SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE ZNANOSTI
Ak. god. 2009./ 2010.

Ante Kranjčević

APLIKACIJA ZA RAČUNANJE N-GRAMA

Diplomski rad

Mentor: dr. sc. Kristina Vučković

Zagreb, 2010.

Sadržaj

1. Uvod

Nemoguće je u današnje doba zamisliti obradu teksta bez pomoći računala. Velika količina teksta već se nalazi zapisana u digitalnom obliku što ga čini pogodnijim za računalnu obradu, za razliku od teksta koji se nalazi u pismenom obliku kojeg je prethodno potrebno digitalizirati. Računalna, kao i ručna obrada teksta, vrši se na mnogim razinama ovisno o korisničkim potrebama. Ovaj rad opisuje mrežnu aplikaciju nazvanu *bigtri* čija zadaća je obrada teksta na način da iz njega izdvaja bigrame[1], trigrame[2], bigrame riječi[3] i trigrame riječi[4].

Bigrami, trigrami, bigrami riječi i trigrami riječi lingvističke su jedinice koje su sastavni dio jezika kojim se služimo. U informacijskom riječniku za njih koristimo skupnu riječ n-grami. Na osnovu n-grama jezici se uspoređuju na lingvističkoj bazi, te se izrađuje tablica učestalosti pojavljivanja istih lingvističkih elemenata u svakom pojedinom jeziku.

Misao vodilja, te svrha izrade ovoga rada bila je nepostojanje takvoga alata na hrvatskom jeziku prema mojim saznanjima, te pojednostavljivanje rada korisnicima koji se koriste ovakvom analizom jezika. Aplikacija je postavljena na mrežni server tako da joj je moguće pristupiti u bilo kojem trenutku, dostupna je svima bez restrikcija, što povećava količinu teksta u bazi. Veća količina spremljenog teksta daje nam preciznije izlazne podatke.

Korisnici u radu s aplikacijom nisu ograničeni obrađivanjem samo hrvatskog jezika, nego je moguć izbor 38 svjetskih jezika koji su prikazani u tablici 1.

Ovaj je rad podjeljen u dva dijela. U prvom dijelu detaljno su pojašnjene metode izrade aplikacije, te svaki dio aplikacije koji korisnik koristi u radu. U drugom dijelu rada napravljena je usporedba sa sličnim postojećim alatima koji se mogu pronaći besplatno na internetu. Prilikom ocjenjivanja i uspoređivanja aplikacija u obzir su uzeti funkcionalnost, brzina obrade unešenog teksta, izgled obrađenih podataka, te mogućnost obrade izlaznih podataka.

[1] *bigram* – niz od dva uzastopna znaka
[2] *trigram* – niz od tri uzastopna znaka
[3] *bigram riječi* – niz od dvije uzastopne riječi
[4] *trigram riječi* – niz od tri uzastopne riječi

Jezik				
Albanski	Filipinski	Japanski	Poljski	Švedski
Armenijski	Finski	Kineski	Portugalski	Talijanski
Bjeloruski	Francuski	Latvijski	Rumunjski	Tibetanski
Bosanski	Grčki	Litvanski	Ruski	Turski
Danski	Hrvatski	Makedonski	Slovački	Ukrajinski
Engleski	Indonezijski	Mongolski	Slovenski	Vijetnamski
Esperanto	Irski	Norveški	Srpski	
Estonski	Islandski	Njemački	Španjolski	

Tablica 1 – Popis jezika u bazi

Podatke dobivene analizom jezika (pojavljivanje pojedinih bigrama, trigrama, bigrama riječi, trigrama riječi) moguće je koristiti kao referentne podatke za daljnje istraživanje jezika.

2. Teorijski dio

2.1. n-grami znakova

N-grami znakova su pod-nizovi *n* predmeta iz danog niza znakova. Predmeti u ovom slučaju mogu biti fonemi[5], slogovi riječi ili riječi. N-gram veličine 1 naziva se *unigram,* n-gram veličine 2 naziva se *bigram,* n-gram veličine 3 naziva se *trigram,* dok se n-grami veličine 4 ili više jednostavno nazivaju *n-grami*. Modeli n-grama su tipovi modela koji se koriste za predviđanje sljedeće stavke u svakom slijedu, te se koriste u različitim područjima statističkih obrada prirodnih jezika. N-grami modela modeliraju sekvence, posebice prirodnih jezika, pri tome koristeći statistička svojstva n-grama. Primjeri korištenja statističkih svojstava n-grama su još u prepoznavanju govora, obradi fonema kao i obrađivanja sekvenci fonema.

Statistička vjerojatnost sljedećeg n-grama bila je ideja za eksperiment Claudea Shannona[6] u radu o teoriji informacija 1948. godine. Shannon je postavio pitanje, kako s obzirom na redosljed slova možemo izračunati vjerojatnost sljedećeg slova u riječi. Promatrajući podatke došao je do odgovora da može izvesti distribuciju vjerojatnosti za sljedeće slovo koje ima zbroj vjerojatnosti svih sljedećih slova 1.0. Npr. za sekvencu "for ex" sljedeći je n-gram po njegovom izračunu imao vrijednost a = 0,4, b = 0,00001, c = 0,... i tako sva ostala slova abecede dok zbroj vjerojatnosti njihovih pojavljivanja nije 1.0. Ovakvi modeli n-grama često su kritizirani u literaturi jer im nedostaje eksplicitni prikaz dalekometne ovisnosti[7]. Najžešći kritičar ovoga modela je *Noam Chomsky*[8], čija kritika je proizašla iz razloga što modeli n-grama koji su spomenuti ranije u tekstu imaju raspon ovisnosti (n-1), te se kao takvi ne mogu koristiti kao rješenje za neograničene ovisnosti koje stvaraju prirodni jezici. Iz tog razloga n-gram modeli nisu imali veći utjecaj na razvoj lingvističke teorije nego se koriste samo u praktičnim aplikacijama za obradu jezika, glasovno prepoznavanje jezika i generiranje slučajnih nizova fonema (Manning, Schütze, 1999.).

[5] *Fonem* – najmanja jedinica govora koja je bitna za značenje
[6] *Claude Shannon (1916. – 2001.)* – Američki matematičar i ing. elektrotehnike, naziva se ocem informacijske teorije
[7] *Dalekometna ovisnost* – fenomen koji se javlja u analizama prostornih i vremenskih nizova podataka, a odnosi se na stopu raspadanja statističke ovisnosti
[8] *Noam Chomsky (1928.)* – Američki filozof, lingvist i politički aktivist, otac generativne gramatike

2.2. Markovljev model

Markovljev lanac, nazvan po Andreyu Markovu[9], predstavlja niz stanja sustava koji se koristi za statističko modeliranje. Sustav koji je opisao Markov takav je da u svakom trenutku može prijeći u neko novo stanje ili može ostati u istome stanju što ovisi samo o trenutnom stanju u kojemu se sustav nalazi.

Markovljevi lanci imaju širok spektar uporabe u statistici, književnosti, informacijskoj znanosti. Primjer Markovljevog lanca je sljedeći: ako pratimo prehrambene navike osobe koja jede samo grejp, sir i grašak i prati sljedeća pravila da jede samo jednom dnevno i ako je jučer jeo sir, neće ga jesti danas, dok će grašak i grejp jesti uz istu vjerojatnost. Ako je jučer jeo grejp, danas će jesti grejp uz vjerojatnost 1/10, sir 4/10 i grašak 5/10. Konačno, ako je jučer jeo grašak, neće ga jesti danas, no jesti će grejp uz vjerojatnost 4/10 ili sir 6/10. Prehrambena navika iz primjera može se prikazati Markovljevim lancem, jer odluka što će danas jesti, ne ovisi o tome što je osoba jela prije nekoliko dana, nego jedino ovisi o tome što je jučer jela (Rabiner, Juang, 1986.).

Ovakvu proceduru koristio je i Claude Shannon, prilikom predviđanja mogućnost pojavljivanja n-grama, a to je upravo Markovljev lanac. Shannon je eksperimentom pokazao da i je bez poznavanja cijelog sustava moguće pribaviti mnogo statističkih pravilnosti istog tog sustava.

2.3. Frekvencijska analiza

Frekvencijska analiza koristi se kod prikazivanja učestalosti znakova, slova, bigrama, trigrama i ostalih komponenti teksta u jeziku. Određena pravilnost se pojavljuje u nabrojanim slučajevima, te je pogodna za analizu jer postoji karakteristična distribucija bigrama ili trigrama karakteristična za svaki pojedini jezik. Kao primjer navest ćemo engleski jezik u kojemu je slovo „e“ veoma često, dok se s druge strane slovo „x“ ne pojavljuje tako često u jeziku. Što se bigrama tiče, u engleskom jeziku, bigrami 'st', 'ng', 'th' veoma su česti, za razliku od bigrama 'nz' ili 'qj' *(*Wikipedia*)*.

[9] *Andrey Markov (1856. – 1922.)* – Ruski matematičar najpoznatiji po radu na teoriji stohastičkih procesa koji su poslije nazvani markovljev lanac

Analize pokazuju kako frekvencija pojedinih znakova, odnosno bigrama i bigrama riječi, varira ovisno o autoru, kao i o temi o kojoj autor piše. Kao primjer možemo uzeti tekst u kojemu se piše o „x-zračenju“, te u tom tekstu možemo očekivati veliku frekvenciju znaka 'x', odnosno bigrama koji u sebi sadrži znak 'x'. Znakovi, odnosno slova, te, bigrami, trigrami, kao i bigrami, odnosno trigrami riječi mogu se koristiti kao dokaz autentičnosti nekoga teksta, te dokazati da li je pojedino djelo napisao autor za kojeg se to tvrdi. Preciznu prosječnu frekvenciju pojavljivanja bigrama, trigrama, bigrama riječi i trigrama riječi možemo dobiti samo analizom velike količine reprezentativnog teksta (Manning*, 1999.).*

Frekvencija pojavljivanja znakova imala je velik utjecaj na dizajn pojedinih tipkovnica, tj. rasporeda tipki na njima. Kao primjer pojavljuje se takozvana *AZERTY* tipkovnica, koja je specifična po tome što su najfrekventnija slova stavljena u prvi red tipkovnice (*AZERTY),* za razliku od tipkovnice kakvu mi koristimo kojoj je raspored slova u gornjem redu *QWERTZ (*Wikipedia*)*.

3. *bigtri* aplikacija

U sljedećem poglavlju pokazat ću na koji je način ustrojena aplikacija, kako funkcioniraju pojedine komponente programskog kôda, te funkcionalnost i snalaženje u korištenju same aplikacije. Aplikacija je pojednostavljena i organizirana tako da bude što jednostavnija za uporabu, korisniku pruža maksimalnu funkcionalnost i omogućava što jednostavnije snalaženje prilikom korištenja.

3.1. Izrada aplikacije

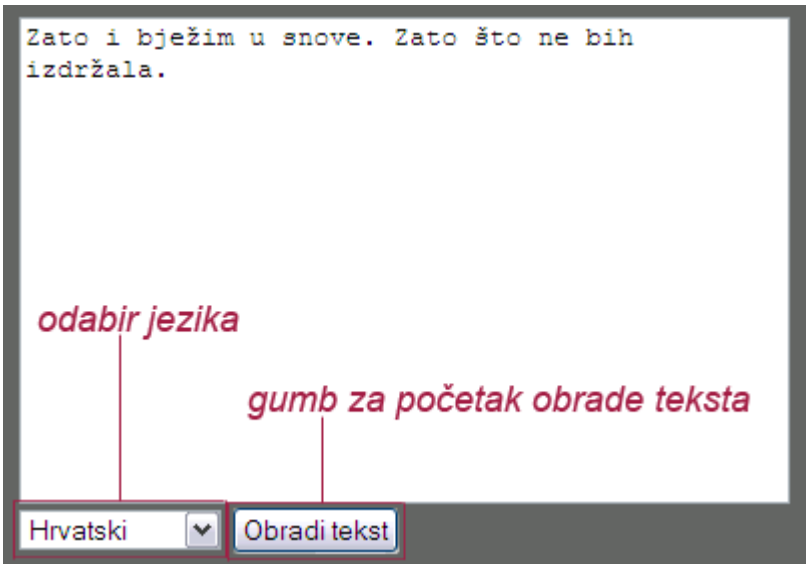
Aplikacija *bigtri* izgrađena je u C# programskom sučelju koristeći Microsoft-ovu ASP. net tehnologiju servera kako je već spomenuto u uvodu. Koristeći programski jezik C# dobio sam jednostavnost i veliku mogućnost implementacije s ostalim komponentama koje su korištene u izradi ovog alata. Brzina obrade jezika, korištenje potencijala za rad nad bazom podataka, te velika baza korisničke podrške elementi su koji su prevagnuli u izboru programskog alata za izradu *bigtri* aplikacije.

