FILOLOGIJA 22-23, Zagreb 1994

UDK 808.62-3 : 808.62-1
Izvorni znanstveni članak
Primljeno XII/1993.

Maja CVITAŠ
Muzejski dokumentacioni centar, Zagreb

# PROBLEMI PRAVOPISNOGA RJEČNIKA

Pravopisni rječnik može se vidjeti kao skup pravopisno signifikantnih riječi. Takav rječnik dio je svega jezičnoga sustava. Korištenje elektronskih medija potiče raspravu o normiranu modelu rječničkih podataka. Jedan hipotetički podatkovni model testiran je na Babić–Finka–Moguševu pravopisnom rječniku.

## Okviri pravopisnog rječnika

Pravopis kao dogovor o pismenoj realizaciji jezika artikulira se na dva načina, skupom pravila te pravopisnim rječnikom. Pravila će dati odgovor na sustavna pitanja (na primjer prvo slovo u rečenici). Bez obzira na potpunost skupa pravila postoje direktna pitanja na koja bismo teško dobili odgovor primjenom odgovarajućih pravila (momče ili momće?). U slučajevima takvih pitanja koristimo pravopisni rječnik koji brže, uz manje jezičnoga znanja daje odgovor.

Pravopisni rječnik treba obuhvaćati skup "pravopisnih natuknica" koji sačinjavaju svi pravopisno karakteristični oblici riječi. Tako će se u rječniku pojaviti vokativ jednine momče imenice momak dok je nominativni oblik pravopisno neinteresantan, pa se neće pojaviti kao pravopisna natuknica.

## Izbor medija i reprezentacija jezičnog sustava

U sveukupnom životnom okolišu elektronski se medij sve češće nameće kao nužnost. Mnoge teoretski već poznate metode postaju pomoću novih tehnologija moguće i u praksi. Upotrebom računala vrijeme potrebno za njihovo provođenje smanjuje se na veličine koje su savladive u realnom ljudskom vijeku. Poznato je da su metode multivarijatne analize podataka, bogate operacijama matričnog računa, svoju upotrebnu vrijednost dobile tek pojavom računala.

Obrada jezika jedna je od zahtjevnijih disciplina koja također upotrebu računala mora shvatiti kao jednu od pretpostavki u svojoj realizaciji. Pravopisni je rječnik dio općega jezičnog sustava. On je samo jedan od načina kojim izražavamo naše dogovore o pismenoj jezičnoj realizaciji, a podliježe svim zakonitostima koje sveu-

kupan sustav propisuje. U tim okvirima nasljeđuje i model svoje realizacije.

Da bi se jezik lakše izučavao, definirat će se model podataka koji ga što vjernije i sveobuhvatnije reprezentira. S obzirom na svoju kompleksnost i veličinu, te moguće načine iskorištavanja, logično je da taj model svoju realizaciju doživi na magnetskom mediju.

Svjesni smo činjenice da jezik postoji bez obzira na naše dogovore i propise o njegovoj realizaciji. Jednako tako i podaci koji ga reprezentiraju ne smiju zavisiti o procesima za njihovo korištenje. Model podataka mora biti spreman da odgovori na zahtjeve znanih, ali i budućih procedura i primjena.

Osim nezavisnosti podataka o načinu njihove upotrebe treba uspostaviti i njihovu unutarnju strukturu. Prvenstveno se moraju identificirati latentne nezavisne podatkovne cjeline.

Svaka od tih cjelina neka je zatvoren skup parametara sustava koji gradimo. Tako normaliziranim podacima, principom "jednog podatka na jednom mjestu" izbjegava se redundantnost informacija u sustavu, a omogućava se njegova modularna izgradnja i promjena. Svaku cjelinu možemo zamišljati kao popis podataka oblikovan u skladu s nekim kriterijem. Osigurava se nezavisno promatranje i izučavanje tih pojedinih dijelova, a mogućnost višestrukoga povezivanja podataka unutar istog popisa te podataka iz različitih popisa otvara prostor reprezentaciji različitih jezičnih aspekata i podsustava.

