Savka N. Blagojević[1]
Univerzitet u Nišu
Filozofski fakultet
Centar za strane jezike

Miljana K. Stojković-Trajković
Visoka poslovna škola strukovnih studija
Leskovac

Pregledni rad
UDK 371.3:81'276.6
004.4:81'374
81'322.2:81'33
Primljeno 13. 2. 2016.

PRIMENA AUTOMATSKE EKSTRAKCIJE TERMINA KOD IZRADE GLOSARA[2]

Učenje i usvajanje vokabulara predstavlja jedan od ključnih ciljeva u učenju stranog jezika. Upravo zbog toga nastavnici stranih jezika moraju prepoznati koji je vokabular neophodan za razumevanje određene tematske jedinice, a koji se vokabular najčešće koristi u govornom jeziku. Za tu svrhu, nastavnici su, do sada, koristili pristup koji je podrazumevao konsultovanje udžbenika, rečnika i liste vokabulara. Danas je ovaj pristup značajno unapređen i olakšan razvojem informacione tehnologije. Metoda 'automatske ekstrakcije termina' (Automatic Term Extraction), poznata i pod nazivom 'automatsko prepoznavanje termina' (Automatic Term Recognition), omogućava nastavnicima stranih jezika da putem upoređivanja datog teksta sa korpusom dostupnim na mreži i na osnovu frekventnosti pojavljivanja termina u njima, utvrde najrelevantniju listu reči, tj. vokabulara za određenu tematsku jedinicu. Ovaj rad ima za cilj da objasni metodu automatske ekstrakcije termina i da uporedi neke od softvera koji su dostupni nastavnicima stranih jezika, prevodiocima i lingvistima i koji se u tu svrhu mogu koristiti.

Ključne reči: automatska ekstrakcija termina, strani jezik, vokabular, softver

1. Uvod

Za nastavnike stranog jezika struke jedno od važnih pitanja jeste kako doći do ključnog glosara[3] koji je neophodan za razumevanje jezičkog ma-

[1] savka.blagojevic@filfak.ni.ac.rs

[2] Rad je urađen u okviru projekta 17814: *Dinamika struktura srpskoga jezika*, koji finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

[3] Mada su reči vokabular, rečnik i glosar sinonimskog karaktera, u radu će se ove reči koristiti tako što će 'vokabular' označavati fond stranih reči kojim neko raspolaže, pod 'rečnikom' će

terijala u okviru kurseva stranih jezika. Postoji nekoliko pristupa u odabiru vokabulara koji će se obrađivati na času, a metod skaliranja koji predlažu Pacienca, Penačioti i dr. (PACIENCA, PENAČIOTI i dr. 2005) i koji se odnosi na klasifikaciju termina u odnosu na to koliko su oni bliski određenom predmetu, odnosno domenu, jedan je od pouzdanijih metoda, kako navodi Farel (FAREL 1990), koji se za tu svrhu može koristiti.

Kada je reč o jeziku struke, godinama je bilo zastupljeno korišćenje tzv. 'ručne ekstrakcije termina', gde nastavnik u tekstu obeležava termine na osnovu konsultovanja rečnika iz određene oblasti, ili izdvaja termine koje su eksperti iz određenog polja označili kao najbitnije. Naravno, pri tome je nastavnik stranog jezika morao biti upućen na saradnju sa stručnjacima i ekspertima iz domena iz kojeg je vokabular da bi i sam razumeo uskostručno značenje tih termina. Međutim, rad nastavnika može biti uveliko olakšan ako pred sobom ima udžbenik u kome su njegovi autori, kao i eksperti iz određenih oblasti, kurzivom ili podebljanim slovima već označili određene termine u tekstovima koji se obrađuju[4], te nastavnik ne mora sam da analizira tekst i donosi odluku koji su termini najrelevantniji za datu oblast. Ovo je način rada za koji se nastavnici jezika struke, tradicionalno, najčešće odlučuju, smatra Nejšon (NEJŠON 2001).

Treći pristup odabira vokabulara koji će se koristiti u kursu stranog jezika struke vrši se u odnosu na frekventnost upotrebe datog termina. Frekventnost upotrebe datog termina određuje se na osnovu učestalosti njegovog pojavljivanja u različitim domenima. Ukoliko se nastavnik odluči za ovaj pristup, neophodno je da pri svom odabiru vokabulara koristi liste termina poznatije kao „Vocabulary Lists", a koje su dostupne na internetu i nastavniku mogu biti od velike pomoći.

Do sada su nastavnici stranih jezika, lingvisti i prevodioci isključivo koristili navedenu 'ručnu' ekstrakciju termina (Manual Term Extraction – MTE), koje se ogleda u iščitavanju korpusa i izdvajanju termina za koje nastavnik ili stručnjak smatra da su ključni za razumevanje teksta. U skorije vreme, sa razvojem kompjuterističke lingvistike, razvio se metod 'automatske ekstrakcije termina' (Automatic Term Extraction – ATE), o kome možemo naći podatke u radovima nekoliko autora, kao što su Kagoura i Umino (KAGOURA, UMINO, 1996). Isti termin se u literaturi javlja i pod drugim imenima, poput 'automatskog prepoznavanje termina' (Automatic Term Recognition), kao kod Gampera i Stoka (GAMPER, STOK 1998), kao i 'automatsko pronalaženje

se podrazumevati sistematizovane reči nekog stranog jezika date sa prevodnim ekvivalentima, dok će se reč 'glosar' koristiti samo za skup stručnih termina i fraza karakterističnih za jednu naučnu ili iskustvenu oblast.

