

Uporaba XML-a u hrvatskim korpusima

Marko Tadić

Odsjek za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu
Ivana Lučića 3, 10000 Zagreb
marko.tadic@ffzg.hr

Abstract

Article discusses the usage of mark-up languages in annotating language resources within Human Language technology framework. It concentrates on usage of XML for annotating Croatian corpora (Croatian National Corpus, Croatian-English and Croatian-Slovene parallel corpora) which are being compiled in Institute of linguistics at the Faculty of Philosophy, University of Zagreb.

U uvodnom dijelu rada iznose se i omeđuju temeljni termini vezani uz područje primjene korpusa u lingvistici. Cilj je tome polaznom određivanju prezentacija vrste podataka o kojima će biti riječi. U nastavku se eksplicira područje jezičnih tehnologija, njihov opseg i razdioba te se u središnjem dijelu razrađuje problematika obilježavanja jezičnih resursa na primjeru hrvatskih korpusa koji se sastavljaju u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu: Hrvatskoga nacionalnog korpusa, Hrvatsko-engleskoga paralelnog korpusa i Hrvatsko-slovenskoga paralelnog korpusa. Zaključno se iznose mogući nedostaci sadašnjih sustava i planovi za njihovu daljnju razradu.

1. Korpus u lingvistici

Korpus je u lingvistici definiran kao »skup tekstovnih odsječaka koji su odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak«. ¹ Još preciznije: računalni korpus je »korpus koji je kodiran na standardan i dosljedan način s nakanom da bude računalno pretraživan«. ² Kako se jezik ne može istraživati niti izravno niti u svojoj potpunosti, potrebno je uzeti jezični uzorak i postaviti metodološki postulat da istražujući taj uzorak zapravo istražujemo jezik u cjelini. ³ Tako ovisno o istraživačkim potrebama nailazimo na jednojezične ili višejezične korpusne, zatvorene (u slučaju mrtvih jezika ili umrlih pisaca) i otvorene korpusne (bili oni uzorkovani i reprezentativni ili ne). Nadalje, korpusi se mogu sastavljati o pisanim ali i govorenih tekstova. Raznim se metodama, ponajprije statističkim, ali i ne samo statističkim, dolazi do lingvistički relevantnih podataka o tekstu i jeziku u kojem je taj tekst ostvaren. Područje koje se bavi korpusima dio je lingvistike ali izrazito interdisciplinarnan (nalazi se na dodiru s informatikom) a naziva se kompjutorska lingvistika ili računalno jezikoslovlje. ⁴ Kako sam lingvist po zvanju, pristupam ovome području s tipičnim zahtjevima istraživača jezika koji treba jezično relevantne podatke iz golemih količina teksta.

¹ EAGLES (1996).

² *ibid.*

³ Situacija je nadasve slična istraživanjima u društvenim znanostima (sociologiji ili psihologiji) u kojima je postupak uzorkovanja i strukturiranja reprezentativnih uzoraka ispitanika standardna metoda istraživanja.

⁴ Za razliku od dijela informatike koji se bavi obradom prirodnoga jezika (*natural language processing*).

2. Jezične tehnologije

Podatci dobiveni analizom korpusa temeljni su kako za znanje o konkretnom jeziku tako i za razvitak jezičnih tehnologija. Ako je tehnologija »znanost o tehničkim postupcima prerade sirovina u proizvode« ⁵ tada je u ovom slučaju sirovina jezik a proizvodi su sustavi koji korisniku omogućuju jednostavniju uporabu prirodnoga jezika u računalnome okružju.

Jezične su tehnologije jezično specifične tj. ovise izravno o strukturi obrađivanoga jezika i moraju se iznaći zasebno od drugih jezika. ⁶

2.1 Jezične tehnologije prema FP5 EU

Područje jezičnih tehnologija (*Human Language Technologies*) definirano je u okviru 5. Framework Programa Europske Unije kao sastavni dio (Key Action III — KA III: *Multimedia Content and Tools* (MC&T)) njegove glavne teme *Information Society Technologies Programme* (IST). ⁷ Kad je riječ o jezičnim tehnologijama za hrvatski, ukoliko se Republika Hrvatska sama odmah ne pobrine za njihov razvitak, nitko to drugi neće učiniti, a hrvatski će jezik u znatnoj mjeri postati funkcionalno »nepismen« zbog nedostatka alatâ za njegovu uporabu u digitalnim komunikacijskim kanalima (Internet, GSM itd.).