3.2. Unos i statistika teksta

3.2.1. Metoda unosa i obrade teksta

Tekst koji se obrađuje i analizira unosi se preko komponente za unos teksta koje se nalazi na početnoj mrežnoj stranici www.ante.sibinj.com. Nakon unosa teksta, bilo ručnim unosom ili *copy*[10] - *paste*[11] metodom unosa, potrebno je odabrati jezik na kojemu je tekst napisan (korisnik ima na izbor 38 jezika kao što je prikazano u tablici 1), te pritisnuti gumb *Obradi tekst, kako* je prikazano na slici 1. Pritiskom na gumb aktivira se dio programskog kôda koji preko jednostavne petlje iz teksta izdvaja jedan po jedan bigram, trigram, bigram riječi, te trigram riječi. Niz znakova koji dobijemo i u koji su spremljeni bigrami, trigrami, bigrami riječi i trigrami riječi spremaju se u listu koja se koristi tijekom daljnjeg rada aplikacije (izrada statistike, spremanje u tekstualnu datoteku).

[10] *copy* – kopirati
[11] *paste* – zalijepiti



Slika 1 – Forma za unos teksta

Prilikom pisanja petlji za obradu teksta potrebno je bilo voditi brigu o slučajevima kada se u tekstu umjesto jednog razmaka između riječi nalazi više razmaka. Provjera razmaka veoma je bitna, jer inače ne bih dobio pravilne bigrame, trigrame, bigrame riječi i trigrame riječi nakon obrade teksta. Programska petlja prikazana je na slici 2.

U ovom dijelu koda jasno se vidi kako naredba *Replace* kroz While petlju provjerava cijeli tekst. Petlja se izvršava onoliko puta koliko je potrebno dok ne ukloni sve višestruke razmake između riječi.

```
while (victim.IndexOf("  ") >= 0)
{
        victim = victim.Replace("  ", " ");
}
victim = victim.Replace(Environment.NewLine, "");
```

Slika 2 – Petlja za provjeru razmaka

Primjer formatiranja, te pravilnog i nepravilnog računanja bigrama prikazan je u tablici 2. Kao primjer naveo sam rečenicu „**Zato i bježim.**“ U rečenici iz primjera se ispred slova 'i' nalazi dvostruki razmak, umjesto pravilnog jednostrukog razmaka.

Zato i bježim.	
Nepravilno	**Pravilno**
'Za'	'Za'
'at'	'at'
'to'	'to'
' '	' i'
' i'	'i '
'i '	'bj'
'bj'	'je'
'je'	'ež'
'ež'	'ži'
'ži'	'im'
'im'	'm.'
'm.'	

Tablica 2 – Primjer pravilne i nepravilne obrade teksta

Iz tablice se vidi da je razlika između pravilno obrađenog teksta i nepravilno obrađenog teksta u jednom bigramu. Razlika nastaje zbog jednog znaka viška, tj. zbog dodatnog razmaka koji se pojavljuje u rečenici. Jasno je da takvi slučajevi nikako ne dolaze u obzir ukoliko želimo točan broj obrađenih leksičkih jedinica.

Petlja koja obrađuje tekst uklanja sve višestruke razmake i pretvara ih u jedan razmak kako bi tekst bio pravilno obrađen. Osim dvostrukih razmaka, petlja obrađuje slučajeve kada se između susjedna dva reda teksta nalazi više od jednog praznog reda pretvarajući ih u jedan red. Ručni unos jednog ili više uzastopnih praznih redova spriječen je onemogućavanjem pritiska tipke *enter* prilikom ručnog unosa teksta u komponentu za unos teksta. Ovo ograničenje se vidi iz sljedećeg koda koji je prikazan na slici 3.

```
<asp:TextBox ID="txtText" runat="server"
Height="239px" TextMode="MultiLine"
Width="381px"
onkeydown = "return (event.keyCode!=13);">
</asp:TextBox>
```

Slika 3 – Ograničenje pritiska tipke *enter*

Slika 4 prikazuje kôd kojim biramo jezik iz padajućeg izbornika.

```
<asp:DropDownList ID="ddlJezik" runat="server"
DataSourceID="AccessDataSource1"
DataTextField="Jezik" DataValueField="ID">
</asp:DropDownList>
```

Slika 4 – Izbor jezika iz padajućeg izbornika

Kôd koji obrađuje tekst pokreće se pritiskom na gumb *Obradi tekst* kako je i navedeno ranije u poglavlju. Izgled kôda prikazan je na slici 5.

```
<asp:Button ID="btnObradi" runat="server"
onclick="btnObradi_Click" Text="Obradi tekst"/>
```

Slika 5 – Kôd obrade teksta

3.2.2. Stuktura podataka u bazi

Obrađene podatke potrebno je spremiti u bazu podataka kako bi se mogli koristiti u daljnjem radu. Aplikacija koristi *Microsoft Access* bazu podataka za spremanje podataka. *Access* baza odabrana je zbog jednostavnosti rada. U ovom je slučaju baza sačinjena od jedne datoteke koja u sebi sadrži šest povezanih tablica navedenih u tablici 3.

Tablice		
Jezik	BigramChar	TrigramChar
Tekst	BigramWord	TrigramWord

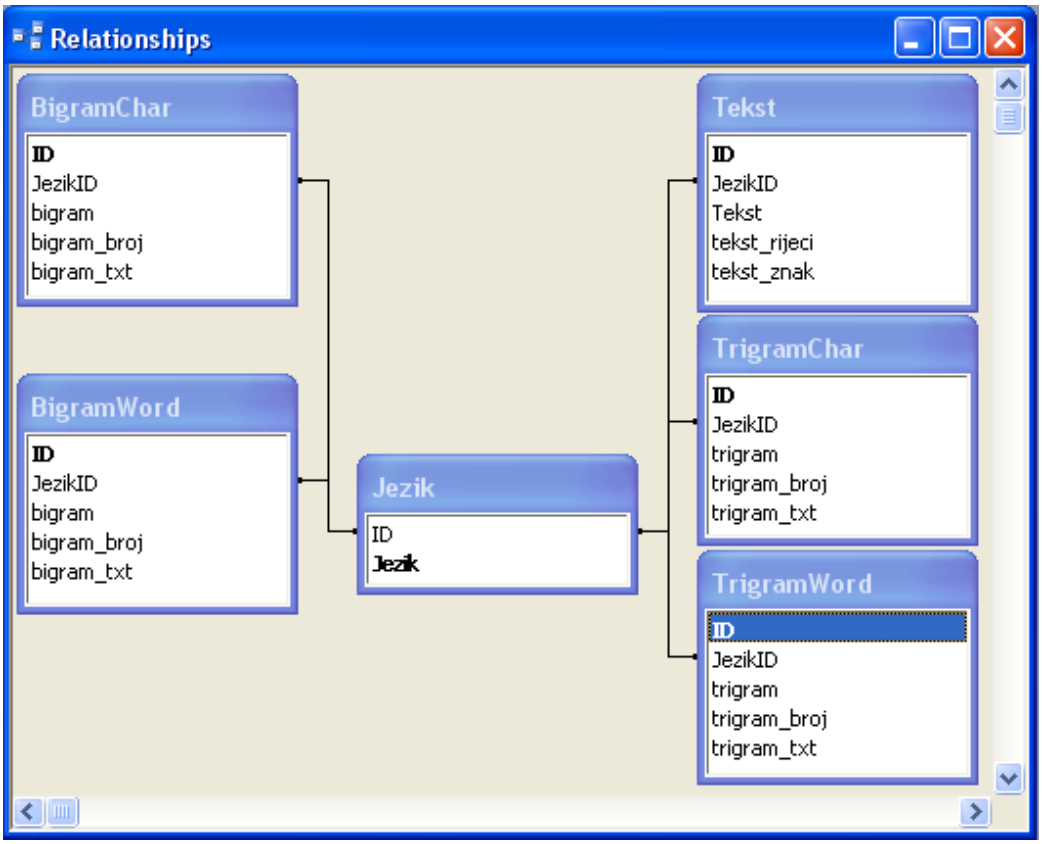
Tablica 3 – Popis tablica u bazi podataka

Alternativa *Access*-u, je *SQL*[12] baza podataka, no ona zahtijeva posebni server za rad s njom i iz tog razloga nije korištena u aplikaciji. Odabirom *Access* baze podataka niti u jednom trenutku nije narušena funkcionalnost. Jedini mogući

[12] *SQL* – kratica za *Structured Querry Language* – računalni jezik dizajniran za rad nad bazama podataka

nedostatak je brzina rada s bazom, jer se *SQL* baza podataka odlikuje većom brzinom pristupa i obrade podataka u tablicama.

Tablica *Jezik* glavna je tablica u bazi. Ona integrira sve ostale tablice preko jedinstvenog ključa. Jedinstveni ključ polje je naziva *ID* (jedinstevni identifikacijski broj svakog pojedinog jezika) koje je brojčanog tipa i u kojemu se ne mogu pojavljivati dvostruke vrijednosti. Nemogućnost pojavljivanja dvostruke vrijednosti u polju znači da svaki jezik u tablici ima jedinstveni identifikacijski broj pomoću kojega identificiramo ostale podatke u bazi. Jezici iz ove tablice vidljivi su korisnicima u padajućem izborniku na početnoj stranici aplikacije.



Slika 6 – Integritet tablica unutar baze podataka

Tablica *Tekst* služi za spremanje svih tekstova koji su upisani u komponentu za unos teksta i nakon toga obrađeni. Na taj način, ova tablica služi kao kontrolna tablica, kako bi se u svakom trenutku mogli ponovno provjeriti rezultati dobiveni nakon obrade teksta. Ostale četiri tablice (*BigramChar, BigramWord, TrigramChar, TrigramWord*) služe za spremanje bigrama, trigrama, bigrama riječi i trigrama riječi pojedinačno u svaku od navedenih tablica. Logička struktura i integritet tablica unutar baze podataka prikazani su na slici 6.

3.2.3. Unos obrađenih podataka u bazu

Nakon unosa i obrade podataka, sljedeća zadaća aplikacije je spremiti podatke u bazu. Spremanje započinje na način da se lista podataka koja je dobivena nakon obrade teksta, kao što je opisano u poglavlju 3.2.1., uspoređuje s podatcima koji se već nalaze u svakoj od pojedinih tablica za odabrani jezik. Ukoliko se u tablicama već pojavljuju bigrami, trigrami, bigrami riječi i trigrami riječi, koji su dobiveni i nakon obrade teksta, ne unose se u nova polja u tablice, nego se njihov broj pojavljivanja uvećava za broj pojavljivanja dobiven u obrađenom tekstu. S druge strane, ukoliko se radi o prvom pojavljivanju istih (bigrama, trigrama, bigrama riječi i trigrama riječi), unose se u tablice u nova prazna polja. Kao početni broj pojavljivanja postavlja se broj dobiven iz obrađenog teksta kako je prikazano na slici 7.

```
if (iNgramID != 0)
        {
                using (OleDbCommand cmd = new OleDbCommand("UPDATE " + tableToSaveTo + " SET " + column +
"_broj =  @ngram_broj, " + column + "_txt = @ngram_txt WHERE ID = @ID", conn))
                {
                        cmd.Parameters.Add(new OleDbParameter("@ngram_broj", iBigramBroj + ngram.Count));
                        cmd.Parameters.Add(new OleDbParameter("@ngram_txt", iBigramTxt + BrojZnakova));
                        cmd.Parameters.Add(new OleDbParameter("@ID", iNgramID));
                        cmd.ExecuteNonQuery();
                }
                ngram.CountBefore = iBigramBroj;
        }
        else
        {
                using (OleDbCommand cmd = new OleDbCommand("insert into " + tableToSaveTo + " (JezikID, " + column
+ ", " + column + "_broj, " + column + "_txt) values (@JezikID, @ngram, @ngram_broj, @ngram_txt)", conn))
                {
                        cmd.Parameters.Add(new OleDbParameter("@JezikID", jezikID));
                        cmd.Parameters.Add(new OleDbParameter("@ngram", ngram.Znakovi));
                        cmd.Parameters.Add(new OleDbParameter("@ngram_broj", ngram.Count));
                        cmd.Parameters.Add(new OleDbParameter("@ngram_txt", BrojZnakova));
                        int i = cmd.ExecuteNonQuery();
                }
                ngram.CountBefore = 0;
        }
```

Slika 7 – Provjera postojanja leksičkih jedinica u bazi podataka

Rezultati obrade unešenog teksta vidljivi su ispod komponente za unos teksta na početnoj stranici aplikacije. Izgled rezultata obrade teksta prikazan je na slici 8.

