1(1)#7 2012

UDK 303.7+81'324
Pregledni članak
Review article
Primljeno: 25.2.2012.

**Mira Hercigonja-Szekeres, Nenad Sikirica, Irena Popović**

Veleučilište Hrvatsko zagorje Krapina
Šetalište Hrvatskog narodnog preporoda 13, 49000 Krapina
mira.hercigonja-szekeres@zg.t-com.hr
nsikirica@vhzk.hr
irena.popovic@vhzk.hr

# Statistička analiza tekstnih podataka

***Sažetak***

*Tekstni podaci čest su rezultat u raznim istraživanjima u sociologiji, psihologiji, marketingu, socioekonomskim analizama... Oni su rezultat ili specifičnih anketa ili analize ciljanih tekstova: (socio)političkih govora, literarnih radova... Takvi tekstni podaci sadrže mnoštvo informacija, ali su problematični za statističku analizu, jer se kodiranjem gubi mnogo informacija. Koristeći metode multivarijatne analize podataka, analiza tekstnih podataka (ATP) pojednostavljena je na statističku analizu riječi kao grafičkih formi, obradu leksičkih formi u totalitetu i učinjena nezavisnom o jeziku teksta. Metode ATP-a dijele se na leksikometrijske metode i statističke metode prilagođene tekstu, a obrada ima dva aspekta: statistički i kontekstni. U ovom radu su kao primjena ATP-a analizirani radovi studenata II. godine Veleučilišta Hrvatsko zagorje Krapina, smjer Operativni menadžment, sa zadanim naslovom „Elektroničko poslovanje".*

***Ključne riječi:*** *statistička analiza teksta, tekstni podaci, tekst, evaluacija nastave.*

## 1. Uvod

Za statističku analizu tekstnih podataka u suvremenoj literaturi sve se češće koristi naziv tekstna statistika upravo zbog primjena statističkih metoda koje su prilagođene obradi tekstnih podataka te kombinaciji s ostalim kontinuiranim i diskretnim veličinama koje se vrlo često pojavljuju uz tekstne podatke. Taj će naziv biti dalje korišten i u ovom radu. Zadnjih desetljeća 20. stoljeća tekstna je statistika doživjela niz promjena, kako u samom području istraživanja, tako i u svojim ciljevima i metodologiji koju koristi.

Tekstna statistika jedna je od više znanosti koja se bavi tekstom, odnosno tekstnim podacima. Neke od njih su lingvistika, analiza razgovora, analiza sadržaja, umjetna inteligencija. Sve se one u istraživanjima koriste kombinacijama lingvistike, matematike i računalnih znanosti, ali manje statističkim modelima i metodama. Uspješnom primjenom statističkih metoda u prirodnim i društvenim znanostima, slijedom toga i u područjima tih znanosti gdje se rabe tekstni podaci, povećava se i zanimanje istraživača za statističku analizu tekstnih podataka, osobito pedesetih i šezdesetih godina 20. stoljeća. U svojim počecima tekstna je statistika bila uglavnom leksikometrija te prepoznavanje ključnih riječi i njihovih tekstnih okruženja.

Unapređivanje informacijskih tehnologija (IT) omogućilo je stvaranje velikih baza podataka, njihovo pretraživanje i analiziranje. Tada se u statistici razvijaju multivarijatne metode statističke analize. One se sve više primjenjuju i na tekstne podatke. To je i vrijeme kada se pojavljuje i obrada prirodnog jezika (engl. *natural language processing*) koja postaje osobito zanimljiva za moguću analizu sadržaja odabranih tekstova. Objedinjavanjem svih tih dosega iz raznih područja u statističkoj analizi tekstnih podataka unapređuju se metode tekstne deskriptivne analize (unaprijeđena leksikometrija) i metode korespondencijske analize. U vrijeme prelaska iz drugog u treće tisućljeće informacijske tehnologije su napravile poseban iskorak u prikazivanju grafike tako da su i statističke analize tekstnih podataka dobile kvalitetan grafički prikaz rezultata, lako razumljiv te blizak i istraživačima s manjim predznanjima u statistici[85].

## 2. Metode statističke obrade teksta

Tekst se statistički analizira kao niz diskretnih kvalitativnih varijabli. Kolikogod se čini da takva primjena statističke analize teksta jednostavno rješava sve probleme u obradi teksta, noviji pristupi pokazuju mnoge nedostatke. Osobito je to vidljivo ako se želi u potpunosti automatizirati (statističku) analizu tekstnih podataka. Naime, postoji velika količina informacija koja se odnosi

85 Manning CD, Schutze H. Foundations of Statistical Natural Language Processing. 1999, Cambridge, MA: The MIT Press. str. 680.
Carroll J. Parsing, U The Oxford Handbook of Computational Linguistics, Mitkov R. ur. 2003, Oxford University Press: Oxford. str. 233-48.

na matricu podataka koja se analizira, a te se informacije ne pojavljuju u samoj matrici. To su metainformacije i njima obiluje svaki tekst. One mogu biti razne nejednakosti među podacima, odnosi među vrijednostima pojedinih varijabli, granične vrijednosti, ali i mnogo toga drugoga što se rabi gotovo rutinski ili intuitivno kad se radi o tekstu.