2.2 Jezični resursi i jezični alati

Same se jezične tehnologije sastoje od tri osnovna područja:

1. jezičnih resursa
2. jezičnih alata
3. komercijalnih proizvoda

Jezične resurse čine korpusi i rječnici pohranjeni u digitalnome obliku tj. u obliku e-teksta. Jezični su alati aplikacije koje obrađuju ili se služe bilo postojećim resursima bilo tekstovima koji se upravo stvaraju. Komercijalni su proizvodi nastali na temelju istraživanja jezičnih resursa jezič-

⁵ Leksikon JLZ (1974), str. 974.

⁶ Dakako, to ne znači da ne treba poznavati metode kojima su drugi dolazili do rješenja za svoje jezike — one su dragocjeno iskustvo koje nam izravno pomaže.

⁷ <http://www.cordis.lu/ist/home.html>. Vidi također u Petek (2000), str. 100 koji iznosi podatke kako IST zauzima 26.3% ukupnoga FP5 proračuna dok MC&T sam ima proračun od 564 milijuna eura.

nim alatima a najčešće se primjenjuju na tekstovima kojima se upravo pristupa.

3. Obilježavanje jezičnih resursa

Obilježavanje strukture podataka u jezičnim resursima predstavlja središnji dio interesa ovoga rada.

3.1 Obilježavanje korpusa

Obilježavanje korpusa u lingvistici obavlja se na nelingvističkoj i lingvističkoj razini. Na nelingvističkoj razini obilježeni su elementi strukture teksta tj. dokumenta kao što su poglavlja u romanu, odlomci, kurzivi, navodnici itd.

```
<zbirka_pjesama>
<autor>Dobriša Cesarić</autor>
<naslov>Lirika</naslov>
<pjesma id="p1">
  <naslov>Oblak</naslov>
  <strofa id="pls1">
    <stih> U predvečerje, iznenada,</stih>
    <stih> Ni od kog iz dubine gledan,</stih>
    <stih> Pojavio se ponad grada</stih>
    <stih> Oblak jedan.</stih>
  </strofa>
  <strofa id="pls2">
    <stih>Vjetar visine ga je njih,</stih>
    <stih>I on je stao da se žari,</stih>
    <stih>Al oči sviju ljudi bjehu</stih>
    <stih>Uprte u zemne stvari.</stih>
  </strofa>
  ...
</pjesma>
<pjesma id="p2"> ... </pjesma>
...
</zbirka_pjesama>
```

Slika 1: Primjer nelingvističkoga obilježavanja (obilježavanje strukture pjesničkoga teksta)

Na lingvističkoj razini obilježavaju se jezične jedinice i/li tekstovno/jezično/obavijesni elementi za koje se pretpostavlja da mogu biti zanimljivi istraživačima.

```
<BODY>
<DIV0 type="MAIN">
<HEAD type="NA">
  <S>
    <W type="R">Outsideri</W>
    <W type="R">i</W>
    <W type="R">u</W>
    <W type="R">Zagrebu</W>
  </S>
</HEAD>
<HEAD type="PN">
  <S>
    <W type="R">Istodobno</W>
    <W type="R">s</W>
    <W type="R">petim</W>
    <W type="R">»Sajmom</W>
    <W type="R">outsider</W>
    <W type="R">umjetnosti<</W>
    <W type="R">u</W>
    <W type="R">New</W>
    <W type="R">Yorku</W>
```

```
<W type="R">u</W>
<W type="R">Muzeju</W>
<W type="R">suvremene</W>
<W type="R">umjetnosti</W>
<W type="R">u</W>
<W type="R">Zagrebu</W>
<W type="R">postavljena</W>
<W type="R">je</W>
<W type="R">izložba</W>
<W type="R">djela</W>
<W type="R">hrvatskih</W>
<W type="R">Outsidera</W>
</S>
</HEAD>
<P>
<S>
  <W type="R">Izložba</W>
  <W type="R">»Outsideri«</W>
  <W type="I">,</W>
  <W type="R">postavljena</W>
  <W type="R">u</W>
  <W type="R">Muzeju</W>
  <W type="R">suvremene</W>
  <W type="R">umjetnosti</W>
  <W type="R">u</W>
  <W type="R">Zagrebu</W>
  <W type="I">,</W>
  <W type="R">otvara</W>
  <W type="R">novo</W>
  <W type="R">poglavlje</W>
  <W type="R">u</W>
  <W type="R">proučavanju</W>
  <W type="R">umjetničkih</W>
  <W type="R">pojava</W>
  <W type="R">s</W>
  <W type="R">tzv</W>
  <W type="I">.</W>
  <W type="R">margine</W>
  <W type="I">.</W>
</S>
...
</P>
...
</DIV0>
<DIV0>...</DIV0>
...
</BODY>
```

