Leksikon emocija hrvatskog jezika

Gašparić, Brigita

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2019./ 2020.

Brigita Gašparić

Leksikon emocija hrvatskog jezika

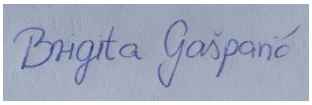
Završni rad

Mentorica: prof. dr. sc. Nives Mikelić Preradović

Zagreb, rujan 2020.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.



_____

(potpis)

Sadržaj

1. Uvod

Nacionalno Istraživačko Vijeće (engl. *National Research Council* – skraćeno NRC) iz Kanade je 2010. godine objavilo projekt pod nazivom *NRC Word-Emotion Association Lexicon* (skraćeno *EmoLex*). Autori EmoLexa, Saif Mohammad i Peter Turney, proveli su istraživanje koje je rezultiralo stvaranjem opširnog leksikona emocija koji sadrži 14 182 natuknice te sentimente i emocije koje su ispitanici povezali s tim natuknicama. Kako bi osigurali da EmoLex ne ostane primjenjiv samo na engleski jezik, 2015. godine preveli su riječi iz leksikona na više od četrdeset jezika, a 2017. su ponovili postupak te preveli EmoLex na više od stotinu jezika, uključujući i hrvatski jezik. Postupak se vršio putem Google Prevoditelja, što znači da je leksikon automatski preveden pomoću internetskog prevodilačkog alata. Kako automatski generirani prijevodi nisu uvijek točni i u većini slučajeva ne dopuštaju višeznačnost, bilo je potrebno ispraviti prijevode na hrvatski jezik i time omogućiti da EmoLex uđe u opću uporabu kao Leksikon emocija hrvatskog jezika. Iako se pokazalo da je ispravljanje ovog velikog leksikona dugotrajan posao, autorica je u ispravljanju EmoLexa pronašla savršenu temu za svoj završni rad na Odsjeku za informacijske i komunikacijske znanosti jer je tema ovog rada spoj interesa s oba studija – informacijskih znanosti s jedne te anglistike s druge strane.

2. Leksikoni emocija i sentimenta

Popisi pojmova koji povezuju određene riječi i izraze s ljudskim doživljajem istih nazivaju se leksikoni emocija i sentimenta (Jakopović i Mikelić Preradović, 2016, str. 64). Takvi leksikoni mogu se izraditi na više načina ovisno o tome na što se stavlja fokus. Sukladno unaprijed određenom fokusu, ovakvi leksikoni imaju mnogo primjena u raznim područjima ljudske djelatnosti. Posebno mjesto zauzimaju u uslužnim djelatnostima koje veliki dio svog marketinga temelje na ispitivanju svojih korisnika i potrošača. Najčešća metoda ispitivanja je ona koja se vrši pomoću kratkih anketa za korisnike koje se temelje na jednostavnim pitanjima o njihovom dosadašnjem iskustvu s određenom uslugom, proizvodima, zaposlenicima, tvrtkom, institucijom, itd. Te ankete daju povratne informacije o stavovima, željama i prigovorima korisnika i pokazuju koliko pozitivno ili negativno korisnici doživljavaju brend neke tvrtke ili institucije te koje emocije vežu uz njih i njihove usluge. S tim saznanjima poduzetnici mogu donositi odluke i mijenjati poslovne planove sukladno rezultatima anketa kako bi udovoljili svojim korisnicima. Ostale primjene leksikona emocija i sentimenta uključuju:

- poboljšanje odnosa s kupcima na način da se uzima u obzir emocionalno stanje kupaca i općenito korisnika te se postupa u skladu s tim stanjem (Bougie *et al.*, 2003);
- praćenje sentimenata prema političarima, filmovima, proizvodima, zemljama i drugim ciljanim pojmovima (Pang i Lee, 2008; Mohammad i Yang, 2011).
- izradu sofisticiranih algoritama za pretraživanje koji su u mogućnosti razlikovati emocije vezane uz neki proizvod ili uslugu (Knautz *et al.*, 2010);
- stvaranje dijaloških sustava koji odgovaraju na prikladan način, sukladno emocionalnom stanju korisnika (Velásquez, 1997; Ravaja *et al*., 2006);
- izradu pametnih sustava za učenje koji uzimaju u obzir emocionalno stanje učenika kako bi učenje bilo učinkovitije (Litman i Forbes-Riley, 2004);
- utvrđivanje rizika od ponavljanja pokušaja samoubojstva analiziranjem oproštajnih pisama (Osgood i Walker, 1959; Matykiewicz *et al*., 2009; Pestian *et al.*, 2008);
- shvaćanje kako osobe različitih spolova komuniciraju kroz poslovne i osobne e-mail adrese (Mohammad i Yang, 2011);
- pomoć prilikom sastavljanja e-pošte, dokumenata i sličnih tekstova kako bi se prenijela željena emocija i izbjegla pogrešna interpretacija (Liu *et al*., 2003);

- prikaz tijeka emocija u romanima i ostalim knjigama (Boucouvalas, 2002; Mohammad, 2011b);
- detektiranje emocija koje naslovi u novinskim člancima pokušavaju izazvati u ljudima (Bellegarda, 2010);
- reklasifikaciju i kategoriziranje informacija/odgovora u internetskim forumima koji funkcioniraju na principu pitanje - odgovor (Adamic *et al*., 2008);
- otkrivanje kako ljudi koriste afektivne riječi i metafore u svrhu uvjeravanja, nagovaranja i prisiljavanja drugih, primjerice u propagandi (Kŏvecses, 2003);
- daljnji razvoj sustava pretvorbe teksta u govor kako bi im jezik postao što prirodniji (Francisco i Gervás, 2006; Bellegarda, 2010);
- razvijanje pomoćnih robota koji očitavaju ljudske emocije i reagiraju u skladu s njima (Breazeal i Brooks, 2004) (Mohammad i Turney, 2013, str. 3).

Ovaj iscrpni popis daje osnovnu predodžbu o važnosti ovakvih rječnika, no treba imati na umu da emocija i sentiment nisu potpuni sinonimi te se leksikoni emocija i leksikoni sentimenata razlikuju ne samo po definiciji, nego i po informacijama koje nam daju o nekoj riječi ili izrazu (više o toj temi u idućem poglavlju).

Mnogi izrazi mogu probuditi različite emocije u različitim kontekstima, no emocija koju izazove neka rečenica ili izraz nije jednostavno zbroj emocija koje izaziva svaka pojedina riječ u toj rečenici ili izrazu. Unatoč tome, leksikon emocija može biti vrlo korisna komponenta sofisticiranog algoritma koji detektira emocije za sve prethodno navedene primjene. Takav leksikon može također biti koristan za evaluaciju automatiziranih metoda koje određuju emocije povezane s određenim izrazom. Nadalje, takav se algoritam može primijeniti na način da automatski generira leksikone emocija za jezike koji još uvijek nemaju takvu vrstu leksikona, primjerice za hrvatski jezik – a upravo to je i cilj ovog rada. Trenutno još uvijek ne postoji leksikon emocija (za bilo koji jezik) koji bi bio visoke kvalitete i širokog opsega, no postoji nekoliko leksikona užeg opsega za nekolicinu jezika, poput *WordNet Affect Lexicona (WAL)* (Strapparava i Valitutti, 2004), *General Inquirera (GI)* (Stone *et al*., 1966) i *Affective Norms for English Words (ANEW)* (Bradley i Lang, 1999). Ovi leksikoni nastali su metodom „nabave iz mnoštva“ (engl. *crowdsourcing*), procesa tijekom kojeg se posao označavanja riječi dijeli na mnogo manjih neovisnih dijelova i zadaje velikom broju anotatora, obično putem interneta. Howe i Robinson (2006), koji su izmislili taj izraz, definiraju „nabavu iz mnoštva“ na sljedeći način:

> „Nabava iz mnoštva je čin tvrtke ili institucije kojim uzimaju funkciju nekoć izvođenu od strane zaposlenika te tu funkciju povjeravaju vanjskim izvršiteljima koji čine nedefiniranu (i obično vrlo opširnu) mrežu ljudi u obliku otvorenog poziva. To može biti u obliku *peer*-proizvodnje (ako se posao odrađuje zajednički, s kolegama), no često se uzimaju i individue, međusobno nepovezane osobe. Ključni preduvjet je korištenje formata otvorenog poziva i velike mreže potencijalnih radnika."

EmoLex je također nastao metodom „nabave iz mnoštva", konkretnije ručnim označavanjem riječi korištenjem Amazonove usluge *Mechanical Turk* (www.mturk.com). U smislu opsega, EmoLex je veći od *WordNet Affect Lexicona*, a usredotočen je na osam temeljnih i prototipnih emocija prema „Indeksu profila emocija (PIE)" Roberta Plutchika (1980) koje čine veselje, tuga, ljutnja, strah, povjerenje, gađenje, iznenađenje i iščekivanje (primjer modela je dostupan na Slici 1.). Model funkcionira na principu polarnih suprotnosti, dakle, osam primarnih emocija je svrstano u oprečne parove: veselje – tuga; povjerenje – gađenje; strah – ljutnja te iznenađenje – iščekivanje. Iz tih se osam emocija „kotač emocija" širi u manje intenzivne emocije koje su zapravo kombinacija više temeljnih emocija ili se sužava u intenzivnije emocije poput bijesa, mržnje i obožavanja (Pico, 2016).

EmoLex u svojoj bazi izraza ima podosta riječi iz prethodno spomenutih resursa - *WordNet Affect Lexicona (WAL)* i *General Inquirera (GI)*, stoga ćemo ih ukratko opisati. WAL je sastavljen od nekoliko stotina riječi te uz svaku riječ stoji napomena u kojoj su navedene emocije koje ta riječ izaziva. Taj popis označenih riječi stvoren je na način da su se ručno označavale emocije za polazišne pojmove (engl. *seed words*) nakon čega su se u WAL-u svi sinonimi tih riječi označili istim emocijama. Za WAL je stvoren i podskup izraza koji je napravljen da odgovara šest emocija koje Ekman (1992) smatra osnovnima: veselje, tuga, ljutnja, strah, gađenje i iznenađenje. GI je sastavljen od 11 788 riječi kojima su dodijeljene oznake za riječi u 182 kategorije, uključujući pozitivnu i negativnu značenjsku orijentaciju. Označene riječi iz oba leksikona su tijekom eksperimenta u *Mechanical Turku* ponovno označene kako bi se ustanovilo koliko se rezultati označavanja volontera bez iskustva razlikuju od rezultata odabranih anotatora koji su označavali polazišne pojmove, no o tome će biti više riječi u poglavlju o EmoLexu (Mohammad i Turney, 2013, str. 2-7).

2.1. Razlika između leksikona emocija i leksikona sentimenta

Prije nego se fokusiramo isključivo na NRC Leksikon (EmoLex) koji sadrži i emocije i sentimente zadanih riječi, potrebno je razjasniti po čemu su ove dvije vrste leksikona razlikuju.

2.1.1. Leksikoni emocija

Jedna od mogućih definicija leksikona emocija glasi:

> „Leksikon emocija je rječnik koji povezuje riječi s kategorijama emocija kao što su ljutnja, strah, iznenađenje, tuga, itd. Takvi leksikoni najčešće se stvaraju kroz proces zvan nabava iz mnoštva ili dobivanje masovne podrške (engl. *crowdsourcing*). Proces nabave iz mnoštva uključuje grupu ljudi od koje se traži da označe skup riječi na način da svakoj riječi pridruže jednu ili više osnovnih emocija. Za svaku riječ se potom izračuna ukupan broj oznaka za neku emociju te se konačan rezultat normalizira kroz osnovne emocije, što predstavlja emocionalnu distribuciju za svaku pojedinu riječ. Konačan rezultat procesa je skup riječi zajedno s njihovom emocionalnom distribucijom što se naziva leksikonom emocija.“ (Ciampaglia *et al*., 2017, str. 432)

Ovo objašnjenje jasno pokazuje da se riječi povezuju s unaprijed određenim *emocijama*, što znači da se od ispitanika tražilo da za svaku riječ pokušaju odrediti kakav osjećaj ona budi u njima. Najčešće korištene emocije su sve ili neke emocije iz već spomenutog Plutchikovog modela na Slici 1., a to su veselje, ljubav, prihvaćanje, iznenađenje, strah, tuga, gađenje, očekivanje, ljutnja i bijes (nazivi emocija često variraju, no smisao modela ostaje isti).

Ovdje je potrebno naglasiti da se emocije mogu prikazati kroz dva različita modela koristeći dva različita pristupa. Prvi (i popularniji) pristup emocije prikazuje kategorički te se emocije sastoje od oznaka poput „dosada“, „frustracija“ i „ljutnja“. Alternativni pristup naglašava važnost temeljnih dimenzija *valencije* i *uzbuđenja* u otkrivanju i dubljem razumijevanju emocionalnog iskustva – poznatiji kao dimenzionalni pristup. Iako postoji mnogo teorija, ona najprihvaćenija pretpostavlja postojanje najmanje dviju temeljnih dimenzija: valenciju (zadovoljstvo/nezadovoljstvo) i uzbuđenje (aktivacija/deaktivacija). Postoji i pretpostavka o postojanju treće dimenzije – *dominantnosti*. U praksi je kategorički model usvojen za izražavanje i otkrivanje emocija u tekstualnim zapisima i videozapisima, dok je dimenzionalni prikaz primjereniji za obradu govora (Mingli *et al*., 2008).