Svaka riječ je nedjeljiv dio teksta i nije ju moguće izostaviti bez posljedica za statističku tekstnu analizu. Posebnu pozornost pri analizi zahtijevaju gramatičke (neleksičke) riječi koje često zovemo prazne ili pomoćne riječi. One svojim velikim brojem i višeznačnošću opterećuju statističku analizu teksta, međutim, njihovo izostavljanje mijenja smisao teksta, a time je i analiza danog teksta nepotpuna.

## 2.1 Posebni tekstovi – odgovori na otvorena pitanja

Statistička tekstna analiza uspješno se primjenjuje na sve vrste tekstova: književne (prozu, poeziju i dramu) i stručne, iz raznih znanstvenih područja, a posebno se koristi u analizi odgovora na otvorena pitanja u raznim upitnicima.

Različiti upitnici za prikupljanje podataka primjenjuju se u mnogim istraživanjima na svim područjima znanosti. Zadnjih desetljeća vrlo se često koriste u socioekonomskim, poslovnim, marketinškim i političkim istraživanjima, ali i u vrednovanju u obrazovanju, medijima, zdravstvenoj i socijalnoj skrbi te javnim istraživanjima općenito. U sve su većem broju to upitnici s tzv. otvorenim pitanjima u kombinaciji s pitanjima na koja su ponuđeni odgovori i/ili se prikupljaju osobni podaci ispitanika te bilo koji drugi podaci.

Otvorena pitanja su pitanja na koja ispitanik odgovara tekstom prema osobnom izboru. Odgovore na otvorena pitanja zovemo otvoreni odgovori.

Otvorene odgovore možemo dobiti kao rezultat triju vrsta upita, a mogu biti i kombinirani u istom upitniku. To su:

- neformalni upiti kroz razgovor
- tematski ciljani upiti kroz razgovor
- upiti u standardiziranim upitnicima sa slobodnim odgovorima.

Slobodni odgovori prema svojem nastanku kao samosvojno izražavanje svakog pojedinog ispitanika nose mnogo informacija. Često su podaci u tim odgovorima redundantni, osobito u slučaju velikog broja ispitanika, pa su i veliki izazov, ali i veliko opterećenje za statističku analizu. To je razlog što se statistička analiza takvih podataka pokušava pojednostavniti naknadnim kodiranjem tih odgovora, što dovodi do niza pogrešaka koje su nepopravljive u daljnjoj statističkoj analizi.

Upitnici koji uključuju i otvorena pitanja imaju dodatnu vrijednost jer zahvaljujući naprednim informacijskim tehnologijama u statističkoj analizi tih upitnika možemo kombinirati sve vrste podataka. Otvoreni odgovori unose se u računalo u izvornom obliku i mogu se kombinirati s demografskim značajkama ispitanika te njihovim odgovorima na zatvorena pitanja. Takvim se postupcima otvoreni odgovori mogu kategorizirati i grupirati, a da se nimalo ne mijenjaju[86].

## 2.2 Postupci pri analizi tekstnih podataka

Analiza tekstnih podataka počinje njihovim unosom u računalo, nazvanim i kompjuterizacija teksta. Najčešće je to unos preko tipkovnice, međutim informacijske tehnologije danas omogućuju relativno jednostavan unos velikih količina teksta.

Tekstni podaci spremljeni u računalo u sljedećem se koraku segmentiraju, što znači podjelu teksta u tekstne jedinice, tj. minimalne dijelove teksta koji se više ne mogu dijeliti. Tekstna jedinica je definirana kao skup znakova između dvaju delimitera. Takvu tekstnu jedinicu zovemo grafička forma ili, češće, riječ. Uobičajeno je da se delimiterom za riječ smatra bjelina (praznina) kojom inače odjeljujemo riječi pri pisanju. Međutim, kao delimiter može se definirati, i obično se definira, čitav skup znakova, najčešće interpunkcija.

Jedno pojavljivanje riječi je niz znakova izvan skupa delimitera koji je s obje strane ograničen delimiterom. Dvije su riječi jednake ako su to dva jednaka niza znakova izvan skupa delimitera koji su s obje strane ograničeni delimiterom.

Vokabular čine sve različite riječi u danom tekstu. Veličina ili duljina teksta je ukupni broj pojavljivanja svih riječi u tekstu.