[4] Isti je princip, bilo da je reč o visoko frekventnim terminima, kada je u pitanju opšti jezik, ili o nisko frekventnim terminima, kada je u pitanju jezik struke.

termina (Automatic Term Detection – ATD), o čemu pišu Kastelvi, Kabre i dr. (KASTELVI, KABRE i dr. 2001).[5]

Kada govorimo o automatskoj ekstrakciji termina, onda se posebna pažnja mora usmeriti ka jednojezičnoj i dvojezičnoj ekstrakciji. Jednojezična ekstrakcija (Monolingual Term Extraction) ima za cilj da analizom teksta ili određenog korpusa identifikuje i ponudi listu termina „kandidata“,[6] kako se popularno kaže u anglističkoj literaturi, odnosno, da sačini listu ključnih termina za razumevanje datog korpusa. Ovakvo izdvajanje termina od velikog je značaja za nastavnike koji se bave stranim jezikom za opštu namenu, akademskim jezikom ili jezikom struke, ali isto tako i za autore udžbenika. Kod dvojezične ekstrakcije termina (Bilingual Term Extraction) neophodno je da softver podržava više jezika jer je njegov zadatak ne samo da identifikuje ključne termine datog korpusa, već i da ponudi njegove ekvivalente na odabranom jeziku. Pored nastavnika i lingvista, za ovaj pristup se najčešće opredeljuju i prevodioci, smatra Fo (FO 2012). Ovde moramo napomenuti da konačnu verifikaciju termina dobijenih bilo kojim od dva navedena pristupa, mora izvršiti sam istraživač. Pre nego što bliže objasnimo metod automatske ekstrakcije termina, neophodno je da se osvrnemo na različite softvere otvorenog koda koji su dostupni nastavnicima i da izvršimo njihovu klasifikaciju u odnosu na metode koje koriste u obradi podataka.

2. Lingvistički, statistički i hibridni metod ekstrakcije termina

U kompjuterističkoj lingvistici, a samim tim i u upravljanju terminologijom (Management Technology), poznata su tri metoda izdvajanja termina: neki od softvera koriste isključivo lingvistički metod, neki statistički, a većina softvera koristi mešoviti tj. hibridni metod koji predstavlja kombinaciju prva dva metoda Pacienca, Penačioti i dr. (PACIENCA, PENAČIOTI i dr. 2005).

Lingvističkim metodom identifikuju se kombinacije reči koje se podudaraju sa određenim morfološkim i sintaktičkim obrascima. Termini koji se smatraju „kandidatima“ uglavnom se identifikuju imeničkom frazom i ovde se koristi sistem fraza IRL (IRL 1970). To je zbog toga jer se smatra da su sintaktički podaci dovoljni da utiču na prepoznavanje ključnih termina (BU-

[5] Mi ćemo u radu koristiti termin ’automatska ekstrakcija termina’ (ATE) zbog toga što se on daleko češće koristi od ostalih termina .

[6] Termin ’kandidati’ biće korišćen i u našem radu da bi označio jedinice koje su posle identifikovanja u korpusu ’spremne’ da uđu u glosar, tj. u konačnu listu termina koji su neophodni za razumevanje datog jezičkog materijala.

RIGOLT 1992), dok nešto kompleksniji sintaktički obrasci mogu biti izdvojeni na osnovu osnovnih formata poput imenica–imenica i pridev–imenica (EVANS, ZAI 1996). U istraživanju koje opisuju Dajli, Habert i dr. (DAJLI, HABERT i dr. 1996) identifikovana su dva formata koja se mogu pronaći u osnovi termina, a to su imenica–imenica i pridev–imenica. Autori smatraju da je za identifikaciju termina neophodno upoređivati dati korpus sa uobičajenim visokofrekventnim izrazima Dajli (DAJLI 1994).

Izdvajanje termina iz korpusa može se izvršiti na osnovu dva modula, prvog koji se odnosi na proces raščlanjivanja, i drugog, koji se odnosi na proces prepoznavanja termina. Za modul raščlanjivanja je značajno da njegov softver koristi površinsku lingvističku analizu koja omogućava identifikaciju oblika poput imenice, glagola, prideva, itd., upotrebom tehnike koja se zove 'označavanje dela govora' (PoS tagging).[7] Kod drugog modula softver vrši upoređivanje termina koji pripadaju svakodnevnoj komunikaciji, a nakon upoređivanja izdvaja iz korpusa samo povezane površinske oblike i izbacuje one forme koje nisu od velikog značaja. Kod ovakve ekstrakcije termina neophodno je postaviti pitanje u vezi sa formom termina koju treba upotrebiti. Veoma je važno da se osnovni termin i njegove forme tretiraju jednakim. Ovde se može primeniti i „stop-lista" kojom se značajno poboljšava obrada podataka. Ona se primenjuje zato što često pojedine jezičke jedinice, kao što su članovi, pomoćni glagoli ili reči koje su u učestaloj upotrebi, mogu 'zbuniti' algoritam prilikom obrade i uticati na relevantnost dobijenih termina. Kada se ovakve jedinice unesu u stop-liste, algoritam ih u obradi podataka više neće koristiti, pa će one, s pravom, biti isključene iz liste relevantnih termina.

