio*** *se nebo nad Dubrovnikom.* Pogreške su čest oblik otklona u korpusnim tekstovima, primjerice, oblik *nečemo* u hrWaC-u je potvrđen 901 put, a *ljep* 3324 puta. Odstupanja su također svojstvena forumskim tekstovima u kojima autori unose idiomatska jezična obilježja: ***Nekak*** *mi se vidi da ti Papu gledaš kroz* ***cvike*** *KPJ. / Zahvalan sam sretnoj* ***zvizdi,*** *što me tako obilno nagradila.* S obzirom na to da su prihvatljive unutar kodificirane norme, dvojnosti su evidentne u svim vrstama korpusnih tekstova: npr. imenica *porijeklo* potvrđena je 34 183 puta, a *podrijetlo* 36 012 puta.

3 Jelaska i Bjedov (2015) razlikuju prijenosna, razvojna, navođena i izvorna odstupanja. Više o tim odstupanjima v. isto: 230–231.

4 S obzirom na to da su svi primjeri preuzeti s hrWaC-a, ne navode se pojedinačni izvori za svaki primjer.

4.2.2. *Tekstne pogreške prema jezičnim razinama*

Netočna proizvodnja jezičnih oblika zastupljena je na svim tekstnim razinama,[5] a ovdje primjenjujemo podjelu prema Pala, Rychlý i Smrž (2003) koji su opisali rad na češkom korpusu *Czech text corpus* (Chyby) koji sadržava različite pogreške, njegovu izradu, tj. kako su pogreške otkrivane, označavane (markirane) i anotirane. Razlikuju: slovne pogreške (*spelling errors*), tipografske pogreške (*typographical errors*), morfološke i sintaktičke pogreške (*morphological and syntactic errors*), prave sintaktičke pogreške (*clear syntactic errors*), interpunkcijske pogreške (*punctuation errors*), semantičke pogreške (*semantic (lexical) errors*) i stilističke pogreške (*stylistic errors*). Svi tipovi takvih pogrešaka zastupljeni su i u hrvatskom korpusu i ovdje ih donosimo odvojeno iako se neke pogreške ne mogu jasno razlučiti, tj. odrediti samo kao jedan tip pogreške, jer bi se prema obilježjima mogle podvesti pod više tipova pogrešaka (primjerice, semantičke i stilističke pogreške te leksičke i semantičke pogreške).

Slovne pogreške rezultat su pogrešne upotrebe slova koja uvjetuje otklone na leksičkoj i morfološkoj razini zbog čega dolazi do pogrešaka u lematiziranju i tagiranju. Primjerice, pretragom leme *ljubimica* dobiven je primjer *Što su sve u stanju kućni* ***ljubimici*** *napraviti dok vas nema te u kojoj mjeri mogu napraviti štetu.* Oblik *ljubimici* tagiran je kao dativ jednine imenice *ljubimica*, iako je iz konteksta jasno da je riječ o zatipku u množinskom obliku imenice *ljubimac*. Sličan rezultat vidljiv je u primjeru *Vi ste zaboravili da u* ***Zagrebi*** *ima najviše potencijalnih birača.* Iz konteksta je razumljivo da je riječ o slovnoj pogrešci u imenu *Zagreb*, a primjer je dobiven pretragom leme *zagrepsti*.

Tipografske pogreške obuhvaćaju pogreške pri zapisu neslovnih znakova kao što su zarez, crtica, spojnica, bjelina, trotočka, navodnici, polunavodnici itd.: *Za Uskrs smo otputovali na kraći " odmor " u Francusku. / Taj put, kao što je rekao Benedikt XVI., moći će izgledati kao putovanje kroz pustinju....* Tipografske pogreške također su uzrok odstupanja na višoj razini, tj. pri tagiranju i lematiziranju. Pretragom prema lemi *sova* potvrđeni su i sljedeći primjeri: ***McDonald ' sova*** *omiljena meta su djeca. / Nova* ***Levi ' sova*** *kolekcija vodi buntovnike na kalifornijsku ljetnu scenu.* U hrvatskom korpusu ključan tip tipografskih odstupanja jest izostavljanje dijakritika zbog čega korpusni alati pogrešno lematiziraju i tagiraju pojedine oblike. Primjerice, u zadanoj pretrazi za lemu *kokoš* dobiveni su sljedeći primjeri: *Sladoled od tajnih sastojaka (****Kokos****, vanilija i kondenzirano mlijeko) / Hladni* ***kokosi*** *su bili doslovno hladni* ***kokosi.*** U pretrazi leme *otac* zbog zanemarivanja dijakritika