U tekstnoj analizi nakon segmentacije teksta slijedi numeričko kodiranje teksta. To je postupak kojim se svakoj riječi pridjeljuje njezin brojčani kôd ili jednostavno broj koji se u statističkoj analizi rabi kao elementarni podatak. Ti su kodovi spremljeni u rječnik riječi koji je jedinstven za svaku pojedinu primjenu. Nakon svih provedenih statističkih analiza s pomoću tog rječnika rezultati se prikazuju s pomoću početnih riječi.

Broj pojavljivanja pojedine riječi je vrlo različit. One riječi koje se pojavljuju samo jedanput posebno su istaknute svojim imenom. To su hapaksi. Taj naziv dolazi iz starogrčkog jezika *hapax legomenon* ( = jedanput izrečeno).

86 Lebart L, Salem A, Berry L. Exploring Textual Data. 1998, Dordrecht / Boston / London: Kluwer Academic Publishers. str. 245.
Lyncoln SY, Guba EG. Naturalistic Inquiry. 1985: SAGE Publication.
Jurafsky D, Martin JH. Speech and Language processing: An Introduction to Natural Language Processing, Computional Linguistics, and Speech Recognition. Prentice Hall series in artificial intelligence. 2009, New Jersey: Pearson education. str. 988.

Uobičajeno je da su analizirani tekstovi različite duljine pa broj riječi u tekstu može biti od nekoliko stotina do više tisuća. Vokabular je strogo vezan uz osobu čije su to riječi. To znači da svaka riječ svojim pojavljivanjem ili nepojavljivanjem karakterizira autora teksta. Istraživanja tekstova vezano uz duljinu i vokabular pokazala su opću karakteristiku vezanu uz učestalost pojavljivanja riječi. Ta je karakteristika nazvana Zipfov zakon (1949.) u čast G. U. Zipfu i glasi:

> Umnožak učestalosti pojavljivanja određene riječi u tekstu i njezina ranga u tablici učestalosti je konstantan.

To znači da će u svakom tekstu biti malo riječi velike učestalosti, a mnogo riječi male učestalosti. To je pravilo i uvjetovalo da se u tekstu posebno istaknu hapaksi, jer su oni najučestaliji.

Tekstna analiza nastavlja se leksikometrijom – prebrojavanjem riječi i stvaranjem indeksa cijelog teksta ili pojedinih njegovih dijelova. Indeksi mogu biti hijerarhijski (po učestalosti pojedine riječi) ili abecedni. U hijerarhijskom indeksu teksta riječi iste učestalosti poredane su prema abecedi.[87]

Ponekad se kod proučavanja samih riječi u danom tekstu provodi i lematizacija (stvaranje lema). Lematizacija je stavljanje svih promjenljivih riječi u osnovni oblik, primjerice sve imenice u nominativ jednine, sve pridjeve u nominativ jednine muškog roda, sve glagole u infinitiv i slično. Takav postupak provodi se uglavnom kod izučavanja književnih tekstova, a u nekim jezicima je razvijena i automatska lematizacija na računalu.

Nakon leksikometrije tekstna analiza nastavlja se uglavnom u dva smjera prema izboru istraživača:

- kontekstne analize
- statističke analize kvalitativnih podataka.

Kontekstne analize su automatizirani postupci kojima se pokušava približiti sadržaju kompjuteriziranog teksta. U tu se svrhu provode postupci konkordancije s odabranom (ključnom riječi) te pronalaženje segmenata koji se ponavljaju i njihova leksikometrija.

Statističke analize najčešće korištene za obradu tekstnih podataka su:

- korespondencijska analiza
- klaster analiza.

87 Lebart L, Salem A, Berry L. Exploring Textual Data. 1998, Dordrecht / Boston / London: Kluwer Academic Publishers. str. 245.
Woods A, Fletcher P, Hughes A. Statistics in language studies. 1996, Cambridge / New York / Melbourne: Cambridge University Press. str. 326.
Sirmakessis S. ur. Text Mining and its Applications. Studies in Fuzziness and Soft Computing. Vol. 138. 2004, Springer. str. 204. 35) Lebart L, Piron M, Morineau A. Statistique Exploratoire Multidimensionnelle (Visualisation et Inference en Fouille de Donnees) 4eme edition. 2006: Dunod. str. 480.

Korespondencijska analiza jedna je od metoda analize glavnih komponenti prilagođena za analizu tablica kontingencije i binarnih tablica. Osniva se na linearnoj algebri. Korespondencijska analiza je primjenjiva i na kategoričke varijable i na binarne varijable. Rezultat je, osim brojčanih vrijednosti, grafički prikaz (položaj točaka u ravninama projekcije) na kojemu geometrijska bliskost između elemenata retka i elemenata stupca prenosi statističku povezanost između redaka i stupaca, odnosno podataka u njima.

