SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET

BOŽO BEKAVAC

Primjena računalnojezikoslovnih alata na hrvatske korpuse

MAGISTARSKI RAD

ZAGREB, 2001.

Sadržaj

1. Uvod

U posljednjim se desetljećima količina podataka zapisanih prirodnim jezikom na elektroničkome mediju znatno povećala. Prvo pitanje koje je iz toga proizišlo bilo je kako pohraniti i organizirati toliku količinu podataka na elektroničke medije. Međutim, drugo, mnogo složenije pitanje, koje se nametnulo nakon prvoga, bilo je kako ekstrahirati i pretraživati takve prirodnim jezikom kodirane zapise. Računala ne mogu "razumjeti" prirodni jezik. Ona mogu samo pohranjivati i pretraživati lingvističke podatke koji odgovaraju opisu podataka što se već nalaze u tekstu. Pretraživanje će biti moguće samo ako se podaci (...) po kojima se pretražuje, već nalaze pohranjeni na računalu.[1] Takvi i slični problemi, ali i razvoj informatike promijenio je tradicionalni pristup proučavanju prirodnoga jezika. Povećana dostupnost velikih količina elektroničkih tekstova, razrađivanje lingvističke metodologije i nagli napredak informatičke tehnologije povoljno su utjecali na razvoj korpusne lingvistike.

Zahvaljujući takvom razvoju proučavanje se jezičnoga fenomena zasnovanoga na računalnome korpusu prirodnoga jezika u posljednjih tridesetak godina premjestilo s margina u sam centar jezičnih istraživanja. **Korpusna lingvistika** u najširem smislu označava istraživanje jezika na osnovi korpusa tekstova, pri čemu se danas obično podrazumijeva – strojno izrađenih korpusa.[2] Tekstovi koji čine korpus zasnovani su na konkretnim jezičnim ostvarajima. Korpusna lingvistika ne bavi se jezikom samo u apstraktnom smislu već se bavi ponajprije konkretnim ostvarenim tekstovima. Predmet opisa u načelu je samo ono što je obuhvaćeno korpusom. **Korpus** podrazumijeva zbir tekstova prirodnoga jezika sastavljen po stanovitu kriteriju.[3]

Iako se nalazi na dodiru računala i lingvistike, korpusnu lingvistiku ne bi trebalo miješati s **računalnom lingvistikom** koja joj je nadređena:

[1] Henderson (1999):1
[2] Bratanić (1991):145

> A branch of linguistics in which computational techniques and concepts are applied to the elucidation of linguistic and phonetic problems. [4]

Tako je unutar računalne lingvistike razvijeno nekoliko područja istraživanja kao što su sinteza govora, strojno prevođenje, korpusna lingvistika, prepoznavanje govora i mnoga druga.

Korpusna lingvistika nije grana lingvistike u istom smislu u kojem su to npr. sociolingvistika, poredbena lingvistika ili psiholingvistika. Ne može ju se odrediti ni prema određenoj razini proučavanja jezika kao npr. morfologiju, sintaksu ili semantiku. Navedene discipline usmjerene su na opisivanje ili objašnjavanje određenih aspekata jezika i definirane su ponajprije svojim jedinicama. Korpusna je lingvistika prije svega *metodologija* koja se zasniva na empirijskom pristupu podacima, te *sredstvo* kojim se može služiti većina navedenih disciplina, ali i mnoge discipline izvan lingvistike.[5]

Pedesetih i šezdesetih godina prošloga stoljeća objavljivanjem niza utjecajnih publikacija Chomsky mijenja smjer lingvistike od empirizma prema racionalizmu. Prema njemu, "jezik je sustav predstavljen u umu/mozgu određenog pojedinca".[6] Jezična je kompetencija (*competence*) najvažnija za lingvističko proučavanje, a smještena je u um govornika nekoga jezika. Konkretni jezični ostvaraji manje su važni jer se na njih može utjecati faktorima koji su izvan jezične kompetencije. Jezične ostvaraje Chomsky smatra siromašnim ogledalom kompetencije,[7] a korpus siromašnim sredstvom za modeliranje kompetencije. Svojim nastupima Chomsky ne samo da umanjuje vrijednost korpusnom pristupu i empirijskom istraživanju jezika, već nastoji odbaciti svaku potrebu za korpusom.[8]

Unatoč kritikama Chomskoga koje su rezultirale sporijim razvojem korpusne lingvistike do sedamdesetih godina prošloga stoljeća, velik je broj istraživača svoj rad i dalje zasnivao na korpusima. Uočili su prednosti konkretnih jezičnih ostvaraja nad introspekcijom, jer je podatke iz korpusa moguće promatrati i provjeravati interpersonalno. Naglim razvojem korpusne lingvistike osamdesetih godina prošloga stoljeća pojavljivalo sve se više argumenata u korist toga pristupa jezičnome opisu.

---

[3] Bratanić (1991):146
[4] CCL (2000): (Računalna je lingvistika…) grana lingvistike koja koristi računalne tehnike i koncepte za rasvjetljivanje lingvističkih i fonetskih problema. (prijevod moj)
[5] McEnery & Wilson (1996):2
[6] Chomsky (1991):38 (prijevod G. Antunović)
[7] McEnery & Wilson (1996):5: (...) *a poor mirror of competence* (...)

Unatoč brojnim previranjima, danas su lingvisti suglasni da korpusni pristup ne odriče vrijednost introspekciji i intuiciji već ih smješta u komplementaran odnos prema istraživanju provjerljivih podataka.[9] Takvo suglasje u nadopunjavanju dvaju pristupa zorno ocrtava Fillmore:

> I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore… [but] every corpus I have had the chance to examine, however small, has taugth me facts I couldn't imagine finding out any other way. My conclusion is that the two types of linguists need one another.[10]

Do sada ne postoji ni jedan magistarski rad ili doktorska disertacija objavljena na hrvatskome jeziku koja obuhvatnije obrađuje računalnolingvističke alate. Iz tog razloga, ali i zbog njihove iznimne važnosti u proučavanju jezika danas, smatram potrebnim u prvom dijelu rada dati pregled općih pojmova ovoga područja (uvod, prvo i drugo poglavlje). Kodiranje korpusa po međunarodnim standardima (poglavlje 3) neizbježno je zbog presudne važnosti pri sastavljanju suvremenih korpusa. Poglavlja 4 i 5 posvećena su računalnojezikoslovnim alatima, njihovoj primjeni na hrvatske korpuse kao i planovima budućega razvoja. U zaključku se daje kritički pregled odnosa hrvatskih korpusa i alata prema suvremenim svjetskim korpusima i prijedlog za njihov daljnji razvoj. Dodaci su korišteni zbog uštede prostora, jer se podaci iz nekoliko poglavlja ponekad referiraju na iste sadržaje.

Korištena terminologija kod stranih autora kad se govori o korpusima (pogotovo o kodiranju korpusa) nije u potpunosti konzistentna. Kako na hrvatskome još ne postoje usuglašeni prijevodi, uz naziv će kad to bude potrebno stajati izvorni termini na engleskome jeziku kojima se služi većina stranih autora.

---

[8] Poznat je njegov primjer kako je rečenica "I live in New York" mnogo češća nego "I live in Dayton, Ohio". Time je htio pokazati da učestalost jedinica u tekstu nema nikakvog značaja.
[9] Bratanić (1991):157
[10] cf. Fillmore prenesen u McEnery & Wilson (1996):18: Mislim da ne može postojati ijedan korpus, bez obzira koliko velik, koji bi sadržavao informacije o svim područjima rječnika i gramatike engleskoga jezika koje bih htio proučavati… [ali] svaki korpus koji sam imao priliku istraživati, bez obzira koliko malen, poučio me činjenicama za koje ne mogu zamisliti da bih ih pronašao na drugi način. Moj je zaključak da dvije vrste jezikoslovaca trebaju jedna drugu. (prijevod moj)

2. Jezični korpus na računalu

2.1. Položaj računalnoga korpusa unutar područja jezičnih tehnologija

Područje se jezičnih tehnologija (*Human Language Technologies*) sastoji od tri osnovna dijela: [11]

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

Jezični su resursi izvori jezičnih tekstova, a sastoje se od korpusa i rječnika pohranjenima u digitalnome obliku tj. u obliku elektroničkoga teksta.[12] Dakle, kada se govori o jezičnim resursima u prvom se redu misli na tekstove nad kojima se obavlja pretraživanje ili služe kao ulazni podaci za daljnju obradu. Ili još preciznije:

> The term *linguistic resources* refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems. Examples of linguistic resources are written and spoken corpora, lexical databases (…)[13]

Jezični su alati aplikacije koje obrađuju ili se služe postojećim resursima kako su definirani gore, ili tekstovima koji se upravo stvaraju. Razvoj je jezičnih alata u usporedbi s jezičnim resursima znatno sporiji.[14]

Komercijalni su proizvodi nastali na temelju istraživanja jezičnih resursa jezičnim alatima.[15] Najčešće se primjenjuju na tekstovima kojima se upravo pristupa,

[11] Tadić (2000a):132
[12] Tadić (2000a):132
[13] Godfrey & Zampolli (1996): Termin *jezični resursi* odnosi se na (obično velike) skupove jezičnih podataka i opisa u strojno čitljivom obliku koji se koriste u izgrađivanju, poboljšavanju ili vrednovanju algoritama ili sustava prirodnoga jezika ili govora. Primjeri jezičnih resursa su pisani i govoreni korpusi, leksičke baze podataka, (…) (prijevod moj)
[14] Sinclair (1991):20

npr. *spelling-checker*. Ne bi ih trebalo miješati s nekim komercijalnim jezičnim alatima za obrade korpusa o kojima će biti riječi u ovome radu.

2.2. Podjela elektroničkih tekstova

U korpusnoj se lingvistici korpus razlikuje od ostalih tekstova prema lingvističkim kriterijima po kojima je odabran i sakupljen. Prema EAGLES preporukama definicija *korpusa* glasi:

> A *corpus* is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.[16]

Za razliku od korpusa, *zbirka* je *tekstova* (ili arhiv) definirana:

> (...) *collection* and *archive* refer to sets of texts that do not need to be selected, or do not need to be ordered, or the selection and/or ordering do not need to be on linguistic criteria.[17]

Prema istim preporukama *računalni* je *korpus*:

> A *computer corpus* is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks.[18]

Dugi niz godina uporaba termina *korpus* odnosila se samo na tiskani tekst. Međutim, danas se pri spominjanju *korpusa* gotovo podrazumijeva svojstvo strojne čitljivosti.[19]

2.3. Povijest računalnih korpusa

Ideja sastavljanja korpusa i korpusno zasnovanog proučavanja jezika mnogo je starija od pojave računala. Korpus opsega 11 milijuna pojavnica koji se sastoji od tekstova na njemačkom jeziku sastavio je *Kaeding* 1897. godine.[20] Taj je korpus impresivne veličine za svoje vrijeme, osobito u svjetlu činjenice da je obrađivan

---

[15] Tadić (2000a):132
[16] EAGLES (1996a): *Jezični* je *korpus* zbirka jezičnih odsječaka koji su odabrani i sakupljeni prema eksplicitnim lingvističkim kriterijima upravo s ciljem da čine jezični uzorak. (prijevod: M. Tadić u Tadić (1998):337)
[17] EAGLES (1996a): (…) *zbirka* i *arhiv* se odnose na skup tekstova koji ne trebaju biti odabrani ili sakupljeni, ili odabiranje i/ili sakupljanje ne treba biti prema lingvističkim kriterijima. (prijevod moj)
[18] EAGLES (1996a): *Računalni* je *korpus* korpus koji je kodiran na standardan i dosljedan način s nakanom da bude otvoren za računalno pretraživanje. (prijevod: M. Tadić u Tadić (1998):337)
[19] McEnery & Wilson (1996):14

ručno. U pogledu opsega može se mjeriti i s nekim suvremenim računalnim korpusima.

Prvi je računalno sastavljen i podržan korpus **Brown korpus**. Sastavili su ga *Kučera i Francis* na Odsjeku za lingvistiku Sveučilišta Brown. Zgotovljen je tijekom 1963-4. godine na temelju tekstova na američkom engleskom jeziku. Tekstovi su objavljeni tijekom 1961. godine (neki su pisani neznatno ranije), a korpus sadrži 1,014,312 pojavnica.[21] Brown korpus uzima se kao standard za buduće korpuse, pa se može pisati Brown korpus =1. Osnovne karakteristike ovoga korpusa su: [22]

1. sastoji se od oko milijun pojavnica,
2. podijeljen je otprilike ravnomjerno po žanrovima (15 žanrova),
3. sadrži 500 uzoraka teksta,
4. oko 2000 pojavnica u svakom uzorku,
5. pisani izvori objavljeni u 1961. godini.