Postavljanje općega modela podataka jezičnoga sustava vrlo je složen proces. Zahtijeva opsežna znanja o jeziku, ali i o mogućnostima njihove reprezentacije u skladu s izabranom tehnologijom. Učenje međusobnom razumijevanju i komunikaciji za stručnjake dviju struka u takvm interdisciplinarnom zadatku bitan je korak u savladavanju cilja. Prikupljanje elemenata znanja kroz pogled tehnologije i medija rezultira u postavljanju hipotetičkoga modela podataka. Trenutni je cilj proizvodnja pravopisnoga rječnika kao jednog od jezičnih alata. Odabrat će se postojeći pravopisni rječnik kao izvor inicijalnih podataka. On će istovremeno služiti za provjeru općega modela jezičnog sustava u izgradnji i modela pravopisnoga rječnika kao njegova dijela.

Konačna realizacija pravopisnoga rječnika u skladu s konačnim općim modelom podataka krajnji je cilj koji želimo ostvariti.

## O obradi teksta

Jezik se realizira u dva oblika, govornom i pismenom. Metode za obradu teksta i za obradu glasa spadaju među vrlo složene postupke i tehnike, a međusobno se vrlo razlikuju. Odabrat će se samo jedan smjer kojem će se posvetiti pažnja. To je korpus tekstova kao izvora jezičnih podataka.

Uz probleme oko odabira reprezentativnih tekstova potrebno je selektirati tehnike, prilagoditi ih za jezično područje ili pak djelomično razviti modele za obradu korpusa.

Taj se zadatak može promatrati kroz tri pravca djelovanja:

— Prijenos tekstova na magnetski medij,
— Označivanje teksta za obradu,
— Pretvaranje teksta u strukturirane podatke.

U stvaranju korpusa elektroničkih tekstova hrvatskoga jezika treba se suočiti s prijenosom tiskanih materijala na magnetski medij (skaner i OCR programska podrška) te uklanjanjem pogrešaka koje nastaju tom prilikom. Nove su tehnologije više ili manje osjetljive na vrste papira i kvalitetu te oblik tiska. Potrebno je ukloniti pogreške koje nastaju prijenosom da bi se tako dobio tekst u elektronički čitljivu, ispravnom obliku. Uspostava radne stanice za automatizirano ispravljanje unesenih tekstova bila je jedan od međukoraka na tom putu.

Iako se uz tekst mogu koristi i grafičko-likovni elementi izražavanja, ideja autora, na putu transformacije u pisani oblik, gubi jedan svoj dio i uvijek je samo dijelom objašnjena. Poruka čitatelju ostvaruje se strukturom dokumenta pri čemu je tekst sa stajališta obrade jezika dominantna komponenta. Sasvim je jasno da se, načinom kako je tekst otisnut na stranici knjige, izborom pisma i položajem, te omjerom tekstualne i izvantekstualne komponente dokumenta, pomaže u ostvarenju cjelovite zamisli.

Izgled dokumenta, odabrano grafičko rješenje, svako razlikovanje dijela od cjeline, ima svoj razlog. Razni su načini da se nešto naglasi ili objasni, na nešto upozori, prenese određena poruka.

Kako priprema tekstova za kompjutorsku obradu predstavlja velik posao, poželjno je pokušati shvatiti i prepoznati što više poruka koje dokument emitira, bez obzira na primarni interes kod stvaranja korpusa tekstova.

Naglašenost dijelova teksta, njihov raspored i mnoštvo drugih informacija potrebno je zabilježiti putem oznaka. Sustav za označivanje treba odabrati tako da značenje oznaka može biti višestruko, sustavno jednostavno promjenljivo.

Potrebno je osigurati dovoljno općenit i modularan način strukturiranja dokumenta koji ostavlja prostor nesmetanoga dodavanja novih načina gledanja na isti dokument. Ta dimenzija se ostvaruje "slojevitim" pridjeljivanjem konkretnog značenja elementima strukture, njihovim međusobnim vezama te pridruženim atributima.