Klaster analiza je statistička metoda za utvrđivanje relativno homogenih grupa objekata klasificirajući pojedine jedinice analize s obzirom na njihovu sličnost, odnosno različitost prema nekim njihovim mjerenim obilježjima. U analizi tekstnih podataka ta se metoda rabi za određivanja bliskosti među elementima leksičke tablice (za retke ili za stupce) grupirajući ih, odnosno stvarajući klastere bliskosti.

Primjenjuju se dvije glavne metode klasteriranja:

- hijerarhijsko
- direktno sa zadanim brojem klastera.

Kod vrlo velikih skupova podataka koristi se kombinacija obiju metoda[88].

## 2.3 Računalni programi za obradu tekstnih podataka

Razvojem tekstne statistike razvijali su se i računalni programi za obradu podataka. Neki su komercijalni, a neki su iz skupine tzv. *free softwere* programa. Računalni programi se razlikuju i prema tekstnim podacima za čiju su obradi namijenjeni, tako da su neki bolji u primjeni na tekst bez dodatnih podataka, npr. književna djela i rasprave bilo koje vrste: filozofske, sociološke, znanstvene, i ostale. Drugi su prvenstveno namijenjeni analizi upitnika pa sadrže i mogućnost analize diskretnih i kontinuiranih veličina, a što im je osobita vrijednost mogu te analize kombinirati s analizom tekstnih podataka.

U ovom radu za računalnu obradu podataka korišten je program Dtm-Vic engleskog imena *Data and Text Mining – Visualization, Inference, Classification* verzija 5.2 iz 2011. godine. To je softver iz skupine tzv. *free software* programa. Namijenjen je u prvom redu studentima i znanstvenim istraživačima za statističku analizu složenih skupova podataka, koji sadržavaju i numeričke i tekstne podatke. Autor tog programa je Loudovic Lebart, a suautori su Monica Becue i André Salem. Program Dtm-Vic omogućava obradu podataka u skladu s najnovijim teorijskim spoznajama iz područja

88 Greenacre MJ, Blasius J. ur. Multiple Correspondence Analysis and Related Methods. 2006, Chapman-HallCRC.: Boca-Raton, FL. str. 581.
Greenacre, MJ. Correspondence Analysis in Practice, 2nd edition. 2007: Chapman & Hall\CRC. str. 280.
Friendly M. Visualizing Categorical Data. 2001, Cary, NC: SAS Institute. str. 436.

multivarijatne statističke analize te analize podataka dobivenih iz raznih upitnika koji uključuju kvantitativne i kvalitativne podatke, osobito odgovore na otvorena pitanja. Tri su njegove posebne prednosti:

- komplementarno korištenje metoda vizualizacije (analiza glavnih komponenata, korespondencijska analiza) i metoda klasteriranja,
- procjenjivanje metoda vizualizacije: metode ponovnog uzorkovanja
- analiza bliskosti (engl. *contiguity analysis*) i odgovarajuće metode.

Unos podataka moguć je direktno, ali i iz programa Microsoft Excel®, a tekstni podaci unose se u txt formatu. Podaci iz Dtm-Vic programa mogu se prenijeti u Microsoft Excel®, a rezultati su mogući i u formatu txt. Korištenje programom nije jednostavno, zahtijeva dosta predznanja iz teorije statističke analize, posebno analize tekstnih podataka. Međutim, taj nedostatak se kompenzira velikim mogućnostima statističkih analiza svih vrsta podataka i njihovim kombinacijama, te vrlo transparentnim rezultatima, osobito grafičkim.[89]

## 3. Primjer

Metode tekstne analize, osobito analize otvorenih odgovora i to kroz neformalne upite koristimo u Veleučilištu Hrvatsko zagorje Krapina (VHZK)[90] za evaluaciju nekih kolegija s ciljem poboljšanja kvalitete sadržaja pojedinih kolegija, ali i samog izvođenja nastave. Ovaj primjer je korišten u ovom radu kako bi se ilustrirali opisi metode tekstne analize te pokazala jedna od mogućnosti primjene tekstne analize u ne baš sasvim uobičajenoj primjeni.

### 3.1 Ispitanici i metode

U ovom radu su rezultati tekstne analize upita „Što mislite o elektroničkom poslovanju nakon odslušanog kolegija "elektroničko poslovanje"? Ta je anketa provedena nakon završetka predavanja i vježbi (30+30) kolegija Elektroničko poslovanje na drugoj godini studiju Operativnog menadžmenta akademske godine 2010/11. Osim pisanog dijela upita koji je ovdje u funkciji neformalnog upita kroz razgovor, studenti su zamoljeni za sljedeće podatke: dob, spol i završena srednja škola prije upisa na VHZK. Svim studentima koji su sudjelovali u ovoj anketi objašnjeno je da su dani podaci tajni i da se ni na koji način u rezultatima neće moći razaznati autor pojedinog odgovora.