Neki od ovih tipova korpusnih parametara vrijede i danas, iako su se okolnosti od vremena kad se sastavljao Brown korpus znatno promijenile. Dakle, okvirni parametri za sastavljanje korpusa koji i danas vrijede bili bi po točkama: [23]

1. Korpus bi trebao biti velik koliko mu omogućuju tehnološke mogućnosti tog vremena. Kada se pojavio Brown korpus, milijun pojavnica bila je revolucionarna veličina. Sredinom sedamdesetih red veličine naglo raste. Taj se trend nastavlja i kasnije tako da *Birmingham Collection of English Texts* 1985. godine ima oko 20 milijuna pojavnica, a sredinom devedesetih, *Bank of English* ima već oko 200 milijuna pojavnica.
2. Žanrovi trebaju biti klasificirani i stajati u određenom omjeru prema korpusu u cjelini i prema pojedinim potkorpusima.
3. Korpus treba sadržati što je moguće širi raspon jezične građe (uzoraka teksta) u cilju postizanja što bolje reprezentativnosti.
4. Uzorci bi trebali biti otprilike jednake veličine (ovo je još uvijek dvojbeni zahtjev za neke današnje sastavljače korpusa).
5. Svi bi izvori trebali biti deklarirani, tj. eksplicitno navedeni.

U današnjem poimanju termina *korpus* u suvremenoj lingvistici, nužno je zadovoljavanje slijedećih uvjeta:[24]

---

[20] McEnery & Wilson (1996):3; Moguš & Bratanić & Tadić (1999):5
[21] Kučera & Francis (1967a)
[22] EAGLES (1996a)

1. **uzorkovanje i reprezentativnost**: uzorci bi u najvećoj mogućoj mjeri trebali oslikavati stanje jezičnoga varijeteta na koji se odnose,
2. **konačna veličina**: određen i konačan broj pojavnica, iako ne mora uvijek nužno biti tako,[25]
3. **strojno-čitljiv oblik**: ovo se svojstvo danas gotovo samo po sebi podrazumijeva kako bi korpus bilo moguće pretraživati i manipulirati njime na suvremen način,
4. **standardna referenca**: svim bi istraživačima trebao biti omogućen pristup u jednakim uvjetima.

Uz Brown korpus najpoznatiji je korpus toga vremena **LOB korpus** (***L**ancester-**O**slo-**B**ergen*). Cilj LOB-a bio je sastaviti korpus britanskog engleskog koji bi bio ekvivalentan Brown korpusu. Struktura i opseg LOB-a gotovo su identični uzoru po kojemu je rađen, uz neizbježnu razliku u izboru tekstova, a to su tekstovi na britanskom engleskom jeziku.[26] Analogijom prema računalima, a prema kriteriju njihove veličine, korpusi od oko milijun pojavnica (Brown i LOB) spadaju u korpuse **prve generacije**.[27]

Drugoj generaciji korpusa pripadaju korpusi osamdesetih godina prošloga stoljeća opsega do 30 milijuna pojavnica. Tipični su predstavnici te generacije korpusa Sinclairov *Birmingham Collection of English Texts* opsega oko 20 milijuna i *Longman/Lancester English Language Corpus* opsega 30 milijuna pojavnica.[28] Unatoč znatnim povećanjima opsega korpusa kojima je ponajprije doprinijela OCR (*optical character recognition*) tehnologija, korpusnim je projektima tog vremena još uvijek glavna prepreka bila nedostupnost e-tekstova.[29] Većina se tekstova unosila skeniranjem ili ručno, tj. utipkavanjem.[30]

Korpusima bi **treće generacije** pripadali korpusi koji su nastali od sredine devedesetih na ovamo, a veličina im je stotinu, pa i nekoliko stotina milijuna pojavnica. Prvi nacionalni korpus koji je ponio takav naziv i postao referentan za neki

---

[23] EAGLES (1996a)
[24] McEnery & Wilson (1996):21
[25] v. monitor korpus, str. 12
[26] Lawrer & Dry (1998):103
[27] Bratanić (1991):154 ;Leech (1991):10
[28] Leech (1991):10
[29] Tadić (2000a):132, Glossary (1999): E-tekst je skraćeni naziv za tekst koji se nalazi u elektroničkome obliku, tj. onaj tekst koji je strojno čitljiv.
[30] McEnery & Wilson (1996):170

jezik je **BNC** (***B**ritish **N**ational **C**orpus*).[31] Sadrži 100 milijuna pojavnica suvremenoga (pisanoga i govorenoga) engleskoga jezika. Zgotovljen je 1994. godine, i on je uopće prvi sastavljeni 100 milijunski korpus.

Bank of English[32] je *COBUILD-ov*[33] korpusni projekt pokrenut u prvome redu u leksikografske svrhe, gdje se prikupljaju tekstovi ne engleskome jeziku koji su uglavnom nastali nakon 1990. godine. U listopadu 2000. godine *Bank of English* ima opseg 415 milijuna pojavnica s izraženom daljnjom tendencijom rasta.

Ako se uzmu široko, moglo bi se reći da u korpuse treće generacije pripadaju i velike baze tekstovnih podataka, tzv. kompjutorizirani arhivi. Sastavljaju se ponajprije od tekstova koji su lakše dostupni, jer već od 80-tih godina postoji velik broj e-tekstova primarno namijenjenih elektroničkoj distribuciji, a često se u tiskanoj inačici uopće i ne pojavljuju.

Među poznatijim je korpusima treće generacije ***O**xford **T**ext **A**rchive* (**OTA**)[34] gdje se nalazi više od 2500 elektroničkih tekstova na više od 25 različitih jezika. OTA je pokrenut 1976. godine[35] kako bi se osiguralo dugoročno spremište elektroničkih tekstova koji su interesantni za jezikoslovna i književna istraživanja. Jedna od najvećih baza korpusa i drugih elektroničkih zbirki tekstova danas je **LDC** (***L**inguistic **D**ata **C**onsortium*)[36]. Osnovan je 1992. godine i nalazi se na Sveučilištu u Pittsburgu, Pennsylvania. Cilj je LDC-a sastavljanje, prikupljanje i distribucija tekstovnih i govornih baza podataka, korpusa, leksikona i sličnih resursa u svrhu istraživanja i razvoja.

Poznatijim korpusima slavenskih jezika svakako pripada *Češki nacionalni korpus* (**CNC**). Sastavlja se u *Institutu za češki nacionalni korpus* koji je 1994. godine osnovan upravo u tu svrhu. Putem Interneta javno je dostupan 20 milijunski korpus koji je dio 100 milijunskog korpusa.

U Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu sastavlja se **HNK** (***H**rvatski **N**acionalni **K**orpus*).[37] Dvije sastavnice ovoga korpusa imaju za cilj u prvoj fazi obuhvatiti tridesetak milijuna pojavnica koji bi se u drugoj fazi

[31] BNC (1997)
[32] Bank of English (2000)
[33] *COBUILD* je odio izdavačke kuće *HarperCollins* specijaliziran za sastavljanje i obradu korpusa
[34] OTA (2001)
[35] Osnivač OTA-e je *Lou Burnard*.
[36] LDC (1999)
[37] HNK (1999a)

proširili na stotinjak milijuna pojavnica. HNK je slobodno dostupan i pretraživ putem Interneta.

Veličini korpusa pridodavala se znatno veća pozornost u prvim generacijama korpusa nego danas. U tom pogledu, danas pri sastavljanju korpusa prevladavaju dva pristupa: prikupiti u korpus sve dostupne tekstove u strojno čitljivu obliku ili pak pažljivo izabirati tekstove, odrediti im duljinu itd.[38] Kako se dostupnost velikih količina e-tekstova u posljednje vrijeme naglo povećala, stvorena je zamisao[39] o tzv. **monitor korpusu** (*monitor corpus*) koji nije ograničen standardnim korpusnim parametrima (u prvom redu konačnom veličinom i vremenskim rasponom tekstova u korpusu). Takav pristup odražava poznati Sinclairov stav o veličini korpusa koje smatra više "zbirkama tekstova" nego korpusima, pa tvrdi:

> The bigger the better.[40]

Tekstovi se u monitor korpus mogu dodavati tako da mu se opseg stalno povećava, ili se mogu osvježavati u smislu da zastarjeli izlaze, a suvremeni ulaze u korpus određenom vremenskom dinamikom (npr. godišnje, mjesečno ili tjedno).[41] Najpoznatiji je Sinclairov monitor korpus koji sastavlja sa skupinom iz COBUILD-a na Sveučilištu u Birminghamu.

2.4. Sastavljanje korpusa

Pri sastavljanju korpusa ponajprije je potrebno voditi računa o korisnicima kojima je namijenjen, jer opseg, kvaliteta i struktura korpusa najviše ovise upravo o *potencijalnoj* ili *ciljanoj* skupini korisnika. Ciljana skupina korisnika u načelu je manji broj istraživača, često usmjeren na određeno područje proučavanja, pa se može govoriti o specijaliziranim korpusima (npr. korpus djela nekog pisca ili korpus npr. pravnih tekstova). U današnjim komunikacijskim mogućnostima (Internet) potencijalni krug korisnika većeg korpusa (npr. nacionalnog korpusa) može biti cijeli svijet.[42]

---

[38] Bratanić (1991):154
[39] Ideja o monitor korpusu potekla je od *Johna Sinclair*a
[40] cf. Joscelyne (1991): Što veći, to bolji. (prijevod moj)
[41] EAGLES (1996a)
[42] v. statistiku HNK, dodatak F1, osobito F2.

Cijena izrade obilježenoga korpusa može biti vrlo visoka u financijskom smislu, ali i smislu angažiranja kvalificiranih ljudskih resursa.[43] U prošlosti je na troškove izrade najviše utjecala sama računalna obrada jer je računalna oprema bila skupa, a vrijeme obrađivanja znatno sporije. Danas na cijenu izrade korpusa najviše utječe količina uloženog ljudskog rada koja je višestruko skuplja od rada stroja.

Za pohranjivanje jezičnoga korpusa na računalo neophodno je da se tekstovi koji čine korpus nalaze u elektroničkome obliku. Samo prikupljanje i izbor tekstova za unos u korpus izlazi izvan okvira ovoga rada, pa će biti samo navedeni mogući postupci za prikupljanje: korištenje već postojećih e-tekstova (preuzimanje (*download*) putem Interneta), suradnja s nakladnicima, skeniranje i utipkavanje.

Osnovni alat koji se koristi u ovoj fazi je uređivač teksta (*text editor*). **Uređivač teksta** je program s pomoću kojega je moguće upisati (*input*), izmijeniti i ispisati (*output*) neki tekst, koji se najčešće nalazi u ASCII (***A**merican **S**tandard **C**ode for **I**nformation **I**nterchange*)[44] obliku. Tekstovi koji sačinjavaju korpus pohranjuju se u bazu podataka (*database*) i/ili kao tekstovne datoteke (*files*) na tvrdom disku računala. Za pretraživanje ili pregledavanje korpusa može se koristi preglednik (*browser*). **Preglednik** je program namijenjen pregledavanju i prikazivanju raznovrsnih oblika datoteka (kao npr. HTML, XML itd.). Preglednik može koristiti datoteke koje se nalaze pohranjene na lokalnom računalu, ali i one koje se nalaze na poslužniku (*server*).

[43] Ide & Brew (2000):1
[44] Više o ASCII-ju u poglavlju 3.1.1.

3. Kodiranje korpusa

Kodiranjem (*encoding*) se smatra način na koji su informacije zapisane na računalu. Kodiranje pismena[45] (c*haracter encoding*) odnosi se na korištenje sustava za zapis pismena. Kodiranje tekstova (*text encoding*) u prvom se redu odnosi na strukturalno (odlomci, rečenice, naslovi itd.) i analitičko (gramatičke kategorije, sintaktičke kategorije itd.) zapisivanje teksta.

Pri osmišljavanju kodiranja korpusa treba voditi računa o složenim prožimanjima kodiranja tekstova i pismena. Tako su iskustva sastavljača korpusa ponekad toliko dalekosežna da utječu na oblikovanje informacijske arhitekture ostalih tipova kodiranja podataka (npr. *World Wide Weba*).[46]

3.1. Kodiranje pismena

Na najnižoj razini računala mogu operirati samo s brojevima. Ona pohranjuju slova i druga pismena pridružujući neki broj svakome od njih. Dakle u načelu, jednom pismenu odgovara jedan broj.

Pri kodiranju korpusa od velike je važnosti da se tekstovi i oznake nalaze u obliku ASCII teksta. Dva su razloga za uporabu ASCII-ja pri kodiranju korpusa:

1. većina programa (*software*) za obradu teksta obrađuje upravo ASCII tekstovne podatke,
2. prenosljivost: ASCII se zbog svoje jednostavnosti može obrađivati i prikazivati na bilo kojoj računalnoj platformi.