Slika 2: Primjer jednostavnog lingvističkoga obilježavanja (obilježavanje rečenica i riječi)

U načelu vrijedi pravilo — što je više obilježavanja uneseno to se preciznije kasnije može pretraživati, no to, dakako, usložnjuje pretraživanje kao što i umnožava količinu podataka koji se moraju obraditi.

3.2 Obilježavanje rječnika

Rječnicima se, kao drugom tipu jezičnih resursa, također njihova inherentna struktura sastavljena od leksikografskih elemenata može eksplicitirati *mark-up* jezikom. Rječnički je tekst po svojoj strukturiranosti, koja se očituje u češćoj izmjeni kraćih tekstovnih odsječaka pripadajućih mnogovrsnim i naizmjenice ponavljajućim kategorijama,

različit od »tekućega« teksta u kojem prevladavaju dulji, monotoni tekstovni odsječci.

```
<entry key="bezant">
  <form>
    <orth type='hw'>bezant</orth>
    <orth type='variant'>bezzant</orth>,
    <orth type='variant'>byzant</orth>
    <pron>"beznt</pron>,
    <pron>bI"z&nt</pron>
  </form>
  <gramgrp><pos>n</pos></gramgrp>
  <sense>
    <trans>
      <tr>bizantinec</tr>,
      <tr>bizantinski zlatnik</tr>
    </trans>
  </sense>
  <sense>
    <trans><usg type='label'>Archit</usg>
    <tr>medaljon</tr>
    <gloss>ornament v obliki okrogle
      plošče</gloss>
    </trans>
  </sense>
  <sense>
    <trans><usg type='label'>Herald</usg>
    <tr>zlat krog</tr></trans>
  </sense>
</entry>
```

Slika 3: Primjer obilježenog rječnika⁸

Koliko rječnik u digitalno pohranjenom obliku dobiva na vrijednosti kad su mu svi leksikografski elementi eksplicitno obilježeni, te stoga i pretraživi, ne treba posebno argumentirati. Dovoljno je napomenuti da je Oxford English Dictionary u svojoj CD-ROM inačici čitav kodiran u SGML-u, a mnogi se rječnici danas obilježavaju tako ne samo zbog ljudskih korisnika već i zbog mogućnosti da im tako pristupe i aplikacije čime postaju pravi *machine readable dictionaries* (MRD).

4. Dva temeljna rješenja organizacije podataka

Informatičarsko bi oko moglo čitav problem pretraživanja masovne količine teksta kojem se mora moći pristupiti na jednostavan, jeftin i učinkovit način, moglo pokušati riješiti na dva moguća načina uz prateće im osobine/probleme:

1. tekst u bazi podataka
 - 1.1. tablice s desetcima/stotinama milijuna zapisa (npr. jedna riječ = jedan zapis)
 - 1.2. varijabilna duljina polja
 - 1.3. višestruko indeksiranje
 - 1.4. gubitak podatka o tekstovnoj strukturi
 - 1.5. promjene u realnom vremenu (?)
 - 1.6. mogući problemi:

⁸ Englesko-slovenski rječnik cf. Erjavec-Evans-Ide-Kilgariff (2000), str. 360.

1.6.1. konverzija teksta u zapise (što uključiti, a što isključiti: npr. interpunkcija, grafikoni, ilustracije, multimedija itd).

1.6.2. ažuriranje: korektura jednoga slova ili redosljeda riječi rezultira reindeksiranjem jedne ili više tablica

1.6.3. itd.