89 Lebart L. Software DtmVic: Exploratory statistical Processing of complex Data sets Comprising both numerical and textual Data. 2009: Paris. http://www.dtm-vic.com

90 http://www.vhzk.hr/moodle/

### 3.2 Rezultati

Anketom je bilo obuhvaćeno 25 studenata druge godine studija Operativnog menadžmenta akademske godine 2010/11 i to 9 studenata i 16 studentica.

Prema dobi, jedino je bilo moguće podijeliti ih na dvije skupine – mlađi od 22 godine i stariji od 22 godine. Ta je raspodjela u Tablici 1.

| **Dob** | |
|---|---|
| 22 godina i mlađi | 18 |
| Stariji od 22 godina | 7 |
| Ukupno | **25** |

Tablica 1. Studenti prema dobi

Studenti se na VHZK upisuju sa završenim raznim smjerovima u srednjim školama. Pomnom analizom definirali smo tri skupine: komercijalist (ta je skupina i najbrojnija), zatim, završena srednja škola tehničkog smjera gdje smo ubrojili i smjer informatika te skupinu koja je završila ostale smjerove srednjeg obrazovanja (gimnaziju, ekonomsku, hotelijersko-turističku i sl.). Kako to izgleda u brojkama pokazano je u Tablici 2.

| **Završena srednja škola** | |
|---|---|
| Smjer komercijalist | 12 |
| Gimnazija, ekonomska, ... | 8 |
| Tehnički smjer | 5 |
| Ukupno | **25** |

Tablica 2. Studenti prema završenoj srednjoj školi.

Rezultati leksikometrijske analize pokazuju nam sljedeće – Tablica 3.

| | |
|---|---|
| Ukupni broj odgovora | 25 |
| Ukupni broj riječi | 1832 |
| Broj različitih riječi | 740 |
| Postotak različitih riječi | 40,4 |

Tablica 3. Rezultati leksikometrijske analize.

Leksikometrijska analiza pokazuje nam da studenti nisu bili baš „pričljivi“ u opisu svojih iskustava s kolegijem Elektroničko poslovanje, prosječno su koristili 73 riječi u svojim odgovorima i dodatno – svi su pisali uglavnom isto. Naime, koristili su ukupno samo 740 različitih riječi. Kako nije bilo potrebe za prepisivanjem, ostaje činjenica da su svi podjednako opisali svoja iskustva s tim kolegijem.

Najučestalije riječi ispisane su u Tablici 4.

| riječ | učestalost |
|---|---|
| i | 71 |
| sam | 55 |
| u | 53 |
| da | 49 |
| je | 48 |
| se | 32 |
| o | 30 |
| elektroničko_poslovanje | 26 |
| to | 23 |
| na | 20 |
| elektroničkom_poslovanju | 20 |
| sada | 19 |
| više | 18 |
| koje | 18 |
| poslovanja | 17 |
| elektroničkog_poslovanja | 16 |
| za | 15 |
| što | 14 |
| a | 13 |
| znala | 13 |
| prije | 13 |
| smatram | 12 |
| sve | 12 |
| nisam | 12 |
| su | 11 |
| do | 10 |
| nešto | 10 |
| može | 10 |