Dok se lingvistički metod za ekstrakciju termina zasniva na jeziku i na upoređivanju određenih jezičkih elemenata, drugi, statistički metod, funkcioniše nezavisno od jezika. Kod statističkog metoda ne dolazi do upoređivanja određenih jezičkih elemenata već se pretražuje njihova ponavljivost. Najbitnija oznaka jeste frekventnost kao i broj pojavljivanja u datom korpusu. Na osnovu ovih komponenti softver pravi listu termina. Softver vrši upoređivanje korpusa na dva načina: najpre upoređuje frekventnost i pojavljivanje termina u datom korpusu, a zatim se vrši upoređivanje sa korpusom koji je dostupan na mreži. Nakon dobijene liste termina sa naznačenom frekventnošću nastavnik može doneti jasnu odluku kojim terminima treba posvetiti punu pažnju i eksplicitno ih objasniti, a koji termini se mogu razumeti u odnosu na kontekst i studenti mogu sami protumačiti njihovo značenje.

Ukoliko se nastavnik odluči za ovakav pristup, on mora biti upoznat sa vrstom vokabulara koji se koristi za zadovoljenje jezičkih potreba. Tako,

[7] Označavanje dela govora (Part of Speech Tagging – POS) spada u značajnu fazu prirodne obrade jezika (natural language processing – NLP) i predstavlja proces gde se proverava da li neki termin u tekstu ili određenom korpusu odgovara određenom delu govora i to na osnovu definicije termina, njegovog značenja i konteksta u kome se pojavljuje.

Rid (RID 2000) citirajući Čanga, Mihva i dr. (ČANG, MIHVA i dr. 2003) navodi da se pored akademskog i tehničkog tipa vokabulara, vokabular može razvrstati i na visokofrekventni, srednjefrekventni i niskofrekventni vokabular. Visokofrekventni vokabulari obuhvataju 80% našeg pisanog i govornog jezika i njih sačinjavaju termini koji ulaze u opseg od 2000 do 3000 porodica termina.[8] Kod srednjefrekventnog vokabulara (mid frequency), frekventnost pojavljivanja termina je u rangu od 4000 do 9000 porodica termina i tu spadaju oni termini koji su ključni za razumevanje autentičnog teksta i nisu učestani u govornom i pisanom jeziku. Na kraju, za nastavnike jezika struke, najvažniji je vokabular sa niskom frekvencijom (low frequency), koga čine termini koji se javljaju u 9000 porodica termina i ovaj vokabular je specifičan za određenu struku, tj. karakterističan je za jezik struke. Pojedini softveri daju tačan broj pojavljivanja datog termina i ukazuju nastavniku na to da li je reč o niskofrekventnim, srednjefrekventnim i visokofrekventnim terminima.

Najveći broj softvera za analiziranje korpusa i ekstrakciju termina koji bi, kao ključni, trebalo da uđu u glosar, koristi kombinaciju lingvističkog i statističkog metoda. Ovaj metod se još naziva hibridnim ili mešovitim metodom. Pomoću prve komponente ovog metoda određuje se frekventnost termina, nakon čega se lingvističkim metodom preuzimaju određena sintaktička pravila i filteri koji su ključni pri odabiru termina određene sintaktičke strukture. Najznačajniji pristup u analizi jeste analiza C-vrednosti (C-value) koja će biti objašnjena kod prikaza softvera TerMine (odeljak 3.7).

3. Upotreba softvera u identifikovanju glosara

Opšte uzev, softvere možemo podeliti na nekoliko načina. Prvu podelu softvera možemo izvršiti na komercijalne softvere i softvere otvorenog koda koji su dostupni korisnicima i čija je upotreba besplatna. Mi ćemo analizirati one softvere koji spadaju u grupu otvorenog koda, zbog činjenice da su u svakom trenutku dostupni nastavniku i njihova upotreba je potpuno besplatna. Kada je reč o softverima za identifikovanje glosara, onda ih delimo na one koji u analizi termina koriste isključivo lingvistički metod, zatim na one koji koriste statistički metod i na kraju, na softvere koji u analizi termina koriste mešoviti metod.

Softveri koji za analizu koriste lingvistički metod od velike su važnosti za lingviste – istraživače iz različitih polja sintakse, pragmatike, morfologije, kao i svih onih koji se bave analizom korpusa. Međutim, oni neće biti predmet

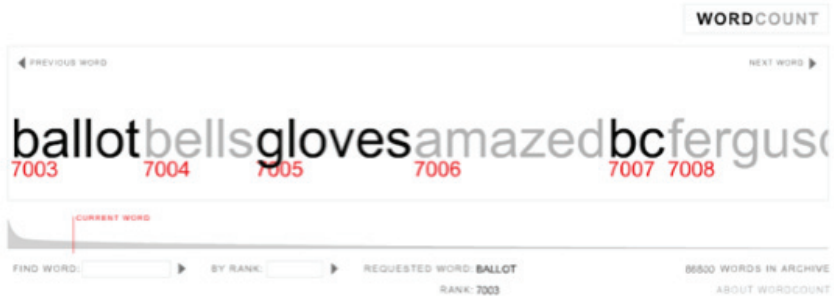
[8] Pojam 'porodica termina' (word family) označava grupu reči, odnosno, grupu termina koji imaju zajednički obrazac ili određenu karakteristiku. U porodicu termina ubraja se osnovni oblik reči sa svim njegovim derivacijama.

daljeg razmatranja u ovom radu, jer je njegov cilj analiza softvera za automatsko izdvajanje termina i priprema glosara za nastavu stranog jezika struke. Iz tog razloga u radu će biti analizirani softveri koji se oslanjaju na statistički i hibridni metod prilikom obrade korpusa.