U ovom trenutku postoje tri pristupa koji dominiraju zadacima fokusiranim na detektiranje emocija, a to su: oni koji se temelje na ključnim riječima, oni koji se temelje na učenju te hibridni pristup koji spaja prethodna dva pristupa što znači da koristi sadržaje i mogućnosti sintaktičkih i semantičkih podataka kako bi se detektirale emocije. Prvi pristup, koji se temelji na detekciji ključnih riječi, ovisi o tome koliko se ključnih riječi nalazi u danom tekstu te često zahtijeva prethodno parsiranje zadanog teksta, tj. dodjeljivanje sintaktičkih obilježja uz uporabu rječnika emocija. Ovaj pristup je jednostavan za primijeniti, intuitivan je i daje vrlo jasne rezultate. Drugi pristup temeljen na učenju koristi kvalificiranog anotatora, osobu koja kategorizira uneseni tekst u klase emocija koristeći se ključnim riječima. Lakše je i brže adaptirati promjene direktno u domeni jer algoritam može vrlo brzo naučiti nove značajke iz korpusa. To se može omogućiti tako da se algoritmu za strojno učenje pruži velik skup podataka za obuku, nakon čega će algoritam izraditi klasifikacijski model. Mana ovog pristupa je nedostatak dostupnih korpusa što rezultira nedefiniranim granicama između klasa emocija i nedostatkom kontekstne analize. Treći, hibridni pristup kombinira primjenu ključnih riječi i strojnog učenja. Najveća prednost ovog pristupa je da može proizvesti točnije rezultate, i to na način da koristi anotatore zajedno s mnoštvom lingvistički obogaćenih informacija iz rječnika i tezaurusa. Ovim pristupom smanjuju se troškovi ljudske radne snage te se minimaliziraju problemi koji nastaju prilikom integriranja više različitih leksičkih resursa (Shivhare i Khethawat, 2012).

Cilj najranijih istraživanja fokusiranih na vezu između teksta i emocija bio je bolje razumijevanje izražavanja emocija ljudi kroz pisanu građu te je glavno pitanje bilo na koji način tekstualni zapisi mogu izazvati različite emocije. Najznačajnija istraživanja u tom području proveli su Osgood i sur. (1975) te Lutz i White (1986). Osgood je koristio višedimenzionalno skaliranje (engl. *multidimensional scaling*, skraćeno MDS) kako bi vizualizirao riječi koje su obilježene emocijom, a taj se postupak temeljio na procjeni blizine riječi koje su dane ljudima iz različitih kultura. Te riječi predstavljale su točke u višedimenzionalnom prostoru, a procjena blizine tih riječi označava udaljenosti između njih. Iz ove vrste skaliranja riječi Osgood pretpostavlja tri dimenzije bitne za procjenu emocija vezanih uz riječi: *evaluaciju*, *potenciju* i *aktivnost*. Evaluacija mjeri u kojem se opsegu riječ odnosi na neki ugodan ili neugodan događaj. Potencija izračunava koliko je riječ intenzivna. Aktivnost se odnosi na to je li riječ u aktivu ili pasivu (primjerice glagol radni ili trpni). Lutz je koristio slične dimenzije, no uočio je razlike u matricama blizine riječi koje su posljedica ispitivanja ljudi iz različitih kultura. Nadalje, Samsonovich i Ascoli (2006) koristili su se

engleskim i francuskim rječnicima kako bi stvorili „konceptualne mape vrijednosti“, vrstu kognitivne mape koja je vrlo slična Osgoodovoj te su se složili oko njegovih triju dimenzija. Cohn, Mehl i Pennebaker (2004), Pennebaker, Mehl i Niederhoffer (2003) te Kahn, Tobin, Massey i Anderson (2007) bavili su se leksičkom analizom tekstova kako bi odredili koje riječi odaju emotivno stanje kod pisaca i govornika. Neki od ovih pristupa oslanjaju se na *Linguistic Inquiry and Word Count (LIWC)* (Pennebaker *et al*., 2001), vjerodostojan računalni alat koji analizira tekst koristeći kategorizaciju rječnika. Metode koje koriste LIWC alat pokušavaju ustanoviti koje su riječi one koje otkrivaju emotivan sadržaj u tekstu (Hancock *et al*., 2007). Istraživači su također primjenjivali pristupe koji se zasnivaju na korištenju korpusa, a ti pristupi pretpostavljaju da ljudi koji koriste isti jezik imaju slične koncepcije različitih suptilnih emocija te su se koristili tezaurusima i emotivnim izrazima, poput već spomenutog *WordNeta* koji se često koristi u istraživanjima u području računalne lingvistike (Miller *et al.*, 1990). Strapparava i Valitutti (2004) su ga proširili informacijama o afektivnim izrazima. *The Affective Norm for English Words* (skraćeno *ANEWs*) (Bradley i Lang, 1999; Stevenson, Mikels i James, 2007) jedan je od nekolicine projekata čiji je cilj razviti skupove normativnih emotivnih procjena za zbirke elemenata koje izazivaju emocije, u ovom slučaju engleske riječi. Te zbirke daju vrijednosti za dimenzije valencije, uzbuđenja i dominacije za svaku stavku (natuknicu), a prosjeci su izvedeni od velikog broja subjekata koji su označavali riječi. Ovdje također valja spomenuti ANET (*Affective Norms for English Text*), koji nudi zbirku normativnih emotivnih procjena za veliki broj kraćih tekstova na engleskom jeziku.

Postoje i sustavi koji u samom tekstu pronalaze afektivne riječi na principu semantičke analize samog teksta. Na primjer, Gill i sur. (2008) su analizirali 200 blogova te izvijestili da su oni dijelovi teksta za koje ljudi smatraju da izražavaju strah i veselje slični emocionalnim konceptima koji predstavljaju te riječi („fobija“ i „panika“ za strah te „blaženstvo“ za veselje). Koristili su se Latentnom semantičkom analizom (*Latent Semantic Analysis* – skraćeno LSA) (Landuer *et al.*, 2007) i hiperprostornom analogijom jezika (*Hyperspace Analogue to Language* – skraćeno HAL) (Lund i Burgess, 1996) kako bi automatski izračunali značenjsku sličnost između tekstova i ključnih riječi za emocije. Iako je ova metoda bila obećavajuća kod tekstova koji su izražavali strah i veselje, pokazala se neadekvatna kod tekstova koji izražavaju šest drugih emocija, poput ljutnje, gađenja i tuge. Pitanje o funkcionalnosti i korisnosti ove metode u području detekcije emocija ostaje otvoreno (El Gohary *et al.*, 2013, str. 2).

2.1.2. Leksikoni sentimenta

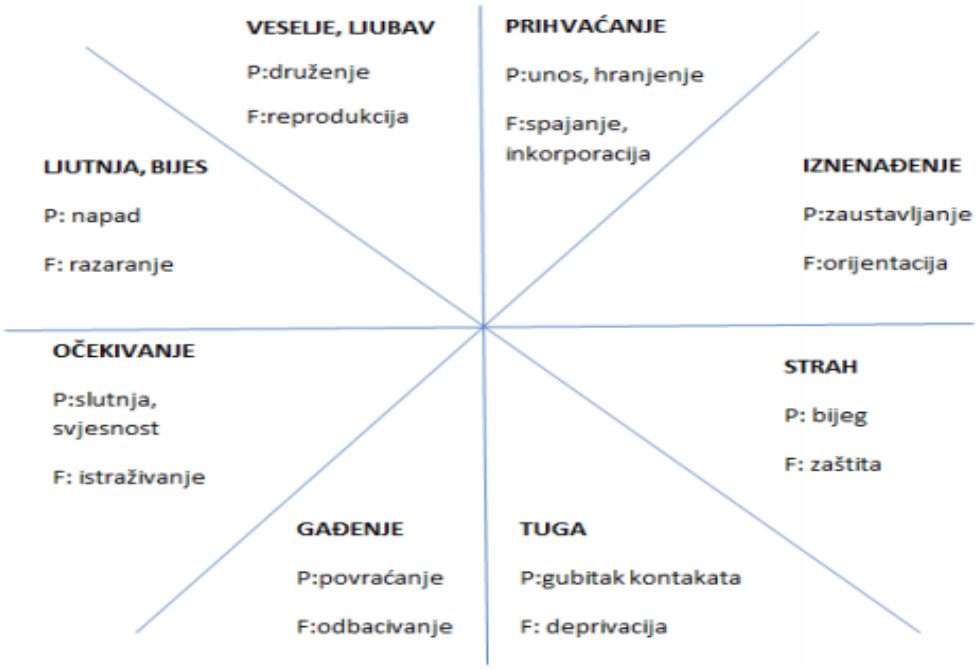
Najkompleksniji pristup tekstualno - afektivnom otkrivanju podrazumijeva sustave koji konstruiraju afektivne modele iz velikih svjetskih korpusa i primjenjuju ih kako bi otkrili afektivne riječi unutar tekstova (Akkaya, Wiebe i Mihalcea, 2009; Breck, Choi i Cardie, 2007; Liu, Lieberman i Selker, 2003; Pang i Lee, 2008; Shaikh, Prendinger i Ishizuka, 2008). Ovaj pristup naziva se analiza sentimenta, otvorena ekstrakcija ili analiza subjektivnosti jer je primarni fokus na *valenciju* nekog teksta (pozitivno ili negativno, dobro ili loše), a ne na dodjeljivanju neke od kategorija emocija (poput ljutnje ili tuge) tekstu (Pang i Lee, 2008). Koristile su se tehnike nadziranog strojnog učenja te nenadziranog učenja kako bi se prepoznale emocije u tekstovima. Najveća mana tehnika nadziranog strojnog učenja jest da su za obuku bile potrebne velike količine označenih tekstova. Emocionalne interpretacije nekog teksta mogu biti vrlo subjektivne, što znači da je potrebno više od jednog anotatora i zbog toga je cijeli proces označavanja dugotrajan i skup. Iz tog razloga su nenadzirane metode zastupljenije u području obrade prirodnog jezika (engl. *Natural Language Processing* – skraćeno NLP) i emocija. Strapparava i Mihalcea (2008) su usporedili jednu nadziranu metodu učenja (Naïve Bayes) i četiri nenadzirane metode (kombinacije LSA i *WordNet Affecta*) za raspoznavanje šest osnovnih emocija. S druge strane, D'Mello i sur. (2010) koristili su se LSA-om kako bi otkrili vrste izgovora i afektivnosti u dijalozima studenata u pametnom sustavu podučavanja (engl. *Intelligent Tutoring System*). Budući da je taj proces zahtijevao kategorizirani model emocija, D'Mello je predložio grupu kategorija koje opisuju afektivna stanja u sustavu studentskih dijaloga. Kort (2001) je integrirao oba modela emocija, stavljajući kategorije u ravninu prema kojoj se mjere valencija i uzbuđenje. Do danas, najviše je znanstvenika koristilo nadziranu metodu koja se temelji na kategoriziranom modelu emocija. Klasifikacija tekstova trenutno je glavna metoda za otkrivanje emocija u tekstu (Read, 2004), no zbog vrlo subjektivne prirode emocija nailazi se na mnogo prepreka. Kao i sur. (2009) utvrdili su tri najveća problema koja su povezana s tehnikama koje se baziraju na ključnim riječima, a to su: „neodređenost u definicijama ključnih riječi“, „nemogućnost raspoznavanja rečenica bez ključnih riječi“ i „nedostatak lingvističkih informacija“ (Bo i Lee, 2008) (El Gohary *et al*., 2013., str. 3).

Nakon uvoda u leksikone sentimenta, vrijeme je da ih detaljnije definiramo i opišemo. Sagar Ahire leksikone sentimenta u svom radu „A Survey of Sentiment Lexicons” (2015) opisuje na sljedeći način:

„Leksički izvor za analizu sentimenta, također poznat pod nazivom *leksikon sentimenta*, je baza podataka koja sadrži leksičke jedinice za određeni jezik zajedno s njihovim pripadajućim sentimentom. To može biti prikazano kao skup n-torki (engl. *tuple*) oblika [leksička jedinica, sentiment]. Ovdje leksička jedinica može biti riječ, jedno od značenja određene riječi, izraz, itd. S druge stane, sentiment se može prikazati na više načina, od kojih su neki:

- fiksna kategorizacija na *pozitivno* i *negativno*;
- konačan broj utvrđenih stupnjeva poput *izrazito pozitivno, donekle pozitivno, neutralno, donekle negativno, izrazito negativno;*
- prava vrijednost koja označava jačinu sentimenta u intervalu poput [-1, +1].“ (str. 2)

Leksikoni sentimenta se, dakle, fokusiraju na dvije suprotnosti – pozitivno i negativno te ponekad na cijeli raspon značenja između ta dva oprečna pojma. Da to prikažemo slikovitije, sentimenti su poput tonova crne i bijele boje, gdje crna boja predstavlja riječ „negativno“, dok bijela boja simbolizira „pozitivno“. Po potrebi se u obzir uzimaju i tonovi sive boje u tom rasponu između crne i bijele koji predstavljaju izraze poput „donekle pozitivno“ i „donekle negativno“. Ako su leksikoni sentimenta raspon između crne i bijele, leksikoni emocija bi u tom slučaju bili paleta od (najčešće) osam boja od kojih bi svaka boja predstavljala određenu emociju. Rezultat takvih rječnika bi tada bila obojenost riječi ovisno o tome kakve sentimente i/ili emocije bude u ljudima, kao što je na Slici 2. prikazano za EmoLex. Budući da ovi leksikoni imaju mnogo prethodno spomenutih primjena, svatko bira tip leksikona ovisno o krajnjoj namjeni. Ako se radi o ocjenjivanju neke usluge poput usluge u kafiću, vrlo vjerojatno će se koristiti crno - bijela paleta sentimenata kako bi se dobila povratna informacija u obliku pozitivno - negativno ili ocjena od jedan do pet. Ako pak želimo saznati kakvi su stavovi ljudi prema nekom društvenom fenomenu ili nekom aspektu trenutne političke situacije u državi, vrlo vjerojatno ćemo im dati širu paletu od osam ili više emocija kako bismo dobili potpuniju sliku o njihovom mišljenju. Iako obje vrste leksikona mogu imati zasebne primjene, potpuniji i šire primjenjivi rezultati svakako se dobivaju ako se riječi označe objema kategorijama: sentiment i emocije. Upravo tako izgleda NRC Leksikon emocija koji se ispravljao kako bi se mogao koristiti za hrvatski jezik.