Tablica 4. Najučestalije riječi

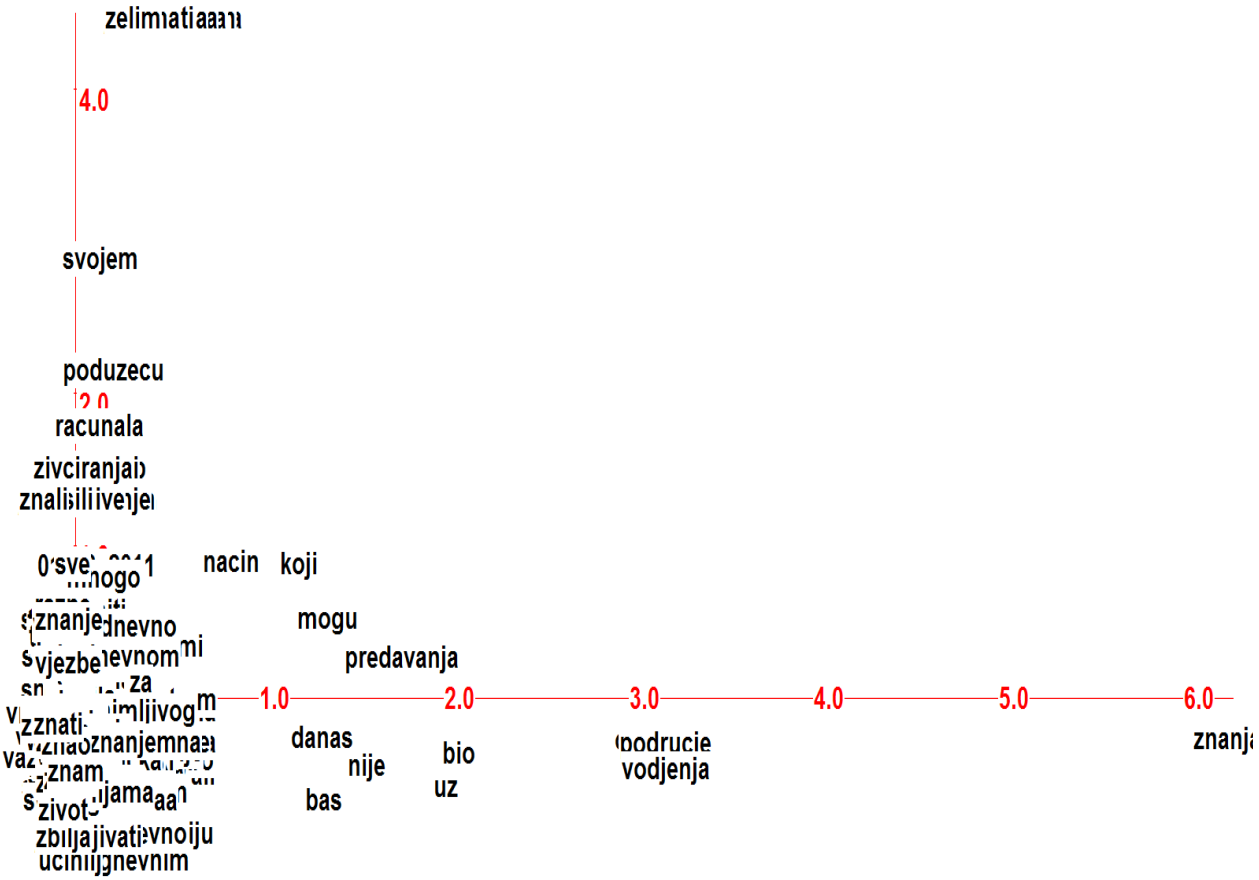
Najveću učestalost imaju tzv. pomoćne ili gramatičke riječi, uglavnom veznici, prijedlozi i oblici pomoćnih glagola. Te nas riječi ne približuju kontekstu odgovora, ali ih ne zanemarujemo. Pri velikom broju odgovora s puno različitih podskupina ispitanika bitan su dio analize. U slučaju malog broja odgovora, kao u ovom ispitivanju, prihvaćamo ih kao činjenicu.

Riječi koje nas približuju kontekstnim razmatranjima su *elektroničko_poslovanje* (u 3 padeža), *poslovanje*, *više*, *znala*, *smatram*, ... To će biti riječi koje će prvenstveno razlikovati ispitanike ili grupe ispitanika.

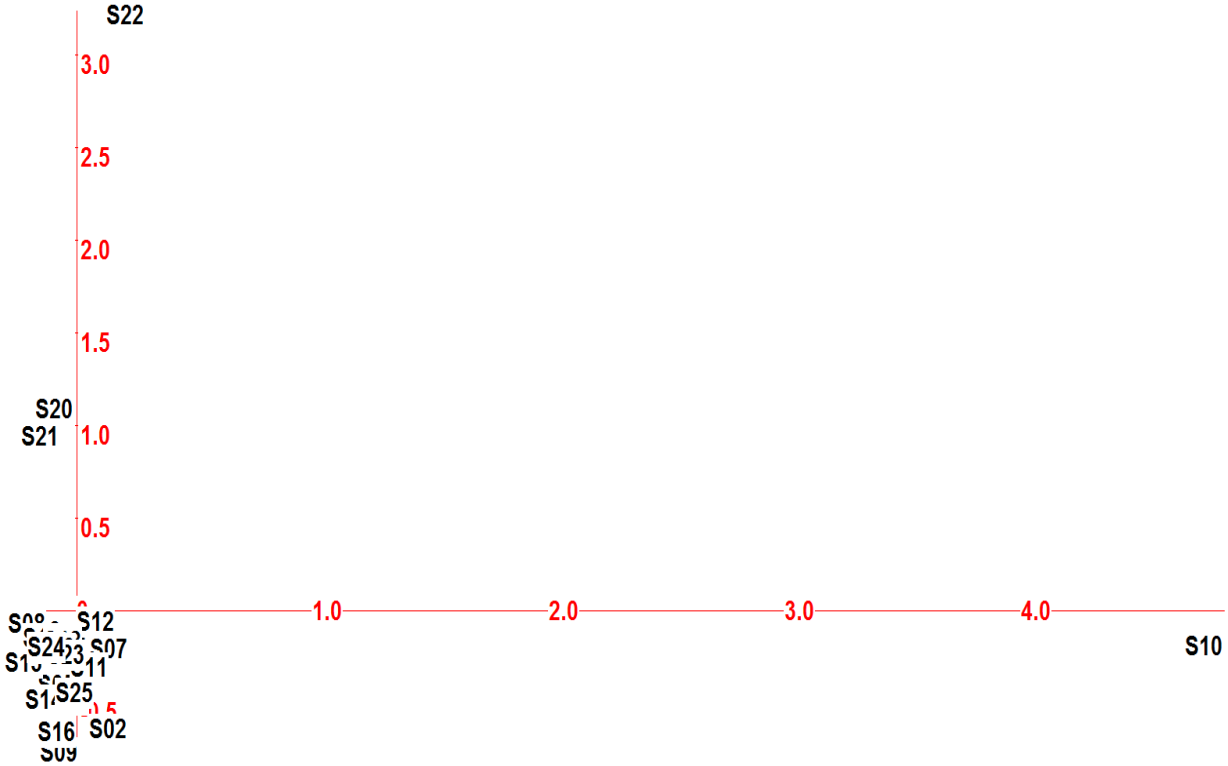
Nakon ovakve „grube“ analize broja riječi korespondencijska analiza treba dati „finiju“ analizu po skupinama ispitanika ili čak pojedinim ispitanicima.

