FILOLOGIJA 59, Zagreb 2012

UDK 81'324:81'374.823
Izvorni znanstveni članak
Primljen 20.VII.2010.
Prihvaćen za tisak 24.I.2011.

Tomislav Stojanov
Zoran Vučić
Institut za hrvatski jezik i jezikoslovlje
Ulica Republike Austrije 16, HR-10000 Zagreb
tstojan@ihjj.hr

KORPUSNOJEZIKOSLOVNA OBRADBA TEKSTOVA *SPORTSKIH NOVOSTI* — *n*-gramsko modeliranje dohvaćanja podataka i vizualizacija

U radu se propitkuje uloga korpusa za jezikoslovna istraživanja i testiraju sučelja dvaju hrvatskih korpusa, *Philologic* i *Bonito*, za jezične upite na razini dokumenta i sadržaja, prikazbe te znakova i forme. Za specijalizirane jezikoslovne pretraživačke upite izgradili smo sportsku novinsku bazu internetskih tekstova *Sportskih novosti* od travnja 2008. do srpnja 2009. godine (3,6 milijuna pojavnica).
Pokazat će se računalni postupci dohvaćanja teksta, *n*-gramski SQL/*regex*-upiti u cilju izvlačenja supojavnih čestotnica i otkrivanja frazema, naziva i stalnijih sintagmema, te njihova vizualizacija u prebirniku (*browseru*) uz pomoć javaskriptne biblioteke *WireIt*.
Ukazali smo da izgrađena metodologija može poslužiti za dobivanje jedinstvenih informacija za jezikoslovna istraživanja, te usporedili rezultate našega pristupa s tražilicom *Google* na osnovi kojih smo istaknuli sedam nedostataka rezultata *Googleovih* pretraživanja za jezikoslovna istraživanja.

1. Motivacija

Uloga korpusâ za jezikoslovna istraživanja (pravopisna, gramatička, leksička, terminološka i ina) važna je i nezaobilazna[1], a posvemašnjom "internetizacijom" jezika *web* se sve više prepoznaje kao "natkorpus" u kojem stručnjaci (a osobito javnost) traže svoje odgovore zaglédajući u

[1] Popis jezikoslovnih radova koji su nastali pretraživanjem korpusâ engleskoga jezika ili korpusnojezikoslovnim istraživanjima nalazi se na *web*-stranicama *http://icame.uib.no/icame-bib2.txt i http://icame.uib.no/icame-bib3.htm* (iz Meyer 2002:11).

103_129_Filologija_59_Stojanov_Vucic_5.indd 103 03.04.2013 19:15:24

izvore i dostupni sadržaj.[2]

Još od početaka korpusnoga jezikoslovlja jezikoslovci su razlikovali *jezičnu potenciju* od *jezične realizacije*, a jezični im je sustav (tj. potencija) bio polazišni teorijski okvir. Tako je i Chomsky u svojoj podjeli kompetentnosti i performantnosti (eng. *competence* i *performance*) iz 1965. istaknuo da jezikoslovna teorija primarno polazi od kompetentnosti.[3]

Zanimljivo, one koje je tada više zanimao jezični ostvaraj od jezičnoga sustava krenuli su putem računalnoga jezikoslovlja te se činilo da je ta podjela čvrsto utvrđena. Međutim, u trenutku kada su računala postala hardverski i softverski široko pristupačna i iskoristiva, jezični upiti[4], jezična istraživanja[5] te ključna jezikopisna (lingvografijska) djela počela su se temeljiti na proučavanju jezične performantnosti, a ne kompetentnosti: rječnici, gramatike i pravopisi dobar su primjer takvih priručnika čija se pravila i činjenice izvode na osnovi proučavanja jezične realizacije.

Mnogi računalno jezikoslovlje smatraju »paralelnim« jezikoslovljem, »vidom računalne obradbe tekstova« koji s jezikoslovljem kao takvim i nema osobite dodirne točke, međutim stava smo da je korpusno jezikoslovlje onoliko jezikoslovno koliko je svako interdisciplinarno područje ujedno i disciplinarno. Drugim riječima, korpusno jezikoslovlje smatramo **i (interdisciplinarnom) jezikoslovnom** disciplinom zaduženom za proučavanje jezične realizacije. To što ono proučava i statističke modele i zahtijeva poznavanje tehnologije samo govori o interdisciplinarnosti i isključuje pritom *jezičnu dimenziju*.

Polazeći od *weba* odnosno korpusâ tekstova kao mjestâ gdje pronalazimo građu za naša jezikoslovna istraživanja, postavlja se pitanje o njihovoj pretraživačkoj funkcionalnosti i stupnju adekvatnosti tražilice za odgovaranje na jezične upite te dohvaćanje informacija.

---

[2] »Language scientists and technologists are increasingly turning to the web as a source of language data, because other resources are not large enough (…)« (Kilgarriff 2003:1).

Primjer korištenja *weba* kao građe za jezikoslovnu tražilicu: *http://www.webcorp.org.uk/*.

[3] »Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-communication, who know its (the speech community's) language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of this language in actual performance.« (Chomsky 1965:3).

[4] Npr. brza pravopisna provjera preko javnih tražilica i utvrđivanje da je nešto vrlo vjerojatno pravopisno točno ako ima više nađenih rezultata.

[5] Npr. pronalaženje toponimâ, provjera sintaktičkih sveza, leksička ovjerenost, itd.

103_129_Filologija_59_Stojanov_Vucic_5.indd 104 03.04.2013 19:15:31

Dva su jezična korpusa hrvatskoga jezika — *Hrvatski jezični korpus* (HJK) u sklopu programa *Riznica* Instituta za hrvatski jezik i jezikoslovlje (*http://riznica.ihjj.hr/*) i *Hrvatski nacionalni korpus* (HNK) Zavoda za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu (*http://www.hnk.ffzg.hr/*) s temeljito različitim pretraživačkim sučeljima (mrežna stranica prilagođena sustava *Philologic* i *desktop* aplikacija *Bonito*).

Oba su dosegla prihvatljivu veličinu srednje velikih korpusa te uporabnu zrelost. Iako različiti — jedan je više digitalizacijski (institutski), a drugi više tehnologijski orijentiran (fakultetski) — i s fundamentalno različitom građom (preklapaju se tek u omanjem dijelu građe), oba korpusa sa svojim pristupima pretraživanju tekstova imaju slična ograničenja.

Osim želje za digitalizacijom tekstova i pohrambenih ciljeva, glavni razlog postojanja svakoga korpusa proizlazi iz potrebe za njegovim pretraživanjem. Ovisno o vrsti našega zanimanja za korpuse izgrađuje se pretraživač s prikladnim stupnjem sofisticiranosti.[6]

Sučelja za pretraživanje korpusâ *Bonito* i *Philologic* ne odgovaraju u našem slučaju proučavanja rezultata visokospecijaliziranih upita na trima razinama:

- na razini dokumenta i sadržaja:
  - pretraživanje ne omogućuje pronalaženje pojavnica koje se pojavljuju u istom dokumentu kao i neke druge pojavnice,
  - nije omogućeno ili nedovoljno razrađeno pretraživanje samo određenih dijelova dokumenata.[7]
- na razini prikazbe:
  - pretraživanje izlistava rezultate (u obliku odlomka odnosno retka ili KWIC-a[8]) ne dajući drugu vrstu informacije osim tekstno-kontekstualne,
  - pretraživanje ne grupira rezultate prema nekoj (statističkoj

[6] Vrlo zanimljivo poglavlje koje govori o specijaliziranim korpusima, ne zbog svoje veličine, teme ili nekim drugim sadržajnim ili formativnim kriterijima, već prema znanstvenom istraživanju nalazi se u Meyer 2002:103 u dijelu s naslovom »Determining whether a corpus is suitable for answering a particular research question«.

7 Značajka softvera Bonito iz izbornika View > Structures koja se omogućuje tipkom F6 mogla bi biti na tragu ove funkcionalnosti, ali u praksi ne funkcionira (npr. traženje samo naslova u Vjesnikovu potkorpusu). Način kako se to ipak može ostvariti jest sintaktički, primjerice, »glava« within <head> (iz *http://www.hnk.ffzg.hr/pretraga.html).* Šteta što upute za pretraživanje sa strukturnim oznakama u ovome dijelu nisu razrađenije i deskriptivnije jer tekstna struktura digitaliziranih djela u XML-u može biti vrlo razvedena u nekim žanrovima te ujedno zanimljiva za korpusno proučavanje.

[8] eng. *key word in context.*

103_129_Filologija_59_Stojanov_Vucic_5.indd 105 03.04.2013 19:15:31

ili drugoj) relevantnosti,
  - ograničeno ili nikakvo pretraživanje o svezama riječi.[9]
- na razini znakova i forme:
  - nije omogućeno pronalaženje nebrojčanih pojavnica koje sadržavaju jednu ili više točaka u sebi bez obzira na svoje medijalno mjesto,[10]
  - nije omogućeno pronalaženje trotočja iza kojih slijede mala slova,[11]
  - nije omogućeno pronalaženje određenoga unikodnog znaka,[12]
  - djelomična ili potpuna nemogućnost pronalaženja određenih razgodaka.[13]

Oba sučelja možemo nazvati "klasičnima" jer su uobičajena u korpusnome jezikoslovlju. Budući da su njihove pretraživačke specifičnosti i funkcionalnosti izrazito različite, držimo da bi oba hrvatska korpusa trebala imati omogućene obje te tehnologije pretraživanja čime bi znatno dobili na svojoj uporabnosti i kvaliteti.

U nastavku opisujemo izgradnju mikrokorpusa koji je poslužio za lingvističke i dohvadbene[14] potrebe. Opisat ćemo kakvi su to naši visokospecijalizirani upiti bili, kako smo izgradili korpus, koji je stupanj korisnosti pretraživanja građe te do kakvih smo rezultata došli.