2. tekst sa strukturom eksplicitno obilježenom na standardiziran način nekim jezikom za obilježavanje (*mark-up language*)

2.1. tekstovna struktura obilježena gustim obilježavanjem elemenata: početci i krajevi (<p> i </p>)

2.2. varijabilna dužina elemenata

2.3. labava ali prisutna i provjerljiva struktura i podudarnost dokumenta sa zadanom strukturom (putem DTD tj. Document Type Descriptiona kojim se propisuje koji se elementi i kako smiju kombinirati)

2.4. lokalne modifikacije osnovnih obilježavanja putem atributa i njihovih vrijednosti (npr. <head type="podnaslov"> ... </head>

Svaki od ova dva pristupa ima svoje prednosti i nedostatke znane i analizirane u mnogo navrata. Kako je tema ovoga rada obilježavanje XML-om sasvim je jasno da je predmet interesa pristup broj 2. No zahtjevi koje lingvistika i jezične tehnologije postavljaju pred obilježavanje npr. korpusa mogu se svesti u nekoliko točaka:

1. mogućnost primjene više različitih vrsta obilježavanja
2. alternativna obilježavanja i/li različite verzije istog obilježavanja
3. različiti prirodni jezici
4. povezanost različitih medija (tj. tekst, govor, signal, audio, video, slike...)
5. moguće složeno povezivanje dokumenata, njihovih dijelova i podataka u netekstovnom obliku⁹

Za očekivati je da je obilježeni tekst kadar dati toliku fleksibilnost i mogućnost stvaranja modela podataka (*data model*¹⁰) koji je potreban za obilježavanje jezičnih resursa.

5. Odabir jezika za obilježavanje

Prvi i najstariji standardizirani jezik za obilježavanje koji redovito bude i prvi kandidat »djed« je svih njih — riječ je o SGML-u.¹¹ No SGML zapravo i nije jezik za obilježavanje već vrsta metajezika kojim se formalno definira generiranje pojedinačnih specifičnih jezika za obilježavanje. Tako je kao podskup SGML-a nastao i HTML kao što je to i XML.¹²

Mnogo se jezičnih resursa još i danas kodira u SGML-u tj. u TEI¹³ sustavu za obilježavanje tekstova no u posljednjih

⁹ cf. Ide (2000) str. 2.

¹⁰ O definiciji modela podataka i potrebe za formalnim opisom podatkovnih objekata (s obzirom na njihov sastav, attribute, pripadnost klasi, primijenljivim procedurama) vidi u Ide (2000), str. 2 i dalje.

¹¹ SGML (1986).

¹² Bray-Paoli-Sperberg-McQueen (1998).

¹³ Sperberg-McQueen, C. M. & Burnard, L. (1990).

je godina znatno ojačao trend »prevođenja« SGML/TEI dokumenata u XML dokumente.¹⁴

5.1 Argumenti za XML umjesto SGML-a

Zašto XML? Nekoliko osnovnih argumenata za XML umjesto SGML-a može se navesti u nekoliko točaka:

1. XML je najnoviji u nizu standarda koji su se koristili u obilježavanju jezičnih resursa (COCOA, SGML, TEI, CES,¹⁵ XML i XCES¹⁶)
2. XML kao i SGML u biti je niz pismena (tj. *ASCII characters*) ali podržava i UNICODE čime je postignuta globalna pismovna i jezična uporabljivost kao i neovisnost o operativnome sustavu
3. uparen sa snažnim jezikom za oblikovanje (*style language*) — XSL-om kao i jezikom za preoblikovanje (*transformation language*) — XSLT-om, omogućuje odabir, preobliku i prikaz podataka koji je potpuno fleksibilan
4. omogućuje obilježavanje kakvo prije sa SGML sustavima nije bilo moguće (razrađen sustav pointera i linkova, *stand-off* obilježavanje, obilježavanje *read-only* dokumenata...)
5. XML kao standardni jezik za obilježavanje podataka nije primjenljiv samo u uskom korpusnolingvističkom području već je njegova specifikacija takva da uključivanjem ostalih formata zapisa dokumenata (bilo izravno ili još bolje neizravno putem pointera) pokriva područje obilježene i dijelom strukturirane multimedije