Slika 2. Shematski prikaz emocija, pripadajućih ponašanja (P) i funkcija (F) (Plutchik, 1962; prema Zelenbrz, 2005)

Slika 1. Interaktivni vizualni prikaz EmoLexa predstavljen bojama (saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm)

3. Opis resursa

Saif Mohammad i Peter Turney su 8. svibnja 2010. godine predstavili *NRC Word-Emotion Association Lexicon* – poznat i pod prijašnjim nazivom *EmoLex*. EmoLex je rječnik sastavljen od 14 182 engleske riječi od kojih je svaka označena s 0 (ne odnosi se na ovu riječ) ili 1 (odnosi se na ovu riječ) za dva sentimenta (pozitivno i negativno) i osam emocija (ljutnja, iščekivanje, gađenje, strah, radost, tuga, iznenađenje i povjerenje).

Da bi se mogao izraditi leksikon koji riječi povezuje s emocijama koje te riječi izazivaju, prvo se morao izraditi popis riječi i izraza za koje je potrebna ručna anotacija. Za izradu EmoLexa kao izvor riječi odabran je *Macquarie Thesaurus* (www.macquarieonline.com.au/thesaurus.html) (Bernard, 1986). Kategorije u tezaurusu bile su grubi indikator značenja riječi (s time da se za riječ koja je navedena u dvije kategorije smatra da ima dva značenja). Bilo koji drugi objavljeni rječnik također bi poslužio svrsi. Osim 57 000 najfrekventinijh pojavnica, *Macquarie Thesaurus* također nudi više od 40 000 najfrekventnijih izraza. S tog popisa odabrani su oni izrazi koji se najčešće pojavljuju u Googleovom korpusu n-grama (engl. *n-grams*) (Brants i Franz, 2006). Prvo je odabrano 200 najčešćih unigrama (izraza sastavljenih od jedne riječi) i 200 najčešćih bigrama (izraza sastavljenih od dvije riječi) koji pripadaju jednoj od četiri vrsta riječi: imenicama, glagolima, prilozima i pridjevima. Prilikom odabira ovih skupova riječi, oni izrazi koji se pojavljuju u više kategorija *Macquarie Thesaurusa* su ignorirani (bilo je samo 187 priloga u obliku bigrama koji su zadovoljili taj uvjet, svi ostali skupovi – za imenice, pridjeve i glagole – sastoje se od 200 izraza svaki). Odabrane su i sve riječi iz Ekmanovog podskupa *WordNet* leksikona koje su imale najviše dva značenja (izraze koji su na popisu u najviše dvije kategorije u tezaurusu) – sveukupno 640 parova u obliku [riječ - značenje]. Uključeni su i svi izrazi iz *General Inquirera* koji nisu bili previše apstraktni (oni izrazi koji su imali najviše tri značenja) – sveukupno 8132 parova [riječ - značenje]. Neki od ovih izraza pojavljuju se u više skupova. Skup ova tri podskupa (Googleovi n-grami, izrazi iz WAL-a i GI-a) broji točno 10 170 parova u obliku [riječ - značenje].

Platforma koja se koristila kako bi se dobilo što više anotacija emocija je Amazonova usluga *Mechanical Turk*. Klijent koji zadaje zadatak *Mechanical Turku* naziva se podnositelj zahtjeva (engl. *requester*). Podnositelj zahtjeva rastavlja zadatak na male, međusobno neovisne dijelove koji se zovu *HIT-ovi (Human Intelligence Tasks* – zadaci koji zahtijevaju ljudsku inteligenciju) i objavljuje ih na stranici *Mechanical Turka*. Podnositelj zahtjeva

određuje: a) neke ključne riječi relevantne za zadatak kako bi pomogli zainteresiranima da pronađu *HIT* na Amazonovoj stranici, b) kompenzaciju koja će biti plaćena po svakom odrađenom *HIT-u,* te c) broj različitih anotatora koji moraju riješiti svaki *HIT*. Ljudi koji rješavaju te zadatke zovu se Turkeri (engl. *Turkers*). Turkeri obično pronalaze poslove i zadatke tako da upišu ključne riječi koje opisuju njihove interese te odrede minimalnu kompenzaciju koju traže za odrađivanje posla (jednog *HIT-a*). Kako bi se označili svi izrazi iz prethodno spomenutih skupova, za svaki se izraz stvorio jedan *HIT* u *Mechanical Turku*. Svaki zadatak imao je popis pitanja, na sva pitanja morala je odgovoriti ista osoba. Tražilo se označavanje pet različitih anotatora za svaki *HIT*. Turker može pokušati riješiti koliko god *HIT-ova* želi.

Iako korištenje *Mechanical Turka* ima mnogo prednosti, poput niske cijene, podjele rada i dobre organizacije te brzog odaziva, također postoje i neki izazovi. Primarni problem je kontrola kvalitete. Zadaci i kompenzacija za te zadatke mogu privući varalice (koji mogu unijeti nasumične informacije), pa čak i zlonamjerne anotatore (koji mogu namjerno unositi netočne informacije). Nadalje, ne postoji kontrola nad obrazovanjem Turkera te se ne može očekivati da će prosječni Turker čitati i slijediti komplicirane i detaljne upute. No ovo ne mora nužno biti nedostatak metode „nabave iz mnoštva“. Vjeruje se da jasne, kratke i jednostavne upute rezultiraju točnim označavanjem i većim slaganjem anotatora oko oznaka za riječi. Još jedan problem je pronalazak dovoljnog broja zainteresiranih Turkera. Ako zadatak ne zahtijeva neke posebne vještine, više će Turkera moći odraditi zadatak. Broj Turkera i broj izraza koje oni označe također ovisi o tome koliko im je zadatak zanimljiv te koliko im je privlačna kompenzacija.

Izvorni i fluentni govornici jezika vrlo su dobri kod određivanja emocija koje neka riječ izaziva. Zbog toga se nije tražilo od anotatora posjedovanje nekih posebnih vještina, izuzev toga da su izvorni i fluentni govornici engleskog jezika. No označavanje emocija, pogotovo kad se koristi metoda „nabave iz mnoštva“, također se suočava s nekim izazovima. Riječi koje se koriste u različitim kontekstima izazivaju različite emocije. Recimo, glagol „vikati“ izaziva jednu vrstu emocija kada je upotrijebljen u kontekstu opomene, a drugu kada ga se koristi u kontekstu poput: „Viči ako nešto zatrebaš.“ Prikupljanje oznaka emocija za natuknice od strane anotatora je otežano jer se brinu oko toga za koje će od mogućih značenja anotirati danu riječ te koliko je to značenje prošireno. S jedne strane nije poželjno odabrati inventar brojnih i detaljnih značenja jer će broj kombinacija [riječ - značenje] biti prevelik i težak za razlikovati. S druge strane, nije poželjno ni raditi samo na razini [riječ - jedno

značenje] jer, kako je već rečeno, riječi u različitim kontekstima izazivaju različite emocije. Još jedan od izazova jest pitanje kako najbolje prenijeti značenje riječi anotatoru. Ako se koriste duge definicije, to znači da će anotatori morati provesti više vremena čitajući pitanje, a budući da je njihova plaća otprilike proporcionalna količini vremena koje provedu na zadatku, smanjuje se broj oznaka koji se može dobiti u zadanom budžetu. Nadalje, poželjno je da anotatori označavaju samo one riječi koje su im otprije poznate i znaju im značenje. Definicije su korisne kod opisa osnovnog značenja riječi, no nisu toliko učinkovite kod prenošenja suptilnih emotivnih konotacija. Iz tog se razloga od Turkera tražilo da ne označavaju riječi s kojima se nisu prije susreli. Također je bilo potrebno uvesti mjere kojima će se zanemariti zlonamjerni anotatori i varalice.

Kako bi se osigurao uspjeh ovog projekta i maksimizirala učinkovitost anotatora, prije nego što se od anotatora tražilo da odrede koje su emocije povezane s ciljanim izrazom postavljeno im je pitanje u obliku odabira riječi (engl. *word choice*). Dane su im četiri različite riječi i pitanje koja riječ ima značenje najbliže ciljanom izrazu. Tri od četiri opcije bile su samo za odvlačenje pažnje. Preostala, četvrta opcija je sinonim jednog od značenja ciljanog izraza. Ovo jednostavno pitanje imalo je više svrha. Ovim pitanjem anotatoru se dalo do znanja za koje točno značenje ciljane riječi se traže oznake, bez da ih se opterećuje dugim definicijama. Drugim riječima, točan odgovor otkriva anotatoru na koje značenje riječi se mora fokusirati prilikom označavanja. Nadalje, ako anotator nije upoznat s ciljanom riječi i svejedno pokuša odgovoriti na pitanje koje se odnosi na ciljani izraz ili nasumično odgovara na upitnik, postoji 75% šansi da će odgovoriti pogrešno te se u tom slučaju zanemaruju odgovori i oznake za emocije tog anotatora za taj ciljani izraz. Ta pitanja s odabirom riječi stvorena su automatski koristeći *Macquarie Thesaurus*. Kako je već spomenuto, objavljeni tezaurusi, poput *Rogetovog* i *Macquarie Thesaurusa*, svrstavaju sav vokabular u otprilike tisuću kategorija koje se mogu protumačiti kao grubo značenje (osnovno značenje riječi). Svaka kategorija ima natuknicu (engl. *head word*) koja najbolje opisuje značenje te kategorije. Pitanje s odabirom riječi za ciljani izraz automatski je stvoreno nakon odabira četiri alternative (izbora): natuknica kategorije u tezaurusu koja se odnosi na ciljani izraz (točan odgovor) i tri natuknice nasumično odabranih kategorija (riječi za odvraćanje pažnje). Četiri izbora prikazana su anotatoru nasumičnim poretkom.

Način na koji su sastavljena pitanja može uvelike utjecati na rezultate istraživanja. Velika se pažnja posvetila tome da pitanja budu kratka i jasna tako da različiti anotatori ne shvate krivo što se od njih traži. Kako bi se ustanovilo koji je primjereniji način izražavanja u pitanjima,

realizirana su dva odvojena procesa označavanja riječi na manjem skupu od 2100 izraza. U jednom se anotatore pitalo je li riječi „povezana“ (engl. *associated*) s određenom emocijom, a u drugom se pitalo „izaziva“ (engl. *evokes*) li riječ određenu emociju. Rezultat je pokazao da se anotatori međusobno više slažu oko oznaka ako je u pitanju glagol „povezati“. Iz tog razloga se od tog trenutka u pitanjima uvijek ispitivala povezanosti neke riječi s određenom emocijom. Primjer jednog *HITa* dostupan je u Mohammad i Turney, 2013, str. 10-11.

Prvi skup izraza (njih 2100) označen je u otprilike tjedan dana. Drugi skup od otprilike 8000 izraza (*HITova*) označen je u približno dva tjedna. Prema tome, količina vremena koja je potrebna za označavanje nije linearno proporcionalna broju *HITova*. Pretpostavlja se da s vremenom, riješenim zadacima i primljenom plaćom, sve više Turkera odlučuje nastaviti rješavati zadatke. Za svaki *HIT* određena je kompenzacija od 0,04 američka dolara (što je u Hrvatskoj otprilike 25 lipa), a Turkeri provedu u prosjeku jednu minutu odgovarajući na pitanja u jednom zadatku, što rezultira satnicom od otprilike 2,40 američka dolara (otprilike 14,90 kuna po satu).

Nakon što su zadaci riješeni i poslani na obradu, koristilo se automatsko skriptiranje kako bi se potvrdile oznake. Neki zadaci su odbačeni jer su pali na određenim testovima (opisanim u nastavku). Dio odbačenih zadataka službeno je *odbijen*, Turkeri nisu bili plaćeni za njih jer nisu slijedili upute. Otprilike 2666 od 50 850 zadataka (10 170 izraza pomnoženo s 5 pitanja) imali su barem jedno neodgovoreno pitanje – ti zadaci su također odbijeni i odbačeni. Iako su riječi koje odvlače pažnju za prvo pitanje odabrane nasumično, ponekad se dogodi da je neka od tih riječi značenjski blizu sinonimu i to rezultira pogrešnim odgovorom na prvo pitanje (odabir riječi koja je značenjem najbliža ciljanom izrazu). Za 1045 izraza, tri ili više anotatora dali su odgovor drugačiji od onog automatski generiranog iz tezaurusa. Ta pitanja su označena kao nepodobna za istraživanje i odbačena. Svi takvi zadaci, njih 5225, su odbačeni. Turkerima su ti zadaci bili plaćeni bez obzira na njihov odgovor na prvo pitanje.

Više od 95% preostalih zadataka imalo je točno odgovoreno prvo pitanje (odabir sinonima). Odbačeni su svi zadaci koji su imali zabilježen pogrešan odgovor na prvo pitanje. Ako je anotator kroz sve riješene zadatke ostvario manje od 66,67% točnih odgovora na prvo pitanje (tj. pogrešno je odgovorio na više od jednog od tri takva pitanja), pretpostavilo se da je, usprkos uputama, anotator pokušao odgovarati na *HITove* za riječi koje mu nisu otprije poznate. Odbačeni su i odbijeni svi zadaci takvih anotatora (ne samo oni zadaci za koje su pogrešno odgovorili na prvo pitanje).

Za svakog anotatora izračunata je maksimalna vjerojatnost njegovog slaganja s većinom na pitanja o emocijama. Izračunata je srednja vrijednost tih vjerojatnosti i standardno odstupanje (devijacija). Odbačene su oznake Turkera koji su bili više od dvije standardne devijacije od srednje vrijednosti, njih 111. Nakon ovog procesa, preostalo je 8883 od početnih 10 170 izraza, svaki sa tri ili više odobrena zadatka. Ova grupa izraza nazvana je glavni skup (engl. *master set*). Stvoren je leksikon s popisom riječi iz glavnog skupa uz koje su označene emocije povezane s tim riječima koji se sastoji od 38 726 zadataka od otprilike 2216 Turkera koji su se okušali u do 2000 zadataka svaki. Otprilike 300 njih je riješilo 20 ili više zadataka (više od 33 000 zadataka ukupno). Glavni skup u prosjeku ima otprilike 4,4 zadatka za svaki od 8883 ciljana izraza. Ukupni trošak označavanja bio je otprilike 2100 američkih dolara (otprilike 13 023 kune). Taj iznos uključuje troškove Amazona (cca. 13% svote koja je plaćena Turkerima) kao i trošak dvostrukog označavanja prvog skupa glagolima „izazvati“ i „povezati“ (Mohammad i Turney, 2013, str 7-12).