5 Težak (1998) i Rosandić (2002) pogreške dijele na jezične i nejezične. Rosandić (2002) jezične pogreške dijeli u nekoliko skupina (podvrsta): gramatičke (morfološke, sintaktičke, tvorbene), leksičke, stilističke i pravopisne, dok Težak (1998) u svojoj podjeli izostavlja leksičke, ali navodi pravogovorne pogreške (koje su tipične u govornom stvaralaštvu). Nejezične pogreške, prema Rosandiću (2002), dijele se na sadržajne, logičke, kompozicijske i interpretacijske.

dobiven je i sljedeći primjer u kojemu se zastupljene dvije vrste tekstnih pogrešaka – odstupanje od kodificirane norme i tipografska pogreška: ... *ja sam joj rekao da ne, a ona je rekla e bas* ***ocu.***

Morfološke i sintaktičke pogreške temelje se na preklapanju završnih oblika riječi, što uzrokuje pogrešku i na sintaktičkoj razini. Moguće je na toj jezičnoj razini izdvojiti dva tipa pogrešaka. Jedna je nerazlikovanje homografnih lema zbog čega se u pretrazi miješaju paradigmatski oblici različitih vrsta riječi. Primjerice, u pretrazi primjera za prijedlog *podno* dobiven je i primjer *podno grijanje*, a u pretrazi primjera za pridjev *djedov* dobiju su rezultati za množinske oblike imenice *djed* (*djedovi, djedova*). Drugi tip morfosintaktičkih pogrešaka temelji se na pogreškama zbog upotrebe nepravilnih paradigmatskih oblika neke leme koji se ne uklapaju u kodificiranu sintagmatsku ili sintaktičku strukturu koje su dio: *Mi* ***bi*** *se trebali malo uljuditi i riješiti te proklete korupcije. / 23. listopada 2012. godine započinje tečaj opće astronomije za učenike osnovnih i srednjih škola, a* ***kojeg*** *mogu polaziti i građani.*

Sintaktičke pogreške pojavljuju se kao otkloni od kodificiranih sintaktičkih obrazaca. Za ovo istraživanje uzimani su primjeri iz administrativnih i publicističkih tekstova u kojima se očekuje primjena standardnojezičnih pravila. Tipične su sintaktičke pogreške nepravilan položaj nenaglašene spone glagola *biti* (*Dolazak na elektroničku evaluaciju* ***je*** *OBAVEZAN za sve redovne studente Zdravstvenog veleučilišta. / Iznosi tih potpora* ***su se*** *posljednjih godina čak smanjivali...*), reducirana upotreba izraza *s obzirom da* umjesto *s obzirom na to da* u zavisnim konstrukcijama (***S obzirom da*** *posljednji pokušaj mirenja 11 sindikata s Vladom danas nije uspio, sindikati od srijede, 5. lipnja kreću u štrajk.*), upotreba upitne konstrukcije *da li* umjesto *je li* (***Da li*** *je ostvarenje vrhunskih ciljeva nemoguće ili organizacijska sposobnost za ostvarivanje vrhunskih ciljeva nije odgovarajuća?*).[6]

Interpunkcijske pogreške nastaju zbog izostavljanja ili pogrešnoga položaja pravopisnih znakova, najčešće zareza u rečenici, osobito o inverziji te ispred vokativa (*Dok se vrte one kuglice po ekranu povremeno se pogledavamo on i ja i sve se čini kao u nekom snu. / Od munje, groma, leda, rata, kuge i gladi oslobodi nas Gospodine.*).

