INFORMATIČKA VREDNOST AUTORSKIH KLJUČNIH REČI U ČLANCIMA PREDSTAVLJENIM U SRPSKOM CITATNOM INDEKSU
INFORMATION VALUE OF AUTHORS' KEYWORDS ASSIGNED TO ARTICLES AVAILABLE THROUGH SERBIAN CITATION INDEX

Tanja Jevremov, *Filozofski fakultet u Novom Sadu*

Sadržaj – *Istraživana je informatička vrednost ključnih reči u člancima predstavljenim u Srpskom citatnom indeksu. Analizirana je proporcija radova sa ključnim rečima na srpskom i engleskom jeziku za različite naučne oblasti. Prikazani su i analizirani brojevi ključnih reči po članku i distribucije dužine i učestalosti ključnih reči u različitim naučnim oblastima. Rezultati upućuju na nedovoljnu zastupljenost ključnih reči društvenim i humanističkim naukama, kao i njihovu nestandardizovanost i nedistinktivnost u svim naučnim oblastima. Uočena je potreba za unapređenjem postupka indeksiranja naučnih radova.*

Abstract – *Information value of keywords in scientific papers available through Serbian Citation Index was investigated. The proportion of papers with Serbian and English keywords was calculated for different scientific fields. The number of keywords per paper as well keywords' length and frequency distributions in different scientific fields were represented and analyzed. Results refer to an unsatisfactory incidence of keywords in social sciences and humanities as well as lack of their standardization and distinctiveness in all scientific fields. A necessity of improving indexing procedure in all scientific fields was noticed.*

1. PROBLEM ISTRAŽIVANJA

Ključne reči u naučnim publikacijama imaju funkciju sažetog predstavljanja njihovog sadržaja. One obezbeđuju osnovu za klasifikaciju dokumenata i formiranje mapa naučnih oblasti postupcima co-word analize. Ovim postupcima određuje se struktura naučnog prostora, koja je uslov za pretragu naučnih informacija i naučnu komunikaciju [1]. Pri tom, kvalitet strukturisanja naučnog prostora, a time i efikasnost pretrage i naučne komunikacije, zavisiće u prvom redu od kvaliteta indeksiranosti publikacija ključnim rečima.

Ključne reči mogu biti određene na više načina. Prvi je indeksiranje radova od strane autora ili indeksera, pri čemu korišćenje rečnika doprinosi standardizaciji ključnih reči određenih na ovaj način. Drugi način je ekstrahovanje ključnih reči iz teksta ili naslova referenci [1] [2]. U Srpskom citatnom indeksu - SCIndeks-u [3] radovi su indeksirani jedino na osnovu ključnih reči datih od strane autora radova. Ovako određeni deskriptori mogu da imaju veći broj nedostataka. Bez konsultovanja rečnika, određivanje deskriptora može biti u velikoj meri subjektivno, što za posledicu ima korišćenje različitih termina za isti pojam i njihovu umanjenu komunikabilnost. Pored toga, u poređenju sa deskriptorima ekstrahovanim iz teksta radova i referenci, ključne reči date od strane autora mogu sadržaj rada opisivati nedovoljno iscrpno i sa premalo termina, što za posledicu ima smanjenu vidljivost rada pri pretrazi [2] [4]. Neki od ovih nedostataka uočeni su kod deskriptora domaćih radova iz oblasti psihologije [5].

Cilj ovog rada je bio da se ostvari sistematski uvid u ove nedostatke. Postavljeno je pitanje koliko ključne reči u domaćim naučnim radovima mogu biti efektivna osnova za pretragu i upotrebljive za opis strukture naučnog prostora. U tu svrhu ispitan je kvalitet ključnih reči na osnovu njihove zastupljenosti, broja, dužine i distribucije u naučnim radovima. Pri tom se pretpostavilo da će radovi iz različitih naučnih oblasti imati različit kvalitet indeksiranosti. Naime, u ranijim istraživanjima je pokazano da je jezik društvenih nauka manje standardizovan u odnosu na jezik prirodnih nauka, što se ispoljava u njihovim različitim distribucijama učestalosti korišćenja termina [6] [7]. Pored toga, u domaćoj nauci se discipline razlikuju s obzirom na dostupnost tezaurusa za indeksiranje. Medicinske nauke, za razliku od ostalih, imaju na raspolaganju takav tezaurus (Medical Subject Headings).