Rezultati korespondencijske analize u grafičkom obliku dani su na Slikama 1-3. Na svim je slikama prikazana samo prva korespondencijska ravnina definirana prvom (horizontalnom) i drugom (vertikalnom) korespondencijskom osi.

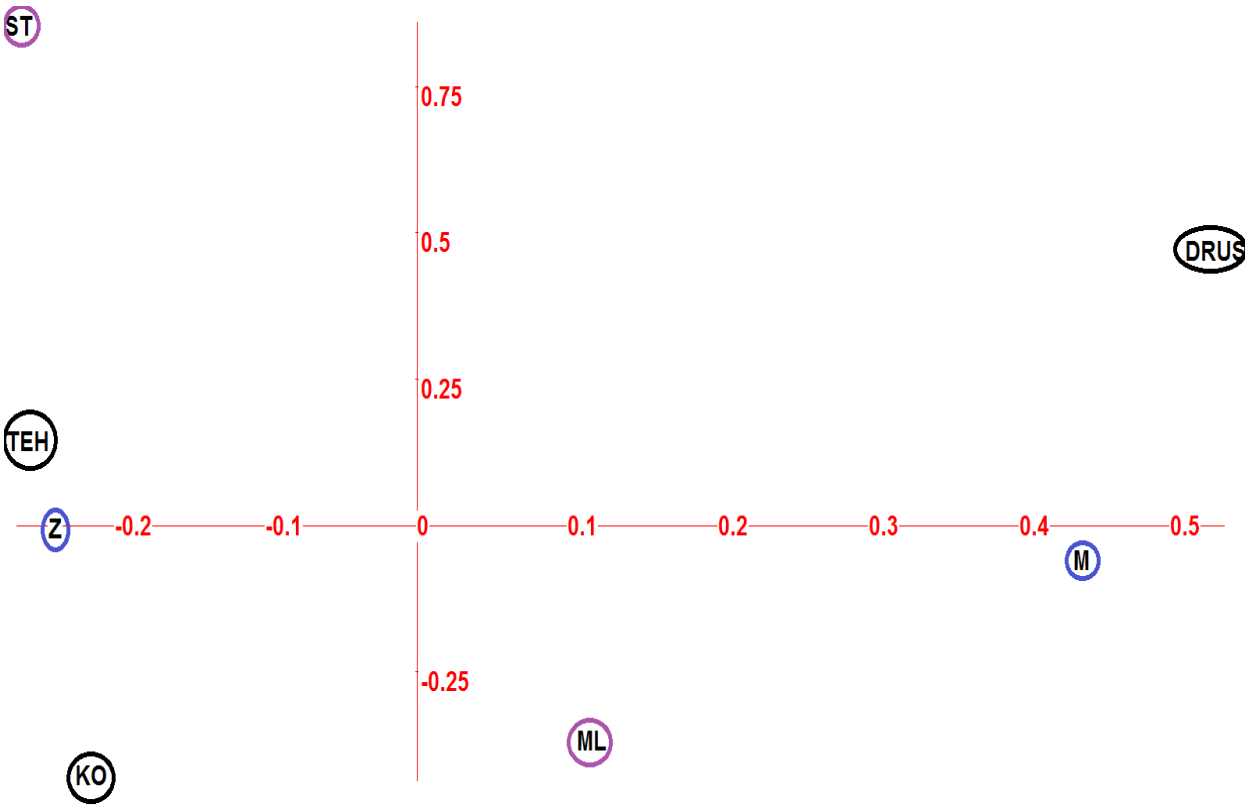
Slika 1. Rezultati korespondencijske analize po riječima



Slika 2. Rezultati korespondencijske analize po studentima




Slika 3. Rezultati korespondencijske analize grupiranje ispitanika prema ostalim podacima: spol, dob i završena srednja škola.