Riječi u leksikonu kasnije su prevedene na više od stotinu jezika pomoću Google Prevoditelja kako bi Emolex postao resurs koji se može koristiti i u drugim jezicima kao rječnik koji povezuje riječi s ljudskim reakcijama i osjećajima koje te riječi izazivaju kod ljudi. Svaka od tih riječi jedan je redak u Excel tablici, dok su stupci označeni od „A“ do „DK“ rezervirani za jezike, sentimente i emocije. Prvi stupac sadrži riječi na engleskom jeziku te se dalje nastavlja niz ostalih jezika po abecednom redu – prvi jezik nakon engleskog je Afrikaans u stupcu „B“, hrvatski jezik je u stupcu „R“, a posljednji jezik je Zulu u stupcu „DA“. Nakon prijevoda riječi na druge jezike slijede sentimenti i emocije. Prvi po redu su stupci za pozitivno i negativno (stupci „DB“ i „DC“), a nakon njih slijedi posljednjih osam stupaca s osam emocija (identičnim redoslijedom kojim su prije nabrojane). Posljednjih deset stupaca je vrlo lako uočiti jer sve ćelije ispod njih sadrže brojke 0 ili 1. Sažetak detalja o NRC Leksikonu emocija također je prikazan na Slici 3., a izgled samog Leksikona na Slici 4.

3.1. EmoLex za hrvatski jezik

Primarni fokus ovog rada su stupci „A“ (koji sadrži engleske riječi), stupac „R“ (koji sadrži hrvatski prijevod engleskih riječi generiran pomoću Google Prevoditelja) te posljednjih deset stupaca koji sadrže emotivne asocijacije s navedenim riječima. Iako Google Prevoditelj postaje sve bolji alat za prevođenje svjetskih jezika s mnogo govornika, još uvijek nije dosegao razinu kvalitete prevođenja koja je potrebna da bi se taj prijevod mogao koristiti u

široj upotrebi bez prethodne ljudske intervencije. Primjerice, Google Prevoditelj često prevodi doslovno i ne dopušta višeznačnost riječi, a vrlo često su prijevodi i pogrešni. Takve greške se događaju i prilikom prevođenja svjetskih jezika, stoga je bilo za očekivati da će prijevodi na hrvatski jezik biti daleko od savršenih. U sljedećem poglavlju bit će detaljno opisan način ispravljanja Googleovih prijevoda te (u rijetkim slučajevima) sentimenata i emocija zbog razlike u značenju određene riječi u engleskom i hrvatskom. Također će se opisati primijećeni obrasci i načini na koje se Google Prevoditelj snalazi kad ne može samostalno ponuditi prijevod neke riječi.

Association Lexicon	Version	# of Terms	Categories	Association Scores	Method of Creation	Papers
Word-Emotion and Word-Sentiment Association Lexicon						
NRC Word-Emotion Association Lexicon (also called EmoLex) README	0.92 (2010)	14,182 unigrams (words) ~25,000 senses*	**sentiments:** negative, positive **emotions:** anger, anticipation, disgust, fear, joy, sadness, surprise, trust	0 (not associated) or 1 (associated) not associated, weakly, moderately, or strongly associated	**Manual:** By crowdsourcing on Mechanical Turk. **Domain:** General	**Crowdsourcing a Word-Emotion Association Lexicon**, Saif Mohammad and Peter Turney, *Computational Intelligence*, 29 (3), 436-465, 2013. Paper (pdf) BibTeX **Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon**, Saif Mohammad and Peter Turney, In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, June 2010, LA, California. Abstract Paper (pdf) Presentation

Slika 3. Sažetak detalja o ERC Leksikonu emocija, verzija 0.92 (saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm)

A1 | English (en)

	A	R	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK
1	English (en)	Croatian (hr)	Positive	Negative	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
2	aback	po krmi	0	0	0	0	0	0	0	0	0	0
3	abacus	računaljka	0	0	0	0	0	0	0	0	0	1
4	abandon	napustiti	0	1	0	0	0	1	0	1	0	0
5	abandoned	napušten	0	1	1	0	0	1	0	1	0	0
6	abandonment	napuštanje	0	1	1	0	0	1	0	1	1	0
7	abate	smanjiti se	0	0	0	0	0	0	0	0	0	0
8	abatement	smanjenje	0	0	0	0	0	0	0	0	0	0
9	abba	Abba	1	0	0	0	0	0	0	0	0	0
10	abbot	opat	0	0	0	0	0	0	0	0	0	1
11	abbreviate	skratiti	0	0	0	0	0	0	0	0	0	0
12	abbreviation	skraćenica	0	0	0	0	0	0	0	0	0	0
13	abdomen	trbuh	0	0	0	0	0	0	0	0	0	0
14	abdominal	trbušni	0	0	0	0	0	0	0	0	0	0
15	abduction	otmica	0	1	0	0	0	1	0	1	1	0
16	aberrant	nenormalan	0	1	0	0	0	0	0	0	0	0
17	aberration	aberacija	0	1	0	0	1	0	0	0	0	0
18	abeyance	stanje neizvje	0	0	0	0	0	0	0	0	0	0
19	abhor	mrziti	0	1	1	0	1	1	0	0	0	0
20	abhorrent	odvratan	0	1	1	0	1	1	0	0	0	0
21	abide	prebivati	0	0	0	0	0	0	0	0	0	0
22	ability	sposobnost	1	0	0	0	0	0	0	0	0	0

NRC-Lex-v0.92-word-translations

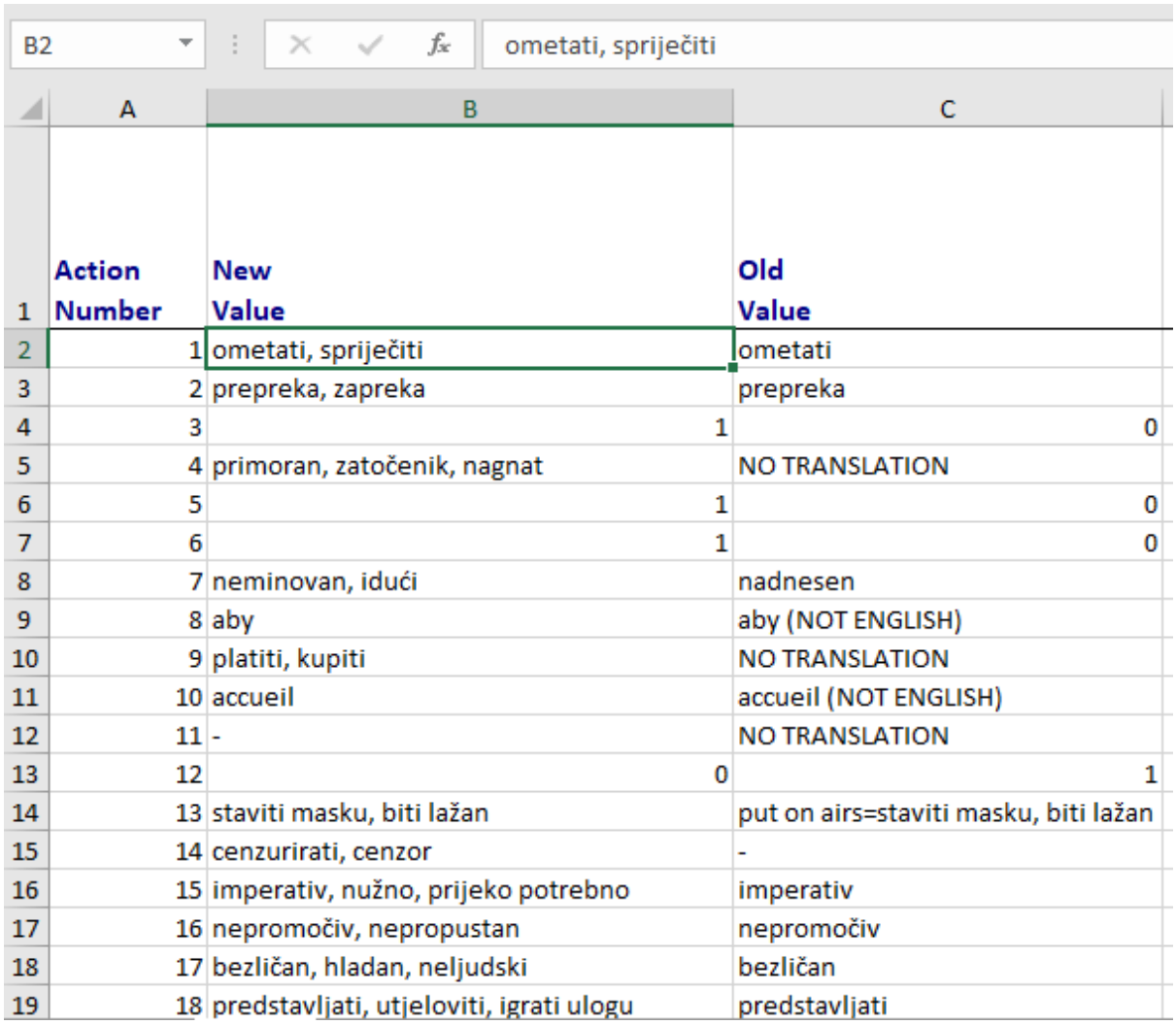
Slika 4. Izgled NRC Leksikona emocija, verzija 0.92 (prije ispravljanja)

4. Opis metode ispravljanja

Ispravljanje prijevoda engleskih riječi na hrvatski jezik u tablici EmoLexa zahtijevalo je mnogo vremena (nešto više od pet mjeseci) i pomoćnih resursa za prevođenje te engleskih i hrvatskih rječnika. Kako bi se osigurao što potpuniji i točniji prijevod, gotovo sve su se riječi upisivale u najmanje jedan, a najčešće u svaki od sljedećih resursa:

1. hrWaC – Hrvatski mrežni korpus koji je prikupljen s .hr internetske domene (reldi.spur.uzh.ch/hr-sr/hrvatski-mrezni-korpus/). Ovaj resurs osigurao je detektiranje prevedenih riječi koje su pomalo zastarjele, nisu više u uporabi ili nisu uopće dio hrvatskog jezika. Nakon što se pronašla adekvatna zamjena za tu riječ (najčešće sinonim), hrWaC je omogućio provjeru i potvrdu da je novi prijevod najčešće korištena riječ u hrvatskom jeziku od svih ostalih predloženih prijevoda.
2. Glosbe (glosbe.com/en/hr) – besplatni online rječnik koji koristi dostupne korpuse i rječnike te daje primjere riječi u rečenici, sve na jednom mjestu. Vrlo koristan resurs za hrvatsko-engleski i englesko-hrvatski prijevod, no preporučljivo ga je koristiti u kombinaciji s drugim rječnicima jer često prikazuje i pogrešne prijevode koje je pronašao u dostupnim korpusima. Također je izuzetno koristan jer vrlo često pruža sva moguća značenja neke riječi na engleskom koja često uključuju i različite vrste riječi (na primjer, engleske riječi koje su zadane u obliku koji se u engleskom naziva *Past Participle* vrlo često imaju isti oblik za pridjev u engleskom, dakle prevode se u glagolskom i pridjevnom obliku na hrvatski jezik).
3. *Merriam-Webster.com* (www.merriam-webster.com) – službena web-stranica jednog od najpoznatijih englesko-engleskih rječnika: Merriam-Webstera. Autorica je odabrala ovaj rječnik zbog pozitivnog prethodnog iskustva, no mnogi se drugi rječnici mogu koristiti umjesto ovog, poput na primjer *Cambridge Dictionary* (dictionary.cambridge.org) ili *Oxford Advanced Learner's Dictionary* (www.oxfordlearnersdictionaries.com). U englesko-engleski rječnik upisivale su se riječi kojima je trebalo provjeriti primarno značenje kao i sva druga moguća značenja. Na taj se način lakše moglo prevesti potpuno značenje jedne engleske riječi na hrvatski, što je najčešće rezultiralo time da je jedna engleska riječ bila prevedena s više hrvatskih riječi i/ili izraza.
4. HJP – Hrvatski jezični portal (hjp.znanje.hr), hrvatska rječnička baza koja pruža uvid u hrvatske riječi te sve njihove oblike i definicije. Ovaj resurs je omogućio provjeru hrvatskih riječi te detekciju posuđenica iz srpskog i bosanskog jezika koje su se tada zamijenile hrvatskom inačicom.

Ovo poglavlje opisat će nekoliko postupaka ispravljanja tablice na primjerima te pružiti detaljnu statistiku ispravaka u **podkorpusu** koji sadrži **7878** riječi. Ove promjene također su evidentirane u ispravljenoj NRC Excel tablici te izvađene u posebnu tablicu pod nazivom „Pogreške“ koja je vidljiva na Slici 5.