2. Baza podataka kao specijalizirani korpus

Naš specijalizirani korpus i nije, zapravo, korpus — on je baza podataka — jer nismo ponudili pretraživanje punoga teksta (eng. *full-text search*), nego smo omogućili parcijalnu, *n*-gramsku pretragu.

Svako je jezično istraživanje posebno i sa specifičnim upitima zbog čega se i razlikuju korpusna pretraživačka sučelja, a područje izgradnje

[9] Značajkom softvera *Bonito* iz izbornika *Concordance* > *Statistics* > *Collocations* koja se omogućuje tipkama CTRL + L moguće je samo dobiti prikaz bigrama i to isključivo gdje je druga bigramska sastavnica ona zdesna. Opcijom proširivanja *n*-grama koju imamo u tom izborniku »*In the range from x to x*« ne dobivamo očekivane i iskoristive rezultate.

[10] Npr. pretraživanje prema regularnome izrazu [a-zA-Z]+\.[a-zA-Z]+.

[11] Npr. pretraživanje \.\.\.\s[a-z].

[12] Regularni izraz: \uxxxx.

[13] Sučelje *Philologic* uopće ne podržava razgotke, a *Bonito* samo neke (npr. sve znakove navodnika).

[14] Dohvadba: predloženi kraći naziv za disciplinu »dohvaćanje podataka« (eng. *information retrieval*).

103_129_Filologija_59_Stojanov_Vucic_5.indd 106 03.04.2013 19:15:31

jezikâ za korpusne upite (eng. *corpus query languages, CQL*) vrlo je živo.[15] Pretraživanje podataka u jezičnim bazama i korpusima temelji se na različitim modelima i algoritmima[16], a mi smo u ovome radu eksperimentirali sa SQL/*regex* pretraživanjem *n*-gramske baze podataka. Drugim riječima, kombinacijom korištenja SQL-a (eng. *structured query language*) i regularnih izraza (eng. *regular expressions*, *regex*) pokušali smo izgraditi jezično uporabive upite na »nižoj« razini želeći ispitati mogućnosti ovoga pristupa i korisnosti za daljnja (korpusno)jezikoslovna istraživanja.

Sintaksa SQL-a izuzetno je moćna u pretraživanju baza podataka, ali je svojom logikom ponajprije usmjerena na pretraživanje tablica i ima ograničenu podršku za tekstnu raščlambu i manipulaciju.[17] Za taj su aspekt specijalizirani regularni izrazi, nevezani za SQL, definirani ekspresivnošću regularnih gramatika u kontekstu formalnih jezika. Budući da je morfologija hrvatskoga jezika regularna (a na morfološkoj razini propitkujemo pojavnice), možemo se koristiti regularnim jezikom (točnije: PCRE provenijencije). Za sintaktičke su upite, pak, regularni izrazi vrlo ograničeni i slabo podržani među različitim sintaksama (prilagodbama, inačicama) regularnih jezika,[18] pa smo se za tu primjenu poslužili SQL-om i *n*-gramskim tablicama. Spajanjem regularnih izraza i upita u SQL-u, tj. kombinacijom »morfološkoga« i »sintaktičkoga« upita dobili smo infrastrukturu za puniju manipulaciju tekstom. Od godine 2005. s prvom bazom koja je ugradila podršku za regularne izraze u SQL sintaksu (*Oracle 10g*), pretraživanje tekstova dobilo je novu dimenziju.

3. Izgradnja testnoga korpusa *Sportskih novosti*

Testni korpus izgrađen je iz tekstova dnevnika *Sportske novosti* (*http://sportske.jutarnji.hr/*) uz pisano dopuštenje uredništva za čin preuzimanja mrežnih stranica za potrebe znanstvenoga rada.[19] Skidanje je tekstova bilo bitno olakšano zbog strukture URL adresa koje se međusobno razli-

[15] *http://citeseerx.ist.psu.edu/search?q=%22corpus+query+language%22.*

[16] Dobar uvod u enciklopedijskome članku *Computers in Lexicography* i *Corpora* u *Encycopedia of Language & Linguistics*, kao i u Biber & Conrad Reppen (1998).

[17] Pretraživanje teksta svodi se na dva osnovna operatora: znak postotka (%) koji zamjenjuje 0 ili više grafema — u regularnoj sintaksi isto značenje ima znak zvjezdice (*), te znak podvlake (_) koji zamjenjuje bilo koji grafem — u regularnoj sintaksi to je znak točke (.).

[18] Tako je, primjerice, sintaksa nama korisne funkcije traženja »gledanja uokolo« (eng. *lookaround*) vrlo slabo podržana.

[19] Na čemu uredništvu Sportskih novosti ovom prigodom i zahvaljujemo.

103_129_Filologija_59_Stojanov_Vucic_5.indd 107 03.04.2013 19:15:31

kuju samo po ID-u bez obzira na rubriku ili vrstu vijesti.[20] S portala je preuzeto 11.657 HTML datoteka koje predstavljaju mrežne vijesti objavljene od 23. travnja 2008. do 2. srpnja 2009. godine i koji veličinom obasežu 67 MB.[21] Kod za preuzimanje i pretvorbu sadržaja pisan je u C#-u kao *desktop* aplikacija *SnParser* za potrebe HJK-a i *Riznice* IHJJ-a. Izgled *SnParsera* moguće je vidjeti u prilogu 1.

U prilozima 2. i 3. vidljiv je put dokumenta od početne *web*-stranice i HTML-a do konačnoga izlaza u format TEI XML. U prilogu 4. vidi se pretraživost toga teksta u HJK-u.

Od cijele HTML datoteke izvučeni su samo tekstovni dijelovi koji se odnose na sadržaj (nadnaslov, naslov, podnaslov i tijelo vijesti), a sve je ostalo filtrirano. Za sada su još uvijek isključeni slikovni podnaslovi, ali njih planiramo vratiti u korisni dio sadržaja *web*-stranice.

4. Izgradnja mrežne i izvanmrežne baze podataka

U sljedećoj su se fazi svi TEI XML-ovi pretvorili u obične tekstualne datoteke radi *n*-gramske raščlambe i spremanja u baze podataka SQLite i MySQL za izvanmrežni i mrežni rad. Baza podataka SQLite[22] ima osobite značajke: riječ je o besplatnoj, otvorenoj, brzoj[23], transakcijskoj, bezinstalacijskoj, raširenoj bazi s velikom softverskom i korisničkom podrškom[24], koja omogućuje rad s bazom podataka izvanmrežno (eng. *serverless*), koja je prenosiva (eng. *portable*)[25], i, konačno, koja omogućuje regularne upite.

[20] Drugim riječima, sustav za upravljanje sadržajem (CMS) *Sportskih novosti* generira uniformne URL adrese.

[21] Preuzimanje sadržaja portala *Sportskih novosti* i pretvorba u TEI XML datoteke koji se koriste u HJK-u napravio je student Bruno Pavec u sklopu svoga diplomskoga rada »Pisanje i pretvorba HTML-teksta s *weba* u TEI XML« kod mentora T. Stojanova na Stručnome studiju informatike pri Tehničkome veleučilištu u Zagrebu, a uspješno je obranjen u svibnju 2009. godine i numeriran brojem 981.

[22] *http://www.sqlite.org/* .

[23] Prema našim izvorima SQLite je brži i od MySQL-a i od PostgreSQL-a, a znajući da ćemo raditi s velikim tablicama, složenijim regularnim upitima i kadšto nedovoljno optimiziranim podatcima, brzina nam je bila važan kriterij pri odabiru.

[24] Količina softvera kojim je moguće upravljati bazom (eng. *database management software*) impozantna je (dobar popis nalazi se na *http://www.sqlite.org/cvstrac/wiki?p=ManagementTools*), a čak se omogućuje i rad s SQLite-om preko prebirnika *Mozilla Firefox* za koji je napravljen dodatak *SQLite Manager Addon for Firefox* (*https://addons.mozilla.org/en-US/firefox/addon/5817/*). Rad s bazama nikada nije bilo lakši i dostupniji, a želja nam je i bila da se zainteresiranim lingvistima približi i pojednostavni izravno propitkivanje jezične baze podataka.

[25] Visoka prenosivost očituje se u neimanju potrebe za ikakvom instalacijom i činjenicom da je cijela baza u jednoj datoteci. Na taj smo način olakšali korisnicima pre-

103_129_Filologija_59_Stojanov_Vucic_5.indd 108 03.04.2013 19:15:31

Sav smo sadržaj pretočili u dvije baze i koristili ih na dva načina — izvanmrežni i mrežni; izvanmrežnim radom željeli smo postići veću brzinu obradbe bez nepotrebnoga opterećenja mrežnoga prometa i procesorske snage poslužioca u IHJJ-u, a što se pokazalo dobrom odlukom, dok nam je MySQL poslužio za potrebe eksperimentalnoga pretraživanja i vizualizacije *n*-grama preko interneta.

Osim dodatka (eng. *add-on*) za Firefox *SQLite Manager Addon for Firefox*, primarno smo rabili SQLiteSpy[26], softver koji nije iziskivao instalaciju na računalu, a koji se pokazao brzim, stabilnim i robusnim.

Za znanstvenoistraživačke potrebe demo-bazu moguće je preuzeti sa stranica IHJJ-a[27], a iznosi 1/11 pune baze, veličine 20,1 MB. Preuzeta baza može se slobodno koristiti i testirati. Jedanaestina predstavlja otprilike tisuću tekstova koje čine 181,8 tisuća pojavnica (eng. *token*) i 46,6 tisuća različnica (eng. *type*).[28]

Puna i indeksirana baza velika je 654 MB koju sadržava 2,4 milijuna pojavnica i 243 tisuće različnica. Demo-baza podataka MySQL obaseže 39 MB s indeksima. U prilogu 5. može se vidjeti struktura baze.

