Pregledni članci i stručni radovi

UDK 81'324
811.163.42'374
81'374
Pregledni članak
Prihvaćeno za tisak 11. 10. 2002.

Božo Bekavac
Zavod za lingvistiku Filozofskog fakulteta, Zagreb

Strojno obilježavanje hrvatskih tekstova – stanje i perspektive

U članku se daje cjelovit pregled radova iz područja strojnog obilježavanja hrvatskih tekstova. Pregled obuhvaća opojavničenje, segmentaciju na rečenice, lematizaciju, POS i MSD označavanje, prepoznavanje naziva i problematiku leksikona. Osim izloženih gotovih radova za svaku cjelinu posebno, upućuje se na slične radove i dodatne izvore, te se daju ideje i smjernice za buduće korake. U zaključku se ističe nužnost poštivanja međunarodnih standarda za obilježavanje tekstova u razvoju jezičnih resursa i računalnojezikoslovnih alata za hrvatski jezik.

Uvod

Ovaj je rad motiviran činjenicom da na hrvatskom ne postoji ni jedan cjeloviti pregled radova iz ovog područja. Premda definirani standardi za suvremeno obilježavanje tekstova na hrvatskom jeziku postoje od 1998. godine,[1] njihova primjena još nije zaživjela. Cilj je ovoga rada odrediti stanje i buduće korake pri strojnom obilježavanju tekstova na hrvatskom jeziku. Uputit će se na radove koji već postoje, a naglasiti oni segmenti koji za hrvatski jezik još ne postoje, ili se tek stvaraju.

Na samom početku valja osvijetliti neka pitanja strojnog obilježavanja tekstova. Potpuna točnost obilježavanja pri strojnoj obradi tekstova danas još nije postignuta stoga što ni jedan računalnolingvistički alat ne radi s potpunom točnošću. Ona je moguća samo ako nakon ili za vrijeme obrade slijedi ljudska intervencija u tekst ili ispravljanje pogrešaka. Drugo je pitanje zašto se u naslovu

1 Erjavec (ur.) ([2]2001), standard na razini POS i MSD obilježavanja

govori o tekstovima, a ne korpusima. Moglo bi se reći da su današnji korpusi (treće generacije) velike baze tekstovnih podataka, pa ih je preciznije nazivati tekstovnim arhivima.

Strojno obilježavanje tekstova nekoga jezika iznimno je složen zadatak: bilo sa stajališta opsega posla koji treba obaviti, interdisciplinarnosti koje ono zahtijeva (lingvistika, informatika i statistika) ili količine znanja/ljudi koji se strojnom obradom bave. No s jednim te istim problemom suočeni su svi jezici: to je mali broj istraživača koji se ovim područjem bave u uglavnom nezavisnim skupinama, najčešće od 5 do 35 istraživača.[2] To je ujedno i odgovor zašto dugo vremena nije postojao zadovoljavajući alat ili integrirana skupina alata koja bi olakšala ili ubrzala ovakve obrade ne samo za jedan nego i za više jezika. U posljednje dvije–tri godine razvijeno je nekoliko alata koji u osnovi zadovoljavaju potrebe strojne obrade teksta. GATE[3] je danas najšire prihvaćen sustav u koji je moguće ugrađivati interaktivne module za jezičnu obradu. No za obradu hrvatskoga potrebno je razviti čitav niz modula,[4] tj. alata specifičnih za hrvatski jezik koji bi bili ugrađeni u GATE. Taj proces, nažalost, zahtijeva suradnju većeg broja jezikoslovaca i informatičara nego što ih se za hrvatski ovim područjem bave.

Danas se strojnom obradom hrvatskoga jezika bavi znatno manje od 35 istraživača, a da situacija bude još nezavidnija, oni su institucionalno odijeljeni u dvije manje–više nezavisne skupine.