Pored ASCII-ja koji je jedan od najjednostavnijih i najrasprostranjenijih skupova pismena, za kodiranje tekstova od iznimne je važnosti i UNICODE, koji je

[45] Termin pisme rabi se onako kako je definirano u László (1993):23
[46] Ide & Brew (2000):1

najsveobuhvatniji do sada uspostavljeni sustav i može kodirati pismena gotovo svih pismovnih sustava jezika svijeta.

3.1.1. ASCII skup pismena

ASCII (***A**merican **S**tandard **C**ode for **I**nformation **I**nterchange*) je jednostavni kôd za prikazivanje pismena engleskoga jezika preko brojčanog sustava. Standardni ASCII skup pismena koristi 7 bitova za kodiranje pismena. Svakom je pismenu pridružen broj između 0 i 127. Pored standardnoga, postoji i prošireni ASCII (*extended ASCII*) koji koristi 8 bita, što otvara mogućnost kodiranja dodatnih 128 pismena.

3.1.2. Unicode

Unicode je standard za zapis pismena kojemu je prvu inačicu u listopadu 1991. godine objavio *Unicode Consortium*[47], osnovan iste godine, a zadužen je za donošenje novih inačica standarda. I prije uspostave Unicode standarda postojalo je nekoliko desetaka skupova pismena, međutim ni jedan od njih nije pokrivao zadovoljavajući broj pismena. Na primjer, za pokrivanje svih pismena jezika Europske unije bilo je potrebno nekoliko različitih sustava za kodiranje. Čak je i pokrivanje svih pismena pojedinih jezika kao što je engleski (slova, interpunkcija, tehnički simboli i sl.) zahtijevalo uporabu nekoliko sustava za kodiranje. Takva praksa često je dovodila do konflikata između sustava, jer dva sustava u nekim slučajevima koriste isti broj za kodiranje različitih pismena. Također, konflikti mogu nastati pri korištenju različitih brojeva za kodiranje istih pismena. Sva računala, pogotovo poslužnici podržavaju mnogo sustava za kodiranje pismena, ali unatoč tome pri razmjeni podataka postoji znatan rizik od konflikta.

Unicode, kao i ASCII, pridružuje jedinstven broj svakome pismenu, ali neovisno o platformi, aplikaciji ili jeziku. Zasnovan je upravo na jednostavnosti i konzistentnosti ASCII-ja, no njegove su mogućnosti bitno veće. Glavni je motiv za

sastavljanje ovoga standarda bio uporaba 16-bitnog kodiranja kako bi se omogućilo kodiranje više od 65.000 pismena, što je dovoljno za pokrivanje pismovnih sustava svih važnijih svjetskih jezika.[48] Međutim, ovisno o tome koristi li 8, 16 ili 32 bita za kodiranje pismena, Unicode definira tri standardna oblika. Sva tri koriste isti zajednički popis početnoga inventara pismena koji mogu jednostavno biti međusobno transformirani bez gubitka podataka. Oblici Unicodea su:[49]

1. UTF-8: namijenjen je u prvom redu HTML (***H**yper **T**ext **M**arkup **L**anguage*) aplikacijama i protokolima koji idu uz njega. Koristi se prednost korespondiranja između ASCII i Unicode brojčanih vrijednosti pismena, pa za transformaciju nije potrebno dodatno nadograđivanje aplikacija,
2. UTF-16: prikladan za uporabu gdje postoji potreba za ravnotežom između učinkovitoga pristupa pismenima i ekonomičnosti za pohranu. Može kodirati više od 1.000.000 pismena.
3. UTF-32: koristi se tamo gdje memorijski prostor ne igra važnu ulogu, tj. nije ga potrebno štedjeti jer je svako pisme kodirano uporabom 32 bita.

Najnovija inačica Unicodea je 3.0.1. i sadrži 49.194 različita kodirana pismena.

3.2. Kodiranje tekstova

Korpus što se sastoji od tekstova kojima nije pridodana nikakva dodatna informacija naziva se **neobilježeni korpus** (*unannotated corpus*). Tekstovi koji čine korpus zapisani su u obliku običnoga ASCII teksta (*plain ASCII text*) i ne sadrže nikakve dodatne oznake.

Nasuprot njemu, **obilježeni korpus** (*annotated corpus*) sadrži različite tipove strukturnih (naslov, odlomak, rečenica itd.) i lingvističkih (gramatičke kategorije, sintaktička struktura itd.) informacija. Bez tako ekspliciranih informacija (tj. oznaka) mnoge osobitosti teksta bilo bi teško odrediti i obraditi računalnim programom. Uporabljivost se korpusa bitno povećava ukoliko je obilježen (*annotated*). Općenito, što je tekst bogatije obilježen, postaje istraživačima korisniji, ali i skuplji za izradu.[50]

---

[47] http://www.unicode.org
[48] Unicode Consortium (2001)
[49] Unicode Consortium (2001)
[50] Lawrer & Dry (1998):108

Ubacivanje dodatnih informacija u tekst zove se obilježavanje (*annotation, mark-up*) ili u širem smislu kodiranje teksta. Korpus se može obilježavati na razini:

1. fonema,
2. riječi-pojavnica,
3. fraza,
4. rečenica,
5. odlomaka,
6. dijelova teksta,
7. teksta.

Na tim se razinama prema istraživačevoj potrebi, ali najčešće prema nekom utvrđenom standardu mogu pridodati različite vrste informacija. Osnovne vrste obilježavanja korpusa pisanoga jezika (postoje i korpusi govorenoga jezika) o kojima će biti riječi u ovome redu bile bi:

1. segmentacija na rečenice,
2. tokenizacija,
3. označavanje vrsta riječi (POS označavanje),
4. lematizacija,
5. parsing,

a iako neće biti problematizirano u ovom radu, tu bi se moglo uvrstiti i:[51]

6. semantičko obilježavanje (obilježavanje značenja riječi, semantičkih kategorija i uspostavljanje veza među dijelovima teksta, npr. agens i pacijens),
7. obilježavanje diskursa (obuhvaća kategorije kao što su *isprike*, *pozdravi*, *oslovljavanje* itd.),
8. problemski-orijentirano obilježavanje (istraživač rabi vlastiti oblik obilježavanja, posebno orijentiran prema vlastitom cilju istraživanja).

[51] McEnery & Wilson (1996a)

3.2.1. Obilježavanje korpusa

Obilježavanje (*annotation, mark-up*) je pridodavanje dodatnih eksplicitnih informacija tekstu za računalnu obradu tamo gdje su one implicitno prisutne osobi koja čita tekst.[52] Pri obilježavanju korpusa oznake se ubacuju iz određenoga skupa oznaka, gdje oznake mogu biti ubačene u elektronički zapis teksta u smislu obilježavanja strukture i drugih osobitosti teksta za koje postoji potreba za obilježavanjem. **Skup** je **oznaka** (*tagset, tag list*) popis svih mogućih oznaka kojima se može obilježavati tekst.

Označavanje (*tagging*) je proces pridruživanja **oznaka** (*tags*) iz skupa ili popisa oznaka dijelovima teksta (pojavnica, rečenica i sl.) koji su delimitirane jezične jedinice.

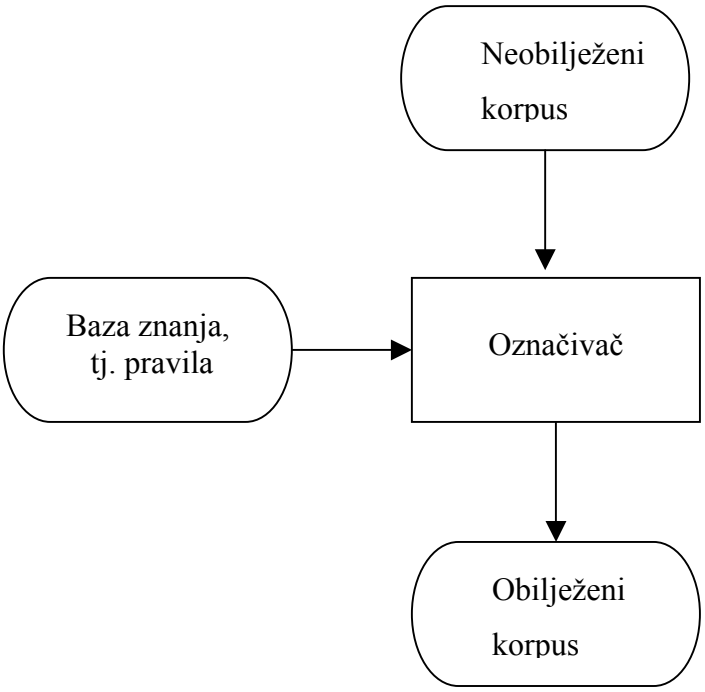
Potrebno je obratiti pozornost na korištenu terminologiju primarne literature korpusne lingvistike kod stranih autora (McEnery & Wilson, Leech, Ide, ali i mnogi drugi). Oni neke slučajeve obilježavanje korpusa, naročito kada je eksplicitno riječ o pridruživanju oznaka pojavnicama nazivaju označavanjem (*tagging*).[53] Ponekad je pri prijevodu na hrvatski nužno pribjeći takvoj terminologiji i žrtvovati dosljednost, jer bi npr. *tagger*, alat koji obavlja obilježavanje bilo krajnje neprecizno prevesti kao "obilježivač". Stoga će se obilježavanje u ovom radu rabiti kao širi pojam, a označavanje će biti samo jedna vrsta obilježavanja. Termin *označavanje* koristit će se kada je eksplicitno riječ o automatskom strojnom dodavanju oznaka.

Obilježavanje se korpusa može obavljati na tri načina:

- ručno: iznimno dugotrajan i skup proces. Velika je mogućnost pogrešaka i nekonzistentnosti u obilježavanju. Današnje je korpuse zbog veličine gotovo nemoguće ručno obilježiti,
- automatski: potreban je alat, tj. program koji treba biti sposoban "prepoznati" na koja se mjesta u tekstu trebaju ubaciti koje oznake,
- poluautomatski: kombinacijom prethodne dvije metode, gdje je rezultat strojne obrade potrebno naknadno ručno uređivati (*post-editing*).[54]

[52] Lawrer & Dry (1998):107: Markup makes explicit for computer processing those features which are implicit for the human reader of a text.
[53] McEnery & Wilson (1996):36
[54] CCL (2000)

Danas se automatsko obilježavanje nastoji rabiti u najvećoj mjeri, tj. koliko je to moguće i kada je to moguće. Alat koji je sposoban automatskim putem ubacivati oznake u tekst naziva se **označivač** (*tagger*). Shematski se automatsko obilježavanje može prikazati:



Slika 1: Shema automatskog obilježavanja[55]

Leech definira sedam pravila kojih bi se trebalo držati pri obilježavanju korpusa: [56]

1. **Trebalo bi biti moguće izbaciti sve oznake iz obilježenoga korpusa kako bi ponovo došli do neobilježenoga korpusa**. U današnje vrijeme to je vrlo jednostavan postupak. Npr. kod "Claire_NP1 collects_VVZ shoes_NN2" izbacivanjem svih pismena iza podcrte (*underscore*) do bjeline lako se dolazi do "Claire collects shoes".
2. **Trebalo bi biti moguće izvući oznake iz teksta** (ovo pravilo je inverzno prvom pravilu). Prva dva pravila trebala bi omogućiti maksimalnu fleksibilnost krajnjem korisniku.
3. **Shema obilježavanja trebala bi biti poznata krajnjem korisniku**. Obilježeni korpus trebao bi imati priručnik u kojemu se nalaze iscrpni detalji o shemi i principima obilježavanja kojima su se vodili autori korpusa. To bi trebalo omogućiti krajnjem korisniku potpuno razumijevanje svih pojedinih oznaka, otklanjanje mogućeg nagađanja i višeznačnosti (*ambiguity*).

[55] cf. Lager (1995):189
[56] McEnery & Wilson (1996):25

4. **Trebalo bi biti jasno kako i tko obilježava korpus**. Korpus može ručno obilježavati jedan ili više ljudi. Alternativno korpus može biti obilježen s pomoću računalnog programa, a rezultat obrade ponovo mogu korigirati jedan ili više ljudi.
5. **Krajnji korisnik treba biti upućen u to da obilježavanje korpusa nije nepogrešivo, nego da je takav korpus potencijalno koristan alat**. Svako obilježavanje korpusa po definiciji je *interpretacija*, bilo strukture bilo sadržaja teksta.
6. **Shema obilježavanja treba biti zasnovana na koliko je to moguće teorijski neovisnim i široko prihvaćenim načelima**. Trebalo bi se povoditi za bazičnim teorijama, a izbjegavati uske i specifične lingvističke teorije.
7. **Nijedna shema obilježavanja nema pravo *a priori* biti prihvaćena kao standard**. Standardi nastaju kroz općeprihvaćene konsenzuse.