Sav je tekst bio podijeljen na nekoliko *n*-gramskih vrsta: unigrame, bigrame, trigrame i tetragrame, smatrajući da nam je kontekst od četiri pojavnice dovoljan. *N*-grami su sljedovi koji se odnose na različite entitete, od grafemâ do pojavnicâ (»riječî«), a u našem su to slučaju pojavnice. Pojavnice su definirane kao nizovi grafema omeđene znakovima razmaka.[29]

N-grami su osnovni statistički entiteti u računalnome jezikoslovlju, a nama su poslužili ne toliko za proučavanje čestotne relevantnosti i dru-

uzimanje demo-baze s naših internetskih stranica.

[26] *http://www.yunqa.de/delphi/doku.php/products/sqlitespy/index* .

[27] *http://ngrami.ihjj.hr/demoSN.sqlite.*

[28] U užemu korpusnom kontekstu, a vezano za grafemski prikaz i regularno pretraživanje, autori se kadšto koriste i nazivima *raznopisnica* za pojavnicu, odnosno *istopisnica* za različnicu.

[29] Mogli bismo za te nazive skovati hrvatske istovrijednice *n-rječje, jednorječje, dvorječje, trorječje, četvororječje*, i tako dalje, kada se entiteti odnose na pojavnice, a *n-slovlje, jednoslovlje, dvoslovlje, troslovlje, četveroslovlje*, i tako dalje, kada se entiteti odnose na grafeme. Njihovi hiperonimi bili bi *n-slovka, jednoslovka, dvoslovka, troslovka, četveroslovka*, i tako dalje.

Pritom naglašavamo razlikovanje značenja tvorbene osnove od semantičkoga polja *slova* i *riječi*. Bez obzira na to što slovo (točnije: slovka) razlikujemo od brojke, za potrebe tvorbe dopuštena je metonimizacija i generalizacija, pa tako *slovo* može podrazumijevati i *brojku* (na isti način kao što dizalica spušta osim što diže).

Nadalje, uvažavajući dihotomiju izgleda (eng. *layout*) i sadržaja (eng. *content*), terminološka opreka »brojci« nije »slovo«, već »slovka«, a koju uzimamo kao tvorbenu osnovu za hiperonimske nazive.

103_129_Filologija_59_Stojanov_Vucic_5.indd 109 03.04.2013 19:15:31

ge statističke raščlambe, već kao minimalne jedinice za pohranu teksta u bazu radi strukturiranih (SQL) i regularnih (*regex*) upita.[30]

N-gramizacijom smo dobili 3,63 milijuna bigrama, 3,55 milijuna trigrama i 3,34 milijuna tetragrama.

Razgodci nisu uklonjeni zato da bi ih se moglo pretraživati i proučavati. Funkcionalne riječi također nisu bile filtrirane. *N*-gramski stop-znakovi, tj. znakovi koji su prekidali *n*-gramske sljedove bili su dvotočje (:), upitnik (?), uskličnik (!), točka (.), trotočje (…) i znak kraja odlomka (eng. *paragraph break,* ¶), dok su znakovi zagradâ, zarezâ i navodnikâ bili neseparatori. Tim smo se jednostavnim pravilima služili u nedostatku kvalitetnoga, besplatnoga i akademskoga rečeničnika (eng. *sentence splitter*) za hrvatski jezik.

Da bismo dobili kvalitetnije izlazne podatke, bilo bi potrebno uključiti leksičku analizu i lematizaciju. Raspolažući bazama hrvatskih imena, prezimena, jednočlanih i višečlanih titula i kratica te pisanjem niza pravila iz nastavka mogli smo dobiti samo polovični rezultat, koji nas nije zadovoljio, te smo ostali na jednostavnom modelu stop-znakova pri izradbi *n*-grama.

Željeni je *n*-gramski model imao sljedeća pravila, a koja bi se daljnje trebala hijerarhizirati, slijedno uspostavljati i dosezno sužavati:

1. ako iza titule, jednočlane ili višečlane, ide razmak, pa bilo koji oblik imena, onda titula *nije* separator.
2. ako iza titule, jednočlane ili višečlane, ide razmak, pa prezime, onda titula *nije* separator.
3. ako iza titule, jednočlane ili višečlane, ide razmak, pa inicijal hrvatske abecede (A-Ž) pisan velikim slovom i s točkom, a nakon toga neko prezime, onda titula *nije* separator.
4. ako iza kratice ide razmak i veliko slovo, onda kratica *jest* separator.
5. ako iza kratice ide razmak i malo slovo, onda kratica *nije* separator.
6. …

Za potrebe rada ukupno smo sakupili 221 jednočlanu/višečlanu titulu i prefiks te 385 kratica (prilog 6.).

Brojidba nam je otkrila da su, očekivano, najčestotnije pojavnice nepu-

[30] Ovdje ćemo napomenuti da je SQL-ov operator LIKE višestruko brži prigodom pretraživanja *n*-gramske baze podataka nego REGEXP, te se preporučuje njegovo korištenje kad god je to moguće i opravdano.

103_129_Filologija_59_Stojanov_Vucic_5.indd 110 03.04.2013 19:15:31

noznačne riječi, a da prvih 20 imeničnih i 20 glagolnih lema (prilog 7.) jednoznačno upućuju na korpus sa sportskim tematskim tekstovima opisanih novinarskim stilom s naglašenom sportskom kompeticijom.[31]

5. Testiranje podataka

Nakon što smo stvorili bazu podataka sačinjenu od različitih *n*-gramskih entiteta, izradili smo niz SQL/*regex* upita kojim smo tražili:

- prvu i/ili drugu pojavnicu bigrama,
- prvu i/ili drugu i/ili treću pojavnicu trigrama,
- prvu i/ili drugu i/ili treću i/ili četvrtu pojavnicu tetragrama.

Oblik sintakse za propitkivanje prve pojavnice u bigramima koja je »HNK« izgleda ovako:

```
SELECT words.w
SUM(freq) AS freqSum
FROM gramsi_2
JOIN words ON gramsi_2.w2 = words.id
WHERE gramsi_2.w1 = (SELECT words.id FROM words WHERE words.w REGEXP
"HNK")
GROUP BY words.w
ORDER BY freqSum DESC
LIMIT 10
```

Taj upit daje nam rezultat u prilogu 8. Preko njega zaključujemo da se u *Sportskim novostima* ni na jednom mjestu ne spominje *Hrvatski nacionalni korpus* ni *Hrvatsko narodno kazalište*, već da se akronim HNK (*Hrvatski nogometni klub*) odnosi (jedino) na klubove Hajduk, Rijeka, Šibenik i Orašje.

Sintaksa za traženje bigrama u kojem je druga pojavnica »Blažević«[32]:

```
SELECT words.w, SUM(freq) AS freqSum
FROM gramsi_2
JOIN words ON gramsi_2.w1 = words.id
WHERE gramsi_2.w2 = (SELECT words.id FROM words WHERE words.w REGEXP
"Blažević")
GROUP BY words.w
ORDER BY freqSum DESC
LIMIT 10
```

[31] Zanimljivo je istaknuti podatak o dvama najčestotnijim glagolima »imati« i »igrati« koji u sportu simbolički preslikavaju poznatu ontološku, gramatičku i povijesnoknjiževnu svezu originala »imati ili biti«.

[32] U *Sportskim novostima* samo je jedan Blažević, i to je bivši hrvatski izbornik. U bigramu je najzastupljeniji u spoju s Miroslav (51), Ćiro (42) i izbornik (3).

103_129_Filologija_59_Stojanov_Vucic_5.indd 111 03.04.2013 19:15:32

Sintaksa za traženje trigrama u kojem je prva pojavnica »trener«:

```
SELECT wo1.w, wo2.w, wo3.w freq
FROM words AS wo3, words AS wo2, words AS wo1, gramsi_3
WHERE w1==wo1.id AND w2==wo2.id AND w3==wo3.id AND wo1.w==’trener’;
```

Isto to pretraživanje samo što tražimo sve oblike pojavnice »hrvatsk«:[33]

```
SELECT wo1.w, wo2.w, wo3.w freq
FROM words AS wo3, words AS wo2, words AS wo1, gramsi_3
WHERE w1==wo1.id AND w2==wo2.id AND w3==wo3.id AND wo1.w REGEXP
"trener";
```

Poznavajući SQL/*regex* možemo stvarati vrlo složene upite s neograničenim brojem *n*-gramskih upitnih kombinacija. Maštovitost kreiranja upita u *n*-gramskoj bazi ima jedino ograničenje u smislu i zanimljivosti odgovorâ do kojih želimo doći, a stvarno je ograničenje procesorska brzina za složenije upite.

Nadalje, zanimljivo je proučavati leksičko-pravopisne i kolokacijske sveze nekih pojavnica, primjerice s kojim svim drugim pridjevnim osnovama »hrvatsk« čini polusloženicu u bazi sportskih tekstova. Tako uočavamo (ručna lematizacija i abecedni poredak):

- američko-hrvatska kombinacija
- anglo-hrvatski odnosi
- ex-hrvatski reprezentativac
- hrvatsko-američka kombinacija
- hrvatsko-austrijska kombinacija
- hrvatsko-bjeloruska kombinacija
- hrvatsko-francuski dvoboj
- hrvatsko-francuski susret
- hrvatsko-južnoafrička kombinacija
- hrvatsko-nizozemski dvoboj
- hrvatsko-peruanski trokut
- hrvatsko-ruski sportski odnosi
- hrvatsko-slovenski korijeni
- hrvatsko-slovensko-bošnjačka snaga
- hrvatsko-srpski okršaj
- hrvatsko-švedski par
- slovensko-hrvatska granica

Za sportsko-leksičku analizu možemo promatrati trigrame čija je prva

[33] Iz takve vrste upita moguće je proučavati pisanje navezaka. Na ovome primjeru zaključujemo da je pisanje navezaka jako slabo zastupljeno u sportskim novinarskim tekstovima. Ukupno je 4984 trigrama s prvom pojavnicom »hrvatsk«, od čega je 801 »hrvatskog«, 258 »hrvatskom«, 44 »hrvatskoga«, 13 »hrvatskome« i 0 »hrvatskomu«.