Hrvatski jezik u ovom području nažalost zaostaje za mnogim europskim jezicima kao npr. za češkim, slovenskim, bugarskim, madžarskim, a kad je riječ o mnogoljudnijim jezicima (engleski, njemački, španjolski), zaostatak je još veći. Taj je zaostatak moguće pretvoriti u prednost jer danas postoje čvršći standardi, razvijeniji alati i metodologije za obrade teksta, te stoga možemo zaobići probleme koje su imali drugi.

1. Obilježavanje tekstova

Obilježavanje (*annotation, mark–up*) je pridodavanje dodatnih eksplicitnih informacija tekstu za računalnu obradu tamo gdje su one implicitno prisutne osobi koja čita tekst.[5] Pri obilježavanju korpusa oznake se odabiru iz određenoga skupa oznaka i ubacuju u elektronički zapis teksta u smislu obilježavanja strukture i drugih osobitosti teksta za koje postoji potreba za obilježavanjem.[6]

Redoslijed kojim se obavlja obilježavanje tekstova može biti od iznimne važnosti za uspješnost obilježavanja. Najčešće se rezultati prethodnih koraka obrade koriste u idućim koracima obrade, pa preciznost obrade izravno ovisi o pre-

2 Cunningham (2000): 26

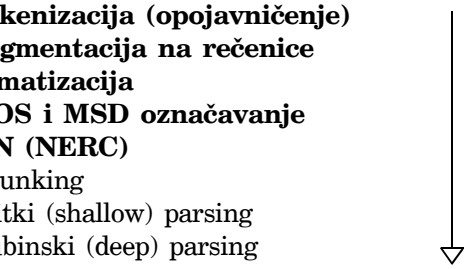
3 Cunningham (2002)

4 O modulima v. iduća poglavlja

5 Lawrer & Dry (1998): 107

6 Više u Žubrinić (1995): 16 i Bekavac (2001): 18

ciznosti rezultata obrade iz prethodne faze. To je glavni razlog zbog kojega je nužna suradnja među istraživačima koji rade na obradi nekoga jezika. Dakle, izbjegavanje ponavljanja na više različitih mjesta istoga mukotrpnog posla trebao bi biti imperativ za maloljudni jezik poput hrvatskoga. Obilježavanje danas nema smisla raditi ručno. Stoga su oznake koje se umeću u tekstove rezultat uzastopne primjene različitih računalnojezikoslovnih alata. Taj bi redoslijed shematski mogao izgledati ovako:



Potrebno je napomenuti da iako čest, ovaj redoslijed nije striktan. Na primjer, pri obradi hrvatskih tekstova lematizacija i POS i MSD označavanje obavljaju se u istom koraku. Opseg ovoga rada obuhvaća faze do PN–a[7] uključujući i nju. Dva su razloga za to: ovaj bi tekst prešao granice članka u časopisu, a još je važniji usredotočenje na aktualne i srednjoročno rješive probleme. Ne bi imalo velikog smisla baviti se fazama koje su za hrvatski u kontekstu strojne obrade još predaleko od izvodivih.

Iz izložene sheme važno je uvidjeti da bi ako ne alati, onda barem obilježeni tekstovi nad kojima se izvodi obrada morali imati svojstvo **višestruke uporabivosti** (*reusability*).[8] To na žalost nije bila ustaljena praksa kod nas do sada. Neprimjenjivanje ovoga načela pridonijelo bi dodatnom zaostajanju hrvatskog jezika u ovom području. Iz tog razloga zadnje poglavlje ovog članaka upućuje na predložene standarde za obilježavanje tekstova.