5.2 Prelazak na XML

Krajem 1998, kad je XML još praktički bio »povojima«, donesena je u Zavodu za lingvistiku odluka da se svi korpusi koji se sastavljaju u okviru zavodskih projekata, kodiraju u XML-u. Ta se odluka danas pokazuje iznimno dalekosežnom jer se recimo tek u prosincu 1999. počeo sastavljati Američki nacionalni korpus koji se u cijelosti kodira u XML inačici CES-a tj. XCES-u.¹⁷

6. Hrvatski korpusi kodirani u XML-u

Za sada postoje tri hrvatska korpusa kodirana XML-om i svi se sastavljaju u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu u okviru projekata Ministarstva znanosti i tehnologije Republike Hrvatske.¹⁸

6.1 Hrvatski nacionalni korpus (HNK)

Najveći i ne samo lingvistički najznačajniji je Hrvatski nacionalni korpus kojemu je ime dano po ugledu na slične velike korpusne u svijetu¹⁹ no njegov su sastav i struktura

¹⁴ XML je postao toliko popularan da je odabran za interni format zapisa dokumenata u MS Office 2000, ono što je u MS Office 97 u neku ruku bio RTF.

¹⁵ Ide (1998) i <http://www.cs.vassar.edu/CES>.

¹⁶ Ide-Bonhomme-Romary (2000).

¹⁷ Iza tog mega-projekta stoje tri američka sveučilišta potpomognuta konzorcijem u kojem se nalaze npr. IBM, Microsoft, Oxford University Press, Cambridge University Press, Harper-Collins itd.

¹⁸ Projekti 130718 i 130821.

¹⁹ npr. Britanski nacionalni korpus (BNC), Češki nacionalni korpus (CNC) itd.

sasvim izvorni.²⁰ Naime, unatoč više od trideset godina tradicije obrade korpusa u Zavodu, zbog nedostatka sustavnog i standardiziranog digitalnog pohranjivanja starijih hrvatskih tekstova kao i nedostatka slobodnoga dostupa do njih, Hrvatski je nacionalni korpus zamišljen i potom sastavljen od dvije osnovne komponente:

1. 30-milijunski korpus suvremenoga hrvatskoga jezika (30m)
2. Hrvatski elektronski tekstovni arhiv (HETA)

Probna je inačica HNK veličine od oko 11 milijuna riječi tekućega teksta, koja još nije bile kodirana u XML-u, dostupna od prosinca 1998. i pretraživa na web-adresi <http://www.hnk.ffzg.hr>.

Trenutačna razina kodiranosti razina je odlomka tj. <P>. Do kraja 2000. trebala bi biti dopunjena 30-milijunska komponenta do ukupnoga opsega i kodirana do <P> razine.

U planu je preciznije obilježavanje na nižim razinama koje omogućuje i preciznije pretraživanje:

1. <S> razina uz bolji algoritam za segmentiranje na rečenice
2. <W> razina (sada već gotovo automatizirana)
3. obilježavanje lematizacije
4. ostale vrste obilježavanja (npr. imena, vremenskih izraza²¹ itd.)

Nakon 2000. očekuje se podrška Ministarstva za rast korpusa do 100 milijuna riječi.

```
<BODY>
<DIV0 type="article" n="v1990311ck01">
  <HEAD type="nn">
    <W type="R">
      <ORTH>POLICIJA</ORTH>
      <LEX>
        <BASE>policija</BASE>
        <MSD>Ncfsn</MSD>
      </LEX>
    </W>
    <W type="R">
      <ORTH>o</ORTH>
      <LEX>
        <BASE>o</BASE>
        <MSD>Spsl</MSD>
      </LEX>
    </W>
    <W type="R">
      <ORTH>DETALJIMA</ORTH>
      <LEX>
        <BASE>detalj</BASE>
        <MSD>Ncmpl</MSD>
      </LEX>
    </W>
    <W type="R">
      <ORTH>VEZANIM</ORTH>
      <LEX>
        <BASE>vezan</BASE>
```

²⁰ Tadić (1996) i Tadić (1998).