3.2.2. Segmentacija na rečenice

Segmentacija teksta na rečenice (*sentence segmentation, sentence boundary disambiguation)* u mnogim je slučajevima prvi korak za brojna područja strojne obrade jezika kao što je npr. označavanje vrsta riječi *(POS tagging)*, sintaktički parsing, sravnjivanje rečenica paralelnoga korpusa ili pak za određivanje čitljivosti teksta.

Segmentacija se rečenice 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, 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 razrješavanje višeznačnosti (*disambiguation*) većina alata rabi posebno namijenjene gramatike regularnih izraza (*special-purpose regular expression grammars*) i pravila iznimaka (*exception rules*).[57] Takvi su algoritamski pristupi najčešće ograničeni na specifičan žanr teksta, i nije ih jednostavno adaptirati na nove

[57] Palmer (1994):2

tipove teksta. Stoga se često rabe alati koji koriste manji, ručno segmentirani korpus za uvježbavanje (*training corpus*), gdje se na "pravom" korpusu mogu postići rezultati i do 98,5 % točnosti (*SATZ*, autor: *David D. Palmer*).[58]

Za hrvatski je jezik testni model alata za segmentaciju na tekstu veličine 2000 rečenica imao točnost od 99,5 %.[59] Točnost sustava znatno je povećana brojnim naknadnim ubacivanjem dodatnih pravila i iznimaka. Tako je npr. u svrhu povećanja točnosti segmentacije na rečenice napravljen opsežan popis hrvatskih kratica, vlastitih imena i prezimena.

3.2.3. Tokenizacija (*tokenisation*)

Ukoliko je riječ definirana kao sve ono što se nalazi između dvije bjeline u tekstu, moglo bi se reći da joj u korpusnoj terminologiji odgovara pojavnica. **Pojavnica** (*token*) bi se mogla definirati kao sve ono što se nalazi između dva pismena koja služe kao graničnici, a ona pismena koja se nalaze između graničnika moraju biti iz abecede kojoj su pridodane znamenke i crtica.[60] U literaturi Brown korpusa pojavnica je definirana kao:

An "individual word" (token) (…) can be simply defined as continuous string of letters, numerals, punctuation marks, and other symbols (i.e. graphemes), uninterrupted by space (…)[61]

Prema tome u nizu: *stol, stola, stol, stola, stol* nalazi se pet pojavnica. Dakle, pojavnica je svako pojedinačno pojavljivanje "riječi" u korpusu, pa bi se pod pojmom milijunski korpus podrazumijevao korpus od milijun pojavnica.

Nasuprot pojavnici, **različnica** (*type*) je jedinstveni oblik pojavnice iz korpusa.[62] Ili u literaturi Brown korpusa različnica je definirana:

A "distinct word" (type) can also be simply defined as a set of identical individual words, as defined above.[63]

Dakle, u gornjem se nizu nalaze dvije različnice: *stol* i *stola*.

---

[58] Palmer (1994):1
[59] Boras (1998):155
[60] Tadić (1991):173; Definicija pojavnice kod jednomiljiunskoga korpusa (Mogušev korpus)
[61] Kučera & Francis (1967b):xx: "Pojedinačna riječ" (pojavnica) (…) može se jednostavno definirati kao niz slova, brojeva, interpunkcije i drugih simbola (tj. grafema) neprekinutih razmakom (...) (prijevod moj)
[62] Glossary (1999)
[63] Kučera & Francis (1967b):xxi: "Različita riječ" (različnica) može se jednostavno definirati kao skup jednakih pojedinačnih riječi, kako su definirane gore. (prijevod moj)

Tokenizaciju (*tokenisation*) bi se moglo definirati kao dovođenje korpusa u stanje u kojem su sve riječi-pojavnice identificirane i eksplicitno obilježene. Potrebno je naglasiti da proces tokenizacije može imati različito značenje u različitim pristupima.

U najjednostavnijem pristupu, pojavnice je lako identificirati, jer pojavnica je sve ono što se nalazi između dva pismena za obilježavanje razmaka, što najčešće odgovara dvjema bjelinama. No to je pravilo vrlo manjkavo: ponekad se između dvije uzastopne bjeline ne nalazi ništa, crta se u rečenici također nalazi između dvije bjeline, a još veći je problem s znacima kao što su +, -. =, % i sl. Primjer teksta na hrvatskome jeziku tokeniziranog ovim pristupom može se naći u dodatku B.

Međutim, tokenizacija je u drugim slučajevima mnogo kompleksnija, jer se pojavnicama mogu smatrati i jedinice koje se sastoje od više riječi (*multy-word units (MWU)*). Npr. datum *20. svibanj* ili *20. 5.* mogao bi biti obrađivan kao jedna jedinica, pa ga u ranijem smislu određenja pojavnice nije moguće tokenizirati.[64] U ovakvom pristupu tokenizacija bi uključivala prepoznavanje **imenovanih entiteta** (*named entity recognition*). Prepoznavanje imenovanih entiteta uključuje obradu teksta pri kojoj se identificiraju izrazi koji su nazivi za npr. ljude, organizacije, datume i sl. Imenovani entiteti (*named entities*) često nose veliku količinu obavijesti, jer su povezani s izvantekstnim svijetom na koji se referiraju.[65] Identifikacija takvih izraza osim pri tokenizaciji ima iznimnu uporabnu vrijednost i kod automatskoga prosljeđivanja dokumenata, pronalaženja dokumenata (*document retrieval*), izrade knjiških indeksa i sl. Prepoznavanje je imenovanih entiteta najčešće složena radnja koji uključuje dva koraka:

1. identifikaciju imena,
2. kategorizaciju imena.

Ljudsko prepoznavanje imenovanih entiteta doseže točnost između 98-99%, dok kod automatskog prepoznavanja točnost doseže do 94% (npr. TTT, autori *Grover, Matheson, Mikheev*).[66] Imenovani entiteti mogu biti vrlo učestali u tekstu, što se može vidjeti iz sljedećeg primjera jednog članka iz *Večernjeg lista*:[67]

[64] Grover & Matheson & Mikheev (2000)
[65] Tadić (2000b)
[66] Grover & Matheson & Mikheev (2000)

(pr. 1)

```
<XML>
<BODY>
<DIV0 type="MAIN">
<HEAD type="NA">Nagrada zagrebačkim gitaristima</HEAD>
<P><ENAMEX TYPE="ORGANIZATION">Zagrebački gitaristički
kvartet</ENAMEX> osvojio je prvu nagradu na <ENAMEX
TYPE="ORGANIZATION">Međunarodnome gitarističkom natjecanju Simone
Salmaso</ENAMEX> u <ENAMEX TYPE="LOCATION">Viareggiu</ENAMEX> u
konkurenciji 14 komornih sastava (u kategoriji D). Prvo mjesto je kao
solist osvojio i član toga renomiranoga zagrebačkog sastava <ENAMEX
TYPE="PERSON">Darko Pelužan</ENAMEX> u konkurenciji 30 gitarista (u
kategoriji C). Članovi <ENAMEX TYPE="ORGANIZATION">Zagrebačkoga
gitarističkog kvarteta</ENAMEX> (koji je 1990. osnovao profesor
<ENAMEX TYPE="PERSON">Ante Čagalj</ENAMEX>, pretežno od studenata
gitare) sada su još <ENAMEX TYPE="PERSON">Mihaela Pažulinec</ENAMEX>,
<ENAMEX TYPE="PERSON">Krunoslav Pehar</ENAMEX> i <ENAMEX
TYPE="PERSON">Melita Ivković</ENAMEX>. To nije prvi put da <ENAMEX
TYPE="ORGANIZATION">Zagrebački gitaristički kvartet</ENAMEX> osvaja
prvu nagradu na nekome međunarodnom natjecanju u <ENAMEX
TYPE="LOCATION">Italiji</ENAMEX>: pobijedio je i prije dvije godine u
<ENAMEX TYPE="LOCATION">Tarantu</ENAMEX> na 6. međunarodnom
natjecanju <ENAMEX TYPE="ORGANIZATION">Trofeo Kawai</ENAMEX>.</P>
<BYLINE>(<ENAMEX TYPE="ORGANIZATION">Večernji list</ENAMEX>)</BYLINE>
</DIV0>
</BODY>
</XML>
```

3.2.4. Označavanje vrsta riječi *(Part-of-speech* (POS) *tagging*)

***Part-of-speech* (POS) označavanje** je pridruživanje gramatičke kategorije svakoj pojavnici u tekstu (ponekad se naziva gramatičko označavanje ili morfosintaktičko obilježavanje).[68] POS označavanje spada u osnovne vrsta lingvističkog označavanja, i služi kao osnova za više razine analize teksta kao što je sintaktički parsing[69] ili obilježavanje semantičkih polja. Pored toga, POS oznake prvi su korak u razrješavanju istopisnica (homografa), 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 konteksta u kojima se

[67] Preneseno iz Tadić (2000b)
[68] McEnery & Wilson (1996):36
[69] Više o parsingu u poglavlju 3.2.6.

nalaze. POS označivači smatraju se najpouzdanijim i najkorisnijim računalnolingvističkim alatom, a prema načinu rada dijele se na:[70]

1. vjerojatnosne (*probabilisic*) označivače: zasnivaju se na vjerojatnosnom računu i statistici,
2. označivače zasnovane na pravilima (*rule-based*): zasnivaju se na klasičnim, ručno pisanim pravilima.

Većina POS označivača danas koristi prvi pristup, a najčešći vjerojatnosni model koji se koristi kod ove vrste označavanja je **Skriveni Markovljev Model**, **SMM** (*Hidden Markov Model, HMM*).

Jedna od podjela POS označivača zasniva se i na stupnju autonomije označavanja u odnosu na uporabu prethodno obilježena korpusa u uvježbavanju označivača na:[71]

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.

Označivač za ulaznu varijablu uzima pojavnice iz korpusa, te ih uspoređuje s riječima iz leksikona (*lexicon*).[72] **Leksikon** (ili elektronički rječnik) u korpusnoj se lingvistici koristi kao sinonim za rječničku bazu podataka što podrazumijeva strojno-čitljiv oblik.[73]

> Lexical knowledge – knowledge about individual words in the language – is essential for all types of natural language processing.[74]

Leksikon potencijalno može sadržati širok raspon informacija o pojedinoj riječi, ovisno o strukturi i vrsti zadatka obrade kojoj je namijenjen. Osnovni leksikon sadrži informacije o morfologiji, bilo kao popis svih oblika riječi, bilo u obliku koji

[70] Van Guilder (1995)
[71] Van Guilder (1995)
[72] McEnery & Wilson (1996):120
[73] CCL (2000)
[74] Grishman & Calzolari (1996): poglavlje 12.4.1: Leksičko znanje – znanje o pojedinačnim riječima u jeziku – osnova je za sve vrste strojne obrade jezika. (prijevod moj)

omogućuje generiranje svih oblika riječi, ili sadrži oboje od navedenoga.[75] Leksikoni koji se koriste pri označavanju vrsta riječi obično su pisani u obliku:

(pr. 2)

```
<riječ>
<POS 1, POS 2, ..., POS n>
</riječ>
```

Iako su se ranije sastavljali ručno, POS obilježeni korpusi nezamjenjiv su izvor za automatsko sastavljanje pouzdanih i sveobuhvatnih leksikona. Zapravo, taj postupak može biti obosmjeran: sastavljanje leksikona iz POS obilježenoga korpusa, ili POS obilježavanje korpusa iz leksikona. Što je obilježeni korpus veći, veća je i 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 sadrže stotine tisuća natuknica iz razloga što je broj oblika svih riječi u prirodnom jeziku velik, osobito u flektivnih jezika kakav je hrvatski. Idealan bi leksikon trebao sadržati sve oblike.

Jedan od mogućih postupaka pri POS označavanju bio bi slijedeći: ukoliko je pojavnica pronađena u leksikonu pridružuje joj se čitav skup POS oznaka (1, 2,..., n). Međutim, ako pojavnica nije pronađena u leksikonu, ulogu preuzima morfološki analizator. Morfološka analiza ne obavlja se u klasičnom smislu već bi se prije moglo reći da se radi o "prepoznavanju" pojavnice na osnovi njenih prefikasa ili sufikasa. Npr. ukoliko riječ *tulipan* postoji u leksikonu, a pojavnica iz korpusa *tulipani* nije pronađena, morfološki će analizator po nastavku za množinu *–i* "prepoznati" o kojoj se riječi radi. Ukoliko pojavnica nije prepoznata ni u interakciji leksikona i morfološkog analizatora, POS se označivač oslanja na vjerojatnost u smislu: "Koji je najvjerojatniji POS promatrane pojavnice u danom ko-tekstu?".[76] Budući da je u ovoj fazi pojavnicama pridružen skup mogućih POS oznaka, a potrebna je samo ona odgovarajuća, pri otklanjanju se višeznačnosti potrebno osloniti na vjerojatnosni ili pristup zasnovan na pravilima. U posljednjem se koraku pojavnici pridodaje POS oznaka.