B2 | ometati, spriječiti

	A	B	C
1	Action Number	New Value	Old Value
2	1	ometati, spriječiti	ometati
3	2	prepreka, zapreka	prepreka
4	3	1	0
5	4	primoran, zatočenik, nagnat	NO TRANSLATION
6	5	1	0
7	6	1	0
8	7	neminovan, idući	nadnesen
9	8	aby	aby (NOT ENGLISH)
10	9	platiti, kupiti	NO TRANSLATION
11	10	accueil	accueil (NOT ENGLISH)
12	11	-	NO TRANSLATION
13	12	0	1
14	13	staviti masku, biti lažan	put on airs=staviti masku, biti lažan
15	14	cenzurirati, cenzor	-
16	15	imperativ, nužno, prijeko potrebno	imperativ
17	16	nepromočiv, nepropustan	nepromočiv
18	17	bezličan, hladan, neljudski	bezličan
19	18	predstavljati, utjeloviti, igrati ulogu	predstavljati

Slika 5. Izgled tablice “Pogreške” u Excelu

Prije svega valja opisati općeniti postupak kojim su se provjeravali prijevodi riječi. Najčešće je postupak uključivao sljedeće korake: prvo su se engleske riječi utipkavale u Glosbe da se dobije najširi popis mogućih prijevoda, zatim su se prijevodi filtrirali u Merriam-Websteru i HJP-u kako bi se provjerilo koji prijevodi su točni, koja je ispravna hrvatska inačica (ili više njih) te njihov natuknički oblik. Na kraju su prijevodi prošli kroz hrWaC kako bi se provjerila učestalost pojedinih prijevoda. Ako je za jednu englesku riječ bilo potrebno navesti više od

jedne riječi kao prijevod na hrvatski jezik, hrWaC je također bio korišten kako bi se prijevodi naveli na način da je najčešće korištena riječ prva, a nakon nje su navedeni ostali prijevodi.

4.1. Primjeri ispravljanja

U nastavku je navedeno nekoliko konkretnih primjera postupka ispravljanja prijevoda u NRC Leksikonu emocija. Ovi primjeri ne uključuju sve odrađene ispravke, već daju općeniti prikaz vrsta pogrešaka te postupak ispravljanja istih.

1. **Primjer**: promjena vrijednosti sentimenta i emocija

Ispravaka sentimenta i emocija riječi u cijelom leksikonu bilo je svega sedam. Općenito, nije bilo potrebno mijenjati sentimente i emocije jer se pokazalo da su oni gotovo jednaki, ako ne i identični, za većinu jezika, što je autor EmoLexa Saif Mohammad i dokazao prilikom njegove izrade (saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm). Jedna od riječi kojima je bilo potrebno promijeniti sentiment je riječ *impediment* koja se na hrvatski prevodi kao „prepreka“ ili „zapreka“. Primarno su oba sentimenta bila označena s 0, što bi značilo da riječi nije ni pozitivna niti negativna, a to nije točno, stoga je sentiment za negativno iz 0 promijenjen u 1 jer riječi „prepreka“ i „zapreka“ u hrvatskom jeziku imaju negativno značenje.

2. **Primjer**: dodavanje još jednog prijevoda

Kao primjer ove vrste ispravka uzet ćemo riječ *mug*. Engleska riječ *mug* prvo se upisala u Glosbe koji je ponudio mnoštvo rezultata koje su zapravo činile rečenice koje su u sebi imale neki oblik riječi „šalica“ ili „opljačkati“ iz čega zaključujemo da su to dva moguća prijevoda te riječi, iako je prethodno u tablici kao prijevod na hrvatski pisalo samo „šalica“. To ne mora nužno značiti da Google Prevoditelj nije ponudio oba prijevoda, već je moguće da su autori leksikona odredili da se u NRC Leksikon upisuje samo prvi ponuđeni prijevod. Dokaz tome je činjenica da se u originalnoj tablici (koja nije ispravljena za hrvatski jezik) u stupcu „R“ uvijek pojavljuje samo jedna riječ kao prijevod. Sada je to ispravljeno, navedeni su svi mogući prijevodi svih engleskih riječi, a za riječ *mug* kao prijevod sada u ćeliji za hrvatski jezik piše „šalica, opljačkati“. Sentimente i emocije nije bilo potrebno mijenjati.

3. **Primjer**: pravopisna pogreška - veliko slovo

U pravilu su sve riječi u cijeloj tablici kroz sve jezike pisane malim tiskanim slovima, izuzev osobnih imena te naziva mjesta, naroda i slično, koji se po pravopisu i trebaju pisati velikim

početnim slovom. Iz tog su se razloga sve riječi koje ne spadaju u tu kategoriju, a bile su napisane velikim početnim slovom izmijenile tako da su u cijelosti napisane malim tiskanim slovima. Zanimljivo je da su neki „prijevodi“ koji su napisani velikim početnim slovom zapravo samo engleske riječi prepisane u ćeliju za hrvatski prijevod. Nepoznato je zašto te riječi počinju velikim slovom no pretpostavka je da se Google Prevoditelj poslužio dostupnim korpusima i kroz tekstove tražio te riječi nakon čega ih preuzeo kao prijevod s početka rečenice gdje su bile napisane velikim slovom. Neke od tih riječi su: „Ublažavanje“, „Otok“, „Rešetka“, „Litosfera“. Tri riječi su u originalnoj tablici bile u potpunosti napisane velikim slovima, te također pretpostavljamo da su tako pronađene online: „PROTUUPRAVNI“, „PRAVI“ i „XEROX“.

4. **Primjer**: pravopisna pogreška

Riječ *wise* je u originalnoj tablici bila prevedena kao „Mudr“. Prilično je jasno da je točan prijevod „mudar“, no kako je Google Prevoditelj pronašao ovaj oblik riječi koji ne samo da je pravopisno netočan, već i počinje velikim početnim slovom, iako su sve slične riječi pisane malim slovima, nije jasno. Pretpostavka je da se Google ponovo poslužio dostupnim korpusima i pomoću konteksta riječ „Mudr“ prepoznao kao prijevod riječi *wise*.

5. **Primjer**: morfosintaktička pogreška

Morfosintaktičke pogreške uključivale su riječi kojima se ispravljao rod, broj, padež, oblik i vrsta riječi ili samo jedno od navedenog. Cilj je bio da sve riječi budu navedene u natukničkom obliku, kao u rječniku hrvatskog jezika – u infinitivu za glagole te u nominativu jednine muškog roda za pridjeve i nominativu jednine za imenice. Nekoliko primjera riječi koje su vrlo vjerojatno preuzete iz nekog teksta u kojem su bile napisane u nekom drugom obliku glase: „ugrađuju“, „onesposobljavaju“, „neisplativim“, „uputio“, „raspravlja“, „jastučnicu“, „orali“, „griješi“, „šapnula“. Sve riječi su po potrebi izmijenjene u točan prijevod i oblik.

6. **Primjer**: posuđenica iz srpskog

Popriličan broj Googleovih prijevoda je je bio točan, ali ne za hrvatski, već za srpski jezik. Jedan od primjera je riječ *insular* koja je prevedena kao „ostrvski“. Ostrvo je srpsko – bosanska riječ koja znači „otok“, stoga je ćelija izmijenjena te u njoj sada stoji da *insular* na hrvatskom znači „otok“. Također je dodan prijevod „izdvojen“ jer je nakon konzultacije s Merriam-Websterom postalo jasno da se riječ može koristiti u oba značenja.

7. **Primjer**: netočan prijevod

Znatan broj prijevoda bio je preveden djelomično točno ili potpuno pogrešno. Vrijedi ista pretpostavka da je Google Prevoditelj pretraživanjem raznih korpusa pronašao pogrešno označenu riječ ili pogrešan prijevod i preuzeo isti. Dobar primjer za donekle točan prijevod je riječ *misconduct* koja je prevedena kao „nesportsko ponašanje“. Ova se riječ u engleskom koristi u navedenom značenju u sportskom žargonu, no u Merriam-Websteru, Oxfordovom i Cambrigdeovom rječniku to je značenje navedeno kao posljednja, dakle, najrjeđa upotreba. Prema tome, ćelija je promijenjena u prijevod koji su rječnici naveli prvo - „nedolično ponašanje“, izraz koji obuhvaća najšire značenje riječi *misconduct*. Novi prijevod također podrazumijeva takvu vrstu vladanja u sportu, dakle nedolično ponašanje podrazumijeva nesportsko ponašanje, stoga se novi prijevod smatra primjerenijim.

Primjer potpuno pogrešno prevedene riječi možemo pronaći ako u originalnoj tablici pogledamo kako je prevedena riječ *sundry* koja u engleskom jeziku predstavlja svojstvo raznovrsnosti i različitosti. Kao prijevod Google navodi frazu „sušen na suncu“, što nema nikakve veze sa stvarnim značenjem te riječi, no, vrlo je moguće da automatski prevoditelj nije mogao pronaći točan prijevod pa je riječ *sundry* rastavio na dvije riječi i dobio sintagmu *sun dry* koja se vrlo lako može prevesti kao „sušen na suncu“. Ćelija je izmijenjena u „raznovrstan, različit“, a budući da su govornici engleskog jezika koji su označavali sentimente i emocije znali pravo značenje ove riječi, njih nije bilo potrebno mijenjati.

8. **Primjer**: preuzeta engleska riječ kao hrvatski prijevod

Iako bi stupac „R“ trebao sadržavati hrvatske riječi koje značenjem odgovaraju engleskim riječima, alat za prevođenje je određeni broj riječi propustio prevesti, a neke od njih također počinju velikim početnim slovom, iako u stupcu „A“ koji je rezerviran za engleske riječi one pišu malim početnim slovom. Zašto ih Google Prevoditelj nije preveo te iz kojeg razloga im je stavljeno veliko početno slovo nije jasno, no u ovim slučajevima je vidljivo da se Google nije koristio korpusima i drugim rječnicima poput Glosbea jer oni te riječi prevode, i to točno. Neki od primjera glase: „Inquierer“, „tinkering“, „overzealous“, „moulded“, „Surround“. Sve riječi su prevedene na hrvatski te započinju malim slovom kako bi bile u skladu s ostatkom tablice. Sentimente i emocije nije bilo potrebno mijenjati.

9. **Primjer**: nema prijevoda

Za određen broj riječi u tablici EmoLexa za hrvatski je prijevod stajala sintagma „NO TRANSLATION“ što na engleskom jeziku znači da nema prijevoda ili nije dostupan za te riječi. Činjenica da Google Prevoditelj nije pronašao nijednu riječ koja bi mogla proći kao

prijevod, pa makar ta riječ uopće značenjski ne odgovara zadanoj riječi na engleskom, zaista je neobično. Iz prethodnih primjera je jasno da su se kao prijevod za hrvatski jezik navodile razne riječi i izrazi koje izvorni govornici ne bi odobrili kao točne, stoga nije jasno zašto za riječi poput *elan, elapse, ergo, especial, loath, lockup, loin, simile* i ostale prijevod uopće nije ponuđen. Odgovori na ova i slična pitanja zasigurno postoje unutar algoritma Google Prevoditelja, no ovaj rad se ne bavi analizom ovog prevodilačkog alata, stoga se njima ovdje nećemo baviti.

10. primjer: riječ se ne može prevesti na hrvatski standardni jezik

Dvije od ukupno 14 182 riječi u tablici nije moguće prevesti na hrvatski jezik jer se njihovo značenje ne može pronaći ni u jednom englesko-engleskom ili englesko-hrvatskom rječniku. Riječi *accueil* i *roc* nemaju određeno zapisano i definirano značenje u engleskom jeziku, stoga se ne mogu prevesti na hrvatski jezik. Iz tog su razloga u ćeliji za hrvatski prijevod označene minusom („-“) Pretpostavlja se da dolaze iz nekog drugog jezika ili su se slučajno našle na popisu riječi ovog leksikona, no ni za jednu riječ nije označen sentiment ili bilo koja emocija, stoga je vrlo vjerojatno da ni ispitanici nisu znali njihovo značenje.

Ovi primjeri pokazuju samo mali dio riječi koje su ispravljene, a u nastavku će biti predstavljena detaljna i cjelovita statistika ovog projekta.

4.2. Statistika pogrešaka

Od ukupno 14 182 unosa, odabran je podkorpus od 6304. riječi do 14 182. riječi što nam kao rezultat daje **podkorpus** koji se sastoji od **7878** riječi. U tom popisu napravljeno je 3011 izmjena, drugim riječima, ispravljeno je **3011 pogrešaka** te se nad njima vršila analiza i kategorizacija ispravljenih prijevoda. Izmjene napravljene u tablici su evidentirane i izvađene u posebnu Excel tablicu pod nazivom „Pogreške“ te su se riječi ispisivale u kategorije prema svom rednom broju u tablici „Pogreške“ ili u obliku [redni broj (riječ)] kako bi se lakše pronašli primjeri za pojedinu kategoriju. Potpuni popis pogrešaka dostupan je pod naslovom „Prilozi“. Nakon što je svih 3011 izmjena svrstano u kategorije, od kojih neke spadaju u više kategorija jer sadrže više vrsta pogrešaka, sve izmjene su se pobrojile kako bi se dobila detaljna statistika koja opisuje udio određene vrste pogrešaka u ukupnom broju pogrešaka kao što je prikazano u Grafikonu 1.