²¹ Prepoznavanje imena (*named entity detection*) i vremenskih izraza (*temporal expressions*) predstavljaju osnovne oblike *information extractiona* i *data-mininga* na temelju prirodnojezičnih dokumenata.

```

    <MSD>Afpmp1-</MSD>
  </LEX>
</W>
<W type="R">
  <ORTH>UZ</ORTH>
  <LEX>
    <BASE>uz</BASE>
    <MSD>SpSa</MSD>
  </LEX>
</W>
<W type="R">
  <ORTH>NOVE</ORTH>
  <LEX>
    <BASE>nov</BASE>
    <MSD>Afpfpa</MSD>
  </LEX>
</W> ...
</HEAD> ...
</DIV0>
</BODY>

```

Slika 4: Primjer lematiziranoga korpusnoga teksta gdje je svaka riječ popraćena gramatičkim kategrijama s kojima je ostvarena

6.2 Hrvatsko-engleski paralelni korpus

Drugi projekt u kojem je XML kodiranje korpusa primijenjeno je Hrvatsko-engleski paralelni korpus. Za razliku od jednojezičnoga HNK ovdje je riječ o dvojezičnom korpusu sa hrvatskim izvornikom i engleskim prijevodom. Tekst korpusa čini 118 brojeva tjednih novina *Croatia Weekly* koje su izlazile od siječnja 1998. do travnja 2000. Korpus je ukupne veličine 3,5 milijuna riječi. Svoju vrijednost paralelni korpusi dobivaju tek kad su prijevodni ekvivalentni teksta u oba jezika sravnjeni (*aligned*). Tako obrađeni korpusi koriste se za istraživanje prijevoda, višejezičnu leksikografiju i, u konačnici, istraživanja strojno (potpomognutog) prevođenja. Najčešći je oblik sravnjivanja na razini rečenice. U načelu više od 80% pojedinačnih rečenica izvornika prevodi se jednom rečenicom, no moguće su i druge kombinacije (0:1, 1:0, 1:2, 2:1, 2:2 itd.).

```

*** Länk: 1 - 1 ***
<BODY> <DIV0 type="MAIN"> <HEAD type="NA">
<S id="CW014199804160101hr.S1"> Neće biti
prijevremenih izbora </S> </HEAD> .EOS
<BODY> <DIV0 type="MAIN"> <HEAD type="NA">
<S id="CW014199804160101en.S1"> NO EARLY
ELECTIONS </S> </HEAD> .EOS .EOP

```

```

*** Länk: 1 - 1 ***
<HEAD type="PN"><Sid="CW014199804160101hr.S2">
Parlamentarni izbori održat će se u redovnom
roku, što znači za više od godinu dana te su
bez osnova spekulacije o raspisivanju
prijevremenih izbora, poručeno je iz vladajuće
stranke </S> </HEAD> .EOS
<HEAD type="PN"><Sid="CW014199804160101en.S2">
According to the ruling HDZ, parliamentary
elections will be held in due course next
year, so that all speculations on calling
early elections are without basis </S> </HEAD>
.EOS .EOP

```

```
*** Länk: 1 - 1 ***
```

```

<P> <S id="CW014199804160101hr.S3"> »Izbori su
za više od godinu dana i sve špekulacije o
prijevremenim izborima, po raznim formulama, o
kojima se govori u javnosti, potpuno su bez
osnova. </S> .EOS
<P> <S id="CW014199804160101en.S3"> "Elections
will be held in a little over a year, and all
speculations about early elections, which are
being discussed in the media and among the
public, have no grounds whatsoever. </S> .EOS

```

Slika 5: Primjer sravnjenih rečenica iz Hrvatsko-engleskoga paralelnoga korpusa²²

Pri obilježavanju sravnjenja tj. povezivanju prijevodnih ekvivalenata korištena je XML-ova mogućnost definiranja veza među dokumentima putem ID atributa.

Dokument 1 (hrvatski):

```

<DIV0 type="MAIN">
  <HEAD type="NA">
    <S id="CW010199803190201hr.S1">Do 1.
kolovoza zabranjeni skupovi u ...</S></HEAD>
  <HEAD type="PN">
    <S id="CW010199803190201hr.S2">Vlada je
ocijenila kako je provođenje mirne ...</S>
    <S id="CW010199803190201hr.S3">Stoga, treba
izbjeći svaki čin koji ...</S></HEAD>
  <P>
    <S id="CW010199803190201hr.S4">Vlada
Republike Hrvatske obvezala je ...</S> ...</P>...
</DIV0>