[75] Grishman & Calzolari (1996): poglavlje 12.4.1

[76] Lager (1995):34: Potrebno je spomenuti da u terminologiji korpusne lingvistike postoji distinkcija između **ko-teksta**, pod kojim se misli na lijevu i desnu tekstnu okolinu pojavnice i **kon-teksta** (ponekad se naziva i situacijski kontekst) koji se sastoji od jezičnoga ko-teksta i njegova odnosa prema izvanjezičnim situacijama.

POS označavanje može uključiti dvije razine:

1. razina: uključuje prepoznavanje i označavanje *vrsta riječi* (*POS*),
2. razina: uz označavanje *vrste riječi*, određuju se i *gramatičke kategorije*.

Druga se razina označavanja pojavnica naziva i **morfosinataktički opis** (*morphosyntactic description, MSD*). Pri svakoj se razini rabe različiti skupovi oznaka, gdje je skup oznaka na drugoj razini znatno veći. Leksikoni koji se koriste pri drugom koraku POS označavanja obično su pisani u obliku:[77]

(pr. 3)

Pojavnica	Lema	MSD
diskreditirajmo	*diskreditirati*	*Vmmp1p*
...		

gdje je pojavnica navedena u obliku u kojem se pojavljuje u tekstu, npr. pojavnica *diskreditirajmo*. Lema je polazni oblik riječi (definirana u slijedećem poglavlju) a u ovom bi slučaju bila *diskreditirati*. MSD je zapisan kao linearno kodirani niz u skladu s *Multext-East*[78] preporukom za MSD, npr. *Vmmp1p* bio bi skraćeni zapis od: PoS:**V**erb, Type:**m**ain, VForm:i**m**perative, Tense:**p**resent, Person:first (**1**), Number:**p**lural.

Prvi je POS označivač nazvan *TAGGIT* napravljen na *Brown* Sveučilištu na tekstovima *Brown* korpusa. Zasniva se na pravilima i u prvoj je inačici imao točnost od 77 %. U ranim osamdesetima, vjerojatnosni označivač *CLAWS* (***C**onstituent **L**ikelihood **A**utomatic **W**ord-tagging **S**ystem*)[79] napravljen na Sveučilištu u Lancasteru ima točnost od 95 %. Od tada se neprekidno razvija, te danas doseže točnost od 96-97 %, ovisno o tipu teksta. U različitim se fazama razvoja nadograđivao i mijenjao skup oznaka. Prvi je skup oznaka nazvan *CLAWS1* imao 132 osnovne oznake, a trenutni skup oznaka nosi naziv *CLAWS7*. U načelu, broj osnovnih oznaka ovisi o planiranom "bogatstvu", ali i preciznosti označavanja. Što je veći broj oznaka, tekst je bogatije obilježen, ali je veća i mogućnost pogreške pri označavanju. Stoga je za kodiranje BNC-a korišten *CLAWS5* sa "samo" 58 oznaka.[80] U kasnijim inačicama *CLAWS* označivač (npr. u inačicama gdje se koristi *CLAWS4*) kombinira

[77] Erjavec (1998):189
[78] Više o *Multext-East* preporuci na adresi: http://nl.ijs.si/ME/ i http://nl.ijs.si/ME/V2/.
[79] CLAWS (2000)
[80] Leech, Garside, Bryant (1994): 624

vjerojatnosni pristup s pristupom zasnovanim na pravilima, pa je arhitektura i razine djelovanja sustava u gruboj podjeli sljedeća:[81]

1. segmentacija teksta na pojavnice i rečenice,
2. početno (bez uvažavanja konteksta) POS označavanje koristeći leksikon, popis završetaka riječi i pravila za označavanje nepoznatih jedinica,
3. POS označavanje zasnovano na pravilima,
4. vjerojatnosno razrješavanje višeznačnosti koristeći HMM. Nakon toga se ponovo prolazi kroz treći korak,
5. ispis rezultata.

Iako za hrvatski jezik postoje dobri rezultati u strojnom generiranju morfoloških oblika riječi[82], još uvijek ne postoje zadovoljavajući rezultati u POS označavanju. Za hrvatski je jezik probno istraživanje u svrhu izrade automatskoga vjerojatnosnog POS označivača točnost dosegla do 91 %.[83]

3.2.5. Lematizacija korpusa

Lematizacija (*lemmatisation*) je svođenje pojavnica iz korpusa na njihove natukničke oblike, tj. svođenje različitih pojavnica (članova iste paradigme) na zajedničku lemu.[84] 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.[85] Lematizacija se na isti način primjenjuje na morfološki supletivne oblike pa bi npr. *jesam*, *bijah* ili *bila* bili svedeni na leksem *biti*. Lema predstavlja sve oblike određene riječi. Kako se u postupku strojnoga prepoznavanje lema redovito moraju prepoznati i morfosintaktički opisi pojavnica, lematizacija se zapravo obavlja u drugoj fazi POS označavanja. Ipak, lematizacija je nužna kao zaseban postupak jer se pri MSD obilježavanju u pravilu određuje gramatički oblik pojavnice, a ne sama lema.[86] Lematizacija je važan postupak u istraživanjima korpusa osobito za jezike koji imaju bogatu morfologiju. Pri analizi vokabulara, ili u leksikografiji, na primjer, omogućuje

[81] Leech, Garside, Bryant (1994): 622
[82] v. Tadić (1994)
[83] Žubrinić (1995):69: označivač pod nazivom SOLAH napravljen 1995. godine u okviru projekta MZT-a RH "Modeli znanja i komunikacijski obrasci"
[84] McEnery & Wilson (1996):42
[85] Glossary (1999)
[86] CES (1996)

istraživaču ekstrahiranje i istraživanje svih inačica određene leme bez potrebe za pretraživanjem svakog pojedinačnog oblika. Također omogućuje uvid u čestotne i distribucijske informacije za pojavnice određene leme. Iako danas postoji oveći broj programa koji obavljaju lematizaciju, većina korpusa još nije lematizirana. Postupak automatske lematizacije, kao i POS označavanja prilično je spor i zahtjeva naknadnu ljudsku intervenciju. Alat koji obavlja automatsku lematizaciju zove se **program za lematizaciju** (*lemmatizer*). Najveću prepreku postizanju veće točnosti automatske lematizacije također predstavljaju istopisnice.

Iako za engleski (kao i za druge "veće" jezike) postoji mnoštvo takvih alata, još ni jedan ne postiže apsolutnu točnost. Za hrvatski jezika probna inačica programa za lematizaciju doseže točnost do 95 %.[87] Korpus hrvatskoga jezika nad kojim je obavljen dio lematizacije poluautomatskim putem jest *Mogušev korpus*.[88]

3.2.6. Parsing

Nakon što se identificiraju osnovne morfosintaktičke kategorije u tekstu, moguće ih je dovesti u međusobne sintaktičke odnose višega stupnja. Taj se postupak naziva parsing.[89] **Parsing** je postupak odvajanja rečeničnih dijelova i opisivanje relacija između njih.[90] Dakle, parsingom se određuje sintaktička struktura rečenice. Međutim, u širem se smislu parsing ne mora odnositi na rečenice, već se može reći da je parsing postupak koji za ulazne veličine uzima odsječak teksta i neku gramatiku, te kao izlaz ima rezultat koji zadovoljava cilj parsinga, tj. pridruživanje kategorija odsječcima.[91] Alat koji obavlja postupak parsinga naziva se **parser**.

Korpus koji sadrži informaciju o rečeničnoj strukturi često se naziva *treebank*, [92] a najpoznatiji je *Penn Treebank*.[93] Potpunim se parsingom (*full parsing*) nastoji napraviti što detaljnija analiza rečenične strukture. Uz odnose sastavnica rečenične strukture obilježeni su i morfosintaktički opisi pojavnica. Kosturni parsing (*skeleton*

[87] Žubrinić (1995):69: alat pod nazivom SOLAH napravljen 1995. godine u okviru projekta MZT-a RH "Modeli znanja i komunikacijski obrasci"
[88] Više o lematizaciji *Moguševa korpusa* u poglavlju 5.2.
[89] McEnery & Wilson (1996):43
[90] Henderson (1999):21
[91] Lager (1995):7
[92] McEnery & Wilson (1996):179
[93] http://www.cis.upenn.edu/~treebank/

parsing) je manje detaljan pristup rečeničnoj analizi. Za razliku od potpunog parsinga, nije obilježena unutrašnja struktura određenih sastavnica, pa su imenske fraze obilježene samo s oznakom *N*, bez osobina kao što je npr. pluralnost.[94]

[94] McEnery & Wilson (1996a)

3.3. Jezici za obilježavanje elektroničkih tekstova

Računalni bi se jezici prema namjeni mogli svrstati u dvije osnovne skupine:

1. *programirni jezici*: oni kojima se može programirati (npr. C++, VBasic, Perl itd.)
2. *jezici za obilježavanje podataka*: jezici sa skupom pravila koja definiraju ograničenja pri obilježavanju podataka (npr. HTML, SGML, XML itd.)

Jezicima za obilježavanje ne može se programirati već služe isključivo obilježavanju podataka. Međutim, sadržaj obilježenih podataka moguće je obrađivati uporabom proceduralnih jezika. Korpusnu lingvistiku u prvome redu zanimaju jezici za obilježavanje.

Već je napomenuto da se neobilježeni korpus pohranjuje u obliku običnoga teksta. Pri obilježavanju korpusa nužno je odabrati jezik za obilježavanje (*markup language*) kojim će se tekst obilježiti. Bitno je razlikovati metajezik od jezika za obilježavanje, jer se u praksi ta dva termina često preklapaju, pa i pogrešno rabe.

Pod **metajezikom** se podrazumijeva sustav koji opisuje ili definira druge jezike.[95] Ili kako je definirano u CES-u, **metajezik za obilježavanje** (*markup metalanguage*) je skup pravila koja formalno opisuju oblik sintaktičkih pravila sheme obilježavanja.[96] Osobine metajezika (npr. SGML, XML) su:

1. neograničen broj oznaka,
2. neograničen broj DTD-ova,
3. obilježavanje strukture podataka,
4. iz njega se izvode drugi jezici.

Jezik za obilježavanje (*markup language*) skup je konvencija za obilježavanje teksta, gdje je specificirano koje su oznake dopuštene, koje su obvezne, kako se oznake razlikuju od teksta, i što pojedine oznake znače.[97] Osobine jezika za obilježavanje (npr. HTML, WML, CML) su:

1. ograničen, tj. konačan broj oznaka,
2. jedinstveni DTD,
3. obilježavanje izgleda podataka,
4. iz njega se ne izvode drugi jezici.

[95] Spencer (1999):24
[96] CES (1996)

Najpoznatiji je jezik za obilježavanje HTML (***H**yper **T**ext **M**arkup **L**anguage*). HTML se može smatrati jednom instancom metajezika, u ovom slučaju SGML-a. Pri obilježavanju korpusa jezik za obilježavanje definira se i opisuje iz metajezika.

Jezici za obilježavanje izgrađeni su na pretpostavci se tekst može razbiti na manje jedinice (odlomke, imena, poglavlja, naslove itd.) koje se mogu međusobno ugnježđivati. Princip je deskriptivan na način da onaj koji kodira sam odlučuje što je tekstni objekt koji će se obilježiti, a ne što bi računalo trebalo raditi s tim objektom. To znači da različite aplikacije mogu obrađivati isti tekst, te da isti tekst može biti obrađivan na više različitih načina.

Pri obilježavanju se tako određenim objektima pridodaju oznake za početak i kraj, a takav se objekt (u biti niz pismena) naziva **element**. Jednostavan element može izgledati ovako:

(pr. 4)

```
<riječ>Ivan</riječ>
```

Oznaka je kôd pridružen nekoj jedinici teksta i označuje neku osobinu ili skup osobina koje pripadaju toj jedinici. Imena se oznaka nalaze unutar kutnih zagrada "<" i ">" i računalni ih program po tim pismenima razlikuje od teksta. **Sadržaj** se **elementa** (`Ivan`) nalazi između otvorne oznake `<riječ>` i zatvorne oznake `</riječ>`.