Visoka je dostignuća, s ciljem razmjene i analize, na polju označivanja teksta dosegla grupa stručnjaka TEI (*Text Encoding Initiative*). U tom radu intenzivno sudjeluje i široka računalna zajednica koja se bavi obradom teksta i jezika. Oblikovan je apstraktan sustav SGML (*Standard General Markup Language*), na bazi standarda ISO 8879, a mnoga su pojedinačna područja već do te mjere razrađena da čine preporuke u sklopu korisničkoga priručnika.

Preporuke formata odnose se na razmjenu, lokalnu obradu teksta i proizvodnju novih tekstova. Njegova upotreba u lokalnoj obradi teksta ili proizvodnji teksta uvijek je na razini lokalne odluke te se očekuje dosta odstupanja od općih preporuka. Dakako, strože se definicije moraju osigurati u okviru razmjene podataka.

Opća je pojava da se jezični sustavi u izgradnji sastoje od dva dijela. Tekstovni dio, opremljen sustavom za označivanje, služi kao izvor podataka i informacija, a

korespondirajući leksički kao rezultat pohrane proizvoda analize označenih tekstova. Posve je logično zamišljati bazu tekstova povezanu s leksikonima, rječnicima i gramatikama.

Komunikacija unutar kruga suradnika i korisnika odvija se kroz Internet mrežu računala uz bogatu dokumentaciju i prepisku. Označivanje dokumenata rječničkog tipa posebno je složeno i od velikog je interesa. Formirane su zasebne grupe koje se bave općenito strukturama stabala. Neke su načelne preporuke već formirane, ali faza analize i diskusije, uz provjeru u praksi, još traje.

SGML jest sustav oznaka koji je opremljen sustavom atributa i njihovih vrijednosti. Krenuo je u razvoj s nizom konkretnih samogovorećih oznaka ("naslov", "odlomak"...) uz poštivanje dotadašnjih standarda (bibliografski navodi).

To je sustav oznaka koji se bavi označivanjem:

strukture dokumenta
- djelo
- poglavlje
- odlomak
- tablica
- grafikon
- ...

sustava znakova
- hrvatska latinica
- ...

stručnih sadržaja
- lingvistika
- psiholingvistika
- evaluacija i projektiranje udžbenika
- izdavaštvo
- baza podataka
- ...

Želja da se istovremeno udovolji različitim zahtjevima uvjetuje oblikovanje apstraktnog sustava za označivanje. Korisnik odgovarajućom interpretacijom općih oznaka izražava svoj način gledanja na dokument. Različitim shvaćanjem istih oznaka programski je moguće proizvesti različite sadržaje i oblike.

Takav je način razmišljanja u skladu sa zahtjevima zadatka koji smo prethodno predstavili. Očito će sustav oznaka s jedne strane i model podataka s druge omogućiti tu transformaciju teksta u strukturirane podatke pružajući potrebne informacije o jezičnom sustavu koji simuliramo.

Transformacija podataka treba biti automatska. Kroz mreže računala danas se već može pristupiti bibliotekama programa koji provode najrazličitije analize teksta u skladu sa SGML-označivanjem. Jasno je da se pretežno odnose na engleski jezik, ali neka je znanja moguće preuzeti odnosno prilagoditi.

## O općem modelu podataka

Opći model podataka o jezičnom sustavu treba reprezentirati postojeća znanja, ali i ona koja se tek očekuju u istraživačkim procesima. On mora biti podloga u

realizaciji raznih jezičnih alata za proizvodnju jezičnih informacija odnosno izvor informacija u oblikovanju inteligentnih sutava.

Umjesto čvrsto klasificiranih podataka, za čije klasifikacije i međusobne veze trebamo visoko lingvističko znanje predlažemo formiranje apstraktne podatkovne strukture kao ljuske za prihvat konkretnih podataka koji su zaduženi da formiraju sam jezični sustav.

Otvorenost sustava omogućuje da pojedini jezikoslovci ili jezikoslovne škole predoče vlastita viđenja jezičnog sustava odnosno sučele dvije teze s ciljem komparativne analize različitih stavova.

Takav pristup omogućava postepeno svladavanje i učenje postojećih znanja te ugradnju novih spoznaja kao rezultata istraživačkih procesa.