2. Jezikoslovni alati – radovi

2.1 Tokenizacija

Tokenizacija (*tokenisation*) ili opojavničenje mogla bi se definirati kao dovođenje korpusa u stanje u kojem su sve riječi–pojavnice identificirane i eksplicitno obilježene, gdje se razlikuju XML/SGML oznake interpunkcije i znamenaka od riječi–pojavnica.[9] U jednostavnom pristupu pojavnice je lako identificirati jer pojavnica je sve ono što se nalazi između dva pismena za obilježavanje razmaka, što najčešće odgovara dvjema bjelinama. Međutim, tokenizacija je u složenijem slučaju mnogo zahtjevnija, jer se pojavnicama mogu smatrati i jedinice koje se sastoje od više riječi (*multi–word units (MWU)*), a one su sintaktički ili

7 PN stoji za prepoznavanje naziva. Više u poglavlju 2.5

8 Više u Ide & Brew (2000) i Bekavac (2001): 65

9 Bekavac (2001): 22

semantički povezane. Na primjer, datum *20. svibnja* ili *20. 5.* mogao bi biti obrađivan kao jedna jedinica, pa ga u ranijem smislu određenja pojavnice nije moguće tokenizirati.[10] U ovakvom pristupu tokenizacija bi uključivala **prepoznavanje naziva** (*named entity recognition*). Prepoznavanje naziva uključuje obradu teksta pri kojoj se identificiraju izrazi koji su nazivi za npr. ljude, organizacije, datume i sl.

Za jednostavni pristup u ovom trenutku već postoji gotovo rješenje, tj. alat 2XML.[11] Osim tokenizacije ovaj se alat pokazao učinkovit i kod pretvaranja (*conversion*) HTML i RTF datoteka u XML format.

2.2 Segmentacija na rečenice

Segmentacija se rečenica (*sentence segmentation, sentence boundary disambiguation)* obavlja ubacivanjem jedinstvenih nizova pismena, tj. graničnih oznaka na početak, odnosno na završetak rečenica u tekstu (u suvremenim shemama za obilježavanje teksta to su nizovi <S> i </S>). Iako izgleda trivijalno, to najčešće uključuje složene postupke iz razloga što su oznake rečenične interpunkcije često višeznačne (*ambiguous*). Na primjer, točka može stajati uz redni broj, kraticu, kraj rečenice, ili pak kraticu ili redni broj na kraju rečenice. Za hrvatski je jezik testni model alata za segmentaciju rečenice na tekstu veličine 2000 rečenica imao deklariranu točnost od 99,5 %.[12] Postiže li ovaj model i na ostalim tekstovima sličnu točnost, u potpunosti bi mogao zadovoljavati suvremene zahtjeve.

2.3 Lematizacija

Lematizacija (*lemmatisation*) je svođenje pojavnica iz korpusa na njihove natukničke oblike, tj. svođenje različitih pojavnica (članova iste paradigme) na zajedničku lemu.[13] Na primjer, pojavnice *stol*, *stolova* ili *stolu* bile bi svedene na lemu *stol*. **Lema** je onaj oblik pod kojim bismo tražili neku riječ u rječniku. Lematizacija se na isti način primjenjuje i na flektivno »nepravilne« oblike pa se npr. *jesam*, *bijah* ili *bila* svode na leksem *biti*. Lema predstavlja sve oblike određene riječi. Kako se u postupku strojnoga prepoznavanja lema redovito moraju prepoznati i morfosintaktički opisi pojavnica, lematizacija se zapravo obavlja u drugoj fazi POS[14] označavanja. Ipak, lematizacija je nužna kao zaseban postupak jer se pri MSD obilježavanju u pravilu određuje gramatički oblik pojavnice, a ne sama lema.[15] Lematizacija je važan postupak u istraživanjima korpusa osobito za jezike koji imaju bogatu morfologiju. Alat koji obavlja automatsku lematizaciju zove se **lematizator** (*lemmatizer*). Najveću prepreku postizanju veće točnosti automatske lematizacije predstavljaju istopisnice (homo-

10 Grover & Matheson & Mikheev (2000)

11 Više u Tadić (2000b): 524, Tadić (2001): 110, Tadić (2002): 445

12 Boras (1998): 155

13 McEnery & Wilson (1996): 42

14 o POS i MSD označavanju v. poglavlje 2.4

15 Ide (1996)

grafi). Jedini veći korpus hrvatskoga jezika nad kojim je obavljen dio lematizacije (poluautomatskim putem) *Mogušev* je *korpus*.[16] Na osnovi tog korpusa izrađen je *Hrvatski čestotni rječnik*.[17]

Za hrvatski jezik postoji program za lematizaciju koji je dio sustava SOLAH, a deklarirana točnost mu je 95 %.[18] O radovima na strojno potpomognutoj lematizaciji više u Tadić (1997: 391).