Brojčana statistika ispravaka ovog podkorpusa od 7878 riječi unutar kojeg je izmijenjeno 3011 ćelija svrstana je u sljedeće kategorije te glasi:

a) „NO TRANSLATION“: prijevod nije pronađen za ukupno **147** riječi, što čini **4.9 %** ukupnog broja izmjena.

b) PROMJENA VRIJEDNOSTI SENTIMENTA I EMOCIJA: ukupno se izmijenilo **7** sentimenta i emocija što čini **0.2%** ukupnog broja izmjena. Od tih sedam izmjena, promjene su bile sljedeće:

- Iz 0 (ne odnosi se na ovu riječ) u 1 (odnosi se na ovu riječ) = **6** ćelija
- Iz 1 u 0 = **1** ćelija

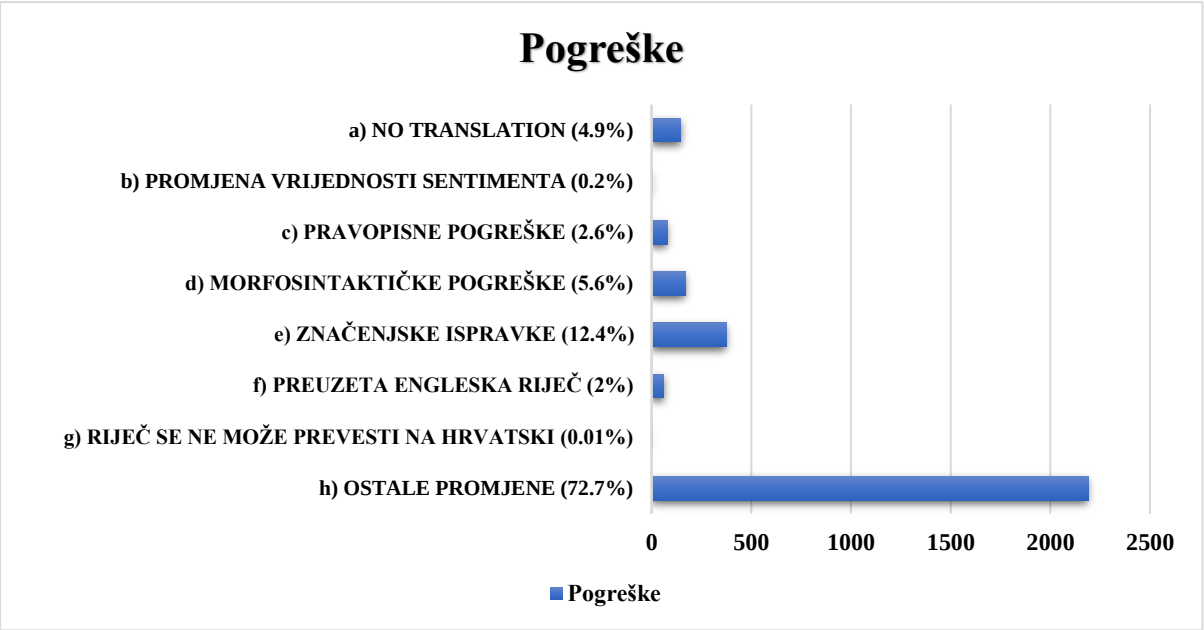
c) PRAVOPISNE POGREŠKE: ukupan broj pravopisnih pogrešaka u ovom podkorpusu iznosi **78** što čini **2.6%** ukupnog broja izmjena. U ovu kategoriju osim riječi s pravopisnim pogreškama spadaju i riječi koje započinju velikim početnim slovom, a nisu vlastito ime, geografska lokacija ili slično, drugim riječima, sve riječi koje nemaju razloga imati početno slovo veliko te riječi koje su cijele napisane velikim tiskanim slovima.

d) MORFOSINTAKTIČKE POGREŠKE: ukupan broj iznosi **170** što čini **5.6%** ukupnog broja pogrešaka. Ova kategorija uključuje riječi koje su u tablici bile navedene u obliku koji nije prikladan rječniku, dakle riječi koje su pisale u rodu, broju i padežu te određenom glagolskom obliku koji nije bio jednina, muški rod, nominativ za pridjeve i imenice ili infinitiv ili glagolska imenica za glagole.

e) ZNAČENJSKE ISPRAVKE: ukupno ih je **374** što iznosi **12.4%** ukupnog broja pogrešaka. Ova kategorija podrazumijeva sve netočne prijevode, kao i prijevode čije je značenje bilo preusko ili preširoko da bi kvalitetno opisivalo značenje engleske riječi. Ovdje također spadaju i srbizmi te ostale posuđenice koje nisu dio hrvatskog standardnog jezika.

f) PREUZETA ENGLESKA RIJEČ KAO HRVATSKI PRIJEVOD: ukupno je **60** engleskih riječi bilo ponuđeno kao prijevod na hrvatski jezik što čini **2%** ukupnog broja pogrešaka.

g) RIJEČ SE NE MOŽE PREVESTI NA HRVATSKI JEZIK: samo **2** (već spomenute) riječi nisu se mogle prevesti na hrvatski jezik te one čine **0.01%** pogrešaka.



Grafikon 1. Vrste pogrešaka i njihov udio u ukupnom broju pogrešaka
Napomena: neke riječi su evidentirane u više kategorija

h) SVE OSTALE PROMJENE: u ovu kategoriju spadaju sve izmjene koje uključuju dodatak još jednog prijevoda ili ispravak prijevoda iz dijalekta u standardni hrvatski jezik. Takvih je promjena bilo **2189** što ujedno čini i najveći udio u ukupnom broju izmjena koji iznosi **72,7%**. Nad nekim riječima vršile su se izmjene koje ulaze u više kategorija, stoga je potrebno napomenuti da su u statistici neke riječi navedene više puta. Na primjer, u kategoriji engleskih riječi koje su preuzete kao hrvatski prijevod, mnogo je riječi pisalo velikim početnim slovom ili velikim tiskanim slovima u cijelosti, stoga spadaju i u kategoriju pravopisnih pogrešaka.

4.3. Najzanimljivije pogreške

Ovaj dio rada predstavit će nekoliko pogrešnih prijevoda koji su odskakali od ostatka pogrešaka zbog svog sadržaja koji ih čini zanimljivima, neobičnima ili čak komičnima. Riječi i prijevodi su ispisani u obliku [*engleska riječ* (točan prijevod) – „netočan prijevod“]:

- *aback* (iznenađen, unatrag) – „po krmi“
- *accomplishment* (postignuće) – „svršavanje“
- *uncooked* (sirov, nekuhan) – „prijesan“
- *unimpressed* (indiferentan, neoduševljen) – „bez otiska“
- *unknowable* (koji se ne može spoznati) – „nepristupačan“

- *machinist* (mehaničar) – „mahaničar“
- *oboe* (oboa) – „oboja“
- *raptors* (grabežljivci) – „Raptorsa“
- *ted* (rasprostirati pokošenu travu radi sušenja) – „rasprostirati pokošenu travu radi sušenje“
- *lesbianism* (lezbijanizam) – „lezbijska ljubav“
- *lordship* (gospodstvo) – „gospodarstvo“
- *mismanagement* (loše upravljanje, nebriga) – „hrđavo upravljanje“
- *prong* (zubac, šiljak, vrh) – „krak“
- *relics* (relikvije) – „mošti“
- *royalty* (kraljevska obitelj) – „kraljevstvo“
- *slough* (bara) – „zmijska košuljica“
- *straits* (vrata, tjesnac) – „moreuz“
- *stud* (pastuh) – „klinac“
- *virology* (virologija) – „virusologija“

5. Rezultat

Nakon pet mjeseci ispravljanja prijevoda, sentimenata i emocija *NRC Word-Emotion Association Lexicona*, poznatog i pod nazivom EmoLex, resurs je za hrvatski jezik sada u potpunosti točan te ne nudi samo jedan prijevod već potpuno prevodi značenje engleskih riječi na hrvatski standardni jezik. EmoLex se sada može koristiti i kao leksikon emocija hrvatskog jezika te ući u opću upotrebu svih koji mu mogu i žele naći primjenu. Prema molbi autora EmoLexa, ispravljena verzija tablice bit će im poslana kako bi upotpunili originalni EmoLex za hrvatski jezik. Rezultat ovog rada je također i detaljan pregled englesko-hrvatskih prijevoda generiranih pomoću Google Prevoditelja te se može koristiti kako bi se poboljšali prevodilački alati za hrvatski jezik. Izgled Excel tablice EmoLexa za engleski i hrvatski jezik nakon ispravljanja dostupan je na Slici 6.

A13254 | uncooked

	A	R	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK
1	Engeski (en)	Hrvatski (hr)	Pozitivno	Negativno	Ljutnja	Iščekivanje	Gađenje	Strah	Radost	Tuga	Iznenađenje	Povjerenje
2	aback	iznenađen, unatrag	0	1	0	1	0	1	0	0	1	0
3	abacus	računaljka	0	0	0	0	0	0	0	0	0	1
4	abandon	napustiti	0	1	0	0	0	1	0	1	1	0
5	abandoned	napušten	0	1	1	0	0	1	0	1	1	0
6	abandonment	napuštanje	0	1	1	0	0	1	0	1	1	0
7	abate	smanjiti se	0	0	0	0	0	0	0	0	0	0
8	abatement	smanjenje	0	0	0	0	0	0	0	0	0	0
9	abba	Abba	1	0	0	0	0	0	0	0	0	0
10	abbot	opat	0	0	0	0	0	0	0	0	0	1
11	abbreviate	skratiti	0	0	0	0	0	0	0	0	0	0
12	abbreviation	skraćenica	0	0	0	0	0	0	0	0	0	0
13	abdomen	trbuh	0	0	0	0	0	0	0	0	0	0
14	abdominal	trbušni	0	0	0	0	0	0	0	0	0	0
15	abduction	otmica	0	1	0	0	0	1	0	1	1	0
16	aberrant	nenormalan	0	1	0	0	0	0	0	0	0	0
17	aberration	aberacija	0	1	0	0	1	0	0	0	0	0
18	abeyance	stanje neizvjesnosti	0	0	0	0	0	0	0	0	0	0
19	abhor	mrziti	0	1	1	0	1	1	0	0	0	0
20	abhorrent	odvratan	0	1	1	0	1	1	0	0	0	0
21	abide	prebivati	0	0	0	0	0	0	0	0	0	0
22	ability	sposobnost	1	0	0	0	0	0	0	0	0	0

Slika 6. Izgled EmoLexa nakon ispravljanja

6. Zaključak

Zadatak ovog rada bio je korekcija hrvatskog dijela EmoLex leksikona, tj. prevođenje i ispravljanje rječnika koji broji više od 14 000 natuknica te provjera emocija i sentimenata vezanih uz te rječničke natuknice. Taj zadatak je odrađen, rječnik je ispravljen te je prikazana detaljna analiza pogrešaka podkorpusa koji obuhvaća više od polovine unosa. Sve pogreške su evidentirane, kategorizirane i oprimjerene te je objašnjen postupak ispravljanja i navedeni alati koji su bili korišteni tijekom postupka ispravljanja. Ovaj rad rezultira popisom hrvatskih riječi uz koje su vezani sentimenti i emocije te se taj rječnik može koristiti kao Leksikon emocija hrvatskog jezika.

Iako je EmoLex ispravljen, to ne znači da je posao gotov. Ovaj projekt ima mnogo potencijala za daljnji rad i napredak na način da se riječi, sentimenti i emocije izdvoje iz tablice EmoLexa te da se stvori zaseban Leksikon emocija za hrvatski jezik koji bi kao bazu imao više od 14 000 natuknica preuzetih iz EmoLexa. Taj Leksikon bi se tada mogao razvijati i povećavati neovisno o EmoLexu te bi se mogle dodavati nove riječi s obilježenim emocijama prikupljene od izvornih govornika hrvatskoga jezika. Takav proširen i upotpunjen Leksikon bi bio konačan rezultat ovog rada koji je zapravo samo početni korak ka stvaranju potpuno neovisnog Leksikona emocija hrvatskog jezika koji bi se mogao koristiti kao temelj za analizu anketa i istraživanja koja uključuju pismeni odgovor korisnika, u obradi prirodnog jezika, za izradu algoritama koji će prepoznavati ljudske emocije, izradu pametnih sustava za učenje, u području računalne lingvistike i strojnog učenja te kod razvijanja robota i sustava pretvorbe teksta u govor.

7. Literatura

1. Adamic, L. A., Zhang, J., Bakshy, E., i Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. *Proceeding of the 17th international conference on World Wide Web*, WWW '08, str. 665-674, New York, NY, SAD. ACM.
2. Ahire, S. (2015). A Survey of Sentiment Lexicons, *Computer Science and Engineering*, IIT Bombay.
3. Akkaya, C., Mihalcea, R. i Wiebe, J. (2009). Subjectivity Word Sense Disambiguation. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, str. 190–199, Singapur.
4. Bellegarda, J. (2010). Emotion analysis using latent affective folding and embedding. *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, Kalifornija.
5. Bernard, J., urednik (1986). The Macquarie Thesaurus. Knjižnica Macquarie, Sydney, Australia.
6. Boucouvalas, A. C. (2002). Real time text-to-emotion engine for expressive internet communication. *Emerging Communication: Studies on New Technologies and Practices in Communication*, 5, str. 305-318.
7. Bougie, J. R. G., Pieters, R., i Zeelenberg, M. (2003). *Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services.* Sveučilište u Tilburgu.
8. Bradley, M. i Lang, P. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. *Technical Report, C-1*, Centar za psihofiziološka istraživanja, Sveučilište u Floridi.
9. Brants, T. i Franz, A. (2006). Web 1t 5-gram version 1. *Linguistic Data Consortium*.
10. Breazeal, C. i Brooks, R. (2004). Robot emotions: A functional perspective. *Who Needs Emotions*. Oxford University Press.
11. Breck, E., Choi, Y., i Cardie, C. (2007). Identifying Expressions of Opinion in Context. *Proc. 20th Int'l Joint Conf.* Artificial Intelligence.
12. Cambridge Dictionary (2016). [online] Dictionary.cambridge.org. Dostupno na: dictionary.cambridge.org [pristupljeno 2.2.2020. – 15.6.2020.].
13. Ciampaglia, G., Mashhadi, A., i Yasseri, T. (2017). *Social Informatics: 9th International Conference*, SocInfo, Oxford, UK.