Semantičke pogreške prema Rosandiću (2002) mogu se ubrojiti u izvanjezične pogreške, dakle uvjetovane su nepoznavanjem sadržaja, a ne gramatike. Najčešće se ostvaruju u pogrešnoj upotrebi pojedinih riječi zbog nerazlikovanja njihova značenja (*Na slikama se lijepo vidi kako je uvaženi odvjetnik Hodak uputio svog klijenta u* ***civilizacijsko*** [umjesto *civilizirano*, op. a.] *ponašanje / Melanom se može razviti iz* ***mladeža*** [umjesto *madeža*, op. a.], *pa ga je potrebno dijagnosticirati na vrijeme kako bi se što uspješnije liječio.*) Također, u skupinu semantičkih pogrešaka mogu

6 Reducirana veznička konstrukcija *s obzirom da* u hrWaC-u je potvrđena 87 027 puta, dok je *s obzirom na to da* potvrđen 24 852 puta. Također, izraz *da li* u hrWaC-u je potvrđen 217 433 puta, a izraz *je li* potvrđen je 321 181 put.

se ubrojiti neki regionalni, razgovorni i neslužbeni oblici koji su u publicistici ili administrativnim tekstovima u upotrebi češće od standarnojezičnoga naziva (*Ravnateljica škole Nada Celestin Jelić potvrdila nam je da je škola od djece* ***karte*** [umjesto *ulaznice*, op. a.] *za kazalište doista naplatila po cijeni od 45 kuna. / Liječnici se slažu da je dobro vrijeme za dijete odvesti* ***zubaru*** [umjesto stomatologu, op. a.] *za njegov prvi rođendan.*

Stilističke pogreške proistječu iz neodgovarajuće primjene oblika koji pripadaju nestandardnoj jezičnoj upotrebi pa se u tekstovima u kojima se očekuje neutralna forma pojavljuju žargonizmi, arhaizmi, regionalizmi ili kolokvijalizmi. Stilističke pogreške manifestiraju se na obličnoj i leksičkoj razini pod utjecajem razgovorne i regionalne upotrebe (*Iako Vlada* ***obećaje*** [umjesto *obećava*, op. a.] *kako se neto plaće zaposlenih neće sniziti, sindikati ogorčeno tvrde da hoće, samo je pitanje koliko. / Škoti bi se mogli ubaciti u igru ali samo ako* ***dobe*** [umjesto *dobiju*, op. a.] *Talijane. / Djeca u sve mlađoj dobi* ***dobijaju*** [umjesto *dobivaju*, op. a.] *prvi mobilni telefon. / Cijepljenje* ***se preporuča*** [umjesto *preporučuje*, op. a.] *djevojčicama i ženama od 9 do 26 godina starosti.*).

4.3. **Pogreške korpusa**

Posljednji tip pogrešaka koji je moguće izdvojiti pri analizi pogrešaka ili odstupanja u primjeni korpusa u istraživačkom radu jest sam korpus, odnosno pozadinski alati koji obrađuju tekstne podatke. Korpusni alati segmentiraju tekstne jedinice na temeljne jezične dijelove: leme (leksičke jedinice) i tagove (morfološke jedinice), a upravo u tim dvama segmentima korisnik može uočiti otklone od očekivanih rezultata pretrage.

4.3.1. *Pogreške zbog sinkretizma*

Temeljne i najčešće pogreške korpusa proizlaze iz sinkretizma na razini leme i tagova. Pri automatskom tagiranju (obilježavanju vrste riječi i morfoloških oblika) ne razlikuju se homografi zbog čega se kao rezultati pretrage prikazuju oblici koji pripadaju istoj vrsti riječi, ali su različita roda, npr. u pretrazi leme *roda* u rezultatima se dobiju primjeri za lemu *rod* (*Otac mu je bio* ***rodom*** *iz Gomirja /* ***Rodu*** *lavanda (lat.* Lavandula*) pripada preko 40 različitih vrsta.*), ili pripadaju različitim vrstama riječi, npr. u pretrazi leme *orao* u rezultatima se dobiju primjeri za lemu *orati* (*Naš traktor mirno je* ***orao*** *zemlju, kada su na njega pripucali Kinezi.*), a u pretragu leme *muž* uvršteni su i oblici leme *mužev* (*... ni moji ni* ***muževi*** *roditelji ne žive u Kaštelima - kazuje nam Valerija.*).