Rezultati korespondencijske analize na Slici 1. pokazuju da su studenti koristili gotovo iste riječi za opis svojeg mišljenja o kolegiju 'Elektroničko poslovanje' – riječi su nagomilane oko sjecišta prve i druge korespondencijske osi. Postoji manji broj riječi koje su upotrijebljene rjeđe i time razlikuju svoje „autore" od ostalih studenata.

Upravo to je potvrđeno na Slici 2. gdje su neki studenti (S10, S22, donekle S20, S21, S9) izvan one skupine nagomilane oko sjecišta prve i druge korespondencijske osi.

Više nam informacija daje Slika 3. na kojoj možemo uočiti razlike u izrečenim riječima prema spolu, dobi i završenoj srednjoj školi. Studenti i studentice razlikuju se prema korištenim riječima samo prema prvoj korespondencijskoj osi. Prema dobi, mlađi od 22 godine i stariji od 22 godine pokazuju različitost u upotrijebljenim riječima i po prvoj i po drugoj korespondencijskoj osi, s tim da je razlika prema drugoj osi nešto veća. Utjecaj podatka o završenoj srednjoj školi daje zanimljiviju različitost među ispitanim studentima s obzirom da su tri kategorije. Vidljivo je da su studenti sa završenom srednjom školom 'ostalih' smjerova (gimnazija, ekonomska i sl.) drugim riječima opisivali svoje mišljenje o kolegiju 'Elektroničko poslovanje' nego studenti koji su završili komercijalni ili neki tehnički smjer srednjeg obrazovanja – to nam pokazuje njihovo razlikovanje prema prvoj korespondencijskoj osi. S obzirom na drugu korespondencijsku os, studenti sa završenim srednjim komercijalnim smjerom razlikuju se od obje ostale skupine.

**Literatura:**

Carroll J. Parsin.: (u) *The Oxford Handbook of Computational Linguistics*, Mitkov R. (ur) 2003, Oxford University Press, Oxford. str. 233-48.

Friendly M.: *Visualizing Categorical Data*, 2001, Cary, NC: SAS Institute. str. 436.

Greenacre M. J., Blasius J (ur): *Multiple Correspondence Analysis and Related Methods*. 2006, Chapman-HallCRC.: Boca-Raton, FL.

Greenacre, MJ.: *Correspondence Analysis in Practice*, 2nd edition. 2007: Chapman & Hall\CRC. str. 280.

Jurafsky D, Martin J.H.: *Speech and Language processing: An Introduction to Natural Language Processing, Computional Linguistics, and Speech Recognition*. Prentice Hall series in Artificial Intelligence. 2009, New Jersey: Pearson Education. str. 988.

Lebart L,; Salem A,; Berry L.; *Exploring Textual Data*. 1998, Dordrecht / Boston / London: Kluwer Academic Publishers. str. 245.

Lebart L.; *Software DtmVic: Exploratory statistical Processing of complex Data sets Comprising both numerical and textual Data*. 2009: Paris. http://www.dtm-vic.com

Lyncoln SY,; Guba E.G.; *Naturalistic Inquiry*. 1985: SAGE Publication.

Manning C.D.; Schutze H.; *Foundations of Statistical Natural Language Processing*. 1999, Cambridge, MA: The MIT Press. str. 680.

Sirmakessis S. (ur).: *Text Mining and its Applications. Studies in Fuzziness and Soft Computing*. Vol. 138. 2004, Springer. str. 204.

Lebart L,; Piron M,; Morineau A.: *Statistique Exploratoire Multidimensionnelle (Visualisation et Inference en Fouille de Donnees)* 4eme edition. 2006: Dunod. str. 480.

Woods A,; Fletcher P,; Hughes A.: *Statistics in Language Studies*. 1996, Cambridge / New York / Melbourne: Cambridge University Press. str. 326.

http://www.vhzk.hr/moodle/

# Statistical Analysis of Textual Data

***Abstract***

*Textual data can be recognized as results of various studies i in sociology, psychology, marketing, socio-economic analysis ... They are the result of specific or target surveys and analysis of texts of (socio)political speeches, literary works ... Such texts contain a wealth of information, but they are problematic for statistical analysis, as coding them we lose a lot of information. Using the methods of multivariate data analysis, analysis of textual data (ATD) is simplified to the statistical analysis of words as graphical forms, the processing of lexical forms in totally and made independent of the language of text. We can recognize two methods: lexicometrical and statistical adapted to the textual data, and process has two aspects: statistical and contextual. In this paper the example of using ADT is analysis of students' answers on question "Electronic Business". Students are on 2nd year of study Operational management at Polytechnic Hrvatsko Zagorje Krapina.*

***Keywords****: statistical analysis of text, textual data, text, evaluation of lecturing.*