Uz početnu oznaku elemenata može stajati **atribut** (ili nekoliko atributa):

(pr. 5)

```
<riječ vrsta="imenica">Ivan</riječ>
```

Ime atributa `vrsta` razdvojeno je od svoje vrijednosti `imenica` znakom jednakosti. Prva je i najčešća uporaba atributa pridodavanje dodatne obavijesti o nekom elementu. Druga je uporaba atributa jednoznačno određivanje bilo kojeg elementa s nakanom da bude naknadno izdvojen.[98] Na primjer:

(pr. 6)

```
<riječ id="ivan05">Ivan</riječ>
```

Vrijednost atributa `ivan05` jednoznačno određuje navedeni element u tekstu. Jednoznačno određeni elementi mogu se na jednostavan način međusobno povezivati (*linking*). Na primjer u tekstu,

---

[97] Erjavec (1997):3
[98] Burnard (1991):7

(pr. 7)

```
Ivan je kuhar.
Marija, njegova supruga, je slikarica.
```

uporabom atributa mogla bi se uspostaviti poveznica (*link*) na slijedeći način:

(pr. 8)

```
<riječ id="ivan05">Ivan</riječ> je kuhar.
<riječ>Marija</riječ>,<relation target="ivan05">
njegova supruga</relation>, je slikarica.
```

Termin *entitet* u kontekstu jezika za obilježavanje ima drugo značenje nego u standardnoj ili informatičkoj uporabi. **Entitet** (*entity*) je bilo koji proizvoljno određen dio teksta kojemu se pridružuje neki proizvoljni naziv. Pri definiranju entiteta združuje se neki sadržaj s nekim nazivom. Taj sadržaj može biti u rasponu od jednog pismena do cijeloga dokumenta. Entiteti imaju iznimnu uporabnu vrijednost u mnogim situacijama, ponajprije kod kodiranja nestandardnih pismena ili simbola koji su specifični za određeno okružje ili aplikaciju. Takvim se pismenima i simbolima pridružuju entiteti koji ih zamjenjuju, a računalo ih je sposobno "razumjeti". Nadalje, uporaba entiteta iznimno je korisna kada se u tekstu više puta nalaze duži dijelovi ili fraze koje se pojavljuju višekratno. Oni se mogu zamijeniti kratkim nazivom entiteta na istom mjestu u tekstu. Na taj se način štede računalni resursi, ali i osigurava konzistentnost teksta jer je ponavljani sadržaj pohranjen samo jednom. Treća je uporaba entiteta pri identificiranju objekata koji se ne mogu izravno prikazati u tekstu (npr. netekstovni objekti kao što su grafikoni ili formule). **Pozivanjem na entitete** (*entity reference*) ubacuje se naziv entiteta na ono mjesto u tekstu gdje se treba nalaziti sadržaj entiteta koji se zamjenjuje. Naziv entiteta razlikuje se od ostalih oznaka po tome što uvijek započinje s "&", a završava s ";".

Budući da obilježeni dokumenti nisu samo linearno nizanje pismena već posjeduju čvrsto definiranu strukturu, treba postojati gramatika koja je propisuje. Dok obilježavanje odlomaka (paragrafa) u tekstu kao npr. `<p>`, `<para>` ili `<paragraf>` za čovjeka nije prepreka pri interpretaciji, za računalnu obradu to bi mogla biti nepremostiva prepreka. Imena oznaka ili atributa kojima se obilježava tekst moraju biti precizno i konzistentno upotrebljavana. Svi elementi, atributi i ostale

karakteristike obilježavanja koje se mogu pojaviti u dokumentu, kao i redoslijed njihovog pojavljivanja definirani su jedinstvenim propisom koji se naziva **DTD** (***D**ocument **T**ype **D**efinition*). On propisuje što se smije, ali i što se ne smije nalaziti na određenom mjestu i u određenom dokumentu. Dakle, DTD je neka vrsta gramatike, skup sintaktičkih pravila za elemente u dokumentu. Budući da su XML i SGML metajezici, DTD kreira sam korisnik i teoretski ih može biti beskonačno mnogo. Kod jezika za obilježavanje postoji samo jedan DTD, kojeg je već unaprijed definirao donosilac standarda. Parser provjerava zadovoljavaju li podaci iz dokumenta pravila koja se nalaze u DTD-u.

Sama se struktura DTD-a u osnovi može podijeliti na:

- deklaracije *elemenata*,
- deklaracije *atributa*,
- specifikacije *entiteta*.

DTD se može nalaziti unutar samoga dokumenta ili u vanjskoj datoteci. Ukoliko se DTD nalazi unutar dokumenta smješta se na sam početak, a ako je u vanjskoj datoteci onda je iz obilježenoga dokumenta potrebno referirati na datoteku u kojoj se DTD nalazi. Referiranje na vanjsku DTD datoteku korisno je ako se ista gramatika koristi za više dokumenata, što je čest slučaj u obilježavanju tekstova. Iako neki parseri (*non-validating parsers*) mogu obrađivati dokumente bez pripadajućeg DTD-a, razlozi za uporabu DTD-a su višestruki:

- provjera slaganja dokumenta s navedenom gramatikom, tj. DTD-om,
- osiguravanje konzistentnosti oznaka i atributa,
- definiranje jedne gramatike (DTD-a) za više korisnika ili grupa,
- provjera usklađenosti s DTD-om dokumenata koji dolaze iz drugih izvora, itd.

Parser u prvom redu koristi DTD da bi provjerio valjanost kodiranja u tekstu. Obilježeni dokument (tekst) je **valjan** (*valid*), ako je u skladu s pripadajućim DTD-om. Obilježeni dokument (tekst) je **ispravan** (*well formed*) ako ga parser korektno obradi, tj. ne javlja pogrešku.

Shema obilježavanja (*markup scheme*) dokumenta sastoji se od tri dijela:[99]

1. skupa pismena (*character set*),
2. sintakse, tj. pravila koja definiraju što čini ispravno (*well-formed*) obilježen tekst,
3. semantike, tj. pravila koja definiraju što čini valjano (*valid*) obilježen tekst.

[99] CES (1996)

Sintaktička pravila definiraju propisane oznake, propisani sadržaj i način njihovoga nizanja. Sintaktički ispravan dokument nije nužno semantički valjan. Npr. niz:

(pr. 9)

```
<word>Novi Zagreb</word>
```

može biti sintaktički ispravan u nekoj shemi obilježavanja, ali "Novi Zagreb" ne mora biti riječ u interpretaciji neke druge sheme obilježavanja. Semantička pravila definiraju koji će sintaktički ispravni tekstovi biti valjani u nekoj shemi obilježavanja.[100] Dva glavna razloga za uporabu sheme obilježavanja su:

1. lokalno obrađivanje (*local processing*), kamo pripadaju različite lingvističke analize, kolokacije[101], oblikovanje i pretraživanje podataka i sl,
2. razmjena podataka (*data interchange*) između pojedinaca ili grupa korisnika.

Više riječi o ispravnosti, valjanosti, propisanim oznakama i sadržaju bit će u poglavlju o XML-u.

Dva metajezika za obilježavanje koja se u praksi najčešće koriste jesu SGML i XML. Prvi je znatno stariji, i može se smatrati pretkom većine suvremenih jezika za obilježavanje (HTML, WML, XML itd.). Budući da je TEI[102] preporučio SGML kao jezik za obilježavanje, većina je korpusa danas kodirana SGML-om. On je izabran kao optimalan metajezik koji je postojao u vrijeme davanja preporuke. Međutim, SGML ima i nedostatke koji su velikim dijelom otklonjeni pojavom XML-a. Glavni nedostaci SGML-a su kompleksnost samoga standarda, ali i cijena programa koji ga koriste. XML je podskup SGML-a, te predstavlja njegovo pojednostavljenje i osuvremenjenje. Iz tog je razloga konverzija jednog oblika u drugi relativno jednostavna. Većina se korpusa posljednjih godina kodira XML-om, uz zamjetan trend konverzije SGML-om kodiranih korpusa u XML oblik. Razvojem digitalnih telekomunikacija i Interneta, gdje zapravo nalazi svoje pravo mjesto jer omogućuje lakšu međuplatformsku razmjenjivost bilo kakvih podataka, XML postaje primarni izbor pri obilježavanju jezičnih korpusa. Stoga će u ovom radu biti posvećeno znatno više pozornosti XML-u.

[100] CES (1996)
[101] Više o kolokacijama u poglavlju 4.2.4.
[102] v. poglavlje 3.4.1.

3.3.1. SGML *(Standard Generalized Markup Language)*

SGML je međunarodni standard za definiranje aplikacijski i platformski neovisnih metoda za zapis tekstova u elektroničkome obliku.[103] Prethodnik SGML-a je GML (*Generalized Markup Language*) razvijen u IBM-ovom istraživačkom centru 1967. godine. GML je zamišljen kao sredstvo za prikazivanje, oblikovanje i razmjenu dokumenata. Prvi radni nacrt (*working draft*) SGML-a nastao je 1980, a kao međunarodni standard prihvaćen je 1986. godine (ISO 8879:1986).

Prema standardu svaki SGML dokument mora sadržati:[104]

1. podatke koji se obilježavaju, tj. tekst,
2. oznake (*markup*),
3. DTD, ili se referirati na određeni DTD.

SGML ima sve navedene osobine metajezika za obilježavanje i smatra se jednim od najsloženijih za uporabu. Na jednostavnom primjeru **deklaracija elementa** u SGML DTD-u izgledala bi ovako:

(pr. 10)

```
<!ELEMENT pjesma    - -  (strofa+)>
```

gdje su pismena:

`<! i >`	otvorne i zatvorne razdjelnice (*delimiters*) oznaka,
`ELEMENT`	ključna riječ vrste deklaracije,
`pjesma`	generički identifikator elementa, tj. njegov naziv,
`- -`	nedopuštanje minimizacije oznaka, tj. početna i završna oznaka su obavezne. SGML u nekim slučajevima dopušta izostavljanje oznaka, pa bi suprotno od – moglo stajati i `o` (*optional*),
`(strofa+)`	model sadržaja (*content model*).

Model sadržaja cjelokupan je skup deklaracija koje se odnose na sve elemente koji se pojavljuju u dokumentu.[105] Moguće vrijednosti koje mogu stajati uz model sadržaja su:

[103] Erjavec (1997):3
[104] Erjavec (1997):11
[105] Spencer (1999):43

+ element `<strofa>` može se pojaviti jednom ili više puta unutar elementa `<pjesma>`,

? element `<strofa>` je opcionalan,

* element `<strofa>` može se uopće ne pojaviti ili se pojaviti više puta unutar elementa `<pjesma>`,

(ništa) element se mora pojaviti samo jednom.

Deklaracija atributa na konkretnom bi primjeru mogla izgledati:

(pr. 11)

```
<!ATTLIST pjesma
id ID #IMPLIED status (draft|revised|published) draft >
```

U samome dokumentu atributi bi prema propisanoj gramatici mogli biti uporabljeni:

(pr. 12)

```
<pjesma id="P1" status="revised"> . . . </pjesma>
```

Neke od mogućih deklaracija atributa mogu biti:

- CDATA: bilo koje pisme,
- SDATA: podaci ovisni o sustavu (*system dependent data*),
- NUMBER: atribut se sastoji samo od brojeva,
- ID: jedinstveni identifikator,
- IDREF: pokazivač (*pointer*) na neki ID,

dok su moguće ključne riječi:

- #REQUIRED: vrijednost mora biti specificirana,
- #IMPLIED: zadana vrijednost (*default*) mora biti postavljena od strane korisnika, u nekim slučajevima može biti postavljena nasljeđivanjem od elementa-roditelja (*parent*),
- #CURRENT: ako vrijednost nije postavljena, bit će korištena posljednja specificirana vrijednost.

Entiteti omogućuju da se proizvoljan naziv pridruži nekom sadržaju, pa bi se SGML **entitet** mogao **deklarirati** na slijedeći način u DTD-u:

(pr. 13)

```
<!ENTITY ffzg "Filozofski fakultet Sveučilišta u
Zagrebu">
```

Nakon pozivanja na entitet bilo gdje u dokumentu:

```
Zgrada preko puta je &ffzg;.
```

sadržaj se entiteta umeće na mjesto gdje je entitet pozvan:

```
Zgrada preko puta je Filozofski fakultet
Sveučilišta u Zagrebu.
```

SGML se za obilježavanje tekstova koristi ASCII skupom pismena. Svako pisme koje se ne nalazi unutar ASCII skupa prikazuje se posebnim entitetima u skladu sa standardom, npr. Tablica 1 za "naših" deset pismena. Takvim pismenima odgovara određeni niz koji predstavlja varijablu čije se ime nalazi unutar pismena & i ;. Dakle, posebna se pismena kodiraju pozivom na već ugrađene entitete. Tako se npr. hrvatsko "ć" u SGML dokumentu zapisuje kao entitet `&cacute;`, a riječ `ćemo` bi u SGML-u bila kodira nizom pismena: `&cacute;emo`. U sljedećoj se tabeli nalaze SGML entiteti kojima je moguće zamijeniti hrvatska pismena kojih nema u u standardnome ASCII-ju:

Pisme	SGML entitet
č	č
ć	ć
đ	đ
š	š
ž	ž
Č	Č
Ć	Ć
Đ	Đ
Š	Š
Ž	Ž

Tablica 1

Glavni nedostatak SGML-a je što još uvijek ne postoji dovoljan broj aplikacija koje bi zadovoljavale pristupačnošću i mogućnostima. Točnije, ne postoji ni jedna

aplikacija koja je implementirala čitav SGML standard. Naime, aplikacije su ili besplatne ali ograničene na uska područja, ili jednostavno preskupe, a opet ne obuhvaćaju potpunu SGML specifikaciju. Drugi je nedostatak složenost SGML-a za običnoga korisnika, pogotovo ako se ima u vidu jednostavnost novijeg jezika za obilježavanje XML-a.