Neki od softvera otvorenog koda koji koriste isključivo statistički metod u pronalaženju ključnih termina jesu WordCount, VocabGrabber, FiveFilters, Text Analyzer, dok hibridni metod koriste WebCrop i Complete Lexical Tutor, TermMine. Kratki prikazi principa na osnovu kojih ovi softveri funkcionišu biće dati u odeljcima rada koji slede.

3.1. WordCount

WordCount spada u softvere otvorenog koda koji u analizi termina koristi statistički metod. Ovaj softver analizira frekventnost zadatog termina u odnosu na korpus od 86.800 termina koji su u upotrebi u engleskom jeziku i dobijen je iz britanskog nacionalnog korpusa. Prednost upotrebe ovog softvera je u tome što nakon unošenja termina nastavnik dobija podatak o njegovoj frekventnosti izražen numeričkim putem. Ovaj podatak pomaže nastavniku da donese odluku u vezi sa tim koliko pažnje treba posvetiti određenom terminu, odnosno, koliko je određeni termin važan za dati korpus. Ukoliko nastavnik predaje strani jezik za opšte potrebe, onda će posebnu pažnju posvetiti visokofrekventnim terminima, a ukoliko je reč o jeziku struke ili akademskom jeziku, veća pažnja će biti usmerena ka niskofrekventnim terminima. Ovaj softver je značajan za određivanje frekventnosti pojedinačnih termina i može ukazati na to da li termin „kandidat“ treba obuhvatiti glosarom ili ne. Međutim, ovaj softver ne daje mogućnost kreiranja liste ključnih termina iz određenog korpusa. Na slici koja sledi (slika 1) prikazana je frekventnost dobijena za zadati termin „ballot“. Kao što možemo videti, njegova frekventnost iznosi 7003, i spada u grupu srednjefrekventnih termina. Na nastavniku je da odluči koliko će vremena posvetiti obradi ovog termina.



Slika 1. Iskazana frekventnost za termin „ballot“

3.2. VocabGrabber

Ovaj softver otvorenog koda omogućava analizu korpusa statističkom metodom i kreira listu ključnog vokabulara za razumevanje korpusa. Takođe ukazuje na to kako su termini upotrebljeni u datom kontekstu. Softver daje mogućnost da se iz dobijene liste termini svrstaju u vokabulare određenih akademskih disciplina, kao što su umetnost i kultura, matematika, geografija i društvene nauke. Nakon unošenja teksta i dobijene liste ključnih termina, nastavnik može izvršiti selekciju termina na osnovu njihove relevantnosti i ponavljivosti, što se može videti na slici 2. Relevantnost se izražava numerički u rasponu od 1 do 5, gde broj 5 označava termine koji su ključni za razumevanje datog teksta, dok broj 1 označava učestale termine koji su već poznati i nisu ključni za razumevanje datog teksta. Prilikom određivanja relevantnosti, softver utvrđuje frekventnost termina u zadatom korpusu, nakon čega se vrši upoređivanje sa pisanim korpusom na zadatom jeziku i iz određenog domena koji je dostupan na veb mreži. Prilikom kreiranja liste nastavnik ima mogućnost da odabere one termine koji su označeni brojevima od 5 do 3 i da odlučili koliko vremena će posvetiti obradi datih termina, ali neće naročitu pažnju usmeravati na termine koji su označeni sa 1 ili 2, zbog činjenice da nisu od velike važnosti za razumevanje zadatog teksta i najčešće su učenicima već poznati.

Slika 2. Selekcija termina na osnovu domena i relevantnosti

Ovaj softver takođe nudi i vizualni tezaurus, odnosno terminološki rečnik koji sadrži sistemsko uređene nazive određenog naučnog domena. Dati rečnik pokazuje dati termin u određenom kontekstu, kao što je to prikazano na slici 3. Pored vizualnog tezaurusa, nastavniku je na raspolaganju i opcija definisanja termina „kandidata“, što može biti od velike važnosti ukoliko je u pitanju jezik struke, uzimajući u obzir specifičnosti termina koje on koristi.