Ovaj tren prepoznati su dijelovi strukture jezičnog sustava i spremni smo dodijeliti radne nazive nekim popisima podataka unutar modela:

TEMA: Način proučavanja i upotrebe jezika
  pravopis
  morfologija
  ...
KATEGORIJA: Kriterij grupiranja jezičnih jedinica
  imenica
  grafem *č*
  ne pripada jezičnom sustavu
  zoologija
  latinsko porijeklo
  od mila
  kratica
  ...
ATRIBUT: Svojstvo jezične jedinice
  N jd.
  komparativ
  3. lice jd.
  ...
VEZA: Veza među jezičnim jedinicama
  sinonim
  osnovni oblik za
  kratica za
  nastavak *-šću*
  osnova za
  ...
JEZIČNA JEDINICA: Jezično značajan niz znakova
  riječ
  osnova riječi
  nastavak
  frazem
  pravopisni znak
  ...
PRAVILO: Svaka jezična zakonitost.
PRIMJER: Prikaz upotrebe jezika.
KOMENTAR: Uputa, napomena...
ZNAČENJE: Značenje jezične jedinice.

Objasnimo otvorenost sustava primjerom. Umjesto da imenice opisujemo nekim fiksnim nizom atributa, koje ovaj tren prepoznajemo, možemo ih opisivati listom elemenata iz popisa relevantnih podataka.

| | |
|---|---|
| JEZIČNA JEDINICA: | vučić |
| | vuk |
| | vuče |
| | ... |
| KATEGORIJA: | imenica |
| | umanjenica |
| | grafem *č* |
| | grafem *ć* |
| | ... |
| ATRIBUT: | N jd. |
| | m. r. |
| | s. r. |
| | ... |

OPIS RJEČNIČKE NATUKNICE:

vučić: grafem č; grafem ć; imenica/N jd., m. r.; umanjenica
vuk: imenica/N jd., m. r.
vuče: grafem č; imenica/N jd., s. r.

Natuknicu iz popisa JEZIČNIH JEDINICA možemo pridruživanjem elemenata s popisa KATEGORIJA klasificirati na po volji mnogo načina. Kao element neke od kategorija formira se opis s proizvoljnim brojem atributa. Popisi se tokom rada jasno proširuju, strukturiraju i mijenjaju. Takva reprezentacija omogućuje da se promjena elementa popisa obavlja jedanput, na jednom mjestu. Obnovljen podatak funkcionira i dalje u svim prethodno uspostavljenim strukturama i vezama u sustava.

U postupku traženja nezavisnih cjelina podataka prepoznalo se do sada nekoliko elemenata ljuske modela podataka. Da bi se pomoglo u odlučivanju kojem popisu pripada pojedini jezični podatak, popise treba opremiti vrlo jasnim, kratkim i nedvosmislenim definicijama sadržaja. Tokom rada treba ih nadopunjavati primjerima i protuprimjerima. Naime, time što je princip rada baziran na potpuno otvorenim strukturama i pojedincu se ostavlja da sam određuje sadržaj popisa podataka, otvara se mogućnost stvaranja "nereda" u sustavu.

Jezični sustav oblikujemo za potrebe raznih načina izučavanja. Onaj dio kojem ovaj tren posvećujemo više pažnje jest pravopis. šelimo ga uklopiti u prepoznati dio jezičnoga modela podataka.

Onoliko koliko do sada seže spoznaja, možemo ga interpretirati na sljedeći način:

| | |
|---|---|
| TEMA: | Pravopis |
| JEZIČNA JEDINICA: | Pravopisna natuknica (PN) |
| KATEGORIJA: | Gramatičke kategorije |
| | Pravopisne kategorije |
| | Ostale iz opisa PN |
| ATRIBUT: | Svojstva gramatičkih kategorija |
| VEZE: | Odnosi PN i ostalih riječi iz opisa |

Popisi podataka ravnopravno prihvaćaju sve podatke. Njihovo kasnije međusobno povezivanje, unutar popisa i među njima, uspostavt će potrebne hijerarhijske strukture.