Pogreškama korpusa pripadaju pogreške koje proistječu iz nerazlikovanja leme u službi imena i u općoj upotrebi te se u rezultatima pretrage s velikom učestalošću pojavljuju primjeri koji imaju obilježja imena, npr. pod *ljevak* često je ime nakladnika *Ljevak* (nakladnik i prezime), pod lemom *bratić* pojavljuje se prezime *Bratić*.

Lema može biti i sastavni dio imena književnoga djela ili filma: npr. kao rezultat pretrage za lemu *kum* najčešće potvrde povezane su s filmom *Kum* (*gledati Kuma*), pod lemom *špijun* često je ime filma *Balkanski špijun*, pod lemom *čarobnjak* često se pojavljuje naslov *Čarobnjak iz Oza*, a pod *dječak* čest je primjer naslova *Zagonetni dječak*. Takvi primjeri ulaze u statistički prosjek pretrage te daju drukčije rezultate od očekivanih ili objektivnih i potvrđuju već izneseno tumačenje da tekstovi u korpusu pripadaju istim izvorima ili istomu tematskom području, što u konačnici rezultira sličnim ili ponavljajućim primjerima.

4.3.2. *Pogreške u skicama riječi*

U ciljanim znanstvenim istraživanjima u Sketch Engineu omogućena je pretraga prema skicama riječi (engl. *word sketches*), što omogućuje odabir velikoga broja kolokacija, tj. najčešćih primjera prema vrsti riječi te morfološkim ili sintaktičkim obrascima (pridjev + imenica, imenica + imenica, imenica u nominativu + imenica u genitivu, predikat + objekt, subjekt + predikat itd.). Zbog svih ranije navedenih odstupanja skice riječi nisu uvijek pokazatelj samo dobrih primjera jer se različiti tipovi pogrešaka i odstupanja mogu pronaći i u toj vrsti pretrage. Korpus, s obzirom na to da se temelji na statističkoj obradi povezivosti riječi, skicama riječi ipak ne može pokazati posebnije semantičke odnose među sastavnicama dobivenih sveza. Primjerice, u pretrazi leme *arhitekt* i njezinih pridjevnih kolokata primjeri su svrstani samo na temelju zadanih morfoloških tagova pridjev + imenica, ali korisnik sam mora prepoznati o kojoj je vrsti sveze riječ (npr. razlikovati slobodne sveze *bogat/slavan arhitekt*, od čvrstih sveza *ovlašteni/diplomirani arhitekt*). S toga se aspekta možemo zapitati koliko korpusni alati uistinu ubrzavaju rad, tj. koliko je vremena potrebno da se nađe odgovarajući primjer. Pritom određenu ulogu imaju i izvori s kojih su preuzeti tekstovi uvršteni u korpus[7] – za neke rjeđe riječi koje su dio sustavne rječničke obrade i tvorbe (npr. pridjevi izvedeni od mocijskih parnjaka) nema potvrda, npr. za *gimnastičarkin*, *gimnastičarov/-ev* korpus ne nudi primjere. Također, s obzirom na iste izvore (najčešće publicistika) pitanje je koliko su pojedine kolokacije relevantni pokazatelji povezivosti riječi, npr. među frekventnijim glagolskim kolokatima uz imenicu *dječak* (prema skicama riječi) jest glagol *preminuti*, što i nije toliko relevantan podatak za upotrebu imenice *dječak*, a upućuje na zaključak da većina tekstova pripada člancima iz crne kronike.

7 Što se može dovesti u vezu s (djelomično) 1. te 2. i 3. tezom iz Hunston (2002) da korisnik korpusa mora biti svjestan i eventualnih nedostataka kako bi korpuse mogao upotrebljavati na pravilan način, da korpus prikazuje svoj sadržaj, a ne jezik u cijelosti te da korpus donosi informacije, a ne dokaze koje treba interpretirati intuicijom i dodatnim jezičnim znanjem.

5. Zaključak

Korpus nam pokazuje stvarno, suvremeno stanje jezične upotrebe, odnos upotrebe i jezične norme i predstavlja suvremeno vrelo jezikoslovnih podataka u skladu s korisnikovim željama i potrebama. Mogućnosti pretrage gotovo su neograničene, osim jednostavne pretrage moguća je i složena pretraga koja uključuje upotrebu regularnih izraza za specifične pretrage, ali takva pretraga od korisnika zahtijeva i specifičnije znanje kako bi u konačnici dobio odgovarajuće rezultate.