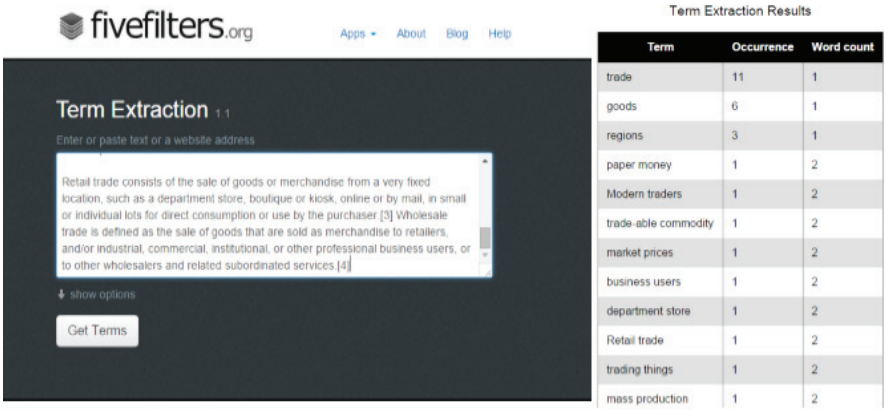
Slika 3. Vizualni tezaurus i termin „kandidat“ prikazan u kontekstu

Da bismo ovo bolje ilustrovali, na slici 4 prikazana je lista termina „kandidata“ za zadati tekst, a u konkretnom slučaju tema je ’trgovina’. Možemo uočiti da lista uključuje relativnost termina i broj njihovog pojavljivanja u korpusu tako da nastavnik može doneti odluku da li da pristupi obradi nekog termina na osnovu njegove relevantnosti ili na osnovu broja njihovog pojavljivanja u tekstu.

Slika 4. Dobijena lista termina „kandidata“ na temu Trgovine

3.3. Five Filters

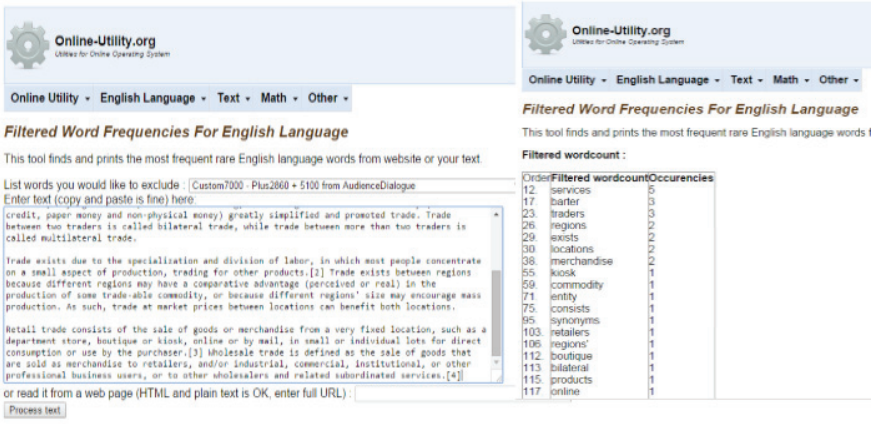
Ovaj softver otvorenog koda koristi statistički metod u automatskom izdvajanju termina i po unošenju teksta daje listu ključnih termina za njegovo razumevanje na osnovu njihovog pojavljivanja kako u samom korpusu, tako i na osnovu upoređivanja sa korpusom na mreži. Na slici 5 prikazan je primer obrade teksta i dobijene liste termina „kandidata“ opcijom Term Extraction. I pored ovakve opcije, ovaj softver ne daje mogućnost definisanja termina u određenom kontekstu, te je na samom nastavniku da odluči da li su termini iz ponuđene liste relevantni za razumevanje jezičkog kursa koji on predaje ili oni to nisu.



Slika 5. Primer obrade teksta i dobijene liste termina „kandidata" opcijom Term Extraction

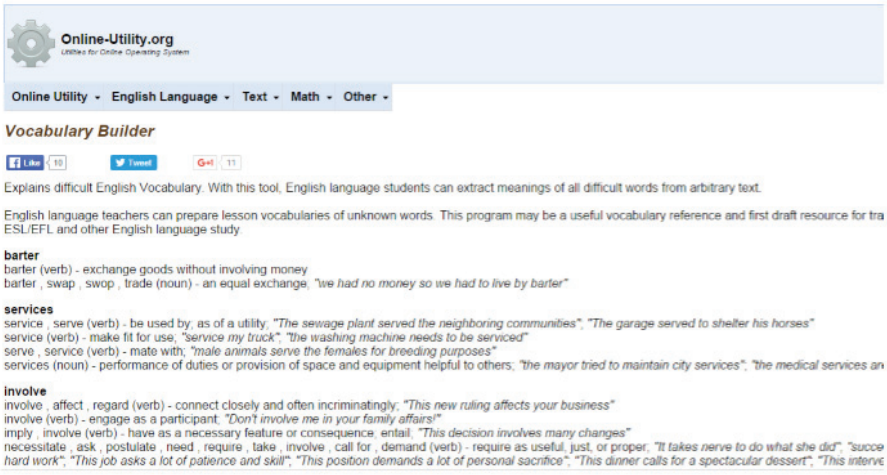
3.4. Text Analyzer

Ovaj softver otvorenog koda koji koristi statističku metodu u analizi termina daje više korisnih opcija nastavnicima stranog jezika u pripremi glosara. Baš kao i softver Five Filters, pored opcije određivanja frekvencije termina i izdvajanja termina iz teksta, on daje nastavniku listu termina na osnovu njihovog pojavljivanja u korpusu. Na slici 6 prikazan je primer ekstrakcije termina opcijom Filtered Word Frequencies For English Language. Sa leve strane prikazan je način unošenja teksta kao i način odabira korpusa za upoređivanje, dok je sa desne strane dat prikaz liste termina „kandidata" i njihov broj pojavljivanja u korpusu.