Slika 8 – Rezultat obrade teksta

Rezultati obrade grupirani su na način da u prvom stupcu imamo ispisan ukupan broj bigrama obrađenog teksta, te svaki pojedini bigram jedan ispod drugoga. Pored svakog od dobivenih bigrama nalazi se broj pojavljivanja istoga u obrađenom tekstu, te broj pojavljivanja ukoliko se bigram od prije nalazi spremljen u bazi. Kao primjer naveo sam bigram 'Za'[13] koji se u obrađenom tekstu pojavljuje 2 puta, dok je u bazu spremljen već 63 puta. Brojevi pojavljivanja bigrama međusobno su odvojeni znakom „/“. Na isti način nabrojani su i ostali elementi teksta i to redom trigrami, bigrami riječi i trigrami riječi.

3.2.4. Vanjsko spremanje rezultata obrade teksta

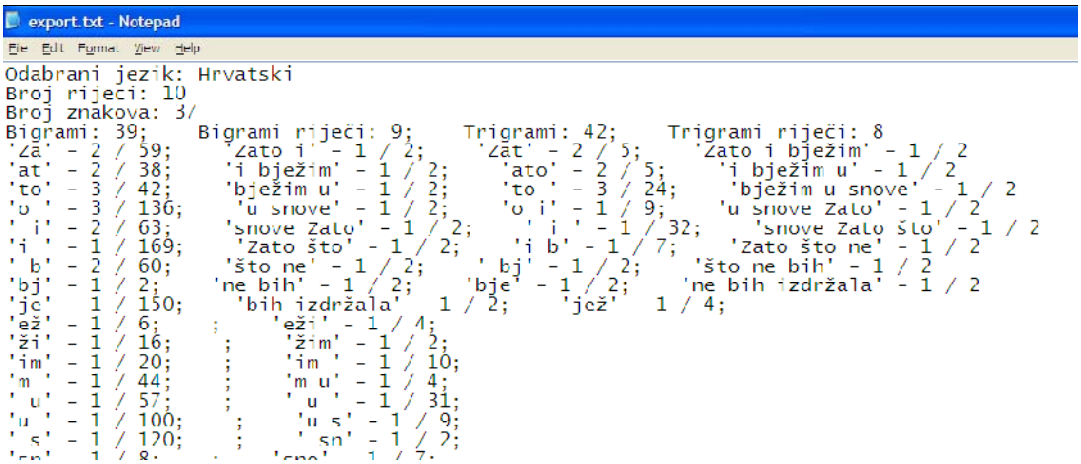
Osim unosa obrađenih podataka u bazu i njihova uspoređivanja, moguće je podatke spremiti i u vanjsku datoteku pritiskom na gumb *Snimi u .txt* koji se nalazi na vrhu stranice. Pritiskom na gumb podatci se na računalo spremaju u datoteku početnog naziva *export.txt.* Kôd je prikazan na slici 9.

```
<asp:Button ID="btnSnimi" runat="server"
Text="Snimi u .txt"
onclick="btnSnimi_Click" />
```

Slika 9 – Spremanje statistike u vanjsku datoteku

[13] Postoji razlika između 'Za' i 'za' bigrama, te ih aplikacija upisuje u posebna polja u tablicu

Datoteka je jednostavna tekstualna datoteka koju je moguće otvoriti u najjednostavnijem računalnom programu za obradu teksta koji dolazi instaliran u sklopu operacijskog sustava na računalima. Mogućnost vanjskog spremanja podataka dodana je jer velik broj korisnika želi nakon obrade teksta dobiti obrađene podatke u tekstualnom formatu kako bi ih mogli koristiti na osobnom računalu. Izgled datoteke nakon otvaranja u tekstualnom *editoru*[14] prikazan je na slici 10.



Slika 10 – Rezultati obrade u tekstualnom obliku

3.2.5. Statistički prikaz rezultata obrade teksta

Nakon obrade vezane uz unos bigrama, trigrama, bigrama riječi i trigrama riječi, moguće je prikazati statistiku i usporedbu istih s bigramima, trigramima, bigramima riječi, te trigramima riječi koji se nalaze spremljeni u bazi. Usporedba je moguća ne samo za odabrani jezik, nego i za preostale jezike iz baze (njih 38).

Do statističkih podataka dolazimo pritiskom gumba *Stats* koji se pojavljuje na početnoj stranici aplikacije nakon obrade teksta (slika 4). Kôd kojim otvaramo statistiku prikazan je na slici 11.

```
<asp:Button ID="btnBigramStats"
runat="server"
onclick="btnBigramStats_Click" Text="Stats" />
```

Slika 11 – Spremanje statistike u vanjsku datoteku

[14] *editor* – aplikacija za uređivanje određenih tipova podataka

Nakon otvaranja stranice pojavljuje se tablica s lijeve strane prozora u kojoj su sadržani usporedni podatci za preostale jezike iz baze. Osim rezultata obade teksta, u tablici se nalaze i nabrojani bigrami, trigrami, bigrami riječi i trigrami riječi koji su u bazu ranije unešeni i obrađeni, ali za druge jezike iz baze. Pored svake leksičke jedinice nalazi se njezin broj pojavljivanja, kao i statistički podatak koji govori koja je frekvencija njezina pojavljivanja u pojedinom jeziku. Izgled stranice koja sadrži tablicu podataka prikazan je na slici 12.

Slika 12 – Statistika obrađenih podataka

Statistika obrađenog teksta izračunata je na način da se broj pojavljivanja pojedinih bigrama, trigrama, bigrama riječi i trigrama riječi podijeli sa zbrojem svih bigrama, trigrama, bigrama riječi i trigrama riječi koje imamo u bazi za odabrani jezik, kako je prikazano na slici 13.

$$X_{st} = \frac{nX_o}{\sum X_o}$$

Slika 13 – Matematička formula za izračun statistike

X_{st} konačni je statistčki izlazni podatak koji dobijemo nakon obrade; nX_o broj je pojavljivanja pojedinih bigrama, trigrama, bigrama riječi, odnosno trigrama riječi u obrađenom tekstu; $\sum X_o$ ukupni je broj bigrama, trigrama, bigrama riječi, odnosno trigrama riječi koje smo dobili nakon obrade unešenog teksta.

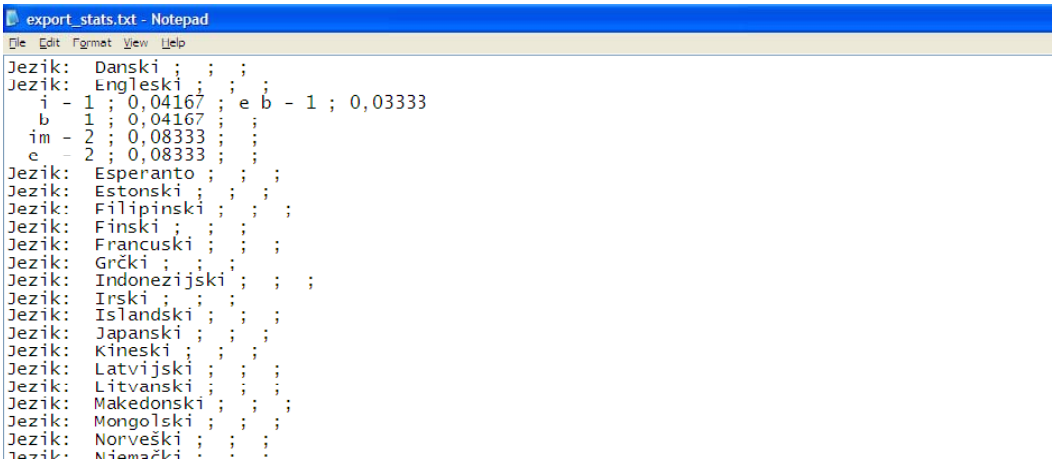
3.2.6. Vanjsko spremanje statističkih podataka

Osim vizualnog pregleda podataka na stranici nakon obrade, dodana je mogućnost spremanja podataka u vanjsku datoteku početnog naziva *export_stats.txt* i to pritiskom na gumb *Spremi u .txt*. Kôd kojim spremamo statistiku prikazan je na slici 14.

```
<asp:Button ID="btnSnimi" runat="server"
Text="Snimi u .txt"
onclick="btnSnimi_Click" />
```

Slika 14 – Spremanje statistike u vanjsku datoteku

U spremljenoj datoteci nalaze se podaci iz tablice koji služe daljnjoj obradi koja se može vršiti lokalno tj. na korisnikovom računalu. Iz tog je razloga ovakvo spremanje podataka nazvano *vanjsko spremanje*. Izgled datoteke nakon spremanja na računalo i otvaranja u tekstualnom editoru prikazan je na slici 15.



Slika 15 – Statistički podaci u tekstualnom obliku

3.3. Statistika teksta

Najvažniji dio nakon obrade podataka je zapravo konačna statistika za svaki od pojedinih jezika koji se nalaze u bazi. U daljnjem tekstu objasnit ću kako i gdje doći do tih podataka, te kako dalje raspolagati dobivenim podatcima.

3.3.1. Statistika *top30*

Na stranici www.ante.sibinj.com/Statistika.aspx moguće je pregledati globalnu statistiku obrađenih podataka koji se nalaze u bazi. Tablica na stranici prikazuje top 30 bigrama, trigrama, bigrama riječi i trigrama riječi za svaki jezik koji se nalazi u bazi poredanih na osnovu frekvencije učestalosti pojavljivanja. Pored svakog obrađenog podatka prikazan je omjer s ukupnim brojem znakova svih tekstova za pojedini jezik. Broj znakova naveden je kako bi korisnik imao referentnu vrijednost i okvir za daljnji rad, jer veća količina znakova u bazi znači veću preciznost podataka.

Izgled stranice koja prikazuje statistiku prikazan je na slici 16. Statistika ocrtava trenutno stanje baze, ovisno o broju i količini teksta koji je unešen i obrađen od početka objavljivanja aplikacije na internetu.

bigtri globalstat - top30

Jezik:	Albanski							
Omjer:		2225 /b./ - 2236 /zn./		372 /b.r./ - 2236 /zn./		2223 /t./ - 2236 /zn./		371 /t.r./ - 2236 /zn./
Statistika:	ë	0,04135	për të	0,01882	të	0,01574	transparencën e zgjedhjeve	0,00539
	P	0,02517	do të	0,01613	në	0,01215	është kulmi i	0,00539
	e	0,02427	nuk do	0,00806	të	0,01125	Balla ka deklaruar	0,00539
	r	0,02022	e zgjedhjeve	0,00538	ër	0,00945	ka deklaruar se	0,00539
	t	0,02022	ketë dhunë,	0,00538	për	0,00900	në transparencën e	0,00539
	të	0,01933	të ketë	0,00538	e	0,00855	e zgjedhjeve të	0,00539
	ar	0,01888	- deklaroi	0,00538	ë p	0,00765	zgjedhjeve të 28	0,00539
	n	0,01708	të gjitha	0,00538	pë	0,00765	të 28 qershorit	0,00539
	d	0,01618	të qenë	0,00538	në	0,00675	në seancën e	0,00539
	ër	0,01483	mori vendimin	0,00538	ar	0,00630	të ketë dhunë,	0,00539
	në	0,01348	i PS	0,00538	Par	0,00630	vendimin për të	0,00539
	sh	0,01213	Parlamentar i	0,00538	r t	0,00585	mori vendimin për	0,00539
	k	0,01213	qoftë për	0,00538	do	0,00585	Parlamentar i PS	0,00539
	i	0,01124	28 qershorit	0,00538	uar	0,00585	nuk do të	0,00539
	en	0,01124	dhunë, por	0,00538	Pa	0,00540	do të ketë	0,00539
	a	0,01124	deklaruar se	0,00538	ka	0,00495	për të qenë	0,00539
	t	0,00989	në seancën	0,00538	po	0,00495	ketë dhunë, por	0,00539
	e	0,00989	seancën e	0,00538	pr	0,00495	- deklaroi deputeti	0,00270

Slika 16 – Prikaz globalne statistike

Slika 17 prikazuje formulu po kojoj se izračunava frekvencija učestalosti pojavljivanja.

$$X_{st} = \frac{nX_o}{\sum X_z}$$

Slika 17 – Matematička formula za izračun frekvencijske učestalosti pojavljivanja

X_{st} je konačni statistčki izlazni podatak koji dobijemo nakon obrade dijeljenjem broja pojedinačnih elemenata sa ukupnim zbrojem znakova, gdje je nX_o broj

pojedinih bigrama, trigrama, bigrama riječi, odnosno trigrama riječi a $\sum X_z$ je zbroj znakova tekstova spremljenih u bazu.