Trenutačno najprikladniji SGML parser koji obavlja provjere ispravnosti i valjanosti dokumenata po mišljenju većine korpusnih jezikoslovaca koji koriste SGML je SP[106] (autor: *James Clark*). Dobre su osobine SP-a: implementiranost najvećeg dijela standarda, izvršavanje na većini platformi, ne sadrži pogreške (*bugs*) i besplatan je.

Primjer SGML-om kodiranog teksta na hrvatskome jeziku (zaglavlje i dio tijela teksta) prema *TEI smjernicama* može se naći u dodatku A.

3.3.2. XML (*eXtended Markup Language*)

XML je metajezik koji posljednjih godina sve češće zamjenjuje SGML, s velikim izgledima da postane općeprihvaćeni standard za obilježavanje korpusa. Glavni razlog koji je pridonio nastanku i širenju XML-a nagli je razvoj Interneta u proteklom desetljeću. Tim se razvojem stvorila potreba za jedinstvenim, prenosljivim i jednostavnim formatom za elektroničku razmjenu podataka.

World Wide Web Consortium (W3C)[107] objavio je prvu preporuku za korištenje XML-a 10. veljače 1998.[108] Činjenica da je XML standard definiran iz jednog izvora (W3C) rezultiralo je *neovisnošću* standarda o pojedinim aplikacijama ili proizvođačima, te osiguralo njegovu *uniformnost*. Izveden je iz SGML-a, te je s njim kompatibilan. Načelno, XML datoteke mogu koristiti alate i aplikacije napravljene za SGML.

Program koji obrađuje XML dokumente zove se *XML parser*. S obzirom da je XML postao široko prihvaćen standard, XML parser je ugrađen i u *Microsoft*

[106] SP is a "object-oriented toolkit for SGML parsing and entity management". SP se može preuzeti s adrese http://www.jclark.com/sp/index.htm bez naknade
[107] W3C je središnje tijelo za propisivanje standarda i preporuka važećih na Internetu, više na adresi: http://www.w3.org
[108] http://www.w3.org/TR/REC-xml

Internet Explorer (inačice IE 5+). Primjeri u ovome radu obrađeni su *Microsoftovim* MSXML 3.0 parserom.

Dobrim osobinama XML-a smatraju se:

1. **prenosljivost**: neovisnost o pojedinom proizvođaču, platformi ili aplikaciji,
2. očuvanje **strukture** podataka: čvrsta i eksplicitno iskazana struktura podataka (ali opet prilagodljiva korisniku) definira odnose među elementima i omogućuje pretraživanje prema obilježenim razinama,
3. XML podaci mogu se razmjenjivati i objavljivati putem **WWW-a** bez ikakve prilagodbe,
4. **odvajanje** podataka od načina prikaza: jednom označene podatke možemo prikazivati na neograničeno mnogo načina ne zadirući u osnovni dokument s podacima (s pomoću proširenja XML-a kao što je XSL),
5. **riješen** je problem kodiranja nestandardnih pismovnih skupova: XML inačica 1.0 u potpunosti podržava UNICODE 3.0,
6. **uniformnost**: standard je definiran na jednom mjestu (W3C) i nema odstupanja među inačicama,
7. XML postaje *de facto* **opći standard** za sve vrste dokumenata: čak i najveći proizvođač aplikacija i operativnih sustava, Microsoft[109] podržava XML i implementira ga u nove inačice svojih aplikacija (*MS Office2000+*, *Internet Explorer 5+, SQL Server 2000* itd.),
8. kada su podaci **jednom** u pregledniku nije ih potrebno ponovo učitavati pri različitim načinima njihova prikazivanja,[110]
9. za XML-om kodirane tekstove postoje odlične tehnike **sažimanja** (*compression*) do mjere da ponekad obilježeni tekst zahtijeva manje diskovnoga prostora nego neobilježeni izvornik.[111]

Budući da je XML metajezik, moguće je prema potrebi definirati neograničeni broj oznaka. Podatak pridružen XML elementu može biti vrlo različite prirode i vrste: npr. naslov, cijena, slika, zvuk, ime, datum itd. XML

[109] Više informacija na adresi: http://www.msdn.microsoft.com/xml
[110] Bekavac (2000):6
[111] Ide & Brew (2000):3

definira logičku strukturu (naslovi, zaglavlja, odlomci itd.), a ne sam grafički izgled podataka (kao što to radi HTML).

3.3.2.1. XML dokument

XML standard organizira podatke u obliku dokumenata. XML dokument mora sadržati najmanje jedan element da bi se mogao smatrati XML dokumentom. Svaki XML dokument mora imati jedinstven polazni element koji se zove **korijen** (*root*) element. Element koji se hijerarhijski nalazi na nižoj razini od promatranoga naziva se **podređeni** (*child*) element. Element koji se hijerarhijski nalazi na višoj razini od promatranoga je **nadređeni** (*parent*) element. Svi elementi osim korijena moraju imati samo jedan neposredno nadređeni element. Svaki element može imati jedan ili više podređenih elemenata. Elementi označavaju tekst kako bi eksplicirali njegovu **strukturu** i **funkciju** (u opreci prema prikazivanju teksta).[112]

Primjer XML dokumenta koji slijedi navedena pravila izgleda:

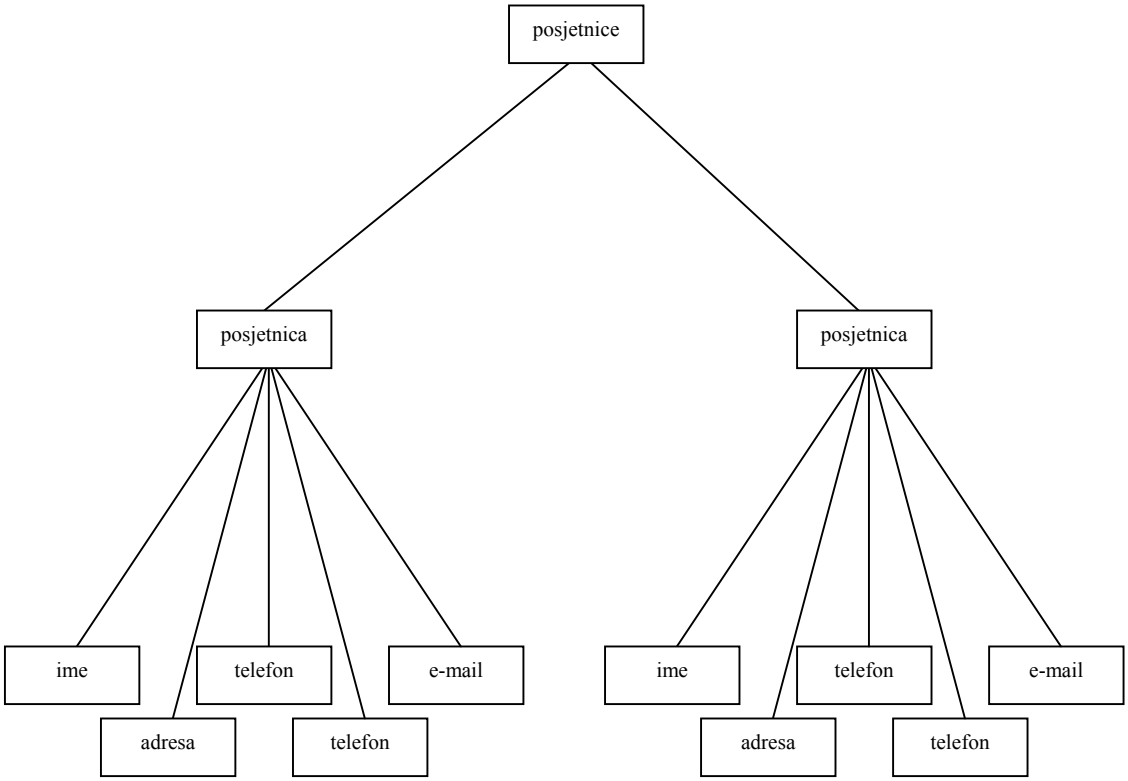
(pr. 14)

```
<?xml version="1.0"?>
<posjetnice>
     <posjetnica>
          <ime>Hrvoje Horvat</ime>
          <adresa>Bauerova 1, 10 000 Zagreb</adresa>
          <telefon>01/5555-555</telefon>
          <telefon>01/4444-444</telefon>
          <e-mail>Hrvoje@yahoo.com</e-mail>
     </posjetnica>
     <posjetnica>
          <ime>Marija Badurina</ime>
          <adresa>Podujina 34, 21 000 Split</adresa>
          <telefon>01/2222-222</telefon>
          <telefon>01/3333-333</telefon>
          <e-mail>Marija@yahoo.com</e-mail>
     </posjetnica>
</posjetnice>
```

[112] Brew (2000):22

Prvi je redak u dokumentu deklaracija XML inačice. U ovom slučaju inačica je 1.0, jedina koja za sada postoji. Pisanje deklaracije nije obavezno, ali je preporučljivo. Ukoliko je deklaracija izostavljena, dokument se obrađuje kao ispravan XML dokument.

Karakteristika je svih XML dokumenata da su strogo hijerarhijski uređeni, tj. elementi su raspoređeni po razinama. Svaki XML dokument može se prikazati i u obliku stabla (*tree*), koje bi u gornjem primjeru izgledalo ovako:



Slika 2: Stablasti prikaz strukture XML dokumenta

Element na najvišoj razini je korijen element <posjetnice>. Neposredno podređeni elementi korijen elementu su elementi <posjetnica>, a neposredno nadređeni element od <telefon> je element <posjetnica>.

XML definira stroga sintaktička pravila koja je nužno poštivati da bi dokument bio ispravan:

1. svaki XML dokument mora imati jedan i samo jedan prvi element koji se zove korijen (*root*) element,
2. svaka otvorna oznaka (*tag*) mora imati pripadajuću zatvornu oznaku,

3. elementi moraju biti uredno ugniježđeni (*nested*), tj. nije dozvoljeno preklapanje otvornih i zatvornih oznaka,
4. XML oznake su *case-sensitive*, tj. osjetljive su na razliku između malih i velikih slova.

3.3.2.2. XML element

Najjednostavniji se XML element sastoji od:

- početne oznake (*tag*), npr. <*ime*>,
- sadržaja elementa, npr. *Hrvoje Horvat*,
- završne oznake, npr. <*/ime*>.

Taj bi jednostavan XML element bio zapisan:

(pr. 15)

```
<ime>Hrvoje Horvat</ime>
```

Imenom oznake u načelu se eksplicira (obilježava) sadržaj elementa.

3.3.2.3. XML atribut

Uz oznaku elementa može stajati jedan ili više atributa. Primjer elementa s jednim atributom:

(pr. 16)

```
<ime prebivalište="Zagreb">Hrvoje Horvat</ime>
```

ili s primjer s dva atributa:

(pr. 17)

```
<ime prebivalište="Zagreb"
JMBG="1111955380335">Hrvoje Horvat</ime>
```

Na je ovaj način sadržaju *Hrvoje Horvat* pridodana informacija o njegovom prebivalištu i JMBG-u. Isti zapis iz prvoga primjera mogao je biti kodiran dodatnim elementima i bez uporabe atributa:

(pr. 18)

```
<ime>
    <prebivalište>Zagreb</prebivalište>
    <JMBG>1111955380335</JMBG>
    Hrvoje Horvat
</ime>
```

Oba su načina ispravna, ali je prvi ekonomičnije kodiran. Koji će se način koristiti odlučuje sam korisnik ovisno o svojim potrebama.