103_129_Filologija_59_Stojanov_Vucic_5.indd 112 03.04.2013 19:15:32

pojavnica glagol, a druga pojavnica »brutal« (ručna lematizacija):

- biti brutalno uništen
- čekati brutalnu borbu
- kažnjavati brutalne nasrtaje
- početi brutalnim porazom
- početi brutalno tući
- pogledati brutalni prekršaj
- zaustaviti brutalnim prekršajem

Dodatnu korisnost dobivamo time što smo u TEI XML datoteke spremali URL adrese HTML-ova s portala *Sportskih novosti* čime rezultate traženja obogaćujemo informacijom o poveznici na mrežni tekst. Na taj način pretraživanje činimo zanimljivim za područje dohvaćanja podataka jer upite možemo slagati na način da uvjetujemo »da nam se pronađu dokumenti koji sadržavaju jedan ili više određenih *n*-grama«.

Primjerice, zanimaju nas svi mrežni tekstovi s portala *Sportskih novosti* koji sadržavaju bigrame *Slaven Bilić* i *Ćiro Blažević*. Upit u nastavku podrazumijeva da će se naći sve obličnice dotičnih riječi, kao što će se potvrditi u daljnjemu tekstu.

```
SELECT files.pn, wo1.w, wo2.w, wo3.w, wo4.w
FROM words AS wo1, words AS wo2, words AS wo3, words AS wo4,
gramsi_2 AS g2_1, gramsi_2 AS g2_2, files
WHERE
g2_1.file == files.id AND
g2_1.file == g2_2.file AND
g2_1.w1 == wo1.id AND g2_1.w2 == wo2.id AND g2_2.w1 == wo3.id
AND g2_2.w2 == wo4.id AND
wo1.w REGEXP "Slaven" AND wo2.w REGEXP "Bilić"
AND wo3.w REGEXP "Ćiro" AND wo4.w REGEXP "Blažević";
```

U prilogu 9. vidljivo je da se obojica spominju u sedam različitih tekstova:

1. 001026.xml[34]
2. 001098.xml[35]
3. 002088.xml[36]
4. 004332.xml[37]

[34] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=1026.*
[35] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=1098.*
[36] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=2088.*
[37] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=4332.*

103_129_Filologija_59_Stojanov_Vucic_5.indd 113 03.04.2013 19:15:32

5. 006486.xml[38]
6. 010641.xml[39]
7. 011918.xml[40]

Za tražene bigrame najrelevantniji je članak pod rednim brojem četiri, a zatim članak pod rednim brojem pet. Valja naglasiti da su sve nađene vijesti tematski povezene s traženim pojmovima i da se oni pojavljuju kao "subjekti" vijesti.

Sljedeći koristni upit jest traženje članaka u kojem se spominju bigram i unigram, primjerice »hrvatsk« i »navijač« te unigram »nered«.

```
SELECT wo1.w, wo2.w, wo3.w, files.pn
FROM words AS wo3, words AS wo2, words AS wo1,
gramsi_2 AS g2, gramsi_1 AS g1, files
WHERE
g1.file == files.id AND
g2.file == g1.file AND
g2.w1 == wo1.id AND g2.w2 == wo2.id AND g1.w1 == wo3.id AND
wo1.w REGEXP "hrvatsk" AND wo2.w REGEXP "navijač" AND wo3.w REGEXP
"nered";
```

Rezultat traženja nalazi se u prilogu 10.

Na isti se način mogu postavljati upiti poput:

- Nađi sve tekstove u kojima se spominju *Dinamovi* ili *Hajdukovi navijači*, a da se u istom članku navode ključne riječi: *sankcije, UEFA* ili *FIFA*.
- Nađi sve tekstove u kojima se spominje *Zdravko Mamić* i izvedenice od *novinar*.
- Nađi sve tekstove u kojima se *Marković* i *Štimac* spominju u istom članku.
- Nađi sve tekstove u kojima se pridjev *vatreni* ne spominje u kontekstu hrvatske nogometne reprezentacije.
- Nađi sve tekstove koji povezuju bigram čija prva sastavnica počinje s *anti…*, a druga s *ispad/incident/ponašanje/dobacivanje/nered/izgred*.[41]

Bilo je zanimljivo pretraživati bazu u potrazi za novim pravopisnim elementima. Uočili smo brojna pravopisna mjesta do sada neopisana i pravopisoslovno neobrađena na kojima bi se hrvatski pravopisni priručnici,

[38] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=6486.*

[39] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=10641.*

[40] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=11918.*

[41] Npr. anti[\w[ćčžšđ]]+ (ispad|incident|ponašanje|dobacivanje |nered|izgred) (a|e|om|u|i|em|ima).

103_129_Filologija_59_Stojanov_Vucic_5.indd 114 03.04.2013 19:15:32

umjesto metodologije pisanja "iz glave" , unaprijedili koristeći računalnu podršku za tekstološka proučavanja jezične realizacije. Teško bi bilo bez korpusnojezikoslovne podrške napraviti pravopisoslovno, gramatikološko ili leksikološko istraživanje o podcrtama, spojnicama, n-crticama i m-crticama,[42] trotočjima, spojevima slovaka i brojaka, korištenju matematičkoga znaka × izvan matematičkoga značenja, samoglasnoj multiplikaciji, pojavljivanju velikih slova unutar pojavnica sadržanih od malih slova, itd. Naša mala sportska baza ukazala nam je na neke pravopisne tendencije i nezabilježene prakse u priručnicima, a o čemu će više biti govora u drugim radovima.[43]

6. Usporedba s *Google searchom*

Želeći usporediti rezultate, iste upite koje smo postavljali našoj bazi postavljali smo i tražilici *Google*.

Znakovito je da napredni upit[44] (eng. *advanced search*) koji bi trebao dati isti učinak kao i pri upitu iz naše baze iz priloga 9. **ne daje** zadovoljavajuće rezultate. Kao što smo vidjeli, očekivali bismo da nam se nađe sedam mrežnih vijesti, a *Google* ih je našao samo četiri[45] (točnost od 57%). Naravno, nađeno ih je puno više — čak 45, ali smo ručno morali eliminirati višak nađenih stranica koji nisu odgovarali našem traženju prema sljedećim kriterijima:

a) nađeni su nazivi izvan traženoga vremenskog perioda od 23. travnja 2008. do 2. srpnja 2009.,
b) nađeni su nazivi izvan teksta članka, a unutar mrežne stranice (u nadglavlju ili zaglavlju vijesti HTML datoteke),
c) nađeni su nazivi indeksirani kao posljedica rezultata traženja na internoj tražilici samoga indeksiranoga portala *Sportskih novosti*.

S druge strane *Google* ne nalazi niti jedan članak koji naš sustav također ne pronalazi.

Na drugome našem primjeru s traženjem bigrama i unigrama, *Google*

[42] Portada—Stojanov (2009).

[43] Tako smo, primjerice, uočili da je 99,4% svih pojavnica koje kombiniraju slovke i brojke po svojoj naravi tipfelera, ali da preostalih 0,6% može poslužiti kao solidan materijal za pravopisoslovni članak o ovoj temi (a osobito pitanje granice pravopisoslovnoga (pr)opisivanja). Istraživanje novovjeke pojave pisanja po načelu »grbave deve« valja početi razmatrati definirajući njegovu vrijednost i korisnost izvan funkcionalnih stilova koji su ga uveli.

[44] allintext: "slaven bilić" "ćiro blažević" site:*http://sportske.jutarnji.hr/*.

[45] Nađeni su članci 1026, 1098, 4332 i 11918, a sljedeća se tri ne nalazi: 2088, 6486, 10641.

103_129_Filologija_59_Stojanov_Vucic_5.indd 115 03.04.2013 19:15:32

niti nema mogućnost dati ikakve rezultate. *Googleovom* sintaksom[46] nije moguće postaviti sličan upit, a do rezultata nije moguće doći ni korištenjem operatora. Jedini mogući način vrlo je kompliciran te nije primjenjiv. Riječ je o doslovnom navođenju svih obličnica hrvatskih pojavnica, a što je zamoran i pogrešiv put.[47]

Zbog *Googleova* nepodržavanja operatora na razini grafema, pristup sa SQL/*regex* sintaksom je i ovdje superioran.

Na isti način kao što nije moguće pretraživati internetske tekstove, nije to moguće niti preko programa *Google Desktop Search* jer dijele zajedničku sintaksu.

U zaključku ističemo sedam nedostataka *Googleove* tražilice u odnosu na SQL/*regex* pristup pretraživanja tekstova uređene *n*-gramske baze podataka:

1. *Preniska vrijednost indeksa odziva (eng. recall).*

Nisu pronađeni svi tekstovi koji su trebali biti nađeni. U terminima struke ovu kvalitetu mjerenja pretraživačkoga algoritma koja se naziva odzivom u vrijednosti 0,57 ocijenili bismo vrlo niskom.

2. *Preopsegovno indeksiranje.*

Ravnopravno se indeksiraju tijelo vijesti i metapodatci na *web*-stranicama čime se raznorodne kategorije podataka miješaju i unose informacijsku zbrku u dohvadbi.

Indeksiranje nadglavljâ (eng. *headers*) i zaglavljâ (eng. *footers*) HTML datotekâ, koji su kadšto izvansadržajni i nevažni podatci neke mrežne stranice (izbornici, zajednički i ponavljajući dijelovi *web*-stranicâ na *web*--mjestu, reklame, poveznice…) moralo bi biti različito evaluirano od teksta kao nosećega i razlikovnoga dijela *web*-stranice.

Svjesni smo da ovo nije kritika samo na *Googleovu* tražilicu ili na druge tražilice koje funkcioniraju na sličan način, već na postojeću SEO metodologiju (eng. *search engine optimization*) i indeksiranje kao takvo. Problem je, dakle, u tome što se indeksira sav tekstualni materijal u HTML-ovima, bez prethodne operacije razlučivanja tekstovnih sadržaja koji čine kvalitativnu razliku jedne *web*-stranice od druge na nekom *web*-mjestu.[48]

[46] Ne funkcionira upit allintext: "hrvatsk navijač" "nered" site:*http://sportske.jutarnji.hr/*.