```

Dokument 2 (engleski):

```

<DIV0 type="MAIN">
  <HEAD type="NA">
    <S id="CW010199803190201en.S1">POLITICAL
RALLIES ...</S> </HEAD>
  <HEAD type="PN">
    <S id="CW010199803190201en.S2">The
Government has assessed that the ...</S>
  </HEAD>
  <P>
    <S id="CW010199803190201en.S3">The Croatian
Government has charged ...</S> ... </P> ...
</DIV0>

```

Dokument 3 (sravnjenja):

```

<link xtargets="CW010199803190201hr.S1 ;
CW010199903190201en.S1">
<link xtargets="CW010199803190201hr.S2
CW010199803190201hr.S3 ;
CW010199903190201en.S2">
<link xtargets="CW010199803190201hr.S4 ;
CW010199903190201en.S3">

```

Slika 6: Primjer kodiranja sravnjenja s pomoću trećeg XML dokumenta u kojem su pohranjene veze između prva dva dokumenta ili njihovih dijelova

²² Sravnjenja iz primjera kao i u cijelom korpusu obavljena su programom Vanilla Aligner v. Danielsson & Ridings (1997).

XML u načelu omogućuje pristupanje dokumentima, njihovim elementima ili dijelovima tih elemenata koji ponovno mogu biti elementi ili pismena i to unutar istoga dokumenta ili unutar drugih XML dokumenata. Ti su mehanizmi povezivanja bitno snažniji od SGML-a:

1. XLink: mehanizam za definiranje veze (jedno- ili višesmjerne) između dva dokumenta ili njihovih dijelova
2. XPath: proširena sintaksa za adresiranje kojom se određuje precizan oblik lociranja u stablu dokumenta i omogućuje pristupanje pojedinim elementima ili njihovim dijelovima
3. XPointer: proširenje XPath sintakse koje omogućuje pristupanje čvorovima ili cijelim nizovima elemenata ili njihovih dijelova²³

Npr. XPath izraz `/div/p[2]/s[3]` adresira svaku treći `<s>` element unutar svakog drugog `<p>` elementa u svim `<div>` elementima unutar XML dokumenta. XPath također omogućuje pristupanje odsječcima teksta unutar elemenata. Tako izraz

```
substring(head[2]/s[2]/text(),8,12)
```

daje »treba izbjeći« iz dokumenta 1. navedenog u primjeru 6.

6.3 Hrvatsko-slovenski paralelni korpus

Treći projekt koji se služi s XML kodiranjem korpusa je Hrvatsko-slovenski paralelni korpus. Riječ je o primjeni iste metodologije kao i u 6.2 na različit jezični par i na korpus ukupne duljine milijun riječi. Partneri na projektu su Filozofski fakulteti u Zagrebu i Ljubljani.

7. Nedostatci XML-a

Trenutačni nedostatci XML-a su prije svega još uvijek nedovršena W3C specifikacija jezika. XML je toliko nov da je još uvijek u stadiju intenzivnog razvitka i novi se dodatci pojavljuju gotovo svaki mjesec.²⁴

Ne toliko za istraživačku, ali ponajprije za komercijalnu primjenu XML-a, jedan od ograničavajućih čimbenika je i sporost XML parsera (bar na Windows platformi). Sporost XSL(T)-a također je prisutna.

Microsoft je vrlo rano u razvitku specifikacije XML-a implementirao XML parser u IE 5.0 no u međuvremenu se specifikacija promijenila tako da se MS implementacija razlikuje od sadašnje službene W3C specifikacije. To je još jedan razlog za moguće probleme oko uporabe XML-a.²⁵

Zbog znatne količine dodatnih podataka kojima se obilježava osnovni sadržaj dokumenta XML datoteke znaju biti vrlo velike što, dakako, utječe na potrebne strojne i vremenske resurse za njihovu obradu.²⁶

Idealna situacija koju bismo rado vidjeli bila bi relacijska baza s ulaznim XML filtrom s pomoću kojeg bi izravno konvertirala XML elemente u RDBMS a istodobno bi se sačuvala fleksibilnost i mogućnost pristupanja svakom elementu XML dokumenta. Najava takve baze pod kodnim imenom Shiloh (MS-SQL 8?) postoji već gotovo godinu dana. Nadajmo se uskoro vidjeti i takvo rješenje.