Slika 6. Primer izvlačenja termina „kandidata" na osnovu frekvencije opcijom Filtered Word Frequencies For English Language

Ovaj softver takođe sadrži opciju Vocabulary Builder koja može biti od velike koristi kako nastavnicima u pripremi glosara tako i učenicima u učenju vokabulara. Ova opcija iz određenog zadatog korpusa izdvaja sve one termine koji su ključni za razumevanje korpusa i definiše „teže“ termina za razumevanje datog teksta jer daje i praktične primere. Na slici 7 prikazan je primer obrade teksta ovom opcijom. Najpre je prikazan unos teksta, a zatim je dat prikaz definisanja termina. Kao što se može videti na ovoj slici, date definicije se odnose i na oblik samog termina kao i na definisanje svih polisemičnih oblika.

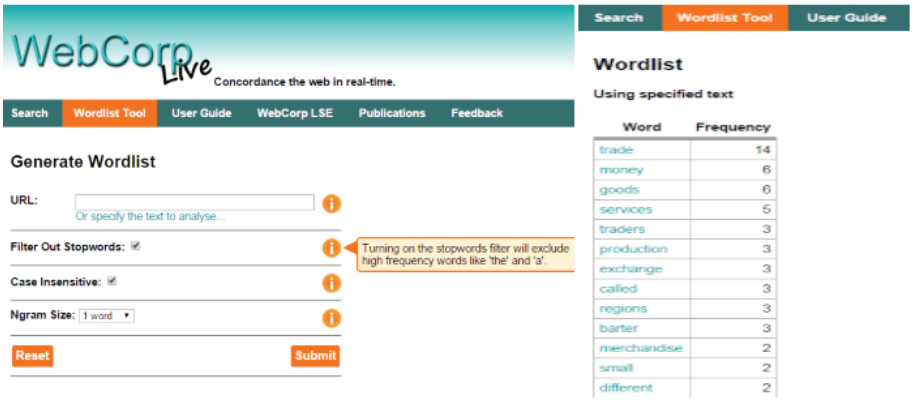


Slika 7. Primer obrade teksta opcijom Vocabulary Builder

U drugu grupu softvera spadaju oni softveri koji koriste hibridnu metodu u automatskom izvlačenju termina i analizi korpusa i koja se temelji na upotrebu analize C-vrednosti (C-value).

3.5. WebCorp

Ovaj softver otvorenog koda koji u obradi podataka koristi hibridni metod sadrži dva pretraživača. Pretraživač Web Corp Live je prvi nastao i omogućava nastavnicima, učenicima i lingvistima pretraživanje korpusa i pronalaženje ključnih termina. Softver koristi hibridni metod i metodom stop liste dolazi do ključnih termina iz zadatog teksta. Nakon obrade teksta, koristeći opciju Wordlist Tool, softver daje pregled termina „kandidata“ na osnovu frekventnosti pojavljivanja u datom tekstu, kao što se može videti na slici 8.



Slika 8. Način obrade teksta upotrebom opcije Wordlist Tool

Kako se pokazalo da opcija pretraživanja i ekstrakcije termina nije bila u potpunosti zadovoljavajuća za lingviste – istraživače, RDEUS (Research and Development Unit for English Studies) na Univerzitetu u Biringemu uporedo je razvio softver namenjen svima onima koji se bave korpusnom lingvistikom, leksikografijom, učenjem stranog jezika i izdavaštvom. Softver WebCorp Linguistics Serach Engine (WebCorp LSE) omogućava pretraživanje više različitih korpusa i proveru zadatog termina. Ovaj pretraživač, za razliku od WebCorp Live, koji se temelji na statističkoj metodi automatskog izdvajanja termina, koristi hibridnu metodu u svojoj analizi. Prilikom obrade podataka najpre dolazi do pronalaženja istih obrazaca i analiziranja delova govora (PoS). Nakon toga se pristupa daljoj analizi u vidu sortiranja i sumiranja dobijenih podataka. Za razliku od njega, WebCorp LSE vrši pretraživanje na osnovu slaganja obrazaca PoS analizom, tj. analizom delova govora. U samoj analizi istraživač može suziti pretragu na određene domene, upotrebu velikih slova, poziciju date reči u rečenici, i izvršiti odabir samo jednog opsega datog termina kako bi izbegli polisemičnost, kao što je to i prikazano na slici 9.

Slika 9. Način obrade termina „kandidata" opcijom WebCorp LSE

Ipak WebCorp LSE ne daje mogućnost obrade celokupnog teksta nego samo pojedinačnih termina „kandidata", tako da se ove opcije međusobno dopunjuju. Prvo je potrebno izvršiti analizu teksta i automatskim izdvajanjem termina doći do ključnog glosara, a zatim dobijene termine „kandidate" proveriti putem WebCorp LSE.