2.4 POS i MSD označavanje

Part–of–speech (POS) označavanje je pridruživanje gramatičke kategorije svakoj pojavnici u tekstu (ponekad se naziva gramatičko označavanje ili morfosintaktičko obilježavanje).[19] POS označavanje spada u osnovne vrste lingvističkog označavanja. Pored toga, POS oznake prvi su korak u razrješavanju istopisnica, tj. pojavnica koje imaju isti lik a različite gramatičke kategorije i/ili značenje. Alat s pomoću kojega se obavlja automatsko POS označavanje naziva se **POS označivač** (*tagger*).

Rezultat automatskoga POS označavanja može biti iznimno precizan. Razlog za to je predvidivost gramatičkih kategorija pojavnica na osnovi ko–teksta u kojima se nalaze. POS označivači smatraju se najpouzdanijim i najkorisnijim računalnolingvističkim alatom, a prema načinu rada dijele se na:[20]

a. vjerojatnosne (*probabilisic*) označivače: zasnivaju se na vjerojatnosnom računu i statistici,
b. označivače zasnovane na pravilima (*rule–based*): zasnivaju se na lingvističkim, ručno pisanim pravilima.

Većina POS označivača danas koristi prvi pristup, a najčešće se koristi u kombinaciji s drugim pristupom. Točnost rezultata povećava se primjenom pravila na rezultat vjerojatnosnog označivača.

Jedna od podjela POS označivača zasniva se i na stupnju autonomije označavanja u odnosu na uporabu prethodno obilježena korpusa u uvježbavanju označivača na:[21]

1. nadgledane (*supervised*): rabe prethodno obilježene korpuse kao osnovu za izradu alata koji će se koristiti u postupku POS označivanja, npr. leksikon, čestote pojavnica i oznaka, vjerojatnosti određenih nizova oznaka itd.
2. nenadgledane (*unsupervised*): umjesto prethodno obilježenih korpusa koriste napredne računalne metode kako bi pronašli automatska grupiranja prema kojima se izračunavaju vjerojatnosti potrebne vjerojatnosnom označivaču, ili pak pronalaze pravila za označivače zasnovane na pravilima.

---

16 *Jednomilijunski korpus hrvatskog književnog jezika* poznatiji je pod nazivom *Mogušev korpus*

17 Moguš, Bratanić, Tadić (1999): 5

18 Žubrinić (1995): 69

19 McEnery & Wilson (1996): 36

20 Van Guilder (1995)

21 Van Guilder (1995)

Označivač za ulaznu varijablu uzima pojavnice iz korpusa te ih uspoređuje s riječima iz leksikona.[22] **Leksikon** (ili elektronički rječnik) u korpusnoj se lingvistici koristi kao sinonim za rječničku bazu podataka što podrazumijeva pohranu leksičke građe u strojno čitljivu obliku. Leksikon potencijalno može sadržavati širok raspon informacija o pojedinoj riječi, ovisno o strukturi i vrsti zadatka obrade kojoj je namijenjen. Osnovni leksikon može sadržavati i informacije o morfologiji, bilo kao popis svih oblika riječi, bilo u obliku koji omogućuje generiranje svih oblika riječi, ili sadrži oboje od navedenoga.