Neovisno o vrsti pretrage korpusni rezultati pokazuju i neke nedostatke (pogreške i odstupanja) s kojima se korisnik suočava tijekom pretrage ili pri analizi rezultata. Ti su nedostatci ovdje razgraničeni prema određenim kriterijima, tj. prema izvorima pogreške ili odstupanja koji mogu biti korisnik, tekst ili korpusni alati.

Iako to nije primarna tema rada, analizom podataka uključenih u ovo istraživanje nameće se pitanje koliko je dopušteno intervenirati u tekst iz korpusa (ako se, primjerice, potvrde iz korpusa navode kao primjeri u rječniku kad se očekuje da se značenje rječničkih natuknica prikaže dobrim primjerima jezične upotrebe) s obzirom na pravilo da se u citirane tekstove ne treba intervenirati.[8] Tekstovi iz korpusa te pogreške i odstupanja od neutralne, kodificirane norme kao sastavni dio takvih tekstova dobar su pokazatelj stvarne jezične upotrebe u odnosu na zadanu, tj. normom propisanu upotrebu. Na nekim je razinama odstupanja od norme lakše uočiti (pravopis, morfologija), dok na sintagmatskoj i sintaktičkoj razini, na kojoj se riječi povezuju u izraze i rečenice, postoji više kombinacija, a nisu sve kombinacije normirane (*prejesti se kruhom / prejesti se kruha*). Korpus nam pomaže da lakše sagledamo tu višu razinu te uočimo neko pravilo ili odstupanje od pravila ako pravilo već postoji.

Znanstvena dosljednost nalaže nam doslovno prenošenje tuđih riječi. Međutim, korpus nam služi kao potvrda neke teze ili misli, on nam treba olakšati rad, a ne otežati ga u potrazi za najboljim primjerom. Ako nam je korpus izvor u leksikografskom radu, važno nam je pronaći dobar primjer za značenje riječi pa minimalne pravopisne intervencije ne bi trebalo shvatiti kao iskrivljavanje podataka. Međutim, ako proučavamo pravopis i poštivanje pravopisne norme odnosno odstupanja od nje, pravopisna intervencija narušit će objektivnost podataka. Također, sinkretizam koji se pojavljuje u lemama i tagovima upućuje nas na to da, primjerice, u sintaktičkim istraživanjima statističke podatke treba uzeti s mjerom opreza.

No bez obzira na sve uočene nedostatke, korpus je u današnje vrijeme nezaobilazan i lako dostupan alat koji nam olakšava mnoga istraživanja na svim jezičnim razinama, od opisa značenja riječi do sintaktičkih obrazaca u sintagmama i rečenicama.

8 Problem intervencije u podatke iz korpusa može se obraditi unutar posebnoga rada.

Literatura

Birtić, Matea i dr. (2012) *Školski rječnik hrvatskoga jezika*. Zagreb: Institut za hrvatski jezik i jezikoslovlje – Školska knjiga.

Blagus Bartolec, Goranka; Matas Ivanković, Ivana (2017) „Kad nam korpus ispunjava želje." *Hrvatski jezik : znanstveno-popularni časopis za kulturu hrvatskoga jezika* 4 (3): 25 – 28.

Blagus Bartolec, Goranka; Matas Ivanković, Ivana (2018) „Corpus analysis of Croatian constructions with the verb *doći* 'to come'." U *Multiword Units in Machine Translation and Translation Technology,* Ur. Mitkov, Ruslan; Monti, Johana; Corpas Pastor, Gloria, Seretan, Violeta, 223–242. Amsterdam: John Benjamins Publishing Company.

Ćavar, Damir; Brozović Rončević, Dunja (2012) „Riznica: The Croatian Language Corpus". *Prace filologiczne* LXIII: 51–65.

Jelaska, Zrinka; Bjedov, Vesna (2015) „Pogrješke ili promjene – ovladanost odabranim hrvatskim morfosintaktičkim sadržajima učenika završnoga razreda osnovne škole." *Jezikoslovlje* 16 (2–3): 227–252.