8. Bibliografija

- Bray, T., Paoli, J., Sperberg-McQueen, C. M. (eds.) (1998) Extensible Markup Language (XML) Version 1.0. W3C Recommendation. (<http://www.w3.org/TR/1998/REC-xml-19980210>).
- Clark, J. (ed.) (1999) XSL Transformations (XSLT), Version 1.0. W3C Recommendation. (<http://www.w3.org/TR/xslt>).
- Clark, J. and DeRose, S., (1999) XML Path Language (XPath), Version 1.0. W3C Recommendation. (<http://www.w3.org/TR/xpath>)
- Danielsson, Pernilla & Ridings, Daniel. (1997). Practical presentation of a "vanilla" aligner. In U. Reyle & C. Rohrer (Eds.), Presented at the TELRI Workshop on Alignment and Exploitation of Texts. Institute Jožef Stefan, Ljubljana (<http://svenska.gu.se/PEDANT/workshop/workshop.html>)
- DeRose, S., Maler, E., Orchard, D. Trafford, B. (eds.) (2000) XML Linking Language (XLink). W3C Working Draft, 2000-02-21. (<http://www.w3.org/TR/xlink>).
- DeRose, S., Daniel, R., Maler, E. (1999) XML Pointer Language (XPointer). W3C Working Draft, 1999-12-06. (<http://w3.org/TR/xptr>)
- EAGLES (1996) Preliminary Recommendations on Corpus Typology, 1996. (<http://www.ilc.pi.cnr.it/EAGLES/home.html>).
- Erjavec, T., Evans, R., Ide, N., Kilgariff, A. (2000) The CONCEDE model for Lexical Databases. In Proceedings of LREC2000, ELRA, Paris-Athens, str. 355-362.
- Ide, Nancy. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In Proceedings of LREC'98., ELRA, Granada, str. 463-470.
- Ide, Nancy (2000) The XML Framework and Its Implications for Corpus Access and Use. In Data Architectures and Software Support for Large Corpora, LREC2000 Workshop Proceedings, ELRA, Paris-Athens, str. 28-32.
- Ide, N., Bonhomme, P., Romary, L. (2000) XCES: An XML-based Encoding Standard for Linguistic Corpora. In Proceedings of LREC2000, ELRA, Paris-Athens, str. 825-830.
- Ide, N. & Brew, C. (2000) Requirement, Tools and Architectures for Annotated Corpora. In Data Architectures and Software Support for Large Corpora, LREC2000 Workshop Proceedings, ELRA, Paris-Athens, str. 1-5.
- Liefke, H. & Suci, D. (1999) XMill: an Efficient Compressor for XML Data, Univ. of Pennsylvania Technical Report MS-CIS-99-26.
- Petek, B. (2000) Funding for Research into Human Language Technologies for Less Prevalent Languages. In Developing Language Resources for Minority

²³ Ide (2000), str. 28.

²⁴ Zainteresirani se za sve novosti oko specifikacije XML-a mogu obratiti na www.w3.org.

²⁵ Premda Microsoft tvrdi da IE 5.5 podržava obje specifikacije.

²⁶ Sažimanje XML dokumenata jedan je od mogućih postupaka koji su se pokušali primijeniti u takvim slučajevima. V. Liefke & Suci (1999)

- Languages: Reusability and Strategic Priorities, LREC2000 Workshop Proceedings, ELRA, Paris-Athens, str. 100-105.
- SGML (1986) ISO 8879: Information processing — Text and office systems — Standard Generalized Markup Language (SGML), ISO, Geneva.
- Sperberg-McQueen, C. M. & Burnard, L. (1990) Guidelines for the Encoding and Interchange of Machine-Readable Texts, Text Encoding Initiative, Chicago-Oxford.
- Tadić, M. (1996) Računalna obradba hrvatskoga i nacionalni korpus. In *Suvremena lingvistika* 41-42, str. 603-612.
- Tadić, M. (1998) Raspon, opseg i sastav korpusa suvremenog hrvatskoga jezika. In *Filologija* 30-31, str. 337-347.
- Tadić, M. (2000) Building the Croatian-English Parallel Corpus. In *Proceedings of LREC2000, ELRA, Paris-Athens*, str. 523-530.
- Thompson, H. & McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents. In *SGML Europe'97*. (<http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>).