Iako su se ranije sastavljali ručno, POS obilježeni korpusi nezamjenjiv su izvor za automatsko sastavljanje pouzdanih i sveobuhvatnih leksikona. Zapravo, taj postupak može biti obostran: sastavljanje leksikona iz POS obilježenoga korpusa, ili POS obilježavanje korpusa iz leksikona. Što je veći obilježeni korpus, veća je mogućnost sastavljanja bogatijega leksikona. Vrijedi i obratno, što je veći leksikon, veća je i mogućnost pronalaženja pripadajućeg POS–a pojavnice iz korpusa. Automatski sastavljeni leksikoni na osnovi POS obilježenih korpusa potencijalno sadržavaju stotine tisuća natuknica iz razloga što broj oblika svih riječi u prirodnom jeziku može biti velik, osobito u flektivnih jezika kakav je hrvatski. Idealan bi leksikon trebao sadržavati sve ovjerene oblike riječi i njima pridružene POS i MSD podatke.

POS označavanje može uključiti dvije razine:

1. *razina*: uključuje prepoznavanje i označavanje *vrsta riječi* (*POS*),
2. *razina*: označava se *vrsta riječi* i određuju *gramatičke kategorije*, tj. njihove vrijednosti.

Druga se razina označavanja pojavnica naziva i **morfosintaktički opis** (*morphosyntactic description, MSD*). Pri svakoj se razini rabe različiti skupovi oznaka, gdje je skup oznaka na drugoj razini znatno veći jer je varijabilnost kategorija i njihovih vrijednosti veća.

Ovaj tip označavanja teksta predstavlja jedan od najvažnijih problema pri obilježavanju hrvatskih tekstova. Jedini cjeloviti rad koji je rezultirao izradom označivača SOLAH magistarski je rad Tomislave Žubrinić. Deklarirana točnost vjerojatnosnog označivača sustava SOLAH izračunata na uzorku[23] bez (za označivač) nepoznatih riječi iznosi 91 %.[24]

S obzirom na to da hrvatski ima bogatu fleksiju, te stoga i velik broj potencijalnih oznaka, valja očekivati i manju preciznost označivača.[25] Stoga bi bilo poželjno sačuvati potencijalne interpretacije pojedinih pojavnica iz leksikona. Na primjer, sačuvane kao atribute elemenata:

Ekonomija je loša.

<s><w lemma=»ekonomija« aa=»Ncfpg; Ncfsn« ta=»Ncfsn«>**Ekonomija**</w><w lemma=»biti« aa=»Vcip3s« ta=»Vcip3s«>**je**</w><w lemma = »loš« aa=»Afpfsnn; Afpfsny; Afpfsvy; Afpmsan–
y; Afpmsgn; Afpnpan; Afpnpay; Afpnpnn; Afpnpny; Afpnpvy; Afpnsgn« ta=»Afpfsnn«>**loša**</w><pt>.</pt></s>

22 McEnery & Wilson (1996): 120

23 Na manjem korpusu srednjoškolskih udžbenika

24 Žubrinić (1995): 69

25 V. Erjavec (1999)

gdje su, *aa* – all analyses (sve moguće interpretacije), *ta* – true analyses (interpretacija označivača, ne nužno točna),[26]

ili po CES standardu[27] kao elemente:

```
<W type=»R«>
  <ORTH>detaljima</ORTH>
  <LEX>
      <BASE>detalj</BASE>
      <MSD>Ncmpd</MSD>
      <MSD>Ncmpl</MSD>
      <MSD>Ncmpi</MSD>
      <MSD>Ncmpi/MSD
  </LEX>
  </W>
```

gdje su, ORTH – pojavnice, BASE – leme, MSD – msd interpretacije iz leksikona.[28]

Iznimno koristan izvor informacija o usporednoj preciznosti nekoliko označivača primijenjenih na srodni (slovenski) jezik nalazi se u Džeroski, Erjavec, Zavrel (2000).

Cjelovitu specifikaciju MSD skupa oznaka (*tagset*) za hrvatski jezik po MULTEXT–East (V.2) standardu sastavio je 1998. Tadić.[29]

2.5 PN (NER)

Prepoznavanje naziva, PN (*Named Entity Recognition, NER*) uključuje obradu teksta pri kojoj se identificiraju izrazi koji predstavljaju nazive za osobe, organizacije, lokalitete, kao i datumi ili valute. Pored PN često se obavlja i prepoznavanje i klasifikacija naziva, PKN (*Named Entity Recognition and Classification, NERC*). Prepoznavanje naziva važno je za nekoliko razina strojne obrade kao što su: složenija tokenizacija, od iznimne je koristi pri segmentaciji na rečenice i na posljetku parsing. Još nema cjelovitih radova[30] iz ovoga područja za hrvatski jezik, postoji najava jednog doktorata.