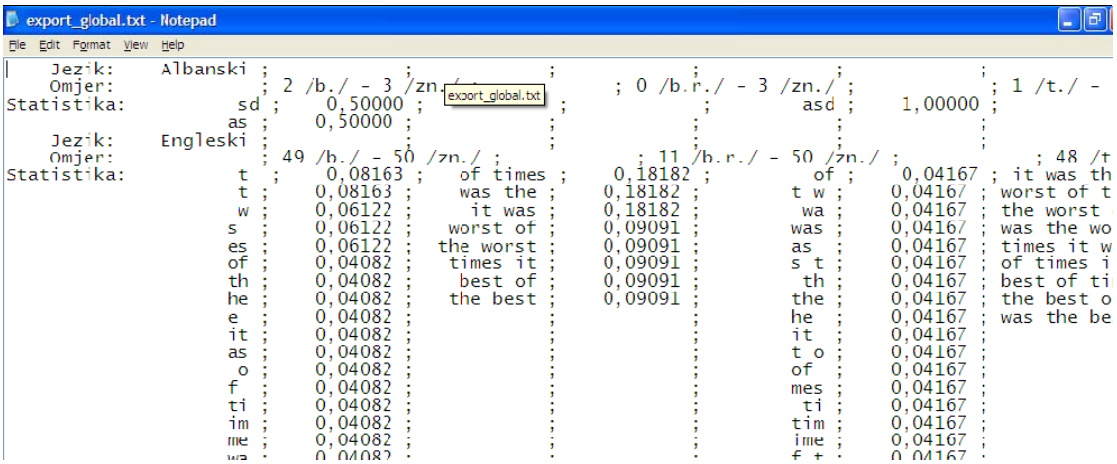
3.3.2. *Top30* vanjsko spremanje rezultata

Statističke podatke sa stranice moguće je spremiti u vanjsku tekstualnu datoteku i to pritiskom na gumb *Snimi u .txt* koji se nalazi na istoj stranici (vidi sliku 19). Nakon pritiska na gumb dobivamo datoteku početnog naziva *export_global.txt* u kojoj se nalaze podatci iz tablice koja je vidljiva na stranici (vidi sliku 18).

```
<asp:Button ID="btnSnimi" runat="server"
Text="Snimi u .txt"
onclick="btnSnimi_Click" />
```

Slika 18 – Spremanje statistike u vanjsku datoteku

Izgled tekstualne datoteke koja sadrži statistiku prikazan je na slici 19.



Slika 19 – Globalna statistika u tekstualnom obliku

Ispod gumba za spremanje u tekstualnu datoteku nalazi se i gumb *Izvezi u .xml* koji korisniku omogućava da cijelu bazu izveze kao vanjsku *xml*[15] datoteku. U na ovaj način izvezenoj *xml* datoteci korisnik ima cijelu bazu dostupnu na osobnom računalu. Bazu može učitati u odgovarajuće računalne aplikacije, te izvršavati željene operacije nad bazom bez potrebe ponovnog spajanja na internet. Ograničenje koje se pojavljuje kod ovakvog rada s bazom je mogućnost da se baza popuni novim podacima nakon što je korisnik spremio *xml* datoteku na računalo. Kako bi se izbjegli

[15] *xml* – kratica za eXtensible Markup Language - odnosno jezik za označavanje podataka

ovakvi slučajevi i kako bi korisnici u svakom trenutku imali posljednje stanje baze spremljeno na računalo, potrebno je što češće spremanje baze na računalo.

4. Usporedba s drugim aplikacijama

Usporedba mrežne aplikacije *bigtri* napravljena je s dvije aplikacije s različitim pristupom pojedinoj aplikaciji. Prva od ostale dvije aplikacije je mrežna aplikacija *Wolfram|Alpha* koja nudi mogućnost obrade bigrama, trigrama, bigrama riječi i trigrama riječi na sličan način. Druga aplikacija je *WordCreator* aplikacija koju korisnik pokreće lokalno s osobnog računala.

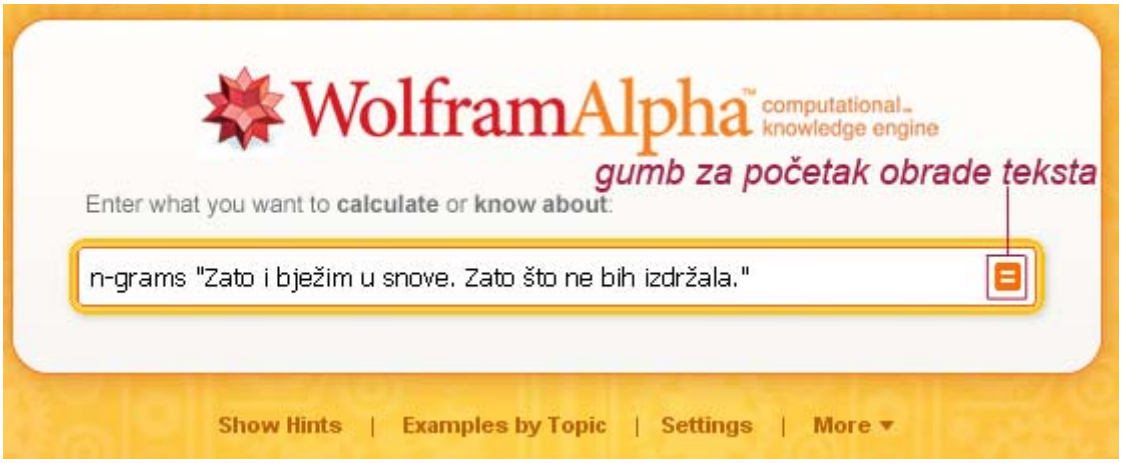
4.1. Wolfram|Alpha

Wolfram|Alpha je aplikacija razvijena vodeći se idejom kako bi cijelokupno sistematizirano znanje trebalo biti dostupno na što jednostavniji način. Razvoj aplikacije započeo je britanski fizičar i matematičar *Stephen Wolfram* poznat po tome što je razvio računalnu aplikaciju *Mathematica*. Zadaća aplikacije je odgovoriti korisnicima na upite koje unose u tekstualno polje, te prikazati izvore iz kojih su prikupljeni prikazani odgovori. Usporedbu aplikacije *Wolfram|Alpha*, *bigtri i WordCreator* započet ćemo od načina unosa teksta, te nastaviti redom do usporede dobivanja izlaznih podataka obrađenog teksta.

4.1.1. Unos i interpretacija

Aplikacija *Wolfram|Alpha* koju koristimo za usporedbu dostupna je na adresi www.wolframalpha.com. Pristup stranici omogućen je svim internet korisnicima u bilo koje vrijeme. Tekst koji želimo obraditi unosi se na drugačiji način nego što radimo u slučaju aplikacije *bigtri. Wolfram|Alpha* multifunkcionalna je stranica koja odgovara na složena pitanja korisnika, te je potrebno znati točnu formu unosa podataka kako bi ga stranica pravilno interpretirala. Ispred teksta koji želimo obraditi potrebno je napisati naredbu *n-grams,* koju interpreter razumije i na taj način zna u kojem obliku želimo povratnu informaciju. Forma unosa teksta je sljedeća - ***n-grams „*Zato i bježim u snove. Zato što ne bih izdržala.*“*** Ovako formatiran tekst unosi se u

vodoravna komponenta za unos teksta koja se nalazi na vrhu stranice kako je prikazano na slici 20.



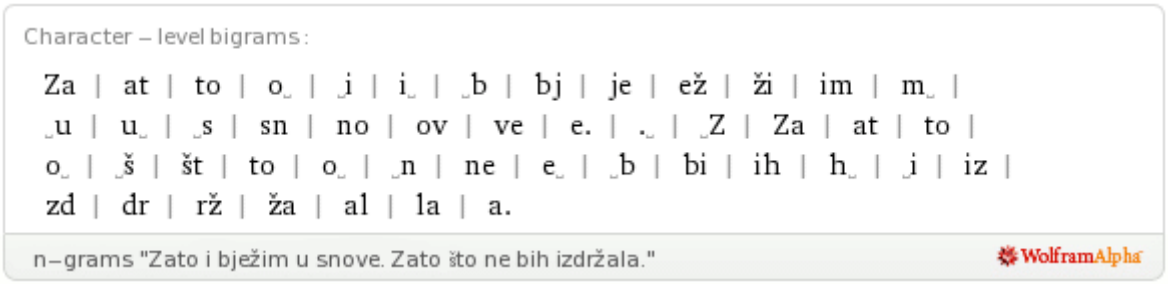
Slika 20 – *Wolfram|Alpha* početna stranica

4.1.2. Prikaz rezultata obrade teksta

Pritiskom na gumb „=“ koji je smješten uz desni rub *tekstualne komponente*, aplikacija izvršava obradu unešenog teksta. Nakon obrade teksta, obrađeni podaci podjeljeni su i grupirani u dva zasebna dijela.

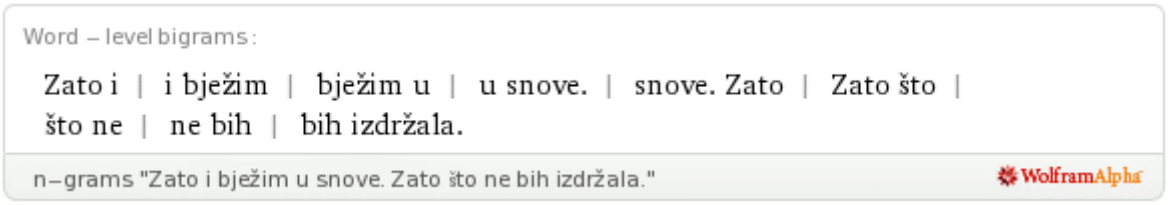
4.1.2.1. Prikaz bigrama

U prvom dijelu prikazuju se samo bigrami i to bigrami znakova i bigrami riječi. Na slici 21 vidljivi su bigrami nakon obrade.



Slika 21 – Bigrami dobiveni nakon obrade teksta

Slika 22 prikazuje ispis bigrama riječi.