14. Cohn, M.A., Mehl, M.R. i Pennebaker, J.W. (2004). Linguistic Markers of Psychological Change Surrounding September 11, 2001. *Psychological Science, vol. 15*, str. 687-693.
15. D'Mello, S. i Graesser, A. (2010). Multimodal Semi- Automated Affect Detection from Conversational Cues, Gross Body Language, and Facial Features. *User Modeling and User-Adapted Interaction, vol. 10*, str. 147-187.
16. Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3), str. 169-200.
17. El Gohary, F.A., Sultan, T.I., Hana, M.A., El Dosoky, M.M. (2013). A Computational Approach for Analyzing and Detecting Emotions in Arabic Text. *International Journal of Engineering Research and Applications (IJERA)*, 3(3), str. 100-107.
18. Francisco, V. i Gervás, P. (2006). Automated mark up of affective information in English texts. P. Sojka, I. Kopecek, i K. Pala, urednici, *Text, Speech and Dialogue*, vol. 4188 *Lecture Notes in Computer Science*, str. 375-382. Springer Berlin / Heidelberg.
19. Gill, R., French, D., Gergle, i Oberlander, J. (2008). Identifying Emotional Characteristics from Short Blog Texts. *Proc. 30th Ann. Conf. Cognitive Science Soc*., izdavači Love, B.C., McRae, K. i Sloutsky, V.M., str. 2237-2242.
20. Glosbe, englesko-hrvatski rječnik (2018). [online] glosbe.com/en/hr. Dostupno na: glosbe.com/en/hr [pristupljeno 2.2.2020. – 15.6.2020.].
21. Hancock, J., Landrigan, C., i Silver, C. (2007). Expressing Emotion inText-Based Communication. *Proc. SIGCHI*.
22. Hollinger, G., Georgiev, Y., Manfredi, A., Maxwell, B. A., Pezzementi, Z. A., i Mitchell, B. (2006). Design of a social mobile robot using emotion-based decision mechanisms. *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, str. 3093-3098.
23. Howe, J. i Robinson, M. (2006). Crowdsourcing: A defnition. *Crowdsourcing: Tracking the Rise of the Amateur*. Weblog.
24. Hrvatski jezični portal (HJP) (2016). [online] hjp.znanje.hr. Dostupno na: hjp.znanje.hr [pristupljeno 2.2.2020. – 15.6.2020.].
25. Jakopović, H. i Mikelić Preradović, N. (2016). Identifikacija online imidža organizacija temeljem analize sentimenata korisnički generiranog sadržaja na hrvatskim portalima. *Medijska Istraživanja*, 22(2), str. 63-82.
26. Kahn, J., Tobin, R., Massey, A., i Anderson, J. (2007). Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *Am. J. Psychology, vol. 120*, str. 263-286.

27. Kao, E. C.-C., Chun-Chieh, L., Ting-Hao, Y., Chang-Tai, H., i Von-Wun, S. (2009). Towards Text-based Emotion Detection. *International Conference on Information Management and Engineering*, str. 70-74.
28. Knautz, K., Siebenlist, T., i Stock, W. G. (2010). Memose: search engine for emotions in multimedia documents. *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, str. 791-792, New York, NY. ACM.
29. Kŏvecses, Z. (2003). *Metaphor and Emotion: Language, Culture, and Body in Human Feeling (Studies in Emotion and Social Interaction)*. Cambridge University Press.
30. Landauer, T., McNamara, D., Dennis, S., i Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Erlbaum.
31. Litman, D. J. i Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Morristown, NJ, USA. Udruženje za računsku lingvistiku.
32. Liu, H., Lieberman, H., i Selker, S. (2003). A Model of Textual Affect Sensing Using Real-World Knowledge. *Proc. Eighth Int'l Conf. Intelligent User Interfaces*, str. 125-132.
33. Liu, H., Lieberman, H., i Selker, T. (2003). A model of textual affect sensing using real-world knowledge. *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, str. 125-132, New York, NY. ACM.
34. Lutz, C. i White, G. (1986). The Anthropology of Emotions. *Ann. Rev. Anthropology, vol. 15*, str. 405-436.
35. Ljubešić, N., Klubička, F. (2014). *hrWaC – Hrvatski web korpus*. [online] reldi.spur.uzh.ch/hr-sr/hrvatski-mrezni-korpus. Dostupno na: reldi.spur.uzh.ch/hr-sr/hrvatski-mrezni-korpus/ [pristupljeno 2.2.2020. – 15.6.2020.].
36. Matykiewicz, P., Duch, W., i Pestian, J. P. (2009). Clustering semantic spaces of suicide notes and newsgroups articles. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, str. 179-184, Stroudsburg. Udruženje za računsku lingvistiku.
37. Merriam-Webster OnLine (1996). [online] merriam-webster.com. Dostupno na: www.merriam-webster.com [pristupljeno 2.2.2020. – 15.6.2020.].
38. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., i Miller, K. (1990). Introduction to Wordnet: An On-Line Lexical Database. *J. Lexicography, vol. 3*, str. 235-244.

39. Mingli, S., Mingyu, Y., Chun, L Na, Ch. (2008). *A robust multimodal approach for emotion recognition*. Sveučilište u Zhejiangu, Kina.
40. Mohammad S. i Turney P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29 (3), str. 436-465.
41. Mohammad, S. (2013). *NRC Emotion Lexicon.* [online] saifmohammad.com/WebPages/NRC-Emotion-Lexicon. Dostupno na: saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm [pristupljeno 20.6.2020.].
42. Mohammad, S. M. (2011b). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. *Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Portland, SAD.
43. Mohammad, S. M. i Yang, T. W. (2011). Tracking sentiment in mail: how genders differ on emotional axes. *Proceedings of the ACL 2011 Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, Portland, SAD.
44. Osgood, C. E. i Walker, E. G. (1959). Motivation and language behavior: A content analysis of suicide notes. *Journal of Abnormal and Social Psychology*, 59(1), str. 58-67.
45. Osgood, C. E., May, W. H., i Miron, M. S. (1975). *Cross-Cultural Universals of Affective Meaning*. Sveučilište u Illinoisu.
46. Oxford Advanced Learner's Dictionary. [online] oxfordlearnersdictionaries.com. Dostupno na: www.oxfordlearnersdictionaries.com. Oxford University Press [pristupljeno 2.2.2020. – 15.6.2020.].
47. Pang, B. i Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), str. 1-135.
48. Pennebaker, J.W., Francis, M., i Booth, R. (2001). *Linguistic Inquiry and Word Count (LIWC): A Computerized Text Analysis Program. Erlbaum Publishers*.
49. Pennebaker, J.W., Mehl, M.R., i Niederhoffer, K. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Ann. Rev. Psychology, vol. 54*, str. 547-577.
50. Pestian, J. P., Matykiewicz, P., i Grupp-Phelan, J. (2008). Using natural language processing to classify suicide notes. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, str. 96-97, Stroudsburg.
51. Pico, I. (23. ožujka 2016). *Kotač emocija* Roberta Plutchika. PsicoPico. Preuzeto sa psicopico.com/en/la-rueda-las-emociones-robert-plutchik/.

52. Plutchik, 1962; prema Zelenbrz, J. (2005). *Indeks profila emocija (PIE) - Plutchik, selekcija, ličnost, norme, osjetljivost upitnika, demografske razlike | diplomski rad*. Filozofski fakultet u Zagrebu, Odsjek za psihologiju.
53. Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3), str. 3-33.
54. Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., i Kivikangas, M. (2006). Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, 15(4), str. 381-392.
55. Read, J. (2004). *Recognising Affect in Text using Pointwise-Mutual Information*. Sveučilište u Sussexu.
56. Samsonovich, A.V. i Ascoli, G.A. (2006). Cognitive Map Dimensions of the Human Value System Extracted from Natural Language. *Proc. AGI Workshop Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, str. 111-124.
57. Shaikh, M., Prendinger, H., i Ishizuka, M. (2008). Sentiment Assessment of Text by Analyzing Linguistic Features and Contextual Valence Assignment. *Applied Artificial Intelligence, vol. 22,* str. 558-601.
58. Shivhare, S. N., Khethawat, S. (2012). Emotion Detection from Text. *CS & IT 05*, str. 371-377.
59. Stevenson, R.A., Mikels, J.A., i James, T.W. (2007). Characterization of the Affective Norms for English Words by Discrete Emotional Categories. *Behavior Research Methods, vol. 39,* str. 1020-1024.
60. Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., i suradnici (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.
61. Strapparava, C. i Valitutti, A. (2004). Wordnet-Affect: An affective extension of WordNet. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004),* str. 1083-1086, Lisbon, Portugal.
62. Velásquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. *Proceedings of the fourteenth national conference on artifcial intelligence and ninth conference on Innovative applications of artifcial intelligence*, AAAI'97/IAAI'97, str. 10-15. AAAI Press.

Popis slika

Popis grafikona

Prilozi

Prilog 1 – Popis pogrešaka po kategorijama

Ovaj popis koji sadrži 3011 pogrešaka napravljen je kako bi se mogli izračunati postoci količine pogrešaka u svakoj kategoriji. U nekim kategorijama navedene su riječi i redni brojevi u Excel tablici pod nazivom „Pogreške“, dok su u ostalima navedeni samo redni brojevi.

a) „NO TRANSLATION“: ukupno 147 (4.9 %)

b) PROMJENA VRIJEDNOSTI SENTIMENTA I EMOCIJA: ukupno 7 (0.2%)

- Iz 0 u 1 = 6 ćelija (4, 6, 7, 224, 1484, 3010)
- Iz 1 u 0 = 1 ćelija (13)

c) PRAVOPISNE POGREŠKE: ukupno 78 (2.6%)

71 (Ubrajanje), 150 (Inquirer), 241 (Otok), 341 (Rešetka), 398 (Litosfera), 384 (lin), 406 (lama), 429 (Ublažavanje), 431 (Asimetričan), 462 (mahaničar), 470 (Mamut), 511 (Materijalno), 540 (Jelovnik), 541 (Mijau), 607 (Los), 620 (Motte), 671 (Pregovaranje), 705 (Bilješka), 735 (oboja), 876 (Zabava), 910 (Nastanjen), 928 (Feniks), 938 (Golub), 1002 (Otrov), 1131 (mogućni), 1160 (Objavljeno), 1165 (Vuci), 1218 (Radiologija), 1228 (Grablje), 1236 (Udaranje), 1237 (Raptorsa), 1264 (Novak), 1265 (Crvena), 1276 (Sklonište), 1280 (Regionalni), 1281 (Registrirajte se), 1348 (Rezervirano), 1350 (Rezidencija), 1377 (Mrežnica), 1397 (Vraćam), 1421 (Rijeka), 1442 (Rota), 1464 (Pravilo), 1467 (Šuška se), 1597 (Izaberi), 1660 (Pošiljka), 1667 (Pronevjera), 1682 (se smanjiti), 1697 (korak ustranu), 1744 (Ploče), 1778 (Smokey), 1803 (Soc), 1970 (Stop), 2032 (Subdukcijska), 2127 (Opskrba), 2146 (Surround), 2171 (Močvara), 2272 (rasprostirati pokošenu travu radi sušenje), 2397 (Savjet), 2407 (Toby), 2428 (PROTUPRAVNI), 2441 (Gradska kuća), 2470 (Prevedi), 2486 (koča), 2508 (Trending), 2540 (PRAVI), 2578 (Blizanci), 2587 (nezabunjljiv), 2633 (Underwood), 2717 (Gornji), 2802 (Brod), 2826 (Djevica), 2838 (Visa), 2886 (Gledati), 2928 (Wight), 2935 (Wimpy), 2944 (Mudr), 2993 (XEROX)

d) MORFOSINTAKTIČKE POGREŠKE: ukupno 170 (5.6%)

23 (ugrađuju), 45 (onesposobljavaju), 114 (neisplativim), 159 (insistirati), 162 (instaliranje), 170 (uputio), 183 (izmjenjuju), 186 (privremena), 195 (međunarodna), 202 (interventni), 214 (intrigantna), 240 (otoka), 243 (izomorfnosti), 254 (neskladna), 292 (znanje), 296

(ključaonica), 301 (šale), 314 (laboratorija), 315 (laboratorija), 316 (nedostaje), 350 (nagne), 356 (pijavica), 376 (pravo zaloge), 381 (podvezivanja), 411 (lokalne), 426 (nazire), 448 (klade), 452 (vreba), 456 (luksuzno), 458 (lirski), 512 (materijalno), 515 (mature), 523 (izmjerena), 540 (menzes), 546 (mreže), 560 (milijuna), 582 (neprilagođeno), 594 (moderiranja), 600 (molekularna), 605 (jednoslojna), 606 (usidrene), 609 (raspravlja), 623 (kreće), 632 (množe), 645 (mutantni), 646 (unakazio), 652 (mitsko), 664 (mornarički), 673 (neoprena), 682 (novorođen), 691 (plemenite), 766 (izostavljena), 773 (operski), 788 (organizirani), 794 (ukrašena), 797 (ortogonalna), 812 (precijenjena), 815 (obložio), 824 (preplavljeni), 853 (panter), 857 (parabolski), 882 (strasan), 920 (trajna), 922 (vazan), 930 (fosfor), 941 (jastučnicu), 960 (planirani), 983 (orali), 986 (orali), 1014 (popularizirao), 1022 (postavio), 1027 (odgođena), 1038 (prakticira), 1040 (pohvalio), 1063 (pripremi), 1065 (pritiskom), 1085 (privatna), 1090 (poštenost), 1102 (proizvodnju), 1138 (ponos), 1175 (nabave), 1199 (kvarcni), 1200 (kvartarni), 1260 (preporuči), 1262 (rekonstituciju), 1277 (odbijaju), 1281 (Registrirajte se), 1283 (registra), 1292 (pomlađuje), 1304 (olakšavanja), 1309 (sjećanja), 1328 (odbijanja), 1337 (reproduciraju), 1368 (obnovljena), 1380 (prepričavala), 1386 (retrospektiva), 1397 (Vraćam), 1401 (revolucionar), 1417 (obalski), 1420 (ritualna), 1437 (romantičan), 1447 (okupite se), 1448 (probuditi), 1463 (upropastila), 1472 (ruralna), 1504 (sotonskim), 1530 (znanstvena), 1541 (rezultate), 1574 (traži), 1577 (primorsko), 1585 (tajnica), 1586 (sekcijska), 1589 (sigurnosti), 1595 (segregirani), 1597 (Izaberi), 1600 (točka i zarez), 1643 (izoštriti), 1685 (smanjila), 1711 (istodobna), 1716 (griješi), 1718 (gospodine), 1772 (najmanja), 1786 (režao), 1790 (kihanje), 1832 (iskra), 1871 (sportski), 1923 (statora), 1944 (ugušena), 1947 (štulama), 1948 (stimulativan), 1965 (kameni), 1987 (slojeva), 1016 (subatomske), 2017 (potkožna), 2036 (potopne), 2048 (opasti), 2151 (preživio), 2154 (sumnjiv), 2194 (simboliziraju), 2197 (sinkroni), 2198 (sinergističko), 2201 (špric), 2209 (tabličast), 2217 (karamele), 2245 (zakasnio), 2270 (tehnička), 2275 (teflonski), 2317 (privezana), 2318 (tetrahedralnog), 2321 (teksturalna), 2326 (teistička), 2521 (bilijuna), 2593 (neodobrenih), 2624 (potkopalo), 2657 (neinficiranog), 2661 (jedinstvena), 2663 (neopravdana), 2715 (uzvisinama), 2733 (uzurpirala), 2757 (nestala), 2838 (vidni), 2877 (upozorio), 2888 (načine), 2890 (oslabljena), 2899 (dobrodošla), 2913 (hir), 2918 (šapnula), 2965 (radnom mjestu), 2990 (kovanog), 3001 (godina)

e) ZNAČENJSKE ISPRAVKE: ukupno 374 (12.4%)