[47] allintext: "hrvatski navijač" OR "hrvatskih navijača" OR "hrvatske navijače" OR "hrvatskoga navijača" OR "hrvatskome navijaču" OR "hrvatskim navijačima" "nered" site:*http://sportske.jutarnji.hr/*. Kao što se vidi, naveli smo samo dio sintakse izgubivši volju da nastavimo sintaktičku kobasicu.

[48] Drugim bismo riječima mogli reći da bi se indeksiranje trebalo više usmjeriti na

103_129_Filologija_59_Stojanov_Vucic_5.indd 116 03.04.2013 19:15:32

3. *Dvostruko indeksiranje.*

Google se koristi tražilicama na samim *web*-stranicama koje indeksira čime nađene rezultate uvrštava pod svoje duplicirajući nađeni sadržaj. Nije jasan kriterij kako *Googleov* pretraživački robot propitkuje podređene tražilice[49], ali nismo očekivali da se od 45 nađenih stranica njih samo 16 (35,5%) odnosi na *Googleov* indeks, a čak dvije trećine (64,5%) otpada na indeks indeksirane tražilice.

4. *Preniska vrijednost indeksa točnosti* (*eng. precision*).[50]

Pronalazi se više tekstova nego što treba. U terminima struke tu kvalitetu mjerenja pretraživačkoga algoritma koja se naziva točnošću u vrijednosti 0,21 ocijenili bismo jako niskom.

Tako nizak indeks točnosti izravna je posljedica točke 2. Naime, dijelovi *web*-stranica generirani su od strane sustava za upravljanje sadržajem portala *Sportskih novosti* i kao takvi su dinamički. To znači da ih, nakon što *Google* indeksira dotičnu stranicu, tamo više ne mora biti. Takav se scenarij dogodio i s mrežnom vijesti[51] u kojoj se niti na jednom mjestu **ne pojavljuje** bigram *Slaven Bilić*, ali se nalazio u zaglavlju u trenutku indeksiranja na dan 17. 5. 2010. u 14:15 po GMT-u (vidi prilog 11.) te bio ponuđen kao rezultat traženja tekstova s imenima i prezimenima dvojice naših nogometnih izbornika.

5. *Nepodržavanje operatora na razini grafema.*

Nisu podržani operatori na razini grafema pa nije moguće u racionalno ostvarivim okvirima pretraživati hrvatske obličnice. *Google* pronalazi samo jedan morfološki oblik, a ne i druge flektivne oblike. Zbog toga upi-

tekstovni, a ne tekstualni materijal.

Iz analiziranih je primjera više nego vidljivo da kvaliteta indeksiranja bitno ovisi o načinu kako se sadržaj ukodirava u HTML kod. *Google* ni nema drugoga izbora nego da indeksira sve ono što nađe u HTML-u pri čemu dolazi do informacijskoga nereda. Nisu, naime, svi tekstualni podatci okruženi HTML elementima jednake indeksne težine. Posve smo uvjereni da bi sljedeće generacije HTML-a i XHTML-a upravo iz navedenoga razloga podizanja kvalitete SEO-a morale imati na umu ove argumente pri definiranju novih jedinica. Možda je jedno od rješenja ovoga sve ozbiljnijega problema približavanje oznaka (eng. *tags*) iz TEI XML fonda HTML-ovim elementima i atributima. Druga naznaka rješenja odnosila bi se na naglašavanje potrebe za novom generacijom semantičkoga *weba* koji bi upravo SEO stavio u prvi plan.

[49] Bilo je svakojakih upita na tim stranicama, od »az« do »Ćiro bet-a-home«, »izbornik« itd. pored očekivanih »Ćiro«, »Blažević«, »Slaven« i »Bilić«.

[50] U mjerenju indeksa točnosti isključene su sve Googleove nađene stranice koje su izvan traženoga vremenskoga raspona. Ukupno ih je izbačeno 12, nakon čega smo došli do broja od 33 pronađene stranice.

[51] *http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_id=12856.*

103_129_Filologija_59_Stojanov_Vucic_5.indd 117 03.04.2013 19:15:32

te poput onoga iz priloga 10. nije moguće jednostavno i elegantno izvesti.

6. *Filtriranje razgodaka.*

Google tražilica filtrira razgotke i nije moguće pretraživati sadržaj prema njima.

7. *Pogrešna detekcija automatiziranih upita.*

Iz nekoga nepoznatog razloga nakon nekoliko upita *Googleovoj* tražilici korištenjem naprednije sintakse, ona je redovito počela odbijati daljnji rad sumnjajući da preko naše IP adrese zapravo odašiljemo automatske upite (prilog 12).

Iz svega iznesenoga utvrđujemo da za ciljeve lingvističke analize tekstova koje smo postavili našim alatom *Google search* nije podesan.

Googleova tražilica radi s *web*-stranicama, a nas za jezikoslovna propitkivanja zanimaju tekstovi, ne *web*-stranice. U mjeri u kojem su njegovi rezultati zadovoljavajući on je sjajan alat za dohvaćanje podataka (ponajviše zbog obimnosti indeksiranih podataka i brzine), ali za visokospecijalizirane i jezikoslovne upite nije odgovarajući.

Valja na kraju naglasiti da smo se koristili *Googleovim* prvim nađenim rezultatima, te da nismo koristili dodatno nađene stranice koje se uključuju opcijom iz priloga 13. S tim uključenim stranicama nalazi se čak 189 stranica, pri čemu se *recall* vrijednost ne popravlja, dok *precision* rapidno pada u odnosu na prvotne vrijednosti.

7. Vizualizacija podataka

Druga baza, ona u MySQL-u, kao što smo rekli, poslužila je za testiranje vizualizacije tražilice i prikaza rezultata. Postavljena je na stranicama *http://ngrami.ihjj.hr/*.[52] Testirali smo JavaScriptnu biblioteku WireIt za *n*-gramsku vizualizaciju s podrškom za regularne izraze. Ona zadano prikazuje »želatinasti« izgled oblakâ koje je moguće efektno pomicati izazivajući reakciju kao što je izbjegavanje sudara sa susjednim čvorovima i granama. Zbog ubrzanja rada ta je značajka isključena kao i pretraživanje trigrama i tetragrama. Iz istoga razloga MySQL-ova baza kreirana je iz demo-baze SQLitea, a ne iz pune.

Upišemo li ključni pojam »nogomet.«[53], prikazat će nam se dva oblaka — jedan lijevi i drugi desni (prikaz 14). Lijevi oblak predstavlja lijevu stra-

[52] Programiranje i upogonjivanje vizualizacije naručeno je napravio student Stručnoga studija informatike Tehničkoga veleučilišta u Zagrebu Bojan Nemčić koji dotičnu temu obrađuje u svome završnom radu.

[53] Točka u regularnoj sintaksi ima posebno značenje.

103_129_Filologija_59_Stojanov_Vucic_5.indd 118 03.04.2013 19:15:32

nu traženoga pojma, a desni oblik desnu. Debljina i boja crte kojom su grane oblaka vezane za svoja središta upućuje na čestotnost bigrama. Tako uočavamo snažnu vezu između *hrvatski nogometaš, najbolji nogometaš*, *vezni nogometaš* i *brazilski nogometaš* s jedne strane i *nogometaš koji, nogometaš svijeta, nogometaš Ivica* i *nogometaš Manchester* s druge strane.

U padajućem izborniku izlistane su po čestotnome kriteriju slične pojavnice. Pretražujemo li bigram uz pomoć regularnih operatora[54], dobit ćemo vizualizaciju čestotnosti trigrama i uočiti čestotne spojeve *strast za nogometom*, *strast za igrom, strast za gledanjem, strast za košarkom*.

8. Zaključak i daljnji rad

Za potrebe naših korpusnojezikoslovnih istraživanja hrvatskih tekstova, kao i za eksperimentalni rad s dohvaćanjem podataka, mrežno sučelje *Philologic* i *desktop* aplikacija *Bonito* nisu bili adekvatni te smo pristupili izradbi baze podataka i SQL/*regex* upita umjesto klasičnoga korpusa s pretraživanjem punoga teksta čime smo željeli ispitati mogućnosti toga pristupa i korisnost za daljnja korpusnojezikoslovna istraživanja.

Izgradili smo izvanmrežnu (SQLite) i mrežnu (MySQL) *n*-gramsku bazu na osnovi 11,6 tisuća tekstova *Sportskih novosti* objavljenih u rasponu od travnja 2008. do srpnja 2009. godine koji obasežu 67 MB, a baza 654 MB.

Ukazali smo na niz korisnih upita koje je moguće postavljati *n*-gramskoj bazi preko SQL/*regex* upita, a znakovita se korist može očitati u području dohvaćanja podataka gdje smo, uspoređujući rezultate s najpoznatijom svjetskom tražilicom — *Googleom*, uočili vlastite prednosti te istaknuli sedam točaka u kojima se očituju nedostatci tražilice *Google* i koncepta *weba* kao sveopćega korpusa u odnosu na naš eksperimentalni SQL/*regex* pristup: (i) preniska vrijednost indeksa odziva (0,57), (ii) preopsegovno indeksiranje, (iii) dvostruko indeksiranje (64,5%), (iv) preniska vrijednost indeksa točnosti (0,21), (v) nepodržavanje operatora na razini grafema, (vi) filtriranje razgodaka i (vii) pogrešna detekcija automatiziranih upita.

Ovo su dodatni argumenti koji ne idu u prilog *web*-tražilicama u usporedbi sa specijalno izrađenim korpusima/bazama za tekstovna pretraživanja. Unatoč tome, *web* kao "sveopći korpus" može poslužiti svojoj svrsi za osnovnija leksikološka pretraživanja.