Tako će imenica i glagolska imenica biti dva ravnopravna elementa popisa kategorija. a uvođenjem veze među njima, "je podkategorija", daje se mogućnost definicije njihova međusobnog odnosa i položaja.

S obzirom na mogućnost višestrukog povezivanja podataka, potpunu otvorenost, proizvoljno dodavanje podataka i postavljenje njihovih međusobnih odnosa, sustav postaje neosjetljiv na promjenu mišljenja i donošenje novih zaključaka što je svojstveno istraživačkoj sredini. Promjena jednog podatka, njegova opisa i klasifikacije te veze s ostatkom sustava ne narušavaju njegovu uspostavljenu okolinu.

## Prikupljanje podataka

Krećući s pozicija korpusne lingvistike, discipline koju smo prihvatili kao podlogu u prikupljanju podataka, za analizu smo odabrali tekst stručne proze — S. Babić, B. Finka, M. Moguš, Hrvatski pravopis.

Rječnik priručnika dragocjen je materijal jer predstavlja skup već sređenih podataka.

Informacije koje rječnik donosi izražene su:

vrstom pisma (masni tisak, tip slova),
znakovima interpunkcije
ključnim riječima/znakovima
(oznake morfoloških podataka...)

Kod slobodnog teksta sustav za označivanje je dakako nužan i jedan je od osnovnih alata korpusne lingvistike. Sustav SGML *(Standard General Markup Language)* s mnogo pristaša u svijetu, vrlo ozbiljno brine o označivanju teksta tipa rječnika. Bilo da iz označenog rječničkog teksta proizvodimo bazu podataka, ili iz baze podataka stvaramo tekst za tisak za koji tada može varirati izgledom od slučaja do slučaja, upotrebljava se isti tip označivanja. U našem slučaju, zbog postojeće sintakse rječničkog navoda, znakovi interpunkcije i kontrolirani popis ključnih riječi mogli su gotovo u potpunosti preuzeti ulogu oznaka. Dodatno je još trebalo samo označiti pravopisnu natuknicu te rezervirane riječi/kratice za veze među riječima. Njihova istaknuta uloga unutar rječničkog navoda bila je ostvarena grafički, masnim tiskom i kurzivom, te je tu informaciju trebalo zadržati putem oznaka.

Želja je bila takav tekst transformirati u elemente odgovarajućih popisa podataka koje predviđa spomenuti opći model u izradi.

Kao primjer neka posluži rječnički navod:

groteska DL jd. -ski i -sci, G mn. -ska i -ski,

iz kojeg je automatski trebalo proizvesti skup sadržanih elementarnih navoda. Algoritam, oblikovan kao kompjutorski program, trebao je prepoznati eksplicitno i implicitno predočene podatke. Ponegdje je neke od podataka bilo moguće proizvesti

"zaključivanjem". Pod elementarnim navodom podrazumijevamo rječničku natuknicu s onoliko elemenata opisa koji je po volji detaljno i opširno opisuju, ali ne narušavaju sadržajnu jednoznačnost.

groteska – im./N jd.
groteski – im./D jd.; *od* groteska; *i* grotesci
groteski – im./L jd.; *od* groteska; *i* grotesci
grotesci – im./D jd.; *od* groteska; *i* groteski
grotesci – im./L jd.; *od* groteska; *i* groteski
grotesaka – im./G mn.; *od* grotesaka; *i* groteski
groteski – im./G mn.; *od* groteska; *i* grotesaka

Koncept rječnika izražen je definicijom pravopisne natuknice. Upravo to bit će kriterij koji će odrediti ulazak proizvedenih natuknica u takav pravopisni rječnik.

Kao rezultat obrade teksta odabranog rječnika proizvedeni su popisi podataka te neke njihove međusobne veze.