2. METOD

Uzorak radova su činili naučni članci referisani u Srpskom citatnom indeksu koji su objavljeni u domaćim naučnim časopisima u periodu od 2002. do 2008. godine. Iz korpusa su izostavljeni radovi od kojih se ne očekuju ključne reči, kao što su prikazi, osvrti, uvodnici i sl. Radovi su razvrstani u pet naučnih oblasti. Zbog dvojne klasifikacije časopisa u kom su objavljeni, neki radovi su morali biti klasifikovani u više od jedne kategorije, te se javljaju ponovljeno u uzorku. Naučne oblasti u koje su radovi klasifikovani bile su: prirodno-matematičke nauke (6592 rada), medicinsko-biološke nauke (12852 rada), društvene nauke (16895 radova), humanističke nauke i jezik (7223 rada) i tehničko-tehnološke nauke (18174 rada).

Iz radova su ekstrahovane ključne reči date na srpskom i engleskom.

Sprovedene su sledeće analize:

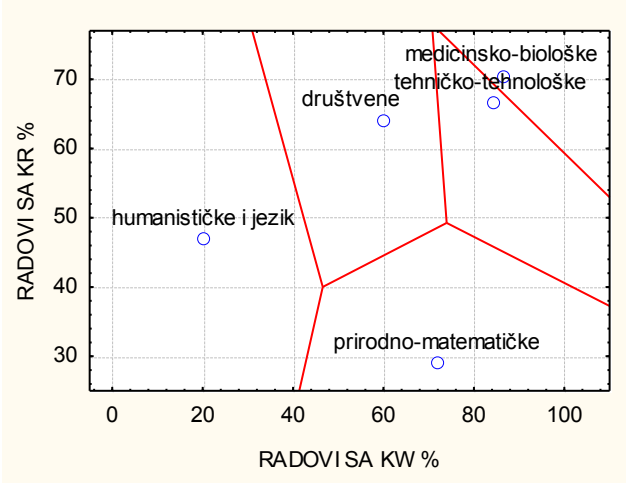
- Izračunat je broj radova koji imaju ključne reči na srpskom i engleskom po posmatranim godištima i naučnim oblastima.
- Izračunate su distribucije broja ključnih reči na srpskom i engleskom u radovima, po naučnim oblastima.
- Izračunate su distribucije broja reči od kojih su ključne reči sastavljene, po naučnim oblastima.
- Izračunate su distribucije frekvencija ključnih reči po naučnim oblastima i upoređene za teorijskom Zipf-ovom distribucijom raspodele reči u tekstu.
- Analiziran je sadržaj najfrekventnijih reči po naučnim oblastima.

3. REZULTATI

Rezultati analize zastupljenosti ključnih reči pokazuju da u posmatranom korpusu čak 22% radova nije indeksirano ni srpskim ni engleskim ključnim rečima. Uočljivo je, međutim, da se u periodu od 2002. do 2008. godine broj radova bez ključnih reči smanjio sa 37% na 11%.

Analiza zastupljenosti ključnih reči u radovima iz različitih naučnih oblasti otkriva različite "profile" ovih oblasti s obzirom na broj indeksiranih radova i jezik na kom su indeksirani (graf. n 1). Radovi iz oblasti madicinskih i tehničko-tehnoloških nauka pokazuju najbolju opremljenost ključnim rečima na oba jezika. Opozit njima su radovi iz oblasti humanističkih nauka i jezika, gde čak 45% radova nema ključne reči ni na srpskom ni na engleskom jeziku.

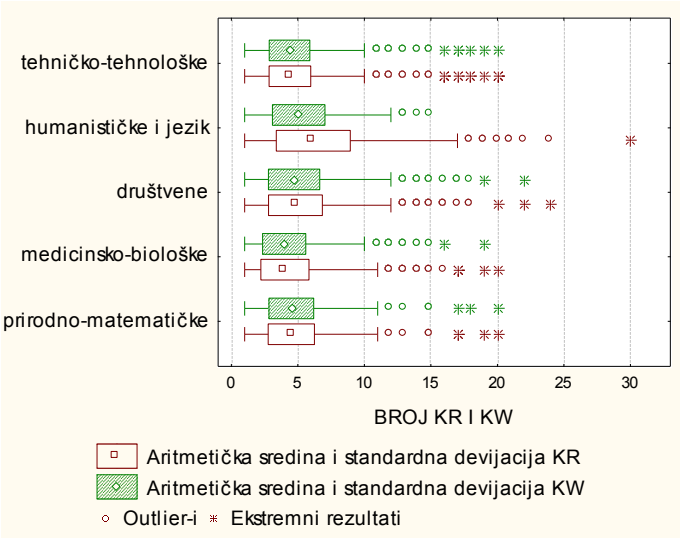
Grafik 1. Zastupljenost ključnih reči na srpskom (KR) i engleskom (KW) u naučnim radovima



Radovi iz oblasti društvenih i prirodno-matematičkih nauka su na suprotnim polovima po pitanju jezika na kom su indeksirani. Kod prirodno-matematičkih nauka ključne reči su većinom date na engleskom, a kod društvenih nauka na srpskom. Ovo je moguća posledica manjeg broja radova na engleskom u društvenim naukama, ali i slabije uređenosti radova u ovoj oblasti.

Osim za oblast humanističkih nauka i jezika, prosečan broj ključnih reči po članku ne prelazi pet, što je kvantitativno manje u poređenju sa prosečnim brojem ključnih reči u inostranim korpusima koji se navode u istraživanjima [1] [8]. Primetan je, međutim, i značajan broj članaka sa prevelikim brojem ključnih reči, većim od deset, za koje se može postaviti pitanje da li imaju funkciju ključnih reči ili možda pre nabrajanja svih pojmova iz teksta[1] (graf n. 2). U oblasti humanističkih nuka i jezika članci imaju nešto veći broj ključnih reči, no veći je i broj radova sa njihovim ekstremno velikim brojem.

Grafik 2. Broj ključnih reči na srpskom (KR) i engleskom (KW) u naučnim radovima



"Preusko" i stoga neinformativno indeksiranje vidno je u slučajevima ključnih reči velike dužine, sa preko deset reči, koje su pre cela rečenica nego pojam. Ovo je posebno uočljivo u oblasti humanističkih nauka i jezika[2]. Većina radova iz posmatranih naučnih oblasti je, ipak, indeksirana terminima sastavljenim od jedne ili dve reči (tab. n. 1). Relativno velika zastupljenost pojedinačnih reči ukazuje, pak, na značajno prisustvo ključnih reči velike opštosti. Naime, za ove reči se može pretpostaviti da su u odnosu na sintagme manje specifične i informativne.

Tabela 1. Zastupljenost ključnih reči u radovima iz različitih naučnih oblasti

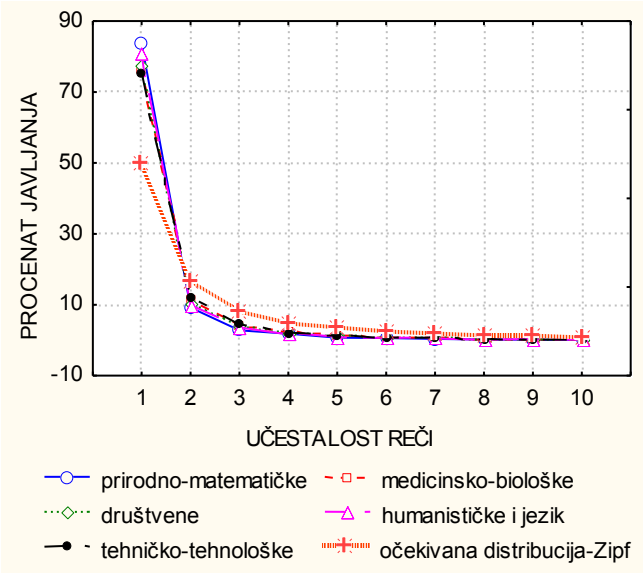
naučna oblast	% KR dužine 1	% KR dužine 2	% KR sa f =1
prirodno-matematičke	36.6	45.9	83.7
medicinsko-biološke	31.6	45.4	76.4
društvene	24.9	48.8	77.6
humanističke i jezik	39.3	43.9	80.8
tehničko-tehnološke	26.5	50.8	75.6

[1] Npr. jedan rad je indeksiran sledećim ključnim rečima: "Bog; Reč; Duh; stvaranje; pranačelo; nebo; zemlja; pratvar; svetlost; dan; noć; dobro; svod; kopno; mora; trava; drveće; svetila; duše; milenje; vreva; letenje; životinje; čovek; muško; žensko; hrana; počinak; rodoslov; predanje"

[2] Npr.: "putovanje sa prekidom u medijalnoj fazi i putovanje koje obrazuje kružnicu celog sižejnog toka"

Distribucije frekvencija ključnih reči određene su izrazitom dominacijom reči jedinične frekvencije (graf. n. 3), što je za raspodele učestalosti reči u tekstu očekivano [6]. Međutim, u oblasti niskih frekvencija dobijene distribucije značajno odstupaju od očekivane Zipf-ove distribucije. Dok teorijska distribucija predviđa 50% reči jedinične frekvencije, u dobijenim distribucijama reči jedinične frekvencije čine čak više od 75% svih reči (tab. n. 1).

Grafik 3. Distribucije učestalosti ključnih reči



Ovako nizak koeficijent ponavljanja ukazuje na nizak stepen standardizovanosti ključnih reči. Pri određivanju ključnih reči koristi se neujednačen vokabular i različiti nazivi za iste pojmove. Ovo je prisutno u svim posmatranim naučnim oblastima, čak i u prirodnim naukama, gde je očekivana veća standardizovanost jezika [6] [7]. Iznenađujuće je da značajno bolja situacija nije ni u medicinskim naukama, koje imaju na raspolaganju tezaurus za indeksiranje radova. Mogući razlozi ovom su širok fenomenološki i terminološki prostor koji medicinske nauke obuhvataju i relativno velik broj odrednica u tezaurusu.

Analiza sadržaja visokofrekventnih ključnih reči pokazuje da se u svakoj od naučnih oblasti mogu prepoznati ključne reči koje ukazuju na njena dominantna istraživačka interesovanja (tab. n. 2). Međutim, pored termina specifičnih za datu oblast i naziva nučnih disciplina, među ključnim rečima se učestalo javljaju termini velike opštosti i geografske odrednice. Zbog svoje višeznačnosti i mogućnosti da se koriste u različitim kontekstima, ovakve ključne reči nisu dovoljno distinktivne za opis sadržaja radova. Takve su, na primer, *Srbija*, *kvalitet* i *razvoj*, prisutne u gotovo svim naučnim oblastima.

Primetno je i da među najfrekventnijima nema ključnih reči koje bi se odnosile na metod istraživanja i opisivale uzorak, način istraživanja i analizu podataka.

Tabela 2. Klasifikacija najfrekventnijih ključnih reči u korpusu*

naučna oblast	pojmovi			
	opšti	geografski	naučne discipline	specifični
prirodno-matematičke	kvalitet, razvoj, životna sredina	Srbija	kinetika	materijali: bakar, zlato...
medicinsko-biološke	ishod, procedure, komplikacije	Srbija	hirurgija, epidemiologija	dijagnostika: ultrasonografija bolesti: infarkt miokarda, osteoporoza
društvene	kultura, kvalitet, razvoj, efikasnost, rizik	Srbija	pravo	obrazovanje: škola, nastava politika: tranzicija, globalizacija
humanističke i jezik	kultura, identitet, vreme	Srbija, Beograd	istorija, lingvistika	književnost: poetika, roman jezik: sintaksa, semantika
tehničko-tehnološke	kvalitet, razvoj, životna sredina	Srbija	ekologija, tehnologija	poljoprivreda: pšenica, kukuruz mehanizacija: traktor

*Navedene ključne reči su frekvencije veće od 50, osim za oblasti prirodno matematičkih nauka (f>30) i humanističkih nauka i jezika (f>40).

4. DISKUSIJA

Istraživanje kvaliteta ključnih reči u domaćim naučnim radovima ukazalo je na njihove nedostatke u opisu sadžaja radova. Uočeno je da, uprkos poboljšanju zadnjih godina, značajan broj radova nema ključne reči i da je prosečan broj ključnih reči po radu relativno mali. Pored toga, uočena je nestandardizovanost i neujednačenost ključnih reči, koja je posledica njihovog relativno arbitrarnog određivanja. Zapaženo je i da je među ključnim rečima prisutan velik broj opštih, nedovoljno preciznih termina.

Nije potvrđena pretpostavka o razlikama u kvalitetu indeksiranja radova iz različitih naučnih oblasti. Osim u opremljenosti ključnim rečima, prirodne, medicinske i tehničko-tehnološke nauke nisu se pokazale značajno boljim od društvenih i humanističkih.

Na osnovu distribucija ključnih reči može se pretpostaviti njihova smanjena upotrebljivost pri pretrazi informacija i utvrđivanju strukture naučnog prostora (co-word analizom). Stoga je potrebno da se sistem dodele ključnih reči radovima bitno unapredi. Ovo bi moglo da se postigne standardizacijom ključnih reči korišćenjem tezaurusa i rečnika za specijalizovane oblasti. Potrebno je povećati broj relevantnih i distinktivnih ključnih reči koje opisuju radove da bi se obezbedila njihova veća vidljivost pri pretrazi [4]. Pored toga, dodeljivanje zasebnih ključnih reči za opis područja istraživanja, metoda, uzorka i geografskog prostora obezbedilo bi matapodatke i osnovu za formiranje uređene strukture naučnih informacije kakvu nudi semantčki web [9].

5. LITERATURA

[1] Y. Ding, G. G. Chowdhury and S. Foo, Bibliometric cartography of information retrieval research by using co-word analysis, *Information Processing & Management*, 37, pp. 817-842, 2001.

[2] E. Garfield, KeyWords Plus: ISI's breakthrough retrieval method. Part 1. Expanding your searching power *on Current Contents on Diskette*, *Current Contents*, 32, pp. 5-9, 1990.

[3] P. Šipka, The Serbian Citation Index: Context and Content, Proc. *ISSI 2005 – 10th International Conference of the Society for Scientometrics and Informetrics*, pp. 710–711, 2005.

[4] T. C. Craven, DESCRIPTION meta tags in public home and linked pages, *LIBRES*, *11*(2), 2001. Preuzeto 22. 04.2009.sa: http://libres.curtin.edu.au/LIBRE11N2/index.htm

[5] T. Jevremov i D. Pajić, Oblasti psiholoških istraživanja u Srbiji opisane na osnovu koincidencije deskriptora naučnih radova, Saopštenje na *XIII naučni skup Empirijska istraživanja u psihologiji*, pp. 70-71, 2007.

[6] C. Kim, Retrieval language of social sciences and natural sciences: A statistical investigation, *Journal of the American Society for Information Science, 33*(1), pp. 3-7, 1982.

[7] M. Milas-Bracović, I. Barany i D. Boras, Zastupljenost ključnih reči iz naslova i teksta članka u njegovom autorskom sažetku, *Informatologia Yugoslavica, 17*(3-4), pp. 243-265, 1985.

[8] K. Mane and K. Börner, Content Coverage on PNAS in 1982-2001., *Proc at Mapping Knowledge Domains, Arthur M. Sackler Colloquium*, Irvine, CA, May 9-11, 2003.

[9] Y. Ding, D. Fensel, M. Klein and B. Omelayenko, The semantic web: yet another hip? *Data & Knowledge Engineering, 41*, pp. 205-227, 2002.