(objavljeno u: Infoteka , 2002, 3, 1-2, 13-21)

Dejan Pajić[1], Pero Šipka[1], Biljana Kosanović[2]

[1]Filozofski fakultet - Katedra za psihologiju, Novi Sad,
[2]Narodna biblioteka Srbije, Beograd

Skriptor - program za raščlanjivanje bibliografskih informacija

Apstrakt: Opisan je Skriptor, program za raščlanjivanje sadržaja i referenci iz periodičnih publikacija, razvijen za potrebe održavanja SocioFakta. S osloncem na pomoćne baze (liste autorskih imena, izdavača i sl.) i jednostavne algoritme za obradu srpskog kao prirodnog jezika, program automatski prepoznaje elemente časopisnih sadržaja i članačkih referenci (ime autora, naslov, izvor, kolaciju itd.) i dodeljuje im standardne labele, čime se obezbeđuje automatski transfer podataka u odgovarajuća polja baze.

Pored osnovnog modula za raščlanjivanje, program sadrži potprograme za konverziju kodnih rasporeda, pretvaranje velikih slova u mala u skladu sa pravopisom, promenu redosleda autorskog imena i prezimena, dopunjavanje nedostajućih informacija, kao i interaktivnu kontrolu i korekciju raščlanjenog materijala.

Skriptor sadrži program za instalaciju i detaljan sistem pomoći, koji operatora upućuje u upotrebu programa i upoznaje ga sa bibliografskim standardima koji se koriste pri izradi SocioFakta. Napisan je u jeziku Visual Basic for Application kao dodatak programu Microsoft Word.

Ključne reči: rašlanjivanje bibliografskih informacija, bibliografske baze podataka, citatne informacije, softver

Izrada bibliografskih baza podataka zahteva odgovarajuću pripremu materijala za unos u bazu. Najvažnija operacija u procesu pripreme je raščlanjivanje bibliografskih jedinica u skladu s nekim od standarda, čime se obezbeđuje smeštaj istovrsnih elemenata referenci (ime autora, naslov rada, izvor, kolacija, itd.) u zasebna polja baze. Velika važnost raščlanjivanja dolazi otuda što ono omogućuje efikasno korigovanje grešaka neizbežnih u procesu unosa, kao i normiranje bibliografskih informacija, nužno zbog velikog broja standarda koji se ravnopravno koriste. Korigovanje i normiranje bibliografskih informacija smatraju se neophodnim uslovom efikasnog pretraživanja.

Zahtev za valjanim raščlanjivanjem postavio se u izradi SocioFakta - jugoslovenske bibliografske baze za društvene nauke, u najoštrijem mogućem obliku [7]. SocioFakt je koncipiran kao citatna baza koja treba da posluži vrednovanju većeg broja naučnih subjekata i generisanju znatno većeg broja scijentometrijskih pokazatelja nego što je to slučaj u klasičnim citatnim bazama (tzv. citatnim indeksima: SCI, SSCI i A&HCI), po uzoru na koje je i nastao. Citatni indeksi u svom sadašnjem obliku predstavljaju kombinovane apstraktno-citatne baze. Pored apstrakata i drugih polja koja služe pretraživanju i trasiranju primarnih izvora (npr. naslovi, adresa autora itd.) citatni indeksi sadrže samo nekoliko polja, odnosno potpolja namenjenih evaluaciji naučnih subjekata, dakle onom što predstavlja posebnost i prednost citatnih baza. Naročito je mali broj polja namenjen evaluaciji na osnovu citiranosti (praktično jedno, s tri potpolja). Raščlanjivanje citatnih informacija na tako mali broj elemenata ne predstavlja naročito složenu aktivnost.

U SocioFaktu se reference daju u punom obliku, tako da je njihovo rašlanjivanje znatno zahtevnije. U svojoj poslednjoj, online verziji (**http://www.nbs.bg.ac.yu/sfakt.htm**), SocioFakt

podržava vrednovanje citiranosti izdavača (profesionalnih i akademskih izdavačkih organizacija), naučnih timova/projekata i naučnih skupova, kao i vrednovanje uredničkog doprinosa pojedinaca. Samo se za kvalitet časopisa generiše desetak nezavisnih pokazatelja koje citatne baze standardno ne nude, a u scijentometrijskoj literaturi se smatraju veoma vrednima. Takvi su npr. jezik citiranog izdanja ili izdavački oblik reference (monografija - zbornik s naučnog skupa - časopis - zbirka radova). Proširenje namene baze imalo je za posledicu bitno povećanje broja polja na koje se raščlanjuje svaka referenca. Taj broj nije standardan jer zavisi od vrste citiranog izvora i broja autora citiranog rada. U mnogim slučajevima broj (pot)polja iznosi više od 10.

S povećanjem broja polja smanjena je njihova diskriminabilnost. Na taj način raščlanjivanje je postalo veoma složen skup operacija. Za razliku od klasičnih citatnih indeksa gde je ono deo procesa unosa i operatorska aktivnost, u pripremi SocioFakta raščlanivanje je zasebna analitička delatnost visokog profesionalnog nivoa, koja objedinjava ekspertizu bibliografskog i naučnog (disciplinarnog) karaktera. S postupnim povećanjem broja časopisa predstavljenih u SocioFaktu (do broja 58, koliko ih se referiše počev od 2000. godine) raščlanjivanje je postalo "usko grlo" u procesu održavanja baze. Uporedo s porastom baze narastala je i potreba za automatizacijom procesa raščlanjivanja.

Automatizacija procesa raščlanjivanja

U elektronskom naučnom izdavaštvu u toku je proces preorijentacije proizvođača na ponudu citatnih baza u formi koja omogućuje njihovu eksploataciju na daljinu i u *web* okruženju, ponekad u vidu tzv. otvorenih arhiva. Paralelno se umnožavaju nastojanja da se automatizuje proces njihove izrade. Jedan od takvih pokušaja je i ResearchIndex (ranije CiteSeer) NEC Research Institute-a - sistem Internet aplikacija za automatsko kreiranje digitalnih biblioteka i autonomnih citatnih indeksa. ResearchIndex prikuplja naučne radove u Postskript, PDF i drugim formatima dostupnim na *World Wide Web*-u, raščlanjuje ih, normira i automatski generiše jednostavne pokazatelje naučnog učinka tipa citiranosti [6]. Mada je za potrebe kreiranja pouzdanog indeksa citiranosti u samom sistemu implementirano nekoliko tehnika za izjednačavanje različitih formata istog citata, polazi se od toga da ResearchIndex kao ulazne podatke najčešće ima formatirane, korigovane dokumente u kojima se literatura navodi u skladu sa nekim od poznatih stilova, odnosno standarda. Sa druge strane, ovde se ne može govoriti o raščlanjivanju u pravom smislu te reči već pre o izdvajanju relevantnih informacija (**information extraction**) pošto je najčešće dovoljno izdvojiti samo ime autora (obično prvoimenovanog), naslov, godinu publikovanja i brojeve stranica citiranog rada [9].

Doslednost i uniformnost u formatiranju članaka i navođenju literature, olakšava izdvajanje relevantnih informacija iz elektronskog dokumenta i omogućava relativno jednostavno raščlanjivanje referenci u skladu sa prepoznatljivim, unapred definisanim modelima, odnosno pravilima (npr: APA ili Vankuver stilovima citiranja). Traganje za podacima koji formiraju prepoznatljive šablone i procesiranje tih podataka u skladu sa uputstvima vezanim za određeni šablon, poznato je kao **template mining**. Ova tehnika ekstrakcije informacija pronašla je svoju primenu i u parsiranju, odnosno raščlanjivanju bibliografskih informacija [3]. Formiranje takvih šablona (templates) podrazumeva i utvrđivanje određenih, jedinstvenih identifikatora ili tzv. **tokena** koji se kasnije koristite kao sintaksički indikatori različitih delova reference, npr: "journal" za naziv časopisa, "eds." za oznaku urednika i zbornika, "i dr." za kraj spiska autora itd. Pored ovih indikatora logičke strukture dokumenta, sugeriše se i potreba za analizom strukture vezane za raspored tipičnih segmenata dokumenta [1].

Objedinjeno korišćenje tokena, formiranje šablona i pravila vezanih za redosled navođenja elemenata reference, identifikovanje čestih interpunkcijskih znakova i formata fonta i konsultovanje lista autorskih imena i naziva časopisa, pokazalo se zadovoljavajućom tehnikom

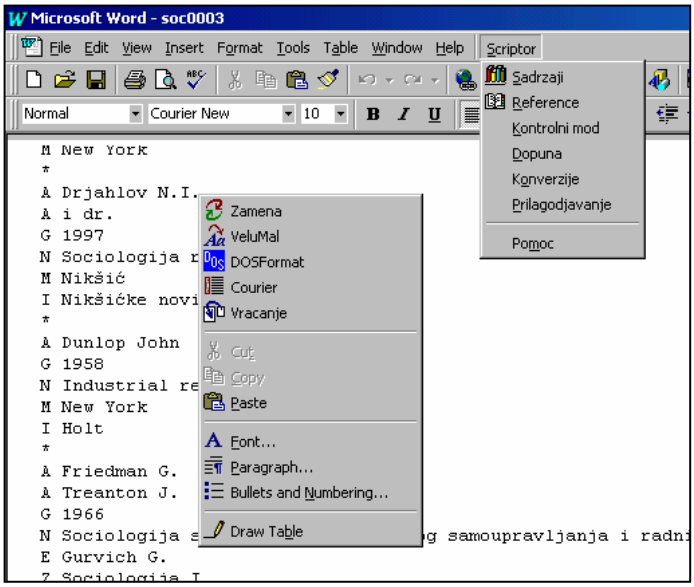
ekstrakcije i raščlanjivanja citatnih informacija [5]. Ipak, ovako formirani modeli predstavljaju najčešće nedovoljno bogat skup obrazaca. Oni ne pokrivaju u potpunosti veliku raznolikost u navođenju citatnih informacija. Svako iole značajnije odstupanje u formatu citata, npr. izostavljanje imena autora, dovodi do toga da se svi naredni elementi reference pogrešno klasifikuju (raščlane). Proces automatskog parsiranja bi se svakako značajno olakšao, ako bi se uredništva časopisa i sami autori doslednije pridržavali zahteva vezanih za pravilno navođenje citirane literature. S druge strane, postoje i predlozi da se pripremljeni šabloni za automatsko generisanje lista referenci učine javno dosupnim ili da se integrišu u često korišćene programe za obradu teksta [3]. Trenutno su, međutim, za ove potrebe autorima dostupne samo demo ili probne verzije komercijalnih programa kakvi su na primer ProCite, Bibliographix ili Citation.

Korak dalje u automatizaciji procesa parsiranja bibliografskih informacija, predstavljaju pokušaji da se kreiraju aplikacije koje će biti sposobne da s osloncem na probabilističke modele, kakvi su npr. Markovljevi lanci, "uče" pravila po kojima su podaci uređeni (10). Pravila se, zatim, u vidu nekog softverskog rešenja primenjuju na elektronske dokumente. Program najpre prepoznaje stil navođenja literature, a potom reference raščlanjuje primenom pravila vezanih za taj stil, čak i bez korišćenja rečnika ili baza imena. Aplikacije ove vrste, međutim, zahtevaju da se pre samog raščlanjivanja obavi "uvežbavanje" na već raščlanjenom materijalu struktuiranom u skladu sa određenim stilom [4]. Broj grešaka u raščlanjivanju značajno se povećava ukoliko se u materijalu javljaju autorska imena, nazivi časopisa ili termini koji nisu postojali u primerima na kojima je obavljeno uvežbavanje, odnosno učenje [9].

Skriptor

Okruženje

Skriptor je dizajniran kao **šablon** (template) za Microsoft Word, odnosno skup tzv. makroa koji sukcesivno izvršavaju veći broj operacija. Napisan je u programskom jeziku Visual Basic for Applications. Program se nakon instalacije integriše u Word 97/2000 i njegove opcije postaju dostupne delom preko novog menija koji se pojavljuje na paleti menija, a delom



Slika 1. Osnovni i priručni meni programa Skriptor

pozivanjem priručnog menija pritiskom na desni taster miša (Slika 1). Ovim je s jedne strane obezbeđen intuitivan interfejs za manje iskusne korisnike, a s druge mogućnost da se koriste

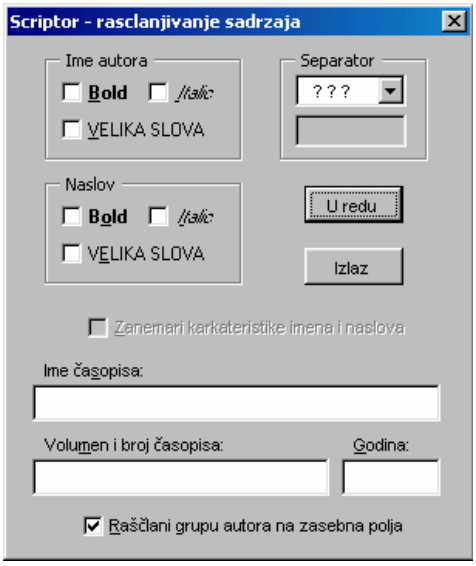
standardne funkcije programa Word za izmenu, brisanje i kopiranje dokumenata i njihovo snimanje u različitim formatima. Stvoreni su i uslovi za efikasniju integraciju s ostalim programima paketa Microsoft Office, prvenstveno programom Access, pošto se rečnici i liste autorskih imena, naziva časopisa, izdavača i gradova koji se konsultuju u toku raščlanjivanja, nalaze u mdb formatu.

Struktura

Osnovu programa Skriptor čine dva modula: modul za raščlanjivanje sadržaja periodičnih publikacija i modul za raščlanjivanje referenci u okviru svakog članka. Nakon raščlanjivanja, podaci se kontrolišu, koriguju, dopunjavaju i pripremaju za čuvanje u formatu koji omogućava automatsko smeštanje podataka u odgovarajuća polja baze. Ova priprema je olakšana velikim brojem pomoćnih rutina za: 1) konverziju različitih kodnih rasporeda uključujući konverziju ćirilice u latinicu i obratno, uz mogućnost transliteracije ruskog alfabeta, 2) pretvaranje velikih slova u mala u skladu sa pravopisom, odnosno uz konsultovanje pratećih rečnika, 3) zamenu mesta autorskog prezimena i imena, odnosno inicijala radi svođenja svake reference na jedinstven format 4) dopunu nedostajućih podataka i razvijanje skraćenih imena časopisa u pun oblik na osnovu liste, 5) interaktivnu kontrolu raščlanjenog materijala u skladu sa zahtevima i 6) upoređivanje svake raščlanjene reference sa njenim originalnim oblikom radi efikasnijeg otklanjanja eventualnih grešaka. Skriptor sadrži detaljan sistem pomoći u formi hiperteksta, tačnije tipičnu Windonjs datoteku pomoći koja omogućava jednostavno i efikasno pretraživanje uputstava u kojima se operateru opisuje način korišćenja samog programa, nude informacije potrebne za pravilno računjivanje i pripremu materijala i opisuju bibliografski standardi korišćeni pri izradi SocioFakta i bibliografskih baza uopšte.

Raščlanjivanje

Modul za raščlanjivanje sadržaja koristi se za formiranje tzv. "kostura" - početne datoteke koja sadrži osnovni bibliografski opis svakog rada u okviru jednog broja časopisa: ime autora, naslov, naziv časopisa, godinu publikovanja i kolaciju. Nakon što korisnik unese naziv časopisa, kolaciju, godinu publikovanja, kao i nekoliko osnovnih smernica vezanih za



Slika 2. Interfejs modula za raščlanjivanje sadržaja

karakteristike fonta i separatore korišćene pri razdvajanju različitih elemenata sadržaja (Slika 2),

rutine u okviru ovog modula, analizirajući prvenstveno izgled, karakteristike i raspored delova teksta, a po potrebi konsultujući i listu imena autora, strukturišu sadržaj periodične publikacije dodeljujući standardne labele odgovarajućim informacijama: AU - autor rada, NA - naslov rada itd. Ovako formiran kostur se nakon toga popunjava pratećim apstraktima, afilijacijama autora i citatnim informacijama koje se raščlanjuju u okviru drugog modula.

Rašlanjivanje članačkih referenci je daleko složenija i zahtevnija operacija koja, da bi se uspešno programerski implementirala, podrazumeva korišćenje većeg broja kriterijuma, sistema procena i pravila čijim se kumulativnim primenjivanjem dolazi do konačne logičke odluke o smeštanju informacije u odgovarajuću kategoriju, tj. polje baze. Algoritmi ovog modula se takođe oslanjaju na različite karakteristike delova teksta (kurziv, podvučen, zadebljan), njihov položaj u okviru rečenice/reference i postojanje tipičnih znakova interpunkcije. Ovo je, međutim, veoma retko pouzdan i dovoljno fleksibilan mehanizam identifikovanja različitih elemenata reference. Stoga su razvijeni dodatni algoritmi za obradu prirodnog jezika čiji je obavezan korak u procesu donošenja odluke konsultovanje baza autorskih imena, naziva časopisa, izdavača i gradova. Istovremeno se ispituje postojanje čestih sintaksičkih pokazatelja i znakova, tj. identifikatora određene vrste informacija. Neki od tih indikatora ili tokena dati su u Tabeli 1.

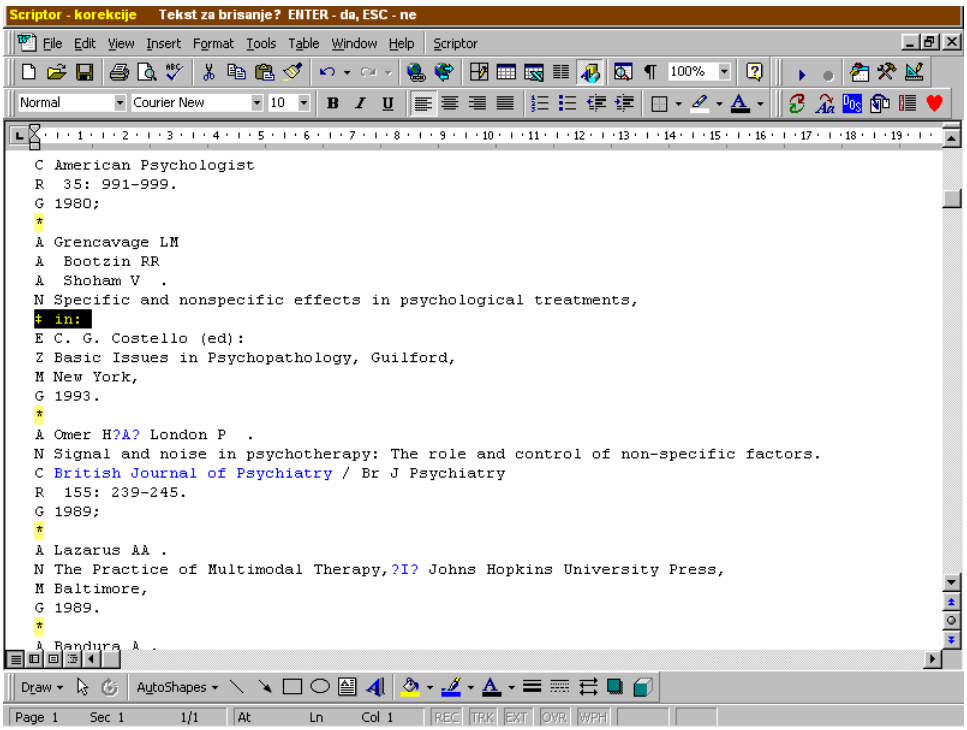
Tabela 1. Česti indikatori prisutnosti određenih bibliografskih informacija

VRSTA INFORMACIJE	**INDIKATOR**
naziv časopisa	ČASOPIS, ANALES, ANALI, ARCHIVES, BULLETIN, GLASNIK, JOURNAL, QUARTERLY, REVIEW, REVIJA, REVUE, ZEITCHRIFT, SERIES, BULL...
ime urednika	RED, EDITOR, EDS, ED, REDACTOR, REDS, EDITORS, HRSG, UR...
naziv izdavača	CENTAR ZA, PRESS, ZAVOD, PUBLISHERS, CO, INC, LTD, VERLAG, FAKULTET, UNIVERZITET, INSTITUT, SAVEZ, IZDATELSTVO, NIP...
kolacija	VOL, VOLUMEN, GODINA, GOD, BROJ, BR, STR, S, SS, P, PP, NUMBER, NR, NO, VOLUME, STRANA/E, STRANICA/E...
naslov zbornika	ZBORNIK, ZBORNIK RADOVA, U MONOGRAFIJI, U, IN, U KNJIZI, OBJAVLJENO U, DEO U...

Prvi korak u raščlanjivanju referenci predstavlja izdvajanje upravo onih polja, odnosno delova reference koji imaju relativno standardizovanu strukturu, stalan i prepoznatljiv položaj u tekstu ili jedinstvene sintaksičke identifikatore (invariants first). Obično se prvo pronalazi podatak o godini iz koje potiče citat i to kao niz od četiri cifre od kojih su prve dve 18, 19 ili 20 [2]. Uobičajeno je da se godina u okviru reference navodi ili neposredno nakon imena autora ili na samom kraju reference i u zavisnosti od toga, algoritam se grana. Ukoliko je godina navedena nakon imena autora, preostaje da se utvrdi sadržaj dela koji sledi nakon godine, kao i da se, ukoliko ih je navedeno više, pojedinačni autori izdvoje u zasebna polja. U drugom slučaju, kada je godina navedena na kraju reference, postupak je nešto složeniji: izdvaja(ju) se autor(i), prvenstveno uz konsultovanje liste autorskih imena i prezimena, a potom se raščlanjivanje nastavlja kao u prethodnom slučaju. Zbog velike raznovrsnosti materijala koji je

potrebno raščlaniti, kao i zbog nestandardnosti u navođenju referenci i neretkih grešaka u štampanju ili optičkom prepoznavanju teksta, praktično na svakom koraku u postupku raščlanjivanja moguća su veća ili manja odstupanja i usložnjavanja procesa donošenja odluke. Već u prvom koraku, kada se iz reference izdvaja polje G (godina), mehanizam odlučivanja se usložnjava ako je, na primer, u referenci navedeno više različitih godina - niski cifara od kojih su neke zapravo deo naslova ili čak predstavljaju brojčanu oznaku stranice.

Postupak raščlanjivanja referenci je u najvećim delu automatizovan i to u zavisnosti od **stepena slobode** koji korisnik definiše pre početka parsiranja. Što je stepen slobode veći to će program češće samostalno, bez konsultacije sa korisnikom, donostiti odluke o tome u koje polje se smešta odgovarajuća informacija. Sa druge strane, u situacijama kada je tekst neuredno formatiran, netipičan ili sadrži dosta grešaka, poželjno je postaviti stepen slobode na nižu vrednost, čime će se korisniku omogućiti da uzme više učešća u procesu interaktivnog parsiranja, prihvatajući ili preinačujući predloge koje program sugeriše. Ovakav način rada je neophodan i zbog postojanja određenih rutina u programu Skriptor koje izmenjuju, upotpunjuju ili brišu delove teksta i tako stvaraju rizik da se relevantni podaci izgube, a greške koje proizvede sam program prikriju.



Slika 3. Modul Korekcije i dopuna - interaktivno kontrolisanje procesa raščlanjivanja

Rutine, odnosno opcije koje omogućavaju interaktivno učešće korisnika programa u procesu raščlanjivanja grupisane su u modul nazvan "Korekcije i dopuna" (Slika 3). Postoje tri osnovna tipa korekcija, odnosno dopuna. Korisniku je omogućeno da prihvati, "overi" početak određenog polja, a time i granice dela teksta koji sadrži neku informaciju, pri čemu je dozvoljeno da se promeni ne samo mesto preloma reference (početak polja), već i naziv (labela) polja. Ovo je posebno korisno kada program, na primer, pogrešno prepozna naziv zbornika kao naziv časopisa. Nakon odluke korisnika, raščlanjivanje se nastavlja u skladu sa njom. Druga vrsta korekcija tiče se brisanja nepotrebnih delova teksta, npr. oznaka "u zborniku:" ili "in:", brojeva stranica kod monografija itd. Treću grupu čine napredniji algoritimi pomoću kojih se

informacije dopunjuju ili usklađuju sa formatom definisanim kao standard u citiranju i posebno kao standard u normiranju podataka u bazi SocioFakt. Ovde, na primer, spadaju rutine kojima se skraćeni nazivi časopisa, korišćenjem tehnika parcijalnog poređenja i izračunavanja međusobne udaljenosti niski, razvijaju u pune, standardne nazive.

Normiranje

Osnovne zamerke koje se upućuju klasičnim citatnim bazama, odnosno citatnim indeksima tiču se, pored očigledne pristrasnosti u izboru časopisa, nedopustivo velikog broja broja grešaka u unosu. Broj grešaka je toliki da značajno umanjuje efikasnost pretraživanja i dovodi u pitanje prikladnost baza za vrednovanje učinka različitih naučnih subjekata, koje je praktično nemoguće bez opsežnog dopunskog korigovanja informacija (data cleaning) i/ili visokih dopunskih troškova za upotrebu tzv. analitičke datoteke proizvođača baza, Instituta za naučne informacije iz Filadelfije.

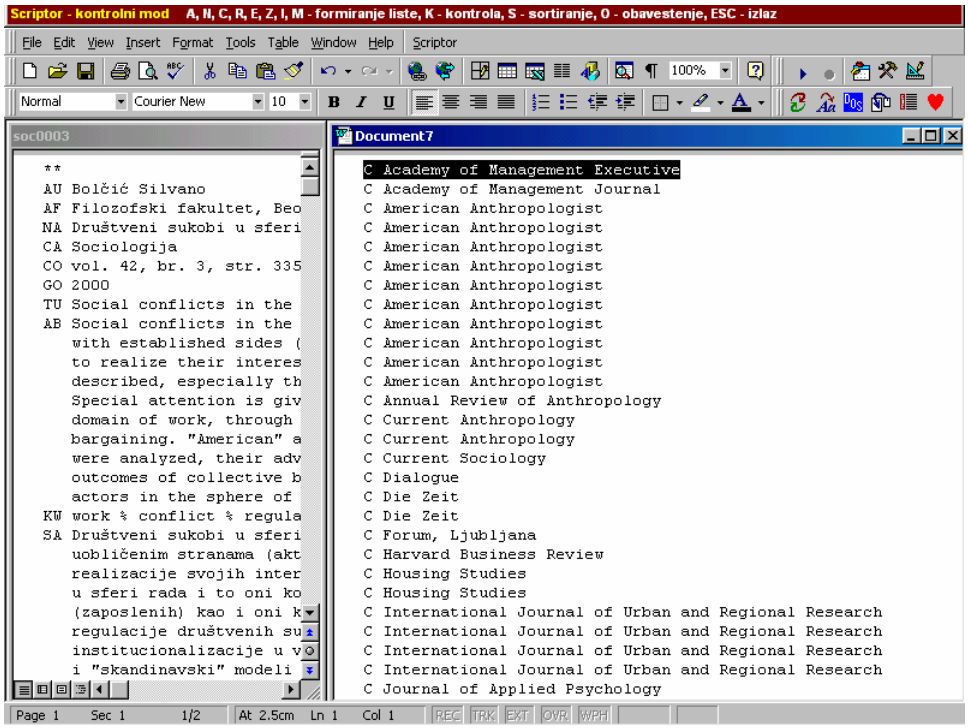
Tom problemu pristupilo se u izradi SocioFakta kao mogućoj ozbiljnoj prepreci za uvođenje SocioFakta u praktičnu upotrebu. Imajući na umu nalaze o distribuciji bibliometrijskih podataka iz baze, naročito nisku učestalost citata nemalog broja domaćih autora, institucija i časopisa, veliki uticaj grešaka u navođenju referenci na citatne pokazatelje, kao i potrebu da SocioFakt posluži vrednovanju što većem broju domaćih naučnih subjekata, zaključeno je da se podaci u bazi moraju korigovati i normirati. U te svrhe razvijen je poseban program **ProKos** kojim se istovrsne, a različito uobličene bibliografske informacije izjednačavaju (svode), a greške nastale, bilo u fazi unosa u bazu, bilo u pripremi i štampanju originalnih dokumenta - referisanih članaka - koriguju i uklanjaju. Prvi deo procesa korigovanja i normiranja, onaj koji zahteva uvid u originalne dokumete, racionalno je obavljati već u fazi raščlanjivanja.

Korigovanje i početno delimično normiranje informacija obavlja se u Skriptorovom Kontrolnom modu. Pozivanjem tog potprograma, korisnik se obaveštava o eventualnim tipičnim greškama u raščlanjenom materijalu. Greške se odnose na neslaganje formata materijala sa normama korišćenim pri izradi baze SocioFakt, npr. postojanje polja C (časopis) i I (izdavač) u okviru iste reference, dupliranost polja, izostanak određenih obaveznih polja, nepotpuno ili nejednako navođenje istih autorskih imena, itd. Podaci kojima je dodeljena pogrešna labela, mogu se lakše uočiti primenom tehnike "vertikalnog poređenja". Ona se obezbeđuje na taj način što se informacije istog tipa (npr. nazivi časopisa) izdvajaju u zaseban (privremeni) dokument i abecedno sortiraju u okviru jedinstvene liste (Slika 4). U toku kontrole korisnik se, zahvaljujući obezbeđenoj "prostornoj" korespondenciji osnovnog i privremenog dokumenta, može u svakom trenutku "vratiti" na raščlanjenu referencu, kako bi uočenu grešku i otklonio.

Evaluacija

Skriptor se intenzivno koristi u održavanju baze SocioFakt, tako da se njegova efikasnost neprestano ocenjuje, pojedinačne rutine veoma često usavršavaju, a baze dopunjavaju. Povremene sistematske provere njegove efikasnosti pokazuju da rezultati u velikoj meri zavise od jezičke i štamparske urednosti, kao i vrste materijala koji se raščlanjuje. U slučaju tipičnih, kompletnih i visoko strukturiranih referenci, odnosno referenci koje pripadaju naučnim oblastima pokrivenih SocioFaktom, a navedene su u skladu sa nekim od standarda u citiranju, uspešnost raščlanjivanja izražena proporcijom pravilnih kategorizacija prelazi 0,97. Taj udeo, međutim, značajno opada onda kada su reference netipične, karakteristike fonta nedosledno primenjivane, raspored elemenata izmenjen u odnosu na većinu standarda ili kada su bibliografski izvori ugrađeni u napomene (opaske, komentare), što je često slučaj kada se reference daju u tzv. fus- ili end-notama, umesto u zasebnom odeljku na kraju teksta. No, s

obzirom na to da je Skriptor alatka koja se koristi u interaktivnom režimu rada i da korisniku-analitičaru pruža veoma komforne uslove za proveru i prepravke automatski generisanih rešenja, tačnost odluka u raščlanjivanju ne mora biti najvažniji kritetrijum njegove efikasnosti. Bar podjednako važan kriterijum je brzina raščlanjivanja koja analitičar postiže u odnosu na rad na klasičan način ("peške"), kakav omogućuju savremeni programi za obradu teksta svojim standardnim rutinama. U vezi s tim, potreba za nekakvom sistematskom evluacijom nije se ni ukazivala, pošto je nesumnjivo da Skriptor višekratno ubrzava postupak raščlanjivanja.



Slika 4. Vertikalno poređenje u okviru Kontrolnog moda

Zaključak

Zbog velikih i teško predvidljivih varijacija u formatima navođenja literature, kao i atipičnosti značajnog broja referenci, automatsko raščlanjivanje bibliografskih informacija domaćeg porekla činilo se idejom koja ne obećava uspešnu realizaciju. Međutim, pokazalo se da je čak i bez primene složenih i resursima zahtevnih statističih metoda, jedan ograničen, ali bogat i fleksibilan skup pravila mogućno pretvoriti u aplikaciju koja proces parsiranja značajno ubrzava i olakšava. Prednost ovakve aplikacije nije samo u ekonomičnosti, već i u tome što obezbeđuje viši kvalitet u raščlanjivanju zahvaljujući tome što ostvaruje određeni nivo "profesionalnosti" (znanja), konsultujući znatno veću količinu korisnih informacija od one kojom raspolaže prosečan obučeni analitičar.

Valja istaći da je efikasnost Skriptora "domen-specifična". Svaki pokušaj njegove primene izvan naučnih oblasti pokrivenih SocioFaktom zahtevao bi ne samo dopunu i ažuriranje pomoćnih baza, već i prilagođavanje algoritama standardima citiranja koji se više koriste u drugim naučnim disciplinama. To jednako važi za druge vidove njegove eventualne upotrebe, kakve su formatiranje referenci u toku pripreme za štampu (knjiga, časopisnih članaka, bibliografija i sl.) ili retrospektivna konverzija.

Literatura

1. Aiello M., Monz C., Todoran L. (2000) Combining linguistic and spatial information for document analysis, u: Mariani J., Harama D. (ur.) *Proseedings of RIAO '2000 Content-Based Multimedia Information Access*, CID, 266-275
2. Bergmark D. (2000) *Automatic extraction of reference linking information from online documents*, Technical Report TR 2000-1821, Cornell University - Computer Science Departement, November, URL: http://www.cs.cornell.edu/cdlrg/Reference Linking/extraction.pdf, preuzeto 25.02.2002.
3. Chowdhury G.G. (1999) Template mining for information extraction from digital documents, *Library Trends*, 48 (1), 182-208
4. Connan J., Omlin C.W. (2000) *Bibliography extraction with hidden Markov models*, URL: http://citeseer.nj.nec.com/294556.html, preuzeto 11.06.2001.
5. Ding Y., Chowdhury G.G., Foo S. (1999) Template mining for the extraction of citation from digital documents, u: *Second Asian Digital Libraries Conference, National Taiwan University, November 8-9*
6. Giles L.C., Bollacker D., Lawrense S. (1998) CiteSeer: An automatic citation indexing system, u: Witten I., Akscyn R., Shipman F.III (ur.) *Digital Libraries - Third ACM Conference on Digital Libraries*, ACM Press, New York, 89-98
7. Kosanović B., Šipka P. (1996) SocioFakt - Jugoslovenska baza za društvene čiweničke nauke, u: Kostić P. (ur.) *Merewe u psihologiji*, IKSI i Centar za primewenu psihologiju, Beograd, 2, 85-95
8. Lawrense S., Giles L.C., Bollacker D. (1999): Digital libraries and autonomous citation indexing, *IEEE Computer*, 32 (6), 67-71
9. Seymore K., Mccallum A., Rosenfeld R. (1999) Learning hidden Markov model structure for information extraction, u: *AAAI '99 - Workshop on Machine Learning for Information Extraction*

Dejan Pajić[1], Pero Šipka[1], Biljana Kosanović[2]

[1]Faculty of Philosophy - Department of Psychology, Novi Sad,
[2]National Library of Serbia, Belgrade

Scriptor: Bibliographic information parsing program

Abstract: Skriptor - a program developed for the use in maintaining SocioFact and designed for parsing journals' contents and references is described. By making use of auxiliary databases (e.g. lists containing authors and publishers' names) and simple algorithms for processing Serbian as natural language, the program recognizes the elements of the journals' contents and articles' references (e.g. author name, book title, journal title, page numbers) and assigns a standardized label to each of those elements, providing automatic transfer of information into the respective database field.

Apart from basic parsing module, the program provides subroutines for conversions of various character sets, word (de)capitalization according to orthographic rules, inversion of author's name and surname position, filling up the missing data, as well as interactive control and correction of the parsed information.

Scriptor comes with an installation program and detailed help file which contains specific instructions for the operators explaining ways to effectively use program itself, and defining bibliographic standards used in the process SocioFact maintenance. Skriptor is written in Visual Basic for Applications as an Microsoft Word template.

Keywords: bibliographic information parsing, bibliographic databases, citation information, software