3.3.2.4. XML DTD (Document Type Definition)

XML DTD je naslijeđen iz SGML-a, a funkcija mu je ostala ista uz neznatne sintaktičke razlike. DTD za prethodno izloženi XML dokument izgledao bi ovako:

(pr. 19)

```
<?xml version = "1.0"?>
<!DOCTYPE posjetnice [
  <!ELEMENT posjetnica      (ime,adresa,telefon+,e-mail?)>
  <!ELEMENT ime             (#PCDATA)>
  <!ELEMENT adresa          (#PCDATA)>
  <!ELEMENT telefon         (#PCDATA)>
  <!ELEMENT e-mail          (#PCDATA)>
]>
```

DOCTYPE je *Document Type Declaration*, i deklarira element koji mora biti korijen u dokumentu (u ovom slučaju `<posjetnice>`). U gornjem DTD-u definirana je posjetnica koja mora sadržati elemente po redoslijedu kojim su navedeni. Na prvom mjestu bezuvjetno mora stajati ime, zatim adresa, dok se element telefon može pojaviti jednom ili više puta. Element e-mail je opcionalan, tj. može se izostaviti, ali i pojaviti. Niz #PCDATA (*parsed character data*) deklarira element koji se sastoji od standardnih pismena. S obzirom da se u praksi najveći broj elemenata najčešće sastoji od standardnih pismena, #PCDATA je najčešća uporabljivana deklaracija. Pisme # je rezervirano pisme, i ne smije se koristiti za deklaraciju vlastitih imena.

Kod XML-a, za razliku od SGML standarda, nije obavezna uporaba DTD-a.

3.3.2.5. CDATA odjeljak

Čest je slučaj da se u XML dokumentu nalaze podaci koji iz određenih razloga nisu prikladni za obrađivanje parserom. Npr. to može bit slučaj kada tekst sadrži pismena koja su rezervirana XML pismena (npr. <, > ili &). Tada se koristi CDATA (*containing character data*) odjeljak koji bi u slučaju da namjeravamo prikazati niz "**a>b&c**" izgledao ovako:

(pr. 20)

```
<primjer>
     <![CDATA[a>b&c]]>
</primjer>
```

Kada parser naiđe na CDATA odjeljak, on ga ne obrađuje već ostavlja tekst koji se nalazi unutar uglatih zagrada u obliku u kojem je naveden. U praksi CDATA odjeljak se koristi kada u tekstu ima mnogo, ili nepoznati broj rezerviranih pismena (u kojima se mogu nalaziti netekstni podaci ili kada se u dokument uključuju binarne datoteke).

*3.3.2.6. Pozivanje na entitete (*Entity References*)*

O **pozivanju na entitete** kao sredstva za zamjenu dugih nizova pismena kratkima već je bilo riječi.[113] Tada je navedeno da se pozivanjem na entitete može riješiti prikazivanje rezerviranih pismena. Niz "3<5" parser bi pogrešno interpretirao jer bi pisme "<" bilo protumačeno kao početak otvorne oznake, pa ga je potrebno kodirati na drugi način. Zbog tog razloga postoji pet unaprijed definiranih internih entiteta koji su prikazani u slijedećoj tablici:

[113] v. poglavlje 3.3.

Pisme	Entitet
<	&lt;
>	&gt;
&	&amp;
‘	&apos;
“	&quot;

Tablica 2

Poseban slučaj pozivanja entiteta je **pozivanje pismena** (*character reference*) koje se koristi za ubacivanje pismena što se ne mogu dobiti izravno preko tipkovnice. Tu se nalaze pismena koja nisu uključena u standardni 7-bitni ASCII. Skup dopuštenih pismena za pozivanje ekvivalentan je UNICODE standardu. Hrvatska slova s dijakritičkim znacima pri pozivanju pismena imala bi sljedeće vrijednosti (u dekadskom i heksadecimalnom obliku):

Pisme	Dekadski	Heksadecimalni
č	&#269;	&#x10D;
ć	&#263;	&#x107;
đ	&#273;	&#x111;
š	&#353;	&#x161;
ž	&#382;	&#x17E;
Č	&#268;	&#x10C;
Ć	&#262;	&#x106;
Đ	&#272;	&#x110;
Š	&#352;	&#x160;
Ž	&#381;	&#x17D;

Tablica 3

Na primjer, rečenica “Jabuke koštaju < $1 u C&A.” bila bi XML-om kodirana:

(pr. 21)

```
Jabuke ko&amp;#353;taju &amp;lt; $1 u C&amp;amp;A.
```

Pozive na entitete jednostavno je identificirati jer uvijek počinju s znakom “&”, a završavaju s “;”.

Primjer XML-om kodiranog teksta (zaglavlje i dio tijela teksta) XCES standardu može se naći u dodatku B.

3.3.2.7. XSL (eXtensible Stylesheet Language)

XSL je tehnologija koja se koristi za manipulaciju, razvrstavanje (*sorting*), filtriranje i oblikovanje XML dokumenata.[114] Kako je XSL čak i u informatičkim terminima mlada tehnologija, još nije u potpunosti standardiziran. Preporuke su standarda u prošlosti doživjele mnogo izmjena, a radni nacrt (*Working draft*) se nalazi na adresi: http://www.w3.org/TR/xsl/

XSL se sastoji od dva nezavisna dijela:

1. **jezika za transformacije, XSLT** (*transformation language*) s pomoću kojega se transformira ispravan XML dokument u novi, drugačiji XML dokument ili neki drugi zapis. XSLT je definiran 16. studenoga 1999. i njegova inačica 1.0 nalazi se na adresi: http://www.w3.org/TR/xslt/ (XSLT Recommendation 1.0),
2. **jezika za oblikovanje** (*formatting language*) koji određuje kako će podaci dobiveni transformacijom biti prikazani. Jezik za oblikovanje još nije u potpunosti standardiziran.

Dok XSLT zadire u sam sadržaj dokumenta, jezik za oblikovanje ima isključivo funkciju prikaza podataka dobivenih transformacijama. S obzirom da se XML podaci mogu oblikovati i s pomoću CSS-a[115] (*Cascading Style Sheet*) koji je starija i standardizirana tehnologija, jezik za oblikovanje nam u ovom radu neće biti od posebnog interesa.

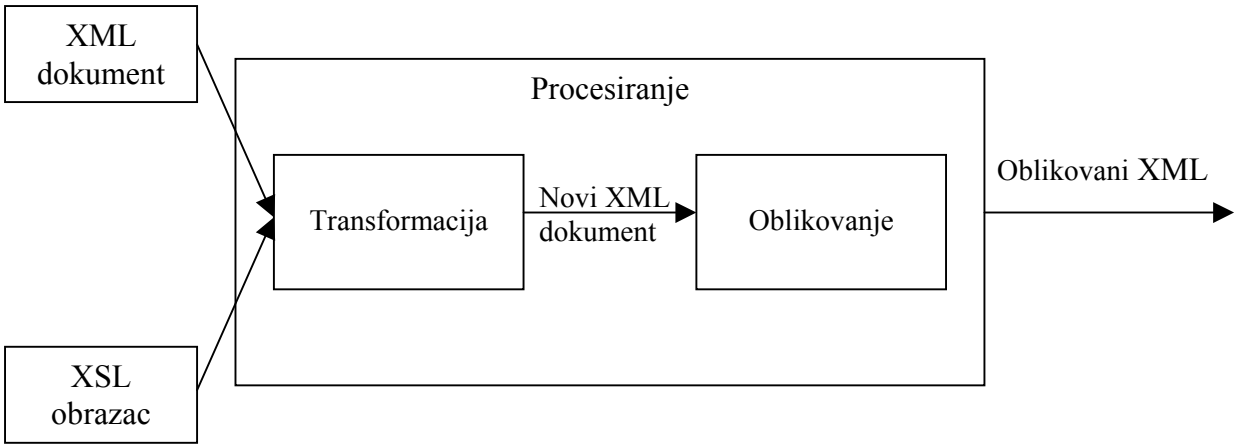
Sadržaj bi se XSL-a mogao podijeliti na:

1. XSL elemente: imaju istu sintaksu kao i XML elementi, i mogu se smatrati naredbama za transformaciju izvornog XML dokumenta. W3C je ograničio i definirao broj XSL elemenata.

[114] Bekavac (2000)
[115] CSS je tehnologija koja je razvijena kao nadopuna HTML-u. Omogućuje lakše oblikovanje i opsežniju kontrolu nad HTML sadržajem. Sa CSS-om kreiramo obrasce koji stilski definiraju kako će izgledati određeni HTML elementi (poveznice, naslovi, e-mail adrese).

2. XSL metode: skup metoda za oblikovanje (*formatting*) različitih tipova podataka (npr. konverzije datuma, brojeva itd.) i brojanje zadanih elemenata.
3. XSL sintaksa uzoraka (*Pattern Syntax*): načini za pretraživanje i filtriranje podataka iz izvornog XML dokumenta.

XML parser za vrijeme obrade obrađuje XML dokument kao stablastu strukturu (v. sliku 3). XSL je obrazac koji je napravljen s ciljem da se iz izvora (XML dokument) preuzmu određeni podaci i preoblikovani ispišu kao novo stablo. Novo stablo kao rezultat transformacije oblikuje se za prikaz s pomoću jezika za oblikovanje:



Slika 3: Transformacija XML dokumenta

Rezultat transformacije XML dokumenta novi je, preoblikovani dokument koji, ovisno o potrebi može imati mnoštvo izlaznih oblika (XML, HTML, WML, običan tekst itd.). Izvorno stablo tj. XML dokument, XSL obrazac i rezultirajuće stablo tri su potpuno odvojena entiteta. Takvo jasno razdvajanje izvornoga XML dokumenta od rezultirajućeg ispunjava jedan od ciljeva XML-a, a to je odvajanje sadržaja od prezentacije dopuštajući i gramatici i strukturi izvornog XML dokumenta da budu neovisni o načinu prikaza.

XSL koristi istu sintaksu kao i XML, pa mora poštivati pravila za ispravnost (*well-formedness*) dokumenata. Sintaktički gledano, XSL je zapravo ispravan XML dokument. XSL dokument u pravilu se sastoji od skupa predložaka i transformacijskih pravila. **Pravilo predloška** (*template rule*) ima **uzorke** (*patterns*) koji specificiraju dio stabla i zatim ga ispisuju ili obrađuju uz pomoć dodatnih instrukcija. Hijerarhija je

od iznimne važnosti, jer se upiti prema XML dokumentu postavljaju prema određenoj razini na kojoj se elementi nalaze. Tako je moguće pristupiti bilo kojem elementu ili atributu (elementi se u tom kontekstu nazivaju i čvorovima) iz XML izvora, te manipulirati njegovim sadržajem. Tehnika kojom se pristupa čvorovima iz izvora naziva se **usporedba s uzorkom** *(pattern matching)*. Usporedba s uzorkom predstavlja jednostavan "**jezik za postavljanje upita**" (*query language*) čija je svrha identifikacija čvorova u XML dokumentu. Pretraživanje se zasniva na imenu, vrijednosti, tipu ili odnosu čvora s obzirom na položaj prema drugim čvorovima u dokumentu. Jednostavni upiti srž su XSL transformacija za dobivanje novog, modificiranog XML dokumenta. Određivanje razine za postavljanje upita vrlo je slično sustavu navigacije kroz direktorije operacijskog sustava. Tako za pristupanje elementu <ime> (primjer 14 str. 40) treba postaviti razinu za postavljanje upita na slijedeći način:

(pr. 22)

```
posjetnice/posjetnica/ime
```

dakle, potrebno je "spustiti" se na treću razinu u hijerarhiji.

XSL omogućuje da definiramo predloške za rezultirajuće stablo u koje se prebacuju podaci iz izvornog stabla odabrani definiranim predlošcima.

3.3.2.8. XML DOM (Document Object Model)

DOM je programsko sučelje za HTML i XML dokumente koje je W3C prihvatio kao standard.[116] XML DOM programsko je sučelje koje omogućuje aplikacijama kretanje po XML stablu i manipulaciju njegovim čvorovima.[117] S pomoću XML DOM-a moguće je:

- kreirati XML dokumente,
- navigati (*navigate*) kroz njegovu strukturu,
- učitavati (*load*) i parsati XML dokumente,
- manipulirati, dodavati ili brisati elemente ili njihov sadržaj.

[116] Specifikacija standarda nalazi se na: http://www.w3.org/DOM
[117] Spencer (1999):77

XML DOM tretira XML dokument kao **stablastu** (*tree*) **strukturu**[118] sastavljenu od **čvorova** (*nodes*). U kontekstu DOM-a čvorovi su pojedini XML elementi. Slično kao kod HTML-a (osobito *dinamičkoga HTML-a*), XML elementima može se pristupiti s pomoću skriptnih jezika. Neki od skriptnih jezika koje XML DOM podržava su JavaScript, VBScript, Perl, Visual Basic, Java, C