Ukazali smo da se *n*-grami mogu rabiti za primarno jezikoslovna istraživanja, te da pristup može dati jedinstvene pravopisoslovne, gramatikološke, leksikološke i dohvadbene podatke.

[54] ^strast za+$.

103_129_Filologija_59_Stojanov_Vucic_5.indd 119 03.04.2013 19:15:32

Ustvrdili smo da stvorena metodologija može poslužiti za daljnji korpusnojezikoslovni rad i da će osobito biti zrela ako se tekstovi prethodno lematiziraju i obilježavaju.

Nastavak rada pretpostavlja daljnju optimizaciju baze, osiguranje procesorske snage za složenije upite, mogućnost filtriranja funkcionalnih riječi, lematizaciju, uvođenje statističke relevantnosti *n*-grama pri upitima[55], veću ergonomičnost izgradnjom vanjskoga sučelja, proširivanje korpusa, razlikovanje tekstova nadnaslovâ, naslovâ i podnaslovâ od tijela vijesti te daljnja lingvistička propitkivanja korpusa.

Sučelje baze podataka i rad sa SQL/*regex* jezicima može biti neergonomično lingvističkim korisnicima, a što bi se izgradnjom vanjskoga sučelja moglo riješiti. Ujedno bi ta sučeona nadstruktura bila mjesto veze s drugim rutinama poput morfološke normalizacije. Na isti način kao što je *BookCAT*[56] softver za katalogiziranje iza kojega stoji obična baza podataka, tako bi i vanjsko sučelje našega alata bio GUI[57] s dodatnom programerskom infrastrukturom i funkcionalnosti.

Za potrebe mrežnoga pristupa i testiranja vizualizacije *n*-gramskih odnosa izgrađena je *web*-stranica na kojem se rezultati pretraživanja baze mogu zornije uočavati (*http://ngrami.ihjj.hr/*).

Bilo bi vrlo zanimljivo kada bismo napravili anketu među novinarima *Sportskih novosti* o tipičnim informacijama do kojih oni moraju doći prigodom pisanja sportskih vijesti referencirajući se na prethodno objavljene članke te izvidjeti mogu li SQL/*regex* upiti zadovoljiti rezultatima upita.

[55] Nalik na način kao što je to učinjeno u aplikaciji *TermeX* (*http://ktlab.fer.hr/termex/*).

[56] *http://www.bookcat.net/*.

[57] eng. *graphical user interface.*

103_129_Filologija_59_Stojanov_Vucic_5.indd 120 03.04.2013 19:15:32

Literatura

Bhogal, J.; A. Macfarlane; P. Smith. 2007. *A review of ontology based query language*, Information Processing & Management, Volume 43, Issue 4, July 2007, Pages 866—886.

Biber, Douglas; Susan Conrad, Randi Reppen. 1998. *Corpus linguistics. Investigating language structure and use*. Cambridbe University Press.

Bratanić, Maja; Ostroški Anić. 2010. *Pedagoški pristup korpusno utemeljenoj izradbi kolokacijskoga rječnika strukovnoga nazivlja*, u postupku recenzije, izloženo na 5. međunarodnom leksikološko-leksikografskom znanstvenom skupu u HAZU, 3.—4. 12. 2009.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge : The M.I.T. Press.

Delač, Davor; Zoran Krleža, Bojana Dalbelo Bašić, Jan Šnajder, Frane Šarić. 2009. *TermeX: A Tool for Collocation Extraction*. Lecture Notes in Computer Science (*Computational Linguistics and Intelligent Text Processing*). 5449 (2009); 149—157.

Doedens, Crist-Jan. 1994. *Text Databases. One Database Model and Several Retrieval Languages.* Language and Computers, Number 14. Editions Rodopi Amsterdam. Amsterdam and Atlanta, GA.

Encyclopedia of Language & Linguistics. 2006. Sedond Edition. Editor Keith Brown, Elsevier.

Kilgarriff, Adam. 2003. *Linguistic Search Engine.* Objavljeno u: Kiril Simov, editor, *Shallow Processing of Large Corpora* : Workshop Held in Association with Corpus Linguistics.

Meyer, Charles F. 2002. *English Corpus Linguistics*. *An Introduction.* Cambridge University Press.

Nogometni leksikon. 2004. Urednici F. Kramar i M. Klemenčić, Zagreb : LZMK.

Pavec, Bruno. 2009. *Pisanje i pretvorba HTML-teksta s weba u TEI XML*. neobjavljeni diplomski rad na Stručnome studiju informatike, Tehničko veleučilište u Zagrebu, br. 981.

Portada, Tomislav; Tomislav Stojanov. 2009. O vodoravnim crticama u hrvatskome pravopisu. *Filologija* 52, 91—120.

103_129_Filologija_59_Stojanov_Vucic_5.indd 121 03.04.2013 19:15:33

Prilog 1.
Izgled aplikacije *SnParser* tijekom rada

Prilog 2.
Izgled stranice na *webu Sportskih novosti*

STRANI NOGOMET Ponedjeljak, 19.01.2009.

Merseyside derbi

Cahill zagorčao život Liverpoolu!

Liverpool je novim skromnim remijem propustio opet zasjesti na vrh ljestvice. Prvi je Manchester United sa 47 bodova, koliko ima i Liverpool na drugom mjestu

Gradski rivali Liverpool i Everton odigrali su 1-1 u Merseyside derbiju i posljednjoj utakmici 22. kola engleskog nogometnog prvenstva. Liverpool je tako nastavio rasprodaju bodova na Anfieldu, dok je Everton nastavio niz odličnih rezultata na gostujućim travnjacima. Ovo je bila 11. Evertonova utakmica u gostima ove sezone u Premiershipu i već 23. bod koji su osvojili izvan Goodison Parka. S druge strane, Liverpoolu je ovo bio čak šesti remi u zadnjih 9 kola, odnosno četvrti u zadnjih pet prvenstvenih utakmica na domaćem terenu. Nakon Fulhama, West Hama i Hulla, a u ranijem dijelu sezone i Stokea, bod na Anfieldu "redsima" je otkinuo i Everton, a s takvim podbačajima se momčad španjolskog menadžera Rafaela Beniteza teško može nadati naslovu prvaka, kojeg navijači čekaju još od 1990. godine.

U gradskom derbiju su domaći nogometaši imali više prilika i bili su blizu pobjedi. Liverpool je poveo u 68. minuti, kada je kapetan "redsa" Steven Gerrard odličnim prizemnim udarcem s više od 20 metara udaljenosti donio veliku radost mnoštvu na Anfieldu. Everton je izjednačio u 87. minuti. Mikel Arteta je izveo slobodni udarac s lijeve strane, a Tim Cahill je neometan iz blizine u velikoj gužvi glavom proslijedio loptu u mrežu. Australski je 29-godišnji nogometaš iskoristio činjenicu da ga prilikom slobodnog udarca nitko od brojnih Liverpoolovih igrača u kaznenom prostoru nije kvalitetno preuzeo.

Liverpool je uputio 9 udaraca prema golu, 5 u okvir, a zanimljivo je da je od 7 udaraca Evertona (prema televizijskoj statistici) čak 6 njih bilo u

103_129_Filologija_59_Stojanov_Vucic_5.indd 122 03.04.2013 19:15:33

Prilog 3.
Kôd iz HTML-a iz priloga 2. pretvoren u TEI XML:

```
   <?xml version="1.0" encoding="UTF-8"?>
   <!DOCTYPE TEI.2 PUBLIC "-//TEI P5//DTD Main Document Type//EN"
«http://www.tei-c.org/Guidelines/DTD/tei2.dtd"[
   <!ENTITY % TEI.prose 'INCLUDE'>
   <!ENTITY % TEI.linking 'INCLUDE'>
   <!ENTITY % TEI.figures 'INCLUDE'>
   <!ENTITY % TEI.analysis 'INCLUDE'>
   <!ENTITY % TEI.XML 'INCLUDE'>
   <!ENTITY % ISOlat1 SYSTEM 'http://www.tei-c.org/Entity_Sets/Unicode/
iso-lat1.ent'>%ISOlat1;
   <!ENTITY % ISOlat2 SYSTEM 'http://www.tei-c.org/Entity_Sets/Unicode/
iso-lat2.ent'>%ISOlat2;
   <!ENTITY % ISOnum SYSTEM 'http://www.tei-c.org/Entity_Sets/Unicode/iso-
num.ent'>%ISOnum;
   <!ENTITY % ISOpub SYSTEM 'http://www.tei-c.org/Entity_Sets/Unicode/iso-
pub.ent'>%ISOpub;]>
   <TEI.2 lang="hr-HR">
    <teiHeader>
    <fileDesc>
    <titleStmt>
    <title>Cahill zagorčao život Liverpoolu!</title>
    <author>TS</author>
    </titleStmt>
    <editionStmt>
    <edition><date>Ponedjeljak, 19.01.2009.</date></edition>
    </editionStmt>
    <publicationStmt><authority/><address/></publicationStmt>
    <sourceDesc>
    <bibl>http://sportske.jutarnji.hr/index.php?cmd=show_clanak&clanak_
id=8019</bibl>
    </sourceDesc>
    </fileDesc>
    <profileDesc>
    <langUsage default="NO"><language id="hr-HR">ISO hr-HR</language></
langUsage>
    </profileDesc>
    </teiHeader>
    <text>
    <body>
    <div type="nadnaslov"><p>Merseyside derbi</p></div>
    <div type="naslov"><p>Cahill zagorčao život Liverpoolu!</p></div>
    <div type="podnaslov"><p>Liverpool je novim skromnim remijem [...]</
p></div>
    <div type="članak">
    <p>Gradski rivali Liverpool i Everton odigrali su 1-1 u Merseyside
derbiju i posljednjoj utakmici 22. kola engleskog nogometnog prvenstva.
Liverpool je tako nastavio [...] </p>
    <p>U gradskom derbiju su domaći nogometaši imali više prilika i bili
su blizu pobjedi. Liverpool je poveo u 68. minuti, kada je kapetan [...]
</p>
    <p>Liverpool je uputio 9 udaraca prema golu, 5 u okvir [...] </p></
div></body></text></TEI.2>
```