Napomena: srbizmi su označeni crvenom bojom

7 (nadnesen), 13 (airs – zrak), 34 (otisak), 72 (neubjedljiv), 73 (nepodesan), 74 (nedosljedan), 94 (gnjev), 97 (neophodan), 110 (dječji), 116 (zaključni veznik), 144 (red), 167 (podstrekivanje), 172 (instrumentalista), 175 (ostrvski), 179 (broj), 183 (izmjenjuju), 198 (isprekidan), 221 (istraga), 213 (prelijeva), 239 (nadražujući), 276 (kalfa), 287(kajzerica), 288 (kengur), 291 (burence), 303 (zabavište), 306 (talent), 336 (reza), 337 (kasnost), 362 (tema), 364 (kreditiranje), 366 (lezbijska ljubav), 371 (pasiva), 389 (lingvista), 398 (odraz litologije), 400 (sporan), 408 (vekna), 410 (lobista), 412 (scena), 419 (loža), 424 (pazi), 425 (razboj), 433 (lordship – gospodarstvo), 437 (glasnost), 463 (ludak), 484 (ubistvo bez predoumišljaja), 487 (klikeri), 490 (prođa), 498 (posrtaljka), 502 (maskarada), 508 (provodadžija), 510 (materijalista), 524 (medaljist), 527 (medicinski), 530 (melanholičan), 535 (monografija), 544 (meza), 552 (mikrokosmos), 554 (midlandski), 555 (sredina leta), 556 (akušerstvo), 559 (milenijum), 561 (milioner), 569 (bjelica), 571 (minuskula), 576 (nesportsko ponašanje), 579 (krivo), 581 (hrđavo upravljanje), 583 (krivo), 593 (modelar), 597 (grba na pisti za skijanje), 603 (monaški), 604 (jednobojna slika), 607 (spokojstvo), 611 (moratorijum), 617 (mučenje), 621 (montiranje), 633 (multiplikator), 636 (ubica), 650 (misterija), 654 (go), 660 (Mala Gospojina), 677 (poletarac), 678 (neto), 687 (nikl), 694 (nodularan), 695 (nominovati), 697 (nPo), 702 (gluho), 703 (nozdrva), 724 (ishrana), 725 (orašasto voće), 728 (gojaznost), 730 (vradžbina), 731 (obit), 738 (posmatrač), 739 (akušer), 740 (akušerstvo), 742 (dobiti), 743 (suprotnost), 745 (pomračiti), 750 (okeanski), 757 (na brzu ruku), 778 (optimista), 780 (optometrista), 813 (rastinje), 816 (koji leže), 833 (paganski), 834 (paganizam), 837 (obilježavanje strana), 841 (paladijum), 847 (čulan), 854 (kič), 858 (stav), 870 (župljani), 899 (platiša), 904 (plava), 918 (ubistven), 925 (pesimista), 932 (pijanista), 945 (javio), 961 (sadilica), 962 (sjetva), 966 (pozlaćen), 973 (saigrač), 974 (prijatan), 984 (zviždovka), 987 (osmjeliti), 996 (plutonijum), 997 (prometovati), 1010 (pompezan), 1058 (povlašćen), 1066 (vajni), 1071 (sveštenstvo), 1072 (sveštenički), 1076 (premazati), 1087 (probni rad), 1103 (skrnaviti), 1104 (vulgarnost), 1110 (strše), 1112 (profilirano), 1117 (sufliranje), 1119 (krak), 1122 (pogon), 1134 (protagonista), 1135 (zaštitu), 1144 (uslov), 1148 (proksimalni), 1154 (psihologija), 1159 (publicista), 1163 (prozračan), 1164 (kučence), 1177 (purista), 1191 (karantin), 1197 (kvartil), 1205 (jorgan), 1216 (ljut), 1217 (temeljito), 1239 (klijenata), 1256 (recitovati), 1273 (refluks), 1286 (rehabilitovati), 1294 (povraćaj), 1301 (protjerivanje), 1303 (mošti), 1308 (sjetio), 1318 (vraćanje), 1324 (reorganizovati), 1330 (odgovor), 1346 (jed), 1349 (stanovi), 1369 (uzdržavanje), 1371 (rezultanta), 1372 (povraćaj), 1373 (potporni), 1385 (ponovni nalazač), 1396 (povraćaj), 1408 (protkana), 1415 (klizati se), 1416 (raskalašan), 1423 (zakovicama), 1425 (brod na sidrištu), 1426 (kolovoz), 1455 (kraljevstvo), 1469

(pobjeći), 1470 (prečaga), 1480 (sobol), 1491 (limuzina), 1494 (svetilište), 1495 (mašina za hoblovanje), 1502 (konvulzivan), 1517 (pripremiti u tavi), 1524 (strvožder), 1536 (ruženje), 1539 (oprljivanje), 1548 (jagma), 1588 (hartije od vrijednosti), 1593 (sablazan), 1594 (kuhalo), 1608 (sljedstven), 1610 (vodnik), 1614 (pilast), 1634 (ispucao), 1639 (plećka), 1665 (Pronevjera), 1684 (slijeganje), 1689 (šant), 1700 (označavanje), 1708 (Ustanove), 1720 (curica), 1736 (zatišje), 1743 (ubica), 1744 (sanke), 1753 (uprta), 1754 (šunjati se), 1762 (zmijska košuljica), 1769 (nipodaštavati), 1770 (djevojčura), 1791 (kikotanje), 1796 (frka), 1797 (grudva snijega), 1803 (socijalista), 1812 (zamjenik), 1819 (ljuto), 1821 (kolo sestara), 1827 (banja), 1830 (šopanje), 1833 (boksovanje), 1840 (specijalista), 1849 (spekulum), 1858 (spirala), 1862 (dugačka), 1867 (govorio), 1870 (avet), 1872 (sportista), 1877 (sisak), 1889 (skvamozne), 1892 (cika), 1903 (stepenik), 1914 (stanovište), 1916 (uzdrhtati), 1917 (izgladneo), 1922 (stacionaran), 1924 (vajarstvo), 1925 (statueta), 1929 (kitica), 1940 (paprikaš), 1942 (položaj upravnika), 1950 (koji žeže), 1964 (kamena roba), 1969 (zapušač), 1974 (urakljiti), 1981 (moreuz), 1993 (tekući), 2006 (strobe), 2013 (gipsani malter), 2015 (klinac), 2016 (okovan klincima), 2029 (odsjek), 2043 (sekvencija), 2073 (odojče), 2081 (zagušljiv), 2082 (ugušenje), 2085 (samoubistvo), 2101 (sunčev zrak), 2102 (sušen na suncu), 2105 (sunce), 2109 (nadčovječanski), 2110 (preklapaju), 2139 (nadmašiv), 2146 (prismotra), 2149 (geometar), 2164 (tampon), 2195 (saosjećati), 2196 (simpatija), 2202 (torokuša), 2206 (žlica), 2208 (stolno posuđe), 2213 (taktika), 2220 (podrezan), 2221 (pljeva), 2225 (talenat), 2231 (raboš), 2234 (morska trava), 2235 (tangens), 2239 (ravan), 2244 (zakasnelost), 2250 (kamenac), 2256 (heklati), 2269 (dosadan), 2277 (reći), 2294 (tense – vrijeme), 2308 (teritorija), 2309 (terorista), 2319 (tetraeadričan), 2324 (raskraviti), 2325 (pretjerano pijenje čaja), 2327 (teokratija), 2329 (terapija), 2330 (otprilike), 2343 (uzan remen), 2350 (vršati), 2358 (povjerenje), 2360 (davilac), 2369 (plimski), 2373 (čarka), 2391 (zujanje u ušima), 2405 (mjera za vunu), 2407 (robijati), 2409 (kulučenje), 2412 (meni), 2431 (totalizator), 2435 (tvrd), 2451 (tramvaj), 2454 (prevazilaženje), 2457 (kopiranje), 2477 (dijagonalan), 2487 (tegliti), 2492 (ergometar), 2523 (oportunista), 2528 (cupkanje), 2530 (tritijum), 2533 (trolejbus), 2539 (mistrija), 2556 (bokorast), 2558 (tulip - lala), 2563 (melodičan), 2572 (mitnica), 2583 (ultimativno), 2602 (neoprezan), 2609 (beskrupulozan), 2620 (student), 2629 (zajamčiti), 2655 (bez otiska), 2664 (nepristupačan), 2672 (neorganizovan), 2680 (neregistrovan), 2683 (nekažnjen), 2702 (netranslatirano), 2726 (uranijum), 2732 (razvoditi), 2742 (upražnjeno mjesto), 2746 (zavjesa), 2752 (vrijednost), 2762 (varijacija), 2765 (PDV), 2777 (komora), 2778 (klijetke), 2801 (kesični), 2809 (vibracioni), 2824 (violinista), 2825 (poskok), 2828 (virusologija), 2880 (odgajivačnica zečeva), 2896 (u crnini), 2907 (god), 2910

(potreban novac), 2926 (poročnost), 2933 (rado), 2940 (osvajanjem), 2949 (straha), 2958 (sviter), 2977 (upropašćen), 2980 (rvanje), 3002 (žute)

f) PREUZETA ENGLESKA RIJEČ KAO HRVATSKI PRIJEVOD: ukupno 60 (2%)

150 (Inquirer), 266 (jimmy), 269 (Jock), 271 (john), 319 (Lacrosse), 377 (lieu), 525 (media), 599 (moulded), 613 (morphism), 616 (mortgagor), 620 (Motte), 771 (open), 826 (overzealous), 923 (pertussis), 927 (petting), 934 (picketing), 934 (ping), 952 (pix), 970 (playa), 991 (plumper), 1013 (pop), 1033 (Pow), 1152 (psalm), 1219 (radius), 1237 (Raptorsa), 1449 (rouser), 1550 (scrapie), 1623 (settlor), 1638 (Shanghai), 1778 (Smokey), 1782 (snags), 1795 (Snoopy), 1932 (Stead), 1946 (stilet), 1963 (ustondiran), 1972 (storming), 2040 (subplot), 2046 (podset), 2104 (sunless), 2114 (supernatant), 2142 (surrendering), 2143 (Surround), 2390 (tinkering), 2402 (titty), 2404 (Toby), 2461 (transfixed), 2505 (Trending), 2603 (uncaring), 2615 (uncorrelated), 2628 (Underwood), 2743 (vascular), 2744 (vacuole), 2773 (vellum), 2817 (vinaigrette), 2832 (Visa), 2928 (Wight), 2959 (Wop), 2969 (wot), 2982 (wright), 2999 (yea)

g) RIJEČ SE NE MOŽE PREVESTI NA HRVATSKI JEZIK: ukupno 2 (0.01%)

11 (accueil – no translation), 1430 (roc - rok)

h) SVE OSTALE PROMJENE: ukupno 2189 (72,7%)

Leksikon emocija hrvatskog jezika

Sažetak

Ovaj rad opisuje prilagodbu leksikona emocija pod nazivom *NRC Word-Emotion Association Lexicon* (skraćeno *EmoLex*) (Mohammad i Turney, 2013) na hrvatski jezik. EmoLex sadrži 14 182 riječi, njihove sentimente, emocije i prijevod s engleskog na više od stotinu jezika, uključujući i hrvatski. Prijevodi su generirani pomoću strojnog prevođenja putem Google Prevoditelja, a obilježja za emocije riječi engleskog jezika izravno preslikana na riječi u ostalim jezicima.

Ovdje će se prikazati prilagodba leksikona iz srebrnog standarda u zlatni standard na način da će se detaljno proučiti obilježja svake riječi i točno prevesti na hrvatski jezik uzimajući u obzir različite kontekste u kojima se pojedina riječ može naći. Također će se preispitati i ispraviti emocije i sentimenti sukladno prijevodima na hrvatski jezik. Ovaj proces vršio se uz pomoć hrvatskog web korpusa hrWaC-a, englesko-engleskog rječnika Merriam-Webster, online rječnika Glosbe i Hrvatskog jezičnog portala, a u radu će se također opisati postupci i metodologija ovog projekta.

Ključne riječi: leksikon emocija, leksikon sentimenta, emocija, sentiment, hrvatski leksikon

Croatian Emotion Lexicon

Summary

This paper describes the adaptation of the *NRC Word-Emotion Association Lexicon* (*EmoLex*) (Mohammad and Turney, 2013) to the Croatian language. EmoLex counts 14,182 words, their sentiments, emotions and their translation from English into over one hundred other languages, including Croatian. The translations were automatically translated by Google Translate while the emotion features for the English words were copied directly to their translated versions in other languages.

The adaptation process of the lexicon from the silver standard to the gold standard will be shown in the following way: the features of every word were scrutinized and translated correctly, taking into consideration different contexts a word may appear in. The process also included reviewing and correcting emotions and sentiments according to their Croatian translation. This adaptation was conducted via the Croatian Web Corpus hrWaC, the Merriam-Webster Dictionary, the Glosbe online dictionary and the Croatian Language Portal. The paper will also describe the methods and methodology of the project.

Key words: emotion lexicon, sentiment lexicon, emotion, sentiment, Croatian lexicon