RIJEČI: potencijalne pravopisne natuknice
NORMATIVNI PODACI:
  popis kategorija
  popis atributa
  popis veza
POVEZIVANJE: parovi riječi prema popisu veza
KLASIFIKACIJA: parovi riječi i kategorija
  vrste riječi
  pravopisne kategorije
  status u jezičnom sustavu
  porijeklo
  pripadnost struci
OPIS: parovi riječi i atributa
  generacija parova riječi kategorija
KOMENTARI: objašnjenja oko upotrebe natuknica

Analiza teksta rječnika dala je doprinos u izgradnji općeg i pravopisnog modela jezičnih podataka. Služila je za dobivanje hipotetičnog leksičkog rješenja koje na bazi korpusa tekstova i specijalističkih znanja treba dopunjavati i modificirati.

Obradom inicijalnih podataka utemeljili smo začetak sustava podataka i označivanja. Nadopuna rječničkoga fonda i odgovarajućeg opisa temeljit će se na potvrdama u korpusu tekstova.

## Daljnji zadaci

U daljnjem radu treba ići u tri pravca. To su potpuna izgradnja općeg modela podataka jezičnog sustava, oblikovanje sustava za označivanje teksta te izrada modela za selekciju i prikaz pravopisnog rječnika.

Radi se o konačnoj identifikaciji i definiciji elementarnih popisa podataka, njihovih međusobnih odnosa i veza. Treba raditi na automatizaciji procedura za popunja-

vanje podataka sustava i odgovarajućih podsustava. Zadatak je definirati analizatore strukturiranih podataka koji će imati mogućnost kakva zaključivanja te prema uzorcima prepoznatih zakonitosti "prizvoditi" srodne podatke.

S obzirom na to da se konačno leksičko rješenje treba dobiti dopunom i korekcijom inicijalno prikupljenih podataka s podacima iz korpusa tekstova, sustav za označivanje teksta treba dovesti na uporabni nivo. Taj sustav treba uskladiti s bibliografskim standardima te sa standardima za prijenos i razmjenu strukturiranih podataka.

Ovisno o načinu korištenja pravopisnih podataka treba razraditi kriterije za njihov odabir. Moguće je varirati u izboru opsega, redoslijeda i prikaza podataka.

Različito se definira rječnički navod ako ograničenje čine dimenzije tiskanoga priručnika i njegove stranice ili ako se prikazuje i koristi na elektroničkom mediju. U tiskanoj realizaciji tipom tiska određujemo prioritete, važnost i pripadnost podataka unutar navoda koji su istovremeno i stalno prisutni na stranici.

Elektronički medij daje mogućnost brzog direktnog pristupa podacima što po potrebi dozvoljava pretraživanje pojedinih dijelova.

Kad nisu predmet našeg radnog zadatka, jezični priručnici i savjetnici su alati za koje želimo da budu u pozadini stalno prisutni. Razvoj tehnologije daje sve veće mogućnosti integracije sustava ili istovremeneo aktiviranja većeg broja paralelnih procesa.

Zbog direktnog pristupa i normalizacije podataka, te nepotrebnosti za njihovom istovremenom, stalnom prisutnosti na ekranu, dijelovi podatkovnih struktura mogu biti opisani do većih detalja. Uz dobru organizaciju ograničenja za prikaz informacija koje prate rječničku natuknicu su sve manja. Hipertekstualni prikaz, nadopunjen funkcijom pretraživanja teksta otvara nove informacijske dimenzije.

Uz brigu o spretnom rječniku, sa dobro strukturiranim navodima i dohvatljivim informacijama, treba posvetiti pažnju i elektroničkim alatima. Sve upućuje na izradu automatiziranih procedura za proizvodnju jezičnih priručnika na oba medija iz iste podatkovne podloge.

## Zaključak

Koje dimenzije daje elektronička realizacija dogovora o pravilnom pisanju? Time što ga gledamo kao dio općega modela za reprezentaciju jezika, uz odgovor na postavljeno pitanje dobit ćemo i informaciju o jezičnoj disciplini koja je odgovorna za potrebno znanje.

Tako će na pitanje "Piše li se ruki ili ruci?" odgovor "ruci" dati gramatika, iako smo pitanje postavili u trenutku pisanja.

S novim mogućnostima tehnologije količina se dostupnih informacija povećava, linearni prikaz zamjenjuje direktnim, a baze podataka i znanja tek su infrastruktura nadolazećih inteligentnih sustava.