3. Jezični resursi

Jezični resursi polazište su svake strojne obrade teksta. O njihovoj veličini, reprezentativnosti i kodiranju zasigurno ovisi kvaliteta obrade.

Korpus koji teži da bude reprezentativan, a kvantitativno zadovoljava zahtjeve suvremene jezične obrade je HNK,[31] sastavlja se u Zavodu za lingvistiku

26 Simov i ost. (u tisku)

27 Više o CES–u u zadnjem poglavlju

28 Tadić (2002): 445

29 Erjavec (ur.) (22001), v. WWW adresu: http://nl.ijs.si/ME/V2/msd/html/node15.html – SECTION03800000000000000000

30 Jedini meni poznat rad koji se dotiče ovog područja je Tadić (2000c)

31 HNK je slobodno pretraživ i nalazi se na WWW adresi: *http: //WWW.hnk.ffzg. hr*

Filozofskog fakulteta Sveučilišta u Zagrebu.[32] S obzirom na opterećenost autorskim pravima, nije ga moguće koristiti za »zajedničku obradu« tekstova, u smislu jezičnog resursa koji bi se obradama nadopunjavao iz više centara. Iz tog je razloga potrebno sastaviti korpus koji će predstavljati etalon (*Golden standard*) i postati slobodan resurs za zajedničku uporabu svih uključenih u strojnu obradu hrvatskoga.[33]

U istom je Zavodu u završnoj fazi izrada morfološkog leksikona[34] prema MULTEXT–East specifikaciji.[35] Od sličnih leksikona postoji i rječnička baza[36] sastavljena na Odsjeku za informacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu.

Važan jezični resurs su i popisi imena (*gazzeteers*). Prvi, i što je još važnije iscrpan popis imena i prezimena s generiranim oblicima napravio je Boras sa suradnicima.[37]

4. Standardi za kodiranje tekstova

Izbor standarda gotovo je od presudne važnosti za višestruku uporabivost jezičnih resursa. Standardni jezik za obilježavanje tekstova/korpusa danas je XML[38] i gotovo da se više ne dovodi u pitanje.[39] Međutim način, tj. shema obilježavanja još nije u potpunosti usuglašena, no većina tekstova ravna se prema TEI–u[40] ili CES–u.[41] TEI standard u potpunosti je usklađen s XML standardom od inačice P4.[42]

5. Zaključak

Velika prepreka razvoju strojne obrade hrvatskoga danas je nedostatak višestruko uporabivog i dostupnog elektroničkog morfološkog leksikona. Leksikon je osnovni jezični resurs za obilježavanje tekstova. Bez njega je nemoguće obavljati označavanje na mnogim razinama tekstova, kao i razvoj računalnojezikoslovnih alata. Jednaki problem koji proizlazi iz nepostojanja dostupnog leksikona je i nedostupan POS označivač za hrvatski jezik. Preduvjet razvoja POS označivača je i pribavljanje velike količine morfosintaktički obilježene jezične

32 Tadić (1997), Tadić (1998), Tadić (2000a)

33 Radi se o tekstovima koji bi bili dostupni svim zainteresiranima za obradu, a nakon obrade obilježeni ponovo pohranjeni na dostupno mjesto.