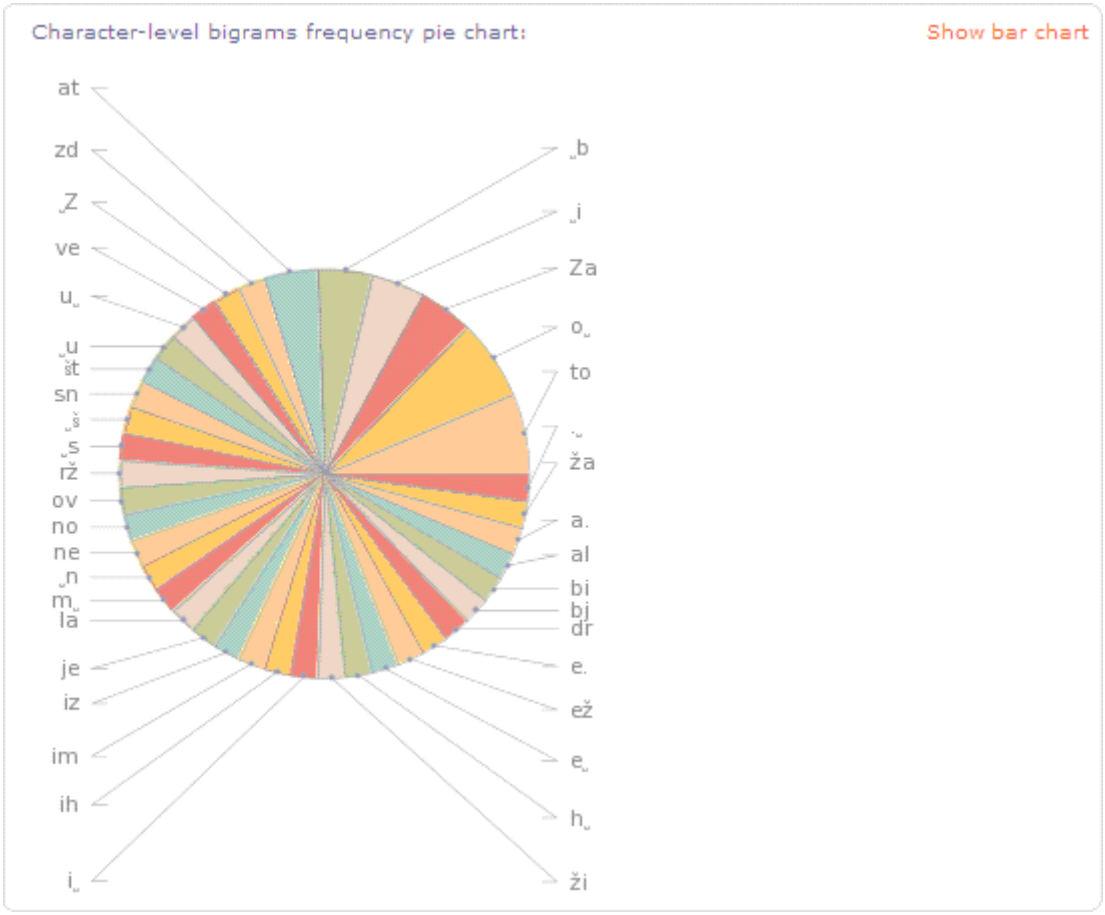


Slika 22 – Bigrami riječi dobiveni nakon obrade teksta

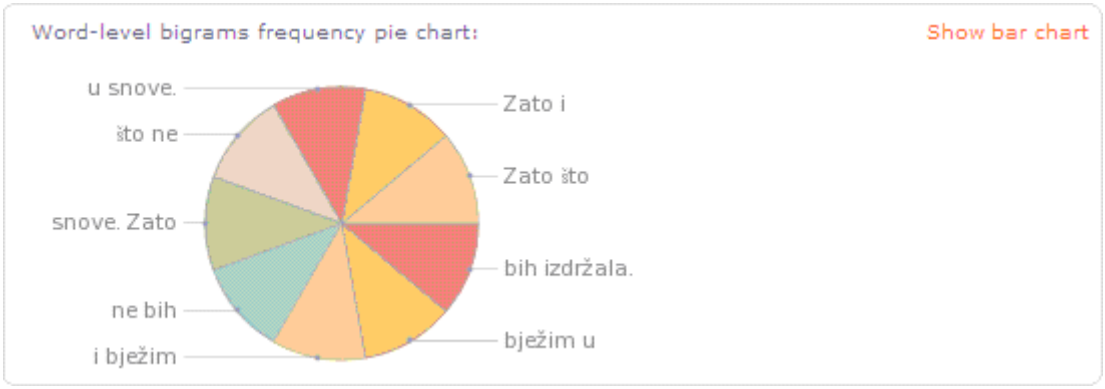
Ispod teksutalnog prikaza bigrama i bigrama riječi nalazi se detaljni grafički prikaz dobivenih podataka u obliku dijagrama. Dijagrami prikazuju broj pojavljivanja pojedinih bigrama, odnosno bigrama riječi u unešenom tekstu (njihov udio u cijelokupnom broju bigrama i bigrama riječi dobivenih iz obrađenog teksta). Gornji dijagram prikazuje broj pojavljivanja bigrama znakova (*Character-level bigrams)* u tekstu. Izgled dijagrama prikazan je na slici 23. Broj pojavljivanja bigrama riječi (*Word-level bigrams)* u tekstu prikazan je na slici 24.

Dijagrami su početno postavljeni u *pie*[16] oblik, no, moguće je iste te dijagrame pregledati u *bar*[17] obliku pritiskom na tekst *Show bar chart*[18].

[16] *pie* dijagram – dijagram u obliku pite
[17] *bar* dijagram – dijagram u obliku štapova (*štapičasti dijagram*)
[18] *Show bar chart* – prikaži *bar* dijagram



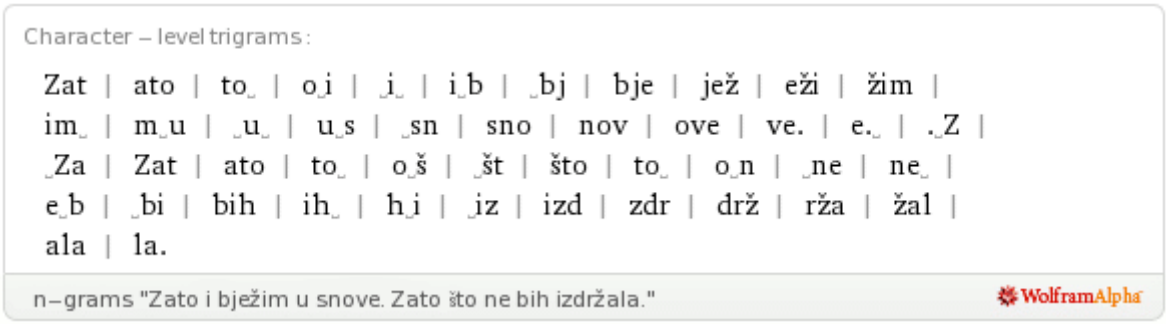
Slika 23 – Dijagram bigrama obrađenog teksta



Slika 24 – Dijagram bigrama riječi obrađenog teksta

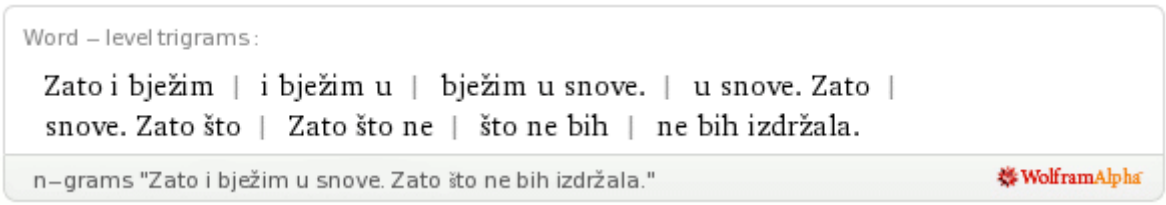
4.1.2.2. Prikaz trigrama

Drugi dio obrađenih podataka, koji se nalazi ispod pregleda bigrama identičan je rasporedom gornjem dijelu u kojemu su prikazani bigrami. Razlika u odnosu na gornji dio je u tome što se ovdje nalaze ispisani obrađeni trigrami i trigrami riječi. Ispis trigrama prikazan je na slici 25.



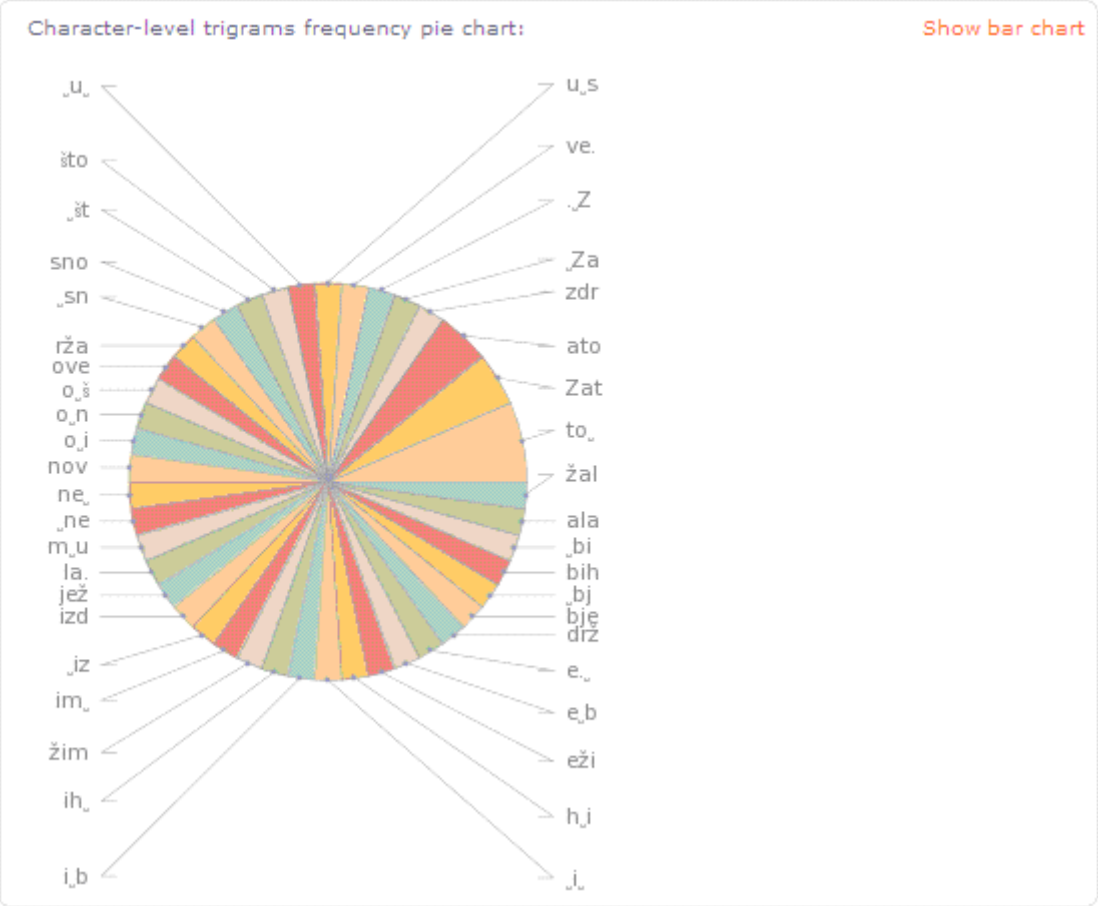
Slika 25 – Trigrami dobiveni nakon obrade teksta

Slika 26 prikazuje ispis trigrama riječi nakon obrade teksta.



Slika 26 – Trigrami riječi dobiveni nakon obrade teksta

Dijagram pojavljivanja trigrama prikazan je na slici 27, dok je na slici 28 prikazan dijagram pojavljivanja trigrama riječi. Kao što se može vidjeti, dijagrami su također početno postavljeni u *pie* oblik, te ih je moguće pregledati u *bar* obliku pritiskom na tekst *Show bar chart*



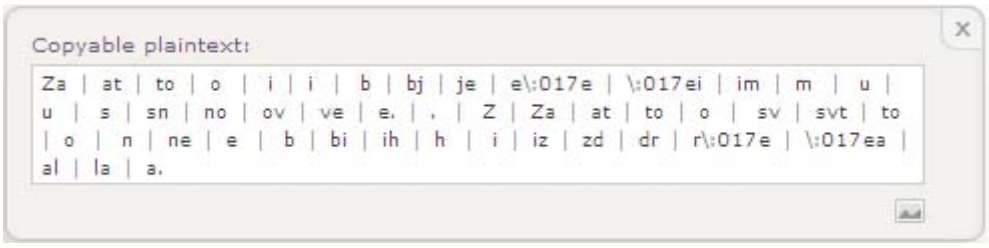
Slika 27 – Dijagram trigrama obrađenog teksta

Slika 28 – Dijagram trigrama riječi obrađenog teksta

4.1.3. Vanjska obrada rezultata obrade teksta

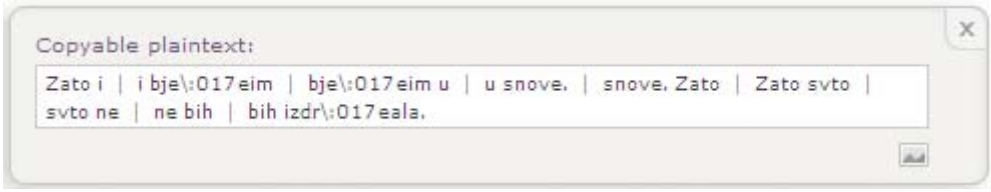
Dobivene izlazne podatke koji su prikazani na stranici nakon obrade podataka može se dobiti i u tekstualnom obliku, te kao takve koristiti lokalno za daljnju obradu. Kako bismo podatake dobili u tekstualnom obliku potrebno je kliknuti na svaku od

slika s ispisanim bigramima, odnosno trigramima. Izgled prozora s bigramima i trigramima identičan je. Razlika je jedino u znakovima koji su ispisani u njima. Tekst, odnosno znakove koje dobijemo posebno u svakom od prozora, jednostavno je označiti i kopirati te ih koristiti u daljnjem radu. Slika 29 prikazuje prozor s bigramima.