Hrvatsko strukovno nazivlje: *Struna*, http://struna.ihjj.hr/, pristup travanj i rujan 2018.

Hunston, Susan (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Klobučar Srbić, Iva (2008) „Obol korpusne lingvistike suvremenoj leksikografiji" *Studia lexicographica: časopis za leksikografiju i enciklopedistiku* 2(3): 39–51.

Ljubešić, Nikola; Klubička, Filip (2014). „{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian." U *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, ur. Bildhauer, Felix; Schäfer, Roland, 29–35. Gothenburg: Association for Computational Linguistics.

Nesselhauf, Nadja (2005) *Corpus Linguistics: A Practical Introduction*. http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf. pristup 30. 4. 2018. i 20. 9. 2018.

Pala, Karel; Rychlý, Pavel; Smrž, Pavel (2003) „Text Corpus with Errors." U *Text, Speech and Dialogue*: *Sixth International Conference*, ur. Matoušek, Vaclav; Mautner, Pavel, 90–97, Berlin – Heilderbeg: Springer Verlag.

Rosandić, Dragutin (2002) *Od slova do teksta i metateksta*. Zagreb: Profil.

Tadić, Marko (1998) „Raspon, opseg i sastav korpusa hrvatskoga suvremenog jezika." *Filologija* 30–31: 337–347.

Tadić, Marko (2003) *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex libris.

Tadić, Marko (2009) „New version of the Croatian National Corpus." U *After Half a Century of Slavonic Natural Language Processing*, ur. Hlaváčková, Dana; Horák, Aleš; Osolsobě, Klára; Rychlý, Pavel, 199–205. Brno: Masaryk University.

Težak, Stjepko (1998) *Teorija i praksa nastave hrvatskoga jezika* (II. izdanje). Zagreb: Školska knjiga.

Utvić, Miloš V. (2013) *Izgradnja referentnog korpusa savremenog srpskog jezika (doktorska disertacija).* Beograd: Univerzitet u Beogradu, Filološki fakultet. https://fedorabg.bg.ac.rs/fedora/get/o:10061/bdef:Content/get. pristup 20. 9. 2018.

About corpus from user's point of view

In contemporary linguistic analysis the corpus has become an unavoidable source of data collection at all language levels. The development of corpus tools for Croatian language follows the development of European corpus and computer linguistics. Croatian can be searched by using three computer corpora: Croatian Language Repository of the Institute for Croatian language and linguistics, Hrvatski nacionalni korpus and hrWaC. The making of these corpora has been described in several works (e.g. Tadić 2009; Ćavar, Brozović Rončević 2012; Ljubešić, Klubička 2014), while works on using the corpus from the user's point of view are rare (Blagus Bartolec, Matas Ivanković 2017, 2018). hrWaC i Hrvatski nacionalni korpus are developed in NoSketch Engine and they are tagged, which enables identification of word classes and morphological forms, and more complex searches (such as syntactic structures). Ultimately, such a search enables individualized processing of corpus data. However, required and obtained results often contain errors at all language levels, which can make statistical interpretation questionable. Source of these disparities can be: 1. corpus user (a user may incorrectly set up a search by improperly setting a regular expression or incorrectly using abbreviations applied in a corpus) 2. texts taken in the corpus (texts in hrWaC have been collected from news portals, forums and web pages of legislative organizations, which means that there are many conversational and unconventional forms, which can be problematic in tagging, and thus in corpus search, e.g. *Citam danas o onoj kokosi iz valjda Duge Rese...*); 3. corpus tagging (in automatic tagging the problem are homographs (e.g. in the search for the noun *orao* 'eagle', corpus gives examples for the verb *orati* 'plow': *Naš traktor mirno je orao zemlju.*). In tagging, the problem can cause forms unusual in Croatian, (e.g. in the search for the noun *sova* 'owl', the corpus gave example: *McDonald'sova omiljena meta su djeca.*). This article presents the language mistakes that can be encountered in using the corpus. Based on typical examples, the errors will be grouped according to their language characteristics. The intention is to contribute to the improvement and quality of the corpus, which will enable further reliable research of a particular language phenomenon.

Key words: Croatian, deviations and errors in the corpus, hrWaC, user perspectives