103_129_Filologija_59_Stojanov_Vucic_5.indd 123 03.04.2013 19:15:33

Prilog 4.
Sportske novosti pretražive u HJK

2. NA. *Vjesnik online* [odlomak | Pod2dio | PodDio | dio]

vjetar je bitno ograničio njegovu igru. Igor Mijić Zagrebu 13. Kup OSIJEK – Rukometaši Zagreba 13. su put pobjednici Kupa Hrvatske. Naslov su obranili pobijedivši u finalu čak 32-26 (17-11) i pokazali peti put ove sezone da im gradski rivali ne mogu blizu. Već su u 14. minuti zagrebaši imali 8-3 prednost, nakratko su agrameri vratili neizvjesnost serijom 3-0, no to je uglavnom bilo sve. Zagreb je u 35. minuti imao deset pogodaka razlike, pa si je dopustio i malo opuštanja do kraja utakmice. Osvajač Kupa je odigrao

3. SN/Totalsport. *Dolaze Pletikosa i Nizozemci,...* [odlomak | PodDio | dio]

Spartak je od 1996. do 2001. godine osvojio šest naslova prvaka Rusije u nizu. Posljednjih godina su ponovno stalni stanovnici gornjeg doma. Naime, 2005., 2006. i 2007. godine su sezonu završavali na drugom mjestu i dvaput su im ispred nosa pehare uzimali gradski rivali CSKA, a jednom Zenit. Premda su najavljivali isti cilj (naslov prvaka) i ove sezone, dosadašnje 24 utakmice su ipak ispisale drugačiji scenarij. Spartak je u prvom dijelu sezone bio u velikoj krizi, koji se najviše odražavao na obrambenu liniju te Pletikosu. U ovom je

4. TS. *Cahill zagorčao život Liverpoolu!* [odlomak | PodDio | dio]

život Liverpoolu! Liverpool je novim skromnim remijem propustio opet zasjesti na vrh ljestvice. Prvi je Manchester United sa 47 bodova, koliko ima i Liverpool na drugom mjestu Gradski rivali Liverpool i Everton odigrali su 1-1 u Merseyside derbiju i posljednjoj utakmici 22. kola engleskog nogometnog prvenstva. Liverpool je tako nastavio rasprodaju bodova na Anfieldu, dok je Everton nastavio niz odličnih rezultata na gostujućim travnjacima. Ovo je bila 11. Evertonova utakmica u

Bibliografija rezultata

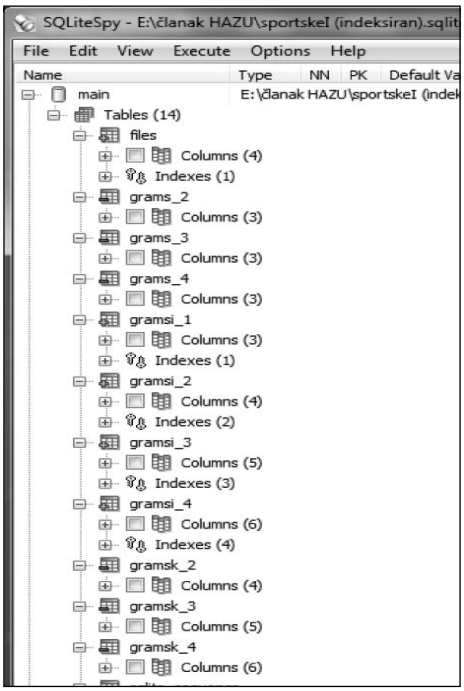
NA **[2001]**, *Vjesnik online* (© 2006, Vjesnik d.d.) [broj pojavnica] **[Vj20010217]**.

NA **[2006]**, *Vjesnik online* (© 2006, Vjesnik d.d.) [broj pojavnica] **[Vj20060520]**.

SN/Totalsport **[n.d.]**, *Dolaze Pletikosa i Nizozemci, Dinamo kod Modrića i u Udinama* () [broj pojavnica] **[005150]**.

TS **[n.d.]**, *Cahill zagorčao život Liverpoolu!* () [broj pojavnica] **[008019]**.

Prilog 5.
Struktura baze



Prilog 6.
Prikaz dijela popisa titulâ i osobnih prefikasa te kraticâ

	A	
1	**titule i osobni prefiksi**	**kratice**
2		
3	adm.	a.
4	agr.	a. a.
5	ak.	a. C.
6	akad.	a. D.
7	art.	a.a.
8	as.	adm.
9	b.	al.
10	bacc.	alb.
11	bacc. admin. publ.	am.
12	bacc. art.	amer.
13	bacc. art.	arcid.
14	bacc. crim.	as.
15	bacc. crim.	av.
16	bacc. inf.	b.
17	bacc. ing.	B. a. B. e.
18	bacc. ing. admin. chris.	B.a.B.e.
19	bacc. ing. aedif.	ban.
20	bacc. ing. aeronaut.	bat.
21	bacc. ing. agr.	batalj.
22	bacc. ing. comp.	bibl.
23	bacc. ing. el.	bilj.
24	bacc. ing. logist.	bisk.
25	bacc. ing. mech.	bl.
26	bacc. ing. nav. arch.	boj.
27	bacc. ing. opt.	bor.
28	bacc. ing. politechn.	borg.
29	bacc. ing. pyrotechn.	bos.
30	bacc. ing. sec.	br.

103_129_Filologija_59_Stojanov_Vucic_5.indd 124 03.04.2013 19:15:33

Prilog 7.
Popis prvih 20 najčestotnijih ukupnih pojavnica te imeničnih i glagolnih lema

1.	je	154960	utakmica	14535	imati	12407
2.	u	125694	minuta	11666	igrati	6516
3.	i	91195	momčad	10485	kazati	3860
4.	na	56013	igrač	10173	moći	3740
5.	se	47738	godina	8091	reći	3370
6.	da	44017	pobjeda	7191	željeti	3135
7.	su	38058	sezona	4712	odigrati	2515
8.	za	35503	igra	6849	pobijediti	2308
9.	s	26997	dinamo	6081	zabiti	1975
10.	a	26154	kolo	5710	znati	1731
11.	od	21661	klub	5461	postići	1609
12.	će	19475	trener	4039	doći	1587
13.	nije	18306	hajduk	5232	uspjeti	1311
14.	ne	18157	dan	3600	izjaviti	1210
15.	to	17155	prvak	3294	trebati	1195
16.	koji	16968	Hrvatska	3191	morati	1189
17.	što	16239	poen	2862	osvojiti	1140
18.	ali	15217	mjesto	2806	moći	1131
19.	sam	14139	euro	2695	izgubiti	999
20.	iz	13306	lopta	2413	htjeti	962 (kao punoznačni glagol)

Prilog 8.
SQL/*regex* upit za bigram gdje je prva riječ »HNK«

SELECT words.w, SUM(freq

```
SELECT words.w, SUM(freq) AS freqSum
FROM gramsi_2
JOIN words ON gramsi_2.w2 = words.id
WHERE gramsi_2.w1 = (SELECT words.id FROM words WHERE words.w REGEXP "HNK")
GROUP BY words.w
ORDER BY freqSum DESC
LIMIT 10
```

w	freqSum
Hajduk	9
Rijeka	4
Hajduk,	3
Šibenik	3
Dubrovnikom,	1
HAJDUK	1
Orašja	1
Remusa	1
Rijeka"	1
Šibenik,	1

103_129_Filologija_59_Stojanov_Vucic_5.indd 125 03.04.2013 19:15:34

Prilog 9.
Rezultat traženja svih dokumenata koji sadržavaju bigrame *Slaven Bilić* i *Ćiro Blažević*

SELECT files.pn, wo1.w,

```
SELECT files.pn, wo1.w, wo2.w, wo3.w, wo4.w
FROM words AS wo1, words AS wo2, words AS wo3, words AS wo4,
gramsi_2 AS g2_1, gramsi_2 AS g2_2, files
WHERE
g2_1.file == files.id AND
g2_1.file == g2_2.file AND
g2_1.w1 == wo1.id AND g2_1.w2 == wo2.id AND g2_2.w1 == wo3.id
AND g2_2.w2 == wo4.id AND
wo1.w REGEXP "Slaven" AND wo2.w REGEXP "Bilić"
AND wo3.w REGEXP "Ćiro" AND wo4.w REGEXP "Blažević";
```

pn	w	w	w	w
.\001098.xml	Slaven	Bilić	Ćiro	Blažević
.\004332.xml	Slaven	Bilić	Ćiro	Blažević
.\004332.xml	Slaven	Bilić	Ćiro	Blažević.-
.\006486.xml	Slaven	Bilić	Ćiro	Blažević
.\006486.xml	Slaven	Bilić,	Ćiro	Blažević
.\011918.xml	Slaven	Bilić	Ćiro	Blažević
.\001026.xml	Slavena	Bilića	Ćiro	Blažević
.\002088.xml	Slavena	Bilića,	Ćiro	Blažević
.\004332.xml	Slavena	Bilića	Ćiro	Blažević
.\004332.xml	Slavena	Bilića	Ćiro	Blažević.-
.\010641.xml	Slavena	Bilića	Ćiro	Blažević
.\004332.xml	Slavenu	Biliću	Ćiro	Blažević
.\004332.xml	Slavenu	Biliću	Ćiro	Blažević.-

Prilog 10.
Rezultat traženja svih dokumenata koji sadržavaju bigram *hrvatsk navijač* i unigram *nered*