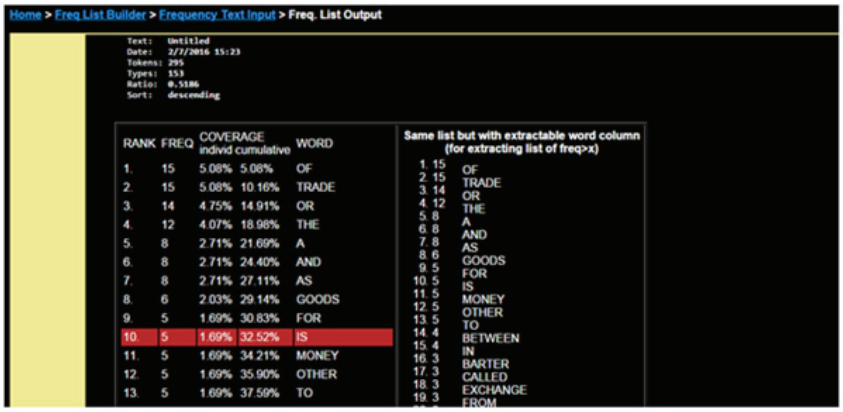
3.6. Complete Lexical Tutor

Ovaj softver otvorenog koda ima više namena. Podeljen je u tri kategorije koje su namenjene učenicima stranog jezika, istraživačima i nastavnicima. Sve opcije i sve kategorije se odnose na učenje, istraživanje i testiranje vokabulara stranog jezika. Koristi hibridnu metodu u analizi termina određenog korpusa, a u pripremi glosara od velike važnosti su dve opcije koje ovaj softver nudi, a to su opcija Key Words Extractor i Web Frequency Indexer. Prvom opcijom se definiše leksika u određenom tekstu ili korpusu statističkom metodom kroz upoređivanje frekventnosti termina iz teksta sa terminima u znatnom većem korpusu. Korpus ovog pretraživača ide i do 14 miliona termina. Ovim metodom nastavnik dobija listu ključnih termina, a ukoliko želi da proveri frekventnost za svaki termin iz samog teksta ili korpusa, onda će koristiti drugu opciju.

(1) 4230.47 barter
(2) 355.29 trade
(3) 242.15 locate
(4) 237.94 goods
(5) 158.93 region
(6) 138.30 exchange
(7) 63.35 service
(8) 59.76 product
(9) 34.29 money

At keyness cut-off of 25, there are 9 keywords from a total of 292 words, for a keyword ratio of 0.031
MEANING : .001 is an extememely high keyword text (many words distinct to this text), .009 is a low keyword text (uses general words), and so on.

Slika 10. Prikaz teksta i dobijene liste termina „kandidata" upotrebom opcije KeyWords Extractor



Slika 11. Prikaz frekventnosti svih termina određenog korpusa upotrebom opcije Frequency List Builder

3.7. TerMine

Ovaj softver otvorenog tipa u analizi termina koristi hibridnu metodu i to metodu analize C-vrednosti (C-value), koja ne zavisi od domena, već vrši analizu celokupnog teksta fokusirajući se na kombinacije termina. Korpus ovog softvera iznosi 9.8 miliona termina „kandidata“ i može se koristiti za analizu svih tekstova iako je softver, prvobitno, kreiran za analizu tekstova iz biomedicine. Ovom analizom se vrši selekcija termina, najpre primenom lingvističke analize kojom se utvrđuju termini „kandidati“ označavanjem dela govora (PoS) i izdvajanjem kombinacija pridev–imenica. U sve to se još uključuju i stop liste. Nakon selekcije, pristupa se statističkoj analizi gde se utvrđuje frekventnost pojavljivanja termina „kandidata“, zatim frekventnost pojavljivanja termina u kombinaciji sa drugim terminima, određuje se dužina termina „kandidata“, i na kraju, utvrđuje se broj tih termina. Nakon toga, dobija se lista najrelevantnijeg glosara za dati tekst. Ovo se može i videti na slici 12.

Slika 12. Prikaz dobijenih termina „kandidata“ analizom C-vrednosti

Zaključak

Nakon izvršene tipologije softvera koji se zasnivaju isključivo na statističkom i onih koji se zasnivaju na hibridnom metodu i nakon njihovog upoređivanja, možemo zaključiti da za automatsku ekstrakciju termina za potrebe korpusne lingvistike, kao i za potrebe učenja jezika struke, bolji pregled daju softveri koji koriste hibridni metod. Kako bi nastavnik bio siguran da je analiza teksta uspešna i postupak ekstrakcije termina u potpunosti validan, najbolje je osloniti se na analizu primenom metoda C-vrednosti. Svi softveri koji koriste ovaj metod pri analizi teksta, gde se, najpre, koristi lingvistički, a zatim statistički pristup, pružaju nastavniku stranog jezika određenu sigurnost u pogledu relevantnosti dobijenih termina.

U odabiru softvera za ekstrakciju termina posebna pažnja mora biti usmerena na unošenje korpusa, budući da samo pojedini softveri imaju mo-

gućnost obrade teksta. Neki od softvera daju mogućnost unošenja celokupnog korpusa, dok drugi omogućavaju analizu svakog pojedinačnog termina. Drugo pitanje koje nastavnici jezika treba da uzmu u obzir kada vrše odabir softera u svrhu ekstrakcije termina jeste da razmotre obim korpusa kojim raspolaže određeni softver, kao i to da li je namenjen lingvistima i da li sadrži dovoljan broj opcija pretraživanja koje su značajne za korpusnu lingvistiku. Međutim, ni u kom slučaju ne treba zanemariti i one softvere koji su zasnovani na statističkom metodu. Ekstrakcija termina i određivanje relevantnosti, kao i frekventnosti, može biti i te kako značajan pokazatelj da li je reč o visokofrekventnim, srednjefrekventnim i niskofrekventnim terminima i time nastavniku jezika ukazati na termine kojima treba posvetiti više pažnje pri obradi datog teksta. Takođe, neki od ovih softvera imaju mogućnost definisanja termina „kandidata“ za izradu glosara i mogu da predstave dati termin u određenom kontekstu, što može biti značajan faktor u učenju stranog jezika struke.