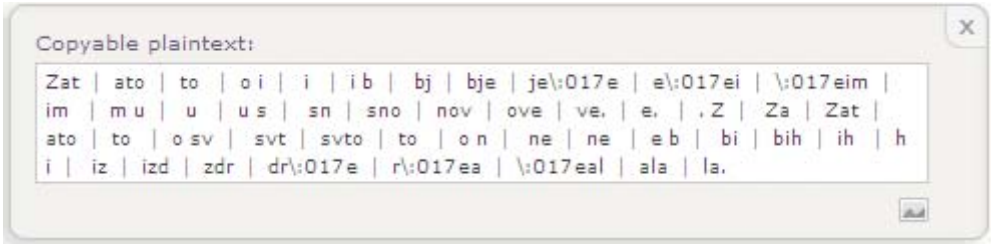
Slika 29 – Prozor s bigramima

Slika 30 prikazuje prozor koji sadrži bigrame riječi.



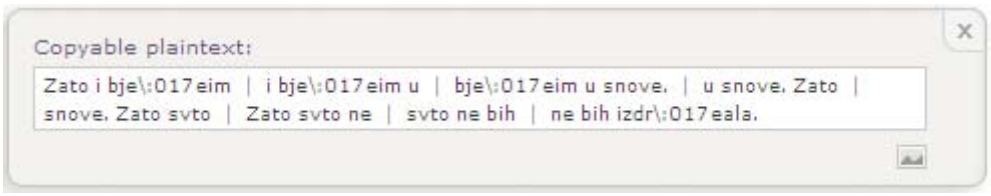
Slika 30 – Prozor s bigramima riječi

Slika 31 prikazuje prozor koji sadrži trigrame



Slika 31 – Prozor s trigramima

Slika 32 prikazuje prozor koji sadrži trigrame riječi.



Slika 32 – Prozor s trigramima riječi

Prikaz hrvatskih znakova (palatala) nije ispravan zbog korištenja UTF tablice koja podržava samo *Central European* kodiranje znakova. Korištenjem druge tablice znakova (fontova) koja u sebi ima sadržane palatale ovakva situacija bila bi izbjegnuta.

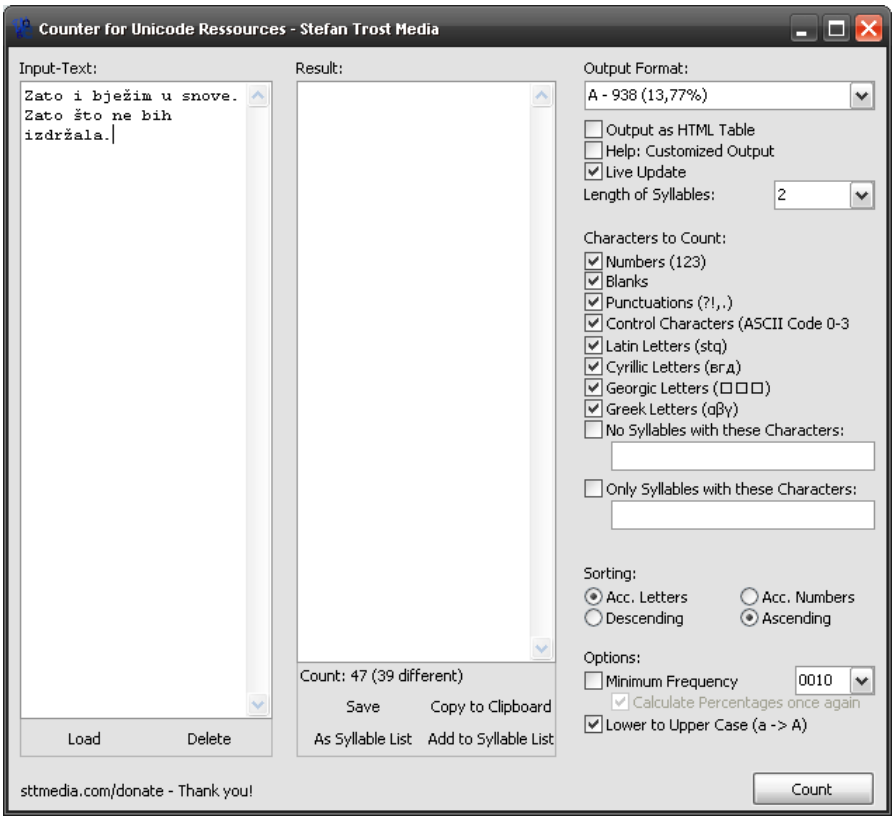
4.2. WordCreator

WordCreator je aplikacija koja se koristi lokalno na računalu u svrhu generiranja slučajnih nizova riječi, rečenica, te kompletnih tekstova sa subjektom i objektom. Osim generiranja, aplikacija pruža mogućnost brojanja riječi i rečenica, te prikazivanja statističkih podataka prebrojanog teksta ovisno o potrebama korisnika.

Aplikaciju *WordCreator* potrebno je snimiti na lokalni disk s internetske adrese http://www.sttmedia.com/*WordCreator*. Nakon snimanja potrebno je samo kliknuti na izvršnu datoteku i pokrenuti aplikaciju. *WordCreator* nije zahtjevna što se tiče konfiguracije računala, te ju je moguće u koristiti bez gubitka funkcionalnosti i na starijim računalima.

4.2.1. Unos i interpretacija

Tekst koji se obrađuje u aplikaciji *WordCreator* unosi se u lijevu komponentu za unos teksta iznad koje piše *Input-Text.* Unos teksta moguć je upisivanjem znaka po znaka putem tipkovnice, ili kopiranjem teksta s bilo koje druge lokacije. Kao primjer obrade teksta korištena je identična rečenica kao i u prethodne dvije aplikacije *„Zato i bježim u snove. Zato što ne bih izdržala.“* Na slici 33 prikazan je pravilno unešen tekst u aplikaciju.



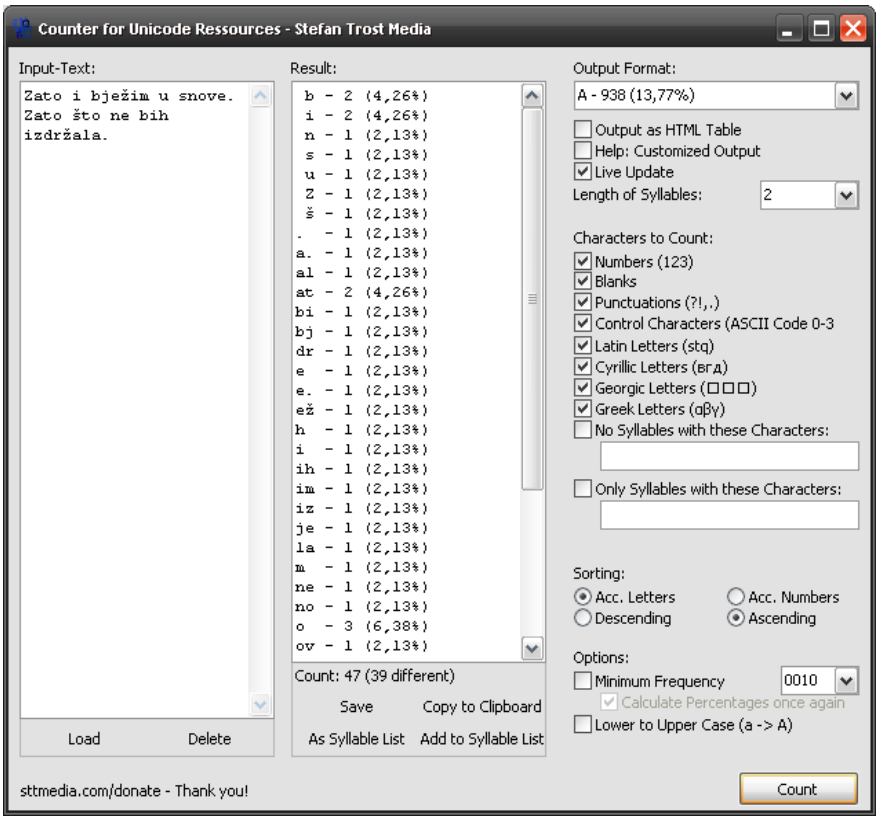
Slika 33 – Unos teksta u *WordCreator* aplikaciju

4.2.2. Prikaz rezultata obrade teksta

Prije ispisivanja rezultata obrade teksta potrebno je izabrati na koji način želimo prikazati podatke, te koje znakove želimo brojiti, te u kojem broju (bigrame ili trigrame). Prilikom prikaza podataka moguće je izabrati želimo li rezultate prikazati samo u brojčanom obliku i to broj pojavljivanja pojedinih elemenata, ili želimo uz taj broj dodati i postotak pojavljivanja jedinice u tekstu. Duljina znakova koju prebrojavamo bira se iz padajućeg izbornika. Sve navedene opcije prikazane su uz desni rub prozora. U primjeru koji je prikazan na slici 34 iz padajućeg izbornika izabran je broj 2, što znači da želimo prikazati bigrame. Nadalje dolazimo do opcija za brojanje znakova teksta, te označavamo sve opcije prebrojavanja i to redom brojeve, prazna polja, rečenične znakove (točka, upitnik, uskličnik...), kontrolne znakove itd.

Rezultate dobivamo pritiskom na gumb *Count* koji se nalazi u donjem desnom dijelu prozora. Rezultati obrade ispisuju se u komponentu za unos teksta koja se nalazi u središnjem dijelu prozora iznad koje piše *Result.* Rezultate je moguće

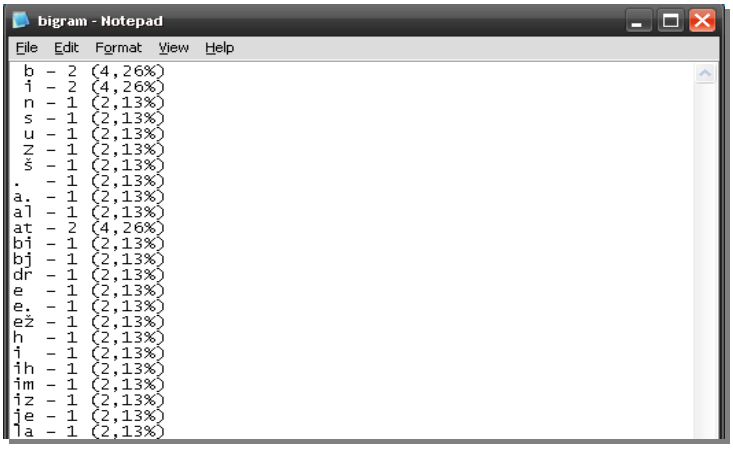
pregledavati, označavati i spremiti u memoriju računala ukoliko ih želimo koristiti u drugim aplikacijama.



Slika 34 – Rezultat obrade teksta *WordCreator* aplikacijom

4.2.3. Vanjska obrada rezultata obrade teksta

Dobivene rezultate obrade teksta moguće je spremiti u vanjsku datoteku, te ih kao takve koristiti u drugim aplikacijama, ili učitati u *WordCreator* aplikaciju. Podaci se spremaju pritiskom na gumb *Save*. Izgled datoteke nakon spremanja na računalo i otvaranja u pripadajućoj aplikaciji prikazan je na slici 35.