34 sastavlja se po načelima iznesenima u Tadić (1994)

35 više o MULTEXT–East specifikaciji u Erjavec (1998): 189

36 Kržak (1985)

37 Boras i ost. (u tisku)

38 *XML (2000)*

39 v. Ide & Romary (2000), Ide (2000), Bekavac (2001): 38

40 TEI (2002); v. i Bekavac (2001): 54

41 Ide (1998); v. i Bekavac (2001): 58 i *http://WWW.xml–ces. org*

42 TEI P4 dokument nalazi se na WWW adresi: *http://WWW.tei–c.org/TEI/P4X/* v. i P5

građe. Ti se prioritetni zadaci trenutno obavljaju u odvojenim i do sad uglavnom nepovezanim skupinama[43] na Filozofskom fakultetu Sveučilišta u Zagrebu, što rezultira neadekvatnom brzinom razvoja osnovnih jezičnih resursa i alata za hrvatski jezik. Povezivanje i usklađivanje rada ovih skupina zasigurno bi ubrzalo razvoj, ali i podiglo kvalitetu i višestruku uporabivost jezičnih resursa.

Za navedene korake razvoja jezičnih resursa i alata danas, za njihovo međusobno usklađivanje i usklađivanje prema ostalim jezicima, neophodno je poštivanje međunarodnih standarda tekstova.

Literatura

Bekavac, B. (2001), *Primjena računalnojezikoslovnih alata na hrvatske korpuse*, magistarski rad, Filozofski fakultet Sveučilišta u Zagrebu

Boras, D. & Mikelić, N. & Lauc, D. (2003), *Leksička flekivna baza podataka hrvatskih imena i prezimena*, Modeli znanja i obrada prirodnog jezika – Zbornik radova, Radovi Zavoda za informacijske studije (knj. 12), 219–237

Boras, D. (1998), *Teorija i pravila segmentacije teksta na hrvatskom jeziku*, doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu

Cunningham, H. (2000), *Software Architecture for Language Engineering*, doktorska disertacija, Department of Computer Science, University of Sheffield

Cunningham, H. (2002), *Developing Language Processing Components with GATE*, (a User Guide), University of Sheffield, GATE se nalazi na WWW adresi: *http://gate.ac.uk/*

Džeroski, S. & Erjavec T. & Zavrel J. (2000), *Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets*, Second International Conference on Language Resources and Evaluation, LREC'00, ELRA, 1099–1104.

Erjavec, T. (1998), *The MULTEXT–East Slovene Lexicon*, Proceedings of the 7th Electrotechnical Conference ERK '98, Portorož, Slovenija, Vol. B, str. 189–192.

Erjavec, T. (1999), *Tagging Slavic Corpora*, pozvano predavanje na Sveučilištu u Tübingenu 15. prosinca 1999, prezentacija predavanja se nalazi na WWW–adresi: http://nl.ijs.si/et/talks/SFB441/tue–slides/

Erjavec, T. (ur.) ([2]2001), *Specifications and Notation for MULTEXT–East Lexicon Encoding V. 2*, Multext–East / Concede edition, specifikacija se nalazi na WWW adresi: http://nl.ijs.si/ME/V2/msd/html/msd.html

Grover, C. & Matheson, C. & Mikheev, A. (2000), *TTT: Text Tokenisation Tool*, Language Technology Group, Human Communication Research Centre, University of Edinburgh, Edinburgh, preuzeto 4. studenoga 2002, sa WWW: http://WWW.ltg.ed.ac.uk/software/ttt/tttdoc.html

Ide, N. & Brew, C. (2000), *Requirement, Tools and Architectures for Annotated Corpora,* u Data Architectures and Software Support for Large Corpora, LREC2000 Workshop Proceedings, ELRA, Paris–Athens, 1–5.

Ide, N. & Romary, L. (2000), *XML Support for Annotated Language Resources*, Proceedings of the Workshop, Web–based Language Documentation and Description, Philadelphia, 148–153.

Ide, N. (1998), *Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora*, Proceedings of the first International Conference on Language Resources and Evaluation, LREC'98, Granada. ELRA, 463–470, *http://WWW.cs.vassar.edu/CES/*.

Ide, N. (2000). *The XML Framework and Its Implications for Corpus Access and Use*, Proceedings of Data Architectures and Software Support for Large Corpora, Paris: European Language Resources Association, 28–32.