I na kraju, treba istaći da je odluka o odabiru softvera, kao i termina „kandidata“ iz liste relevantnog vokabulara, ostavljena samom nastavniku. Isto tako, potrebno je izvršiti dalja istraživanja i uporediti da li se liste dobijenih termina „kandidata“ međusobno podudaraju, a ako se podudaraju, onda u kojoj meri to čine, da bismo sa sigurnošću mogli da utvrdimo koji od softvera daje najrelevantniji glosar koji će se koristiti pri obradi tekstova iz stranog jezika.

Citirana literatura

BURIGOLT 1992: Bourigault, Didier: “Surface grammatical analysis for the extraction of terminological noun phrases.” Proceedings of the 14th conference on Computational linguistics-Volume 3. Association for Computational Linguistics, 1992.

ČANG, MIHVA i dr. 2003: Chung, Teresa Mihwa, and Paul Nation. “Technical vocabulary in specialised texts.” Reading in a foreign language 15, no. 2, 2003: 103–116.

DAJLI 1994: Daille, Béatrice: Approche mixte pour l’extraction de terminologie: statistique lexicale et filtres linguistiques. Diss. 1994.

DAJLI, HABERT i dr. 1996: Daille, Béatrice, Benoît Habert, Christian Jacquemin, and Jean Royauté. “Empirical observation of term variations and principles for their description.” Terminology 3, no. 2 (1996): 197–257.

EVANS, ZAI 1996: Evans, David A., and Chengxiang Zhai. “Noun-phrase analysis in unrestricted text for information retrieval.” Proceedings of the 34th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics. 1996.

FAREL 1990: Farrell, Paul. “Vocabulary in ESP: A Lexical Analysis of the English of Electronics and a Study of Semi-Technical Vocabulary. CLCS Occasional Paper No. 25.” (1990).

FO 2012: Foo, Jody: “Computational terminology: Exploring bilingual and monolingual term extraction.” Linköping Studies in Science and Technology. Thesis, No. 1523. 2012.

GAMPER, STOK 1998: Gamper, Johann, and Oliviero Stock. “Corpus-based terminology.” Terminology 5.2 (1998): 147–159.

IRL 1970: Earl, Lois L: “Experiments in automatic extracting and indexing.” Information Storage and Retrieval 6.4. 1970: 313–330.

KAGOURA, UMINO, 1996: Kageura, Kyo, and Bin Umino. “Methods of automatic term recognition: A review.” *Terminology* 3.2 (1996): 259–289.

KASTELVI, KABRE i dr. 2001: Castellví, M. Teresa Cabré, Rosa Estopa Bagot, and Jordi Vivaldi Palatresi. “Automatic term detection: A review of current systems.” Recent advances in computational terminology 2 (2001): 53–88.

NEJŠON 2001: Nation, I, S, P. Learning Vocabulary in Another Language. Cambridge: CUP. 2001.

PACIENCA, PENAČIOTI i dr. 2005: Pazienza, Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto: “Terminology extraction: an analysis of linguistic and statistical approaches.” Knowledge mining. Springer Berlin Heidelberg, 2005. 255–279.

RID 2000: Read, John: Assessing vocabulary. Cambridge University Press. 2000.

Izvori

http://www.wordcount.org/main.php [20.12.2015]

http://www.webcorp.org.uk/live/wdlistspecified.jsp?text=14525394873740.90888385483429523&sw=on&i=on&n=1[20.12.2015]

http://fivefilters.org/term-extraction/ [20.12.2015]

http://www.visualthesaurus.com/vocabgrabber/#[20.12.2015]

http://www.online-utility.org/text/analyzer.jsp [20.12.2015]

http://www.lextutor.ca/ [20.12.2015]

http://www.nactem.ac.uk/software/termine/cgi-bin/termine_cvalue.cgi [20.12.2015]

Savka N. Blagojević, Miljana K. Stojković-Trajković

APPLYING AUTOMATIC TERM EXTRACTION IN COMPILING A GLOSSARY

Learning vocabulary is one of the key objectives in foreign language acquisition. Due to this, language teachers have to decide which vocabulary is required for understanding specific thematic units, as well as which vocabulary is commonly used in spoken language. For this purpose, language teachers have traditionally used the approach that involves the consultation of textbooks, dictionaries and vocabulary lists. Nowadays, the development of information technology makes this approach much easier. The method of automatic term extraction (Automatic Term Extraction), also known as automatic recognition vocabulary (Automatic Term Recognition), provides teachers with the ability to compare the selected text with the content available on the Web, and on the basis of the frequency of a term occurrence to obtain the most relevant list of vocabulary needed for a particular thematic unit. This method is thoroughly explained in the paper alongside with the description of some other free software that can be successfully used in a process of or compiling glossaries.

Key words: Automatic term extraction, a foreign language, vocabulary, software