SELECT wo1.w, wo2.w, wo3

```
SELECT wo1.w, wo2.w, wo3.w, files.pn
FROM words AS wo3, words AS wo2, words AS wo1,
gramsi_2 AS g2, gramsi_1 AS g1, files
WHERE
g1.file == files.id AND
g2.file == g1.file AND
g2.w1 == wo1.id AND g2.w2 == wo2.id AND g1.w1 == wo3.id AND
wo1.w REGEXP "hrvatsk" AND wo2.w REGEXP "navijač" AND wo3.w REGEXP "nered";
```

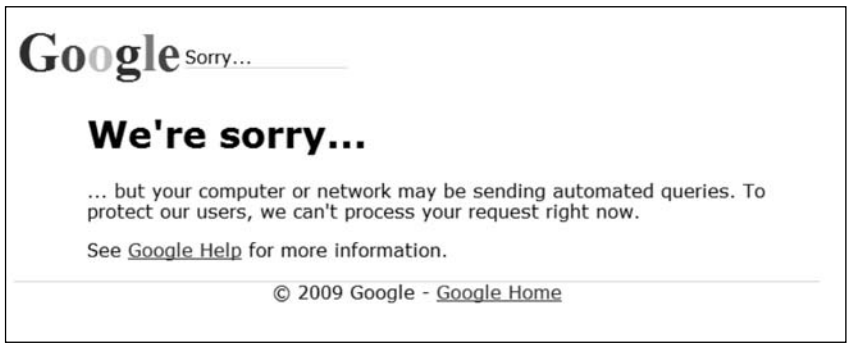
w	w	w	pn
hrvatski	navijač	nerede	.\001006.xml
hrvatske	navijače	nerede	.\001222.xml
hrvatskih	navijača	nerede,	.\001315.xml
hrvatskih	navijača	nereda	.\001384.xml
hrvatskih	navijača	neredi	.\005021.xml
hrvatskih	navijačkih	nereda	.\008199.xml
hrvatskih	navijačkih	nerede",	.\008199.xml
hrvatskih	navijača	nerede	.\009970.xml

103_129_Filologija_59_Stojanov_Vucic_5.indd 126 03.04.2013 19:15:34

Prilog 11.
Prikaz *keširane* stranice s podatkom u zaglavlju koji se više ne pojavljuje na istoj URL adresi

- 14.05.2010 | 17:16 Hrvatska želi Eurobasket 2013., ali kako pobijediti Italiju i Francusku?
- 13.05.2010 | 16:02 Velšani objavili popis za Hrvatsku, Bellamy na njemu, nema Balea
- 12.05.2010 | 17:32 Tony Cottee: Znam pravog čovjeka za West Hama - **Slaven Bilić**
- 11.05.2010 | 19:17 Dunga na SP ne vodi Ronaldinha, Ronalda, Pata, Neymara i Adriana
- 11.05.2010 | 11:47 Biliću otpao i Vedran Runje, pozvan Goran Blažević iz Šibenika
- 05.05.2010 | 12:01 Lovren, Rakitić i Kalinić opet u kadru mladih: "Svi zajedno živimo za Euro!"
- 28.04.2010 | 10:34 Hrvatska preskočila Francusku i došla do devetog mjesta, Brazil prvi
- 27.04.2010 | 17:30 Bilić u inspekciji: "O utakmici protiv Grčke na Poljudu odlučit će igrači!"
- 23.04.2010 | 17:08 Turski Hürriyet Spor: Centar Reala Ante Tomić je velika zvijezda Hrvatske
- 23.04.2010 | 16:23 Hokejaši primili pogodak već u 41. sekundi, izgubili i ispali u treću ligu!
- 19.04.2010 | 22:38 "Svaka čast Bosnichu, ali ne zna on kako je meni odjenuti hrvatski dres"
- 19.04.2010 | 05:19 Gavranović apelira: "Volim Hrvatsku, ali neću čekati poziv deset godina!"
- 14.04.2010 | 10:31 "Ili ćeš nastupati za Sloveniju ili više nećeš trenirati ovdje, preopasan si!"
- 13.04.2010 | 10:05 Eduardo: "Sanjam igrati za Vasco da Gamu poput moga idola Romarija!"
- 12.04.2010 | 13:05 Hrvatska kvalifikacije za Europsko prvenstvo u Srbiji igra protiv Španjolske!

Prilog 12.
Googleovo odbijanje daljnjeg pretraživanja zbog pogrešne procjene da odašiljemo automatizirane upite



Google Sorry...

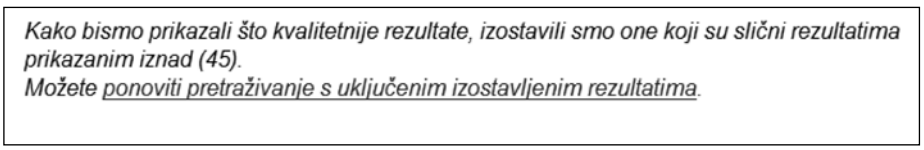
We're sorry...

... but your computer or network may be sending automated queries. To protect our users, we can't process your request right now.

See Google Help for more information.

© 2009 Google - Google Home

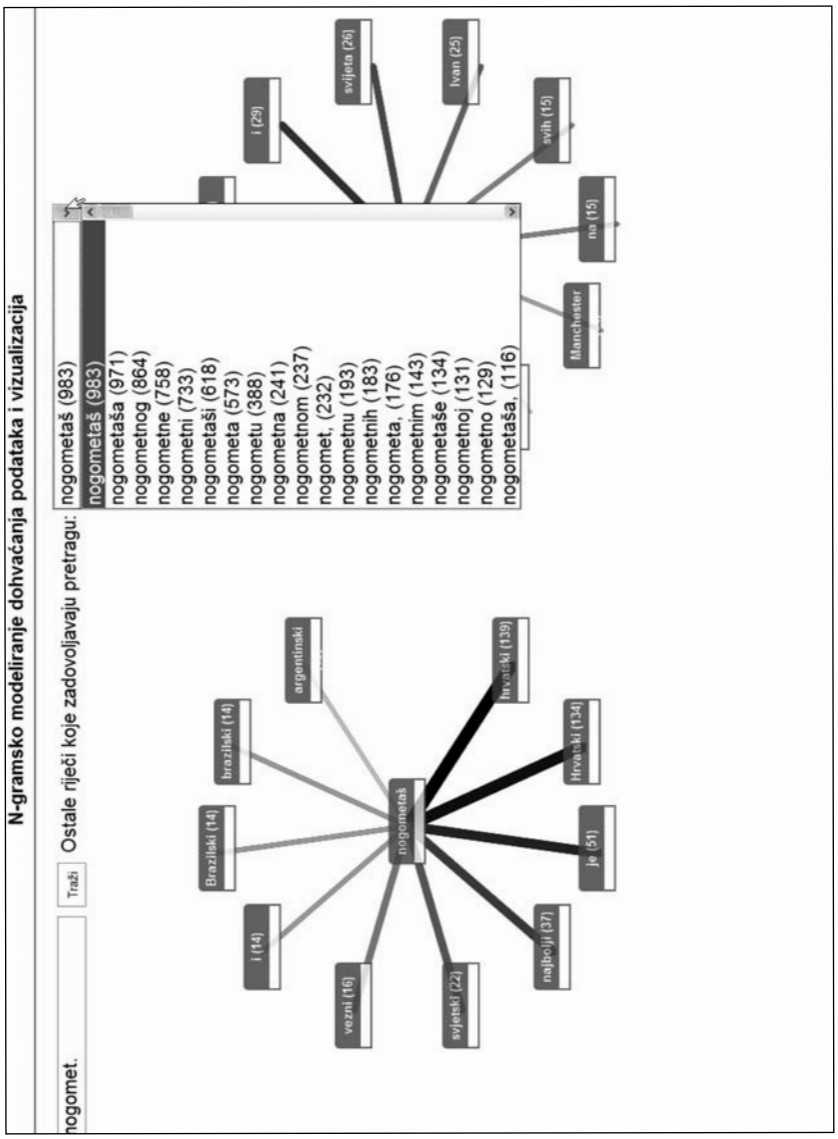
Prilog 13.
Proširenje prikaza nađenih rezultata u pretraživanju Googlea



Kako bismo prikazali što kvalitetnije rezultate, izostavili smo one koji su slični rezultatima prikazanim iznad (45).
Možete ponoviti pretraživanje s uključenim izostavljenim rezultatima.

103_129_Filologija_59_Stojanov_Vucic_5.indd 127 03.04.2013 19:15:34

Prilog 14.
Vizualizacija za traženje pojma *nogometaš*



103_129_Filologija_59_Stojanov_Vucic_5.indd 128 03.04.2013 19:15:34

A Corpus-Linguistic Analysis of *Sportske novosti* — *N*-Gram Model for Information Retrieval and Visualisation

Abstract

The paper examines the role of a corpus in linguistic research on the example of two Croatian language corpora interfaces, *Philologic* and *Bonito*, for language inquires about document and content relation, as well as the level of character and information display. For specialized linguistic search queries we have built the sport newspaper database made of *Sportske novosti* online texts (*http://sportske.jutarnji.hr/*), containing 3,6 mil. of tokens published since April 2008 till July 2009.

The computational procedures of information retrieval and *n*-gram SQL/regex queries will be shown in order to extract token co-frequencies and reveal phrases, collocations and more constant syntagmemes. The JavaScript wiring library *WireIt* is used for a token frequencies visualization in browser.

We have compared the output with Google search results based on which we have pointed out seven Google search shortcomings for linguistic investigations and have concluded that our approach could produce unique results in linguistic research.

Ključne riječi: pretraživanje teksta, SQLite, dohvaćanje podataka, tražilica Google, korpusno jezikoslovlje, Sportske novosti, *n*-gram, kolokacija, hrvatski jezik

Key words: text search, SQLite, information retrieval, Google search, corpus linguistics, Sportske novosti, *n*-gram, collocation, Croatian language

103_129_Filologija_59_Stojanov_Vucic_5.indd 129 03.04.2013 19:15:35

103_129_Filologija_59_Stojanov_Vucic_5.indd 130 03.04.2013 19:15:35