43 Zavod za lingvistiku i Odsjek za informacijske znanosti

Ide, Nancy CES (1996), *Corpus Encoding Standard*, WWW adresa: http://WWW.cs.vassar.edu/CES/

Kržak, M. & Boras, D. (1985), *Lexical Data Base of the Croatian Literary Language*, Informatologica Yugoslavica 17 (3–4), 223–242.

Lawrer, M. J. & Dry, H. A. (1998), *Using Computers in Linguistics*, Routledge, New York

McEnery, T. & Wilson, A. (1996), *Corpus Linguistics*, Edinburgh University Press

Moguš, M. & Bratanić, M. & Tadić, M. (1999), *Hrvatski čestotni rječnik*, Zavod za lingvistiku Filozofskog fakulteta i Školska knjiga, Zagreb

Simov, K. & Kouylekov, M. & Simov, A. (u tisku), *Cascaded Regular Grammars over XML Documents*, u Proc. of the 2nd Workshop on NLP and XML (NLPXML–2002), Taipei, Taiwan, 1 rujna 2002.

Tadić, M. (1994), *Računalna obradba morfologije hrvatskoga književnoga jezika*, doktorska disertacija, Sveučilište u Zagrebu, Filozofski fakultet

Tadić, M. (1997), *Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive,* Suvremena lingvistika 43–44, str. 387–394.

Tadić, M. (1998), *Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika*, Filologija 30–31, str. 337–347. (ISSN 0449–363X)

Tadić, M. (2000a), *Uporaba XML–a u hrvatskim korpusima*, CroInfo2000 – Upravljanje informacijama u gospodarstvu i znanosti, zbornik, Dubrovnik, 16–18. listopada 2000, Nacionalna i sveučilišna knjižnica–Pliva, Zagreb 2000, str. 132–137.

Tadić, M. (2000b), *Building the Croatian–English Parallel Corpus*, LREC2000 zbornik, Atena, 31. svibnja–2. lipnja 2000, ELRA, Pariz–Atena 2000, Vol. I, str. 523–530.

Tadić, M. (2000c), *Information Retrieval Meets Human Language Technology*, CUC2000 Zbornik, CD–ROM, Zagreb, 24–26. rujna 2000, CARNet, Zagreb

Tadić, M. (2001), *Procedures in Building the Croatian–English Parallel Corpus*, International Journal of Corpus Linguistics, poseban broj, 107–123.

Tadić, M. (2002), *Building the Croatian National Corpus*, LREC2002 zbornik, Las Palmas, 27. svibnja–2. lipnja 2002, ELRA, Pariz–Las Palmas 2002, Vol. II, 441–446.

TEI (2002), *Text Encoding Initiative*, WWW adresa: http://WWW.tei–c.org/

Van Guilder, L. (1995), *Automated Part of Speech Tagging: A Brief Overview*, preuzeto 19. ožujka 2001, sa WWW: http://WWW.georgetown.edu/cball/ling361/tagging_overview.html

XML (2000), *Extensible Markup Language (XML) 1.0*, W3C Recommendation, WWW adresa: http://WWW.w3.org/TR/REC–xml

Žubrinić, T. (1995), *Mogućnosti strojnoga označavanja i lematiziranja korpusa tekstova hrvatskoga jezika*, magistarski rad, Filozofski fakultet Sveučilišta u Zagrebu

Automatic Annotation of Croatian Texts – Current state and Perspectives

The article discusses the development of automatic annotation of Croatian e–texts achieved by now. Previous work in the field of language resources and tools for NLP of Croatian language is discussed and further steps proposed. Topics discussed in the paper are tokenisation, sentence segmentation, lemmatisation, POS and MSD annotation, named entity recognition and lexicon. Ideas and possible solutions for current problems are given in the conclusion.

Ključne riječi: strojno obilježavanje, jezikoslovni alati, jezični resursi, korpusna lingvistika, hrvatski jezik

Keywords: automatic annotation, linguistic tools, language resources, corpus linguistics, Croatian language