Slika 35 – Rezultati obrade teksta *WordCreator* aplikacijom u tekstualnom obliku

5. Ocjena razlika

5.1. Unos teksta

Tekst se u sve tri aplikacije unosi preko komponente za unos teksta. Aplikacija *bigtri* ima polja za unos teksta veće nego u slučaju druge dvije aplikacije koje su spmenute u tekstu. Objašnjenje za to veoma je jednostavno.

Aplikacija *bigtri* specijalizirana je za unos veće količine teksta, te je stoga očekivano da će polje u koje se unosi tekst biti veće i preglednije.

Wolfram|Alpha je multifunkcionalna mrežna stranica i aplikacija koja je izrađena kako bi izvršavala mnogo različitih upita, te je stoga limit postavljen na unos teksta. Unos teksta vrši se preko komponente za unos teksa koja može primati samo jedan red znakova, što ga čini nepreglednim ako se radi o dužem tekstu koji želimo obraditi. Prilikom upisivanja teksta moramo još uzeti u obzir i naredbu koja se mora upisati prije samog teksta (*n-grams*), što dodatno može zbuniti korisnika.

WordCreator aplikacija ima izduženo okomito polje za unos teksta koje je također veoma pregledno prilikom unosa teksta, što ga čini preglednijim nego je to slučaj kod *Wolfram|Alpha-e*.

5.2. Prikaz podataka obrade teksta

Prikaz podataka razlikuje se u svakoj od tri aplikacije. U *Wolfram|Alpha* aplikaciji tekst se prikazuje tekstualno i grafički. Grafički prikaz odnosi se na dijagrame koji su prikazani na slikama 23, 24, 27 i 28.

Aplikacija *bigtri* koristi samo tekstualni prikaz podataka kao i *WordCreator,* dok je kod ovog posljednjeg moguće izlazne podatke formatirati i u *html* kôd ukoliko je podatke potrebno objaviti na internetu.

U ovom slučaju prednost je na strani *Wolfram|Alpha*-e, jer grafički prikaz daje jasniji uvid u omjere i količinu pojedinih bigrama, odnosno trigrama u obrađenom tekstu.

5.3. Brzina obrade teksta

Brzinu obrade podataka možemo spomenuti u kontekstu najvažnije komponente prilikom obrade podataka. Pravilna usporedba moguća je samo ukoliko uspoređujemo aplikacije koje su dostupne na internetu, jer je nemoguće ostvariti bržu obradu nego što je to u slučaju aplikacije koja se koristi lokalno na računalu, no usporedbe radi, navest ćemo podatke i za brzinu *WordCreator* aplikacije.

U ovom segmentu aplikacija *bigtri* u velikoj je prednosti ispred *Wolfram|Alpha*–e. Mjerenja su vršena dodatkom za internetski pretraživač *Mozilla Firefox.* Prilikom pritiska gumba za obradu podataka (neovisno radi li se o *bigtri* ili *Wolfram|Alpha* aplikaciji), instalirani dodatak mjeri koliko je vremena prošlo do prikaza obrađenih podataka. Mjerenje je obavljano unutar vremenskog perioda od 60 minuta kako bi se na što bolji način izbjegla mogućnost eventualnog zagušenja internetske linije ili servera na kojemu se nalazi pojedina od aplikacija. U tablici 4 prikazana je prosječna brzina obrade podataka svih triju aplikacija. Usporedba je relevantna samo za aplikacije koje su dostupne na internetu. Jasno je vidljivo kako je brzina obrade 10.5 puta veća ukoliko koristimo *bigtri* aplikaciju.

Aplikacija	1	2	3	4	5	6	7	8	9	10	Prosjek
bigtri	1.208s	0.811s	0.881s	0.897s	0.936s	0.904s	0.88s	0.849s	0.817s	0.931s	0.911s
WAlpha	10.121s	10.31s	9.76s	9.229s	9.127s	9.995s	9.385s	9.104s	9.067s	10.274s	9.638s
WCreator	0.1s	0.1s	0.1s	0.1s	0.1s	0.1s	0.1s	0.1s	0.1s	0.1s	0.1s

Tablica 4 – Rezultati mjerenja brzine obrade podataka

5.4. Vanjska obrada rezultata

Podaci dobiveni obradom teksta koriste se ne samo kao konačna statistika nego i kao podaci kojima se želi nakon obrade koristiti lokalno na računalu.

Kako sam naveo ranije u tekstu, *Wolfram|Alpha* nema ugrađenu mogućnost statističkog pamćenja obrađenih podataka jer nije specijalizirana za takav način obrade teksta. Podatke koji su ispisani na stranici nakon obrade lako postaju dostupni korisniku, no format u kojemu ih stranica ostavlja, te nepotpunost obrade ako se radi o palatalima, uvelike otežava obradu podataka, dok aplikacija *bigtri* nema problema sa prikazivanjem palatala. Razlika je osim palatala, u prikazu leksičkih jedinica.

Kod *Wolfram|Alpha*-e se bigrami, koje sam koristio u ovom primjeru, odvajaju razmacima i znakom "|", dok aplikacija *bigtri* koristi apostrofe prije i poslije bigrama.

WordCreator aplikacija koristi identičan princip vanjskog prikaza podataka kao i *bigtri* aplikacija, jedino su bigrami, odnosno trigrami poslagani okomito svaki u novi red. Primjer izgleda teksta nakon obrade prikazan je u tablici 5.

Wolfram|Alpha	**bigtri**	**WordCreator**
Zato i bježim u snove. Zato što ne bih izdržala.	**Zato i bježim u snove. Zato što ne bih izdržala.**	**Zato i bježim u snove. Zato što ne bih izdržala.**
Za | at | to | o | i | i | b | bj | je | e\:017e | \:017ei | im | m | u | u | s | sn | no | ov | ve | e. | . | Z | Za | at | to | o | sv | svt | to | o | n | ne | e | b | bi | ih | h | i | iz | zd | dr | r\:017e | \:017ea | al | la | a.	'Za' - 2 / 85 'at' - 2 / 64 'to' - 3 / 81 'o ' - 3 / 175 ' i' - 2 / 89 'i ' - 1 / 182 ' b' - 2 / 86 'bj' - 1 / 15 'je' - 1 / 163 'ež' - 1 / 19 'ži' - 1 / 29 'im' - 1 / 33 'm ' - 1 / 57 ' u' - 1 / 70 'u ' - 1 / 113 ' s' - 1 / 133 'sn' - 1 / 21 'no' - 1 / 65 'ov' - 1 / 54 've' - 1 / 50 'e.' - 1 / 24 '. ' - 1 / 62 ' Z' - 1 / 81 ' š' - 1 / 33 'št' - 1 / 36 ' n' - 1 / 127 'ne' - 1 / 59 'e ' - 1 / 214 'bi' - 1 / 57 'ih' - 1 / 36 'h ' - 1 / 35 'iz' - 1 / 34 'zd' - 1 / 18 'dr' - 1 / 24 'rž' - 1 / 21 'ža' - 1 / 18 'al' - 1 / 63 'la' - 1 / 67 'a.' - 1 / 36	b - 2 (4,26%) i - 2 (4,26%) n - 1 (2,13%) s - 1 (2,13%) u - 1 (2,13%) Z - 1 (2,13%) š - 1 (2,13%). - 1 (2,13%) a. - 1 (2,13%)al - 1 (2,13%) at - 2 (4,26%)bi - 1 (2,13%) bj - 1 (2,13%)dr - 1 (2,13%) e - 1 (2,13%)e. - 1 (2,13%) ež - 1 (2,13%)h - 1 (2,13%) i - 1 (2,13%)ih - 1 (2,13%) im - 1 (2,13%)iz - 1 (2,13%) je - 1 (2,13%)la - 1 (2,13%) m - 1 (2,13%)ne - 1 (2,13%) no - 1 (2,13%)o - 3 (6,38%) ov - 1 (2,13%)rž - 1 (2,13%) sn - 1 (2,13%)to - 3 (6,38%) u - 1 (2,13%)ve - 1 (2,13%) Za - 2 (4,26%)zd - 1 (2,13%) št - 1 (2,13%)ža - 1 (2,13%) ži - 1 (2,13%)

Tablica 5 – Izgled teksta nakon obrade u svakoj od tri aplikacije

5.5. Konačna ocjena

Komparacija aplikacija donesena je na osnovu polja za unos teksta, prikaza podataka, brzine obrade, te mogućnosti vanjske obrade dobivenih podataka. Analiza je obavljena iz perspektive kako nisu svi parametri jednake vrijednosti i važnosti za korisnika.

Unos teksta, te mogućnost obrade podataka nakon obrade okarakterizirat ćemo kao najvažnije dijelove procesa obrade teksta, te će se njihov status uzimati kao presudan prilikom davanja završne ocjene. U tablici 6 prikazane su prednosti, odnosno mane koji se odnose na funkcionalnost aplikacija.

Aplikacija	*Wolfram|Alpha*	*bigtri*	*WordCreator*
Unos teksta	-	**+**	**+**
Prikaz podataka	**+**	-	-
Brzina obrade podataka	-	**+**	**+**
Vanjska obrada podataka	-	**+**	**+**
Konačna ocjena	1 / 4	**3** / 4	**3 / 4**

Tablica 6 – Ocjena aplikacija

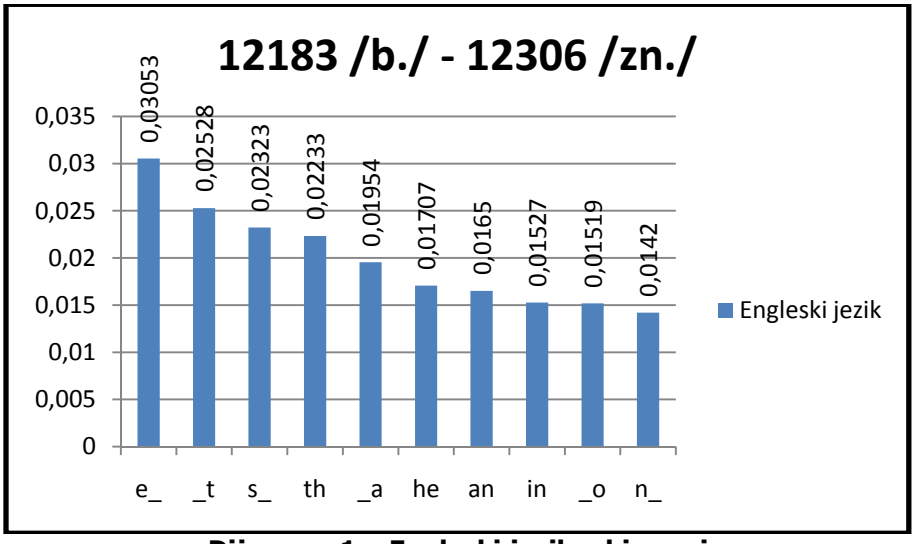
Prikaz podataka jedini je plus koji je *Wolfram|Alpha* dobila kao prednost nad *bigtri* aplikacijom. Pozitivna ocjena rezultat je prikaza grafikona, te pregledniji prikaz teksta koji je vodoravno postavljen. Vodoravni prikaz teksta prvenstveno znači kako nije potrebno dodatno pomicanje stranice prema dolje (skrolanje).

Kategorije u kojima je *bigtri* bolji od *Wolfram|Alpha*-e samorazumljive su i objašnjene ranije u tekstu. Unos teksta pregledniji je i jednostavniji za što je zaslužan veći prostor za unos teksta kao i nepostojanje dodatne naredbe kako bi aplikacija znala što s tekstom koji je unešen. Brzina obrade podataka 10.5 puta je veća kod *bigtri* nego kod *Wolfram|Alpha*-e. Dijelom je to zbog dijagrama i ostalih grafičkih prikaza koji se mogu vidjeti na stranici. Vanjska obrada dobivenih podataka svakako je bolja i kvalitetnija nakon obrade teksta *bigtri* aplikacijom. Obrađeni tekst moguće je jednostavno spremiti na računalo u tekstualnu datoteku koju aplikacija sama generira, što je veoma jednostavno i pregledno, dok kod *Wolfram|Alpha*-e to nije slučaj.

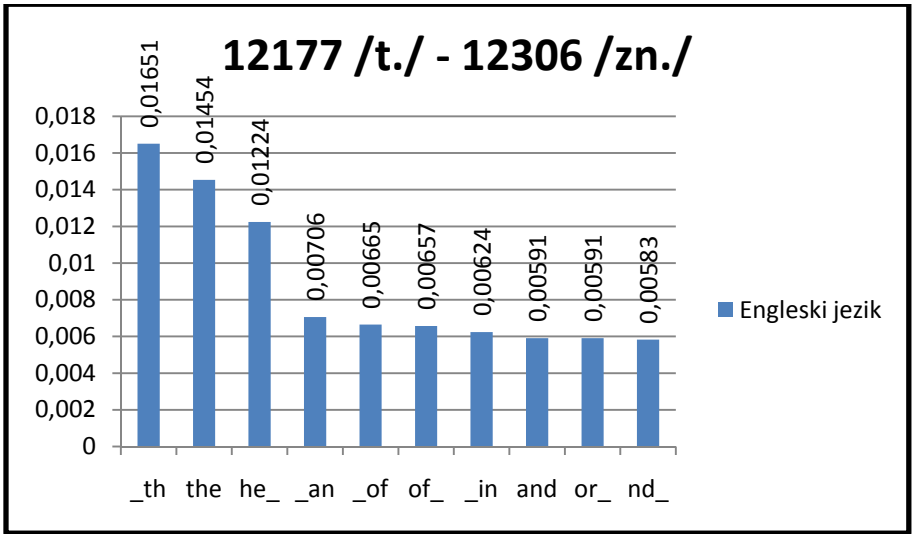
Konačna ocjena pokazuje kako ipak specijalizirana aplikacija kao što je *bigtri* odnosi svojevrsnu pobjedu i pokazuje se kao kvalitetnije rješenje u obradi jezika i raščlanjivanje na bigrame, bigrame riječi, trigrame i trigrame riječi. *WordCreator* aplikaciju nije bilo moguće konačno pozitivno ocjeniti jer nema mogućnost obrade bigrama riječi ili trigrama riječi.

6. Rezultati pojavljivanja bigrama i trigrama

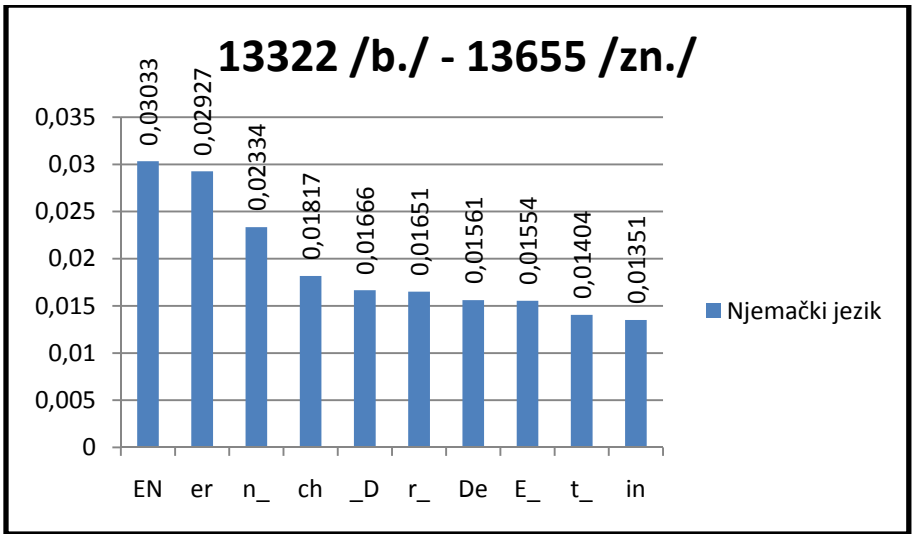
Kao primjer navedeni su rezultati pojavljivanja bigrama i trigrama za hrvatski, engleski i njemački jezik. Tekst koji je referentan za statistički prikaz kopiran je sa mrežnih izdanja novina za svaki jezik. U slučaju hrvatskog jezika tekst je kopiran sa internet stranice www.jutarnji.hr. Tekst na engleskom jeziku kopiran je sa www.thesun.co.uk, dok je tekst na njemačkom jeziku kopiran sa www.bild.de.



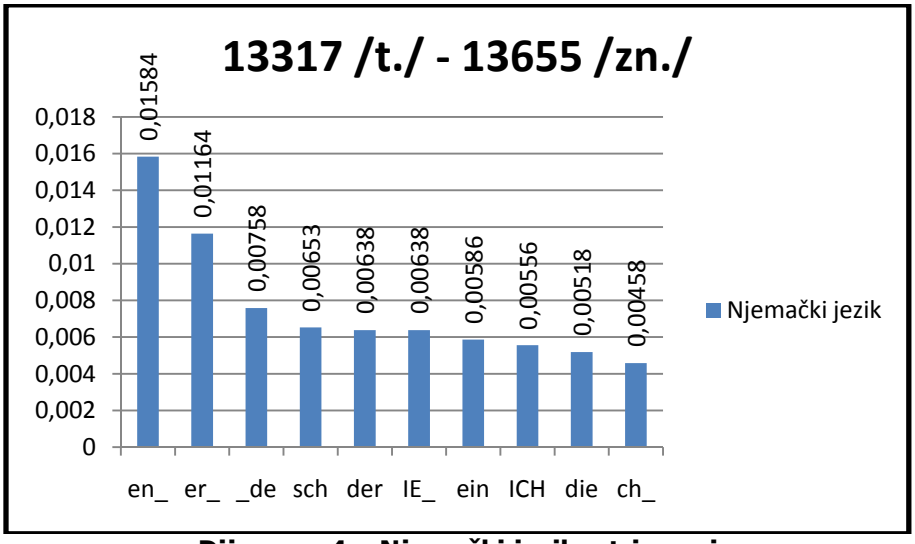
Dijagram 1 – Engleski jezik – bigrami



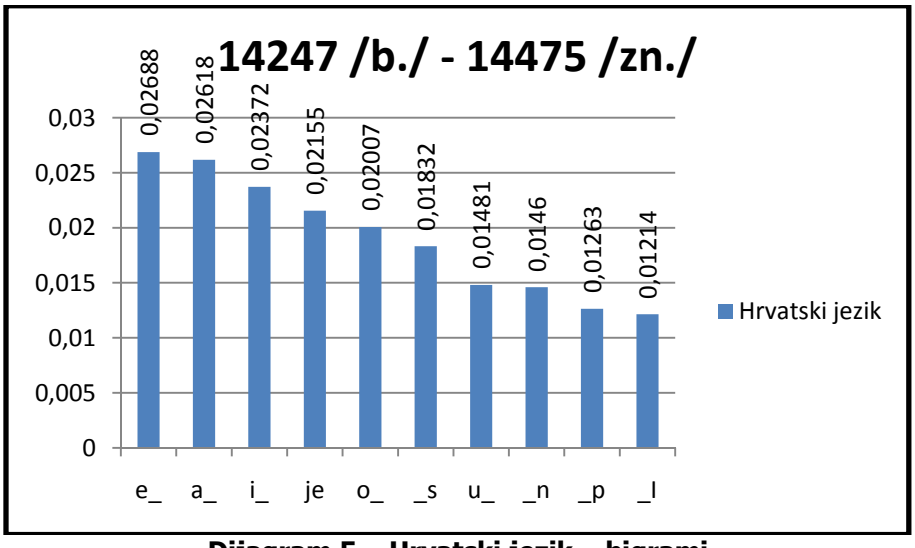
Dijagram 2 – Engleski jezik – trigrami



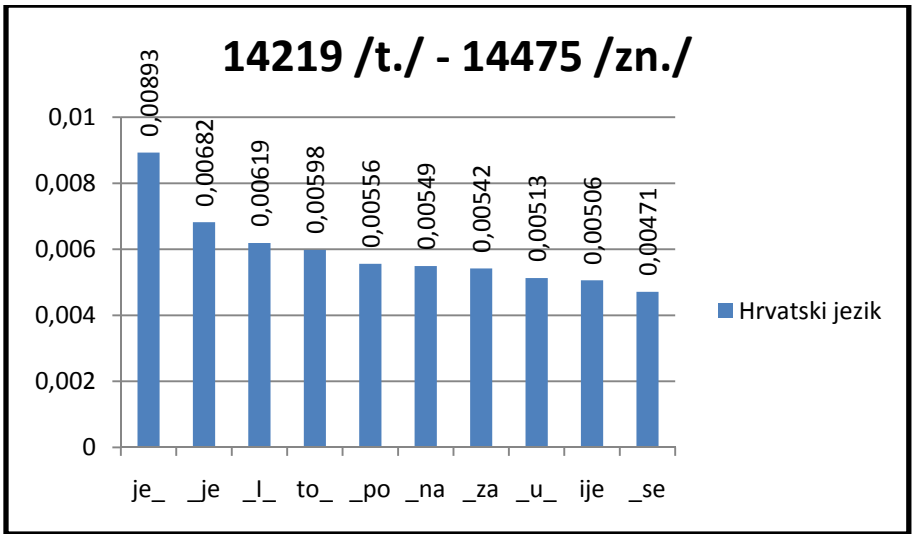
Dijagram 3 – Njemački jezik – bigrami



Dijagram 4 – Njemački jezik – trigrami



Dijagram 5 – Hrvatski jezik – bigrami



Dijagram 6 – Hrvatski jezik – trigrami

7. Zaključak

Izrada aplikacije koja se nalazi pred nama rezultat je, kako sam naveo u uvodu, želje za postojanjem aplikacije ovakvoga tipa na hrvatskom jeziku. Od početka izrade cilj je bio pojednostaviti korištenje i funkcionalnost aplikacije, pri čemu sam naišao na nekoliko problema i poteškoća sa spremanjem podataka u bazu i pravilnim označavanjem, te identificiranjem podataka vezanih uz jezik i tekst koji se spremaju u tablice. Uz pomoć literature i mentora, aplikacija je s vremenom dobila punu funkcionalnost, te je kasnije dodana i mogućnost izdvajanja bigrama riječi, te trigrama riječi, kao komponenti koje su prisutne u jeziku, no ne analiziraju se na adekvatan način.

Rezultati koji su dobiveni postupnom nadogradnjom programa bili su ohrabrujući, te je odlučeno kako bi bilo dobro pružiti korisnicima punu podršku i omogućiti im da mogu kompletnu bazu imati na svom računalu, tako što je omogućen *izvoz* baze u *xml* formatu. Koristeći napredne razvojne alate za izradu internetskih aplikacija kreirana je aplikacija spremna udovoljiti svim zahtjevima internet standarda, kako sigurnosnih, tako i dizajnerskih.

Naposlijetku, statistički rezultati koji su dobiveni obradom bigrama i trigrama govore da je aplikacija u potpunosti referentna i da svi rezultati mogu biti korišteni u analizi jezika kakva je potrebna na najvišem znanstvenom nivou. Vjerujem da će aplikacija koju imamo pred sobom pomoći onima koji se bave ovakvom analizom jezika.

8. Popis literature

1. Jurafsky, Daniel & Martin, James H. : *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Prentice-Hall, 2000.
2. Manning, Christopher D. & Schütze, Hinrich : *Foundations of Statistical Natural Language Processing,* MIT Press, 1999.
3. Rabiner, Lawrence R. & Juang, B. H. : *An Introduction to Hidden Markov Models,* IEEE ASSP Magazine: 4 – 16, 1986.
4. Rabiner Lawrence: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,* Proceedings of the IEEE 77 (2): 257 – 286, 1989.
5. Slonneger Kenneth & Kurtz Barry L. : *Formal Syntax and Semantic of Programming Languages,* Addison – Wesley Publishing Company Inc., 1995.

Popis internet referenci:

- Wikipedia, n-gram, dostupno na Internet adresi http://en.wikipedia.org/w/index.php?title=N-gram&oldid=367088316 08.06.2010.
- Wikipedia, Skriveni Markovljev model, dostupno na Internet adresi http://en.wikipedia.org/w/index.php?title=Hidden_markov_model&oldid=175069346 , 08.06.2010.
- Wikipedia, Claude Shannon, dostupno na Internet adresi http://en.wikipedia.org/w/index.php?title=Claude_Shannon&oldid=366047936 , 08.06.2010.
- Jutarnji List, dostupno na Internet adresi http://www.jutarnji.hr , 14.05.2010.
- The Sun, dostupno na Internet adresi http://www.thesun.co.uk , 14.05.2010.
- Das Bild, dostupno na internet adresi http://www.bild.de , 14.05.2010.