## *Literatura*

Poljak, V. (1980), *Didaktičko oblikovanje udžbenika i priručnika*, Školska knjiga, Zagreb.

ISO. ISO 2788-1986 (E) (1986), *Documentation — Guidelines for the establishment and development of monolingual thesauri.*

Lancaster, F. W. (1986), *Vocabulary Control for Information Retrieval*, Arlington, Virginia: Information Resources Press.

Malić, J. (1986), *Koncepcija suvremenog udžbenika*, Školska knjiga, Zagreb.

Tuđman, M. (1986), *Teorija informacijske znanosti*, Informator, Zagreb

Aitchison, J., Gilchrist, A. (1987), *Thesaurus construction. A practical manual.* London: Aslib. ISBN 0-85142-197-0.

Byrd, R. J., Calzolary, N., Chodorow, M. S., Klavans, J. L., Neff, M. S., Rizk, C. A. (1987), Tools and Methods for Computational Lexicology, *Computational Linguistics*, Vol. 13, No. 3–4.

Martin, J. and McClure, C. (1988), *Structured Techniques: The Basis for CASE*, Prentice Hall, New Jersey.

Brown, H. (1989), Standards for Structured Documents, *The Computer Journal*, Vol. 32, No. 6.

Furuta, R. (1989), An Object-based Taxonomy for Abstract Structure in Document Models, *The Computer Journal*, vol. 32, No. 6.

Ritchie, I. (1989): HYPERTEXT — Moving Towards Large Volumes, *The Computer Journal*, Vol. 32, No. 6.

Willer, M. (ed.) (1989), *Priručnik za UNIMARC*, Nacionalna i sveučilišna biblioteka, Zagreb.

ACH, ACL, ALLC (1990), *Guidelines: For the Encoding and Interchange of Machine-Readable Texts.* Chicago: C. M. Sperberg-McQueen and Lou Burnard.

Tuđman, M. (1990), *Obavijest i znanje*, Zavod za informacijske studije, Zagreb.

Cvitaš, M. i Vedriš, M. (1991), Croatian Literary Language Lexicon: Data Model, XIII International Conference »Information Technology Interface — ITI«, Cavtat.

Maletić, M., Mariani, I. (1991), The Use of Standard Generalized Markup Language on the Croatian Literary Language Corpus, XIII International Conference »Information Technology Interface — ITI«, Cavtat.

Cvitaš, M., Maletić, M. and Gašić, S. (1992), Building an Integrated Lunguistic System, Proc. 14th Int. Conf. »Information Technology Interfaces« ITI '92, Pula.

Garvas Delić, A. (1992), New Representation of Documentation Data Bases, Proc. 14th Int. Conf. »Information Technology Interfaces« ITI '92, Pula.

Maletić, M., Gašić, S. and Cvitaš, M.(1992), Building the Electronic Version of the Croatian Literary Language Corpus, Proc. 14th Int. Conf. »Information Technology ts Interfaces« ITI '92, Pula.

Cvitaš, M., Garvas Delić, A. (1993), About Data Interchange and Standards Involved. Proc. 15th Int. Conf. »Information Technology Interfaces« ITI '93, Pula.

Maja Cvitaš: *Problemi pravopisnoga rječnika*
FILOLOGIJA 22-23 (1994), str. 35-45

# PROBLEMS OF AN ORTHOGRAPHIC DICTIONARY

## Summary

The orthographic dictionary is seen as a set of orthographically significant word-forms. Such a dictionary should be treated only as a part of integral language system. The usage of electronic media induced discussion about normalized lexical data model. The latent lexical structure is described by general data classes and joined working headings. That gives the opportunity to the researcher of forming different subsystems according to his own interpretation of general model.

The hypothetical data model was tested with Babić–Finka–Moguš orthographic dictionary. The manual was analyzed and transformed to structured data by semi-automatic procedures and SGML mark-up system as a tool. The orthographic manual was interpreted as an element of scientific text sub-corpora. Development and adjustment of relevant corpora linguistic methods provided suitable processing tools in automatized production of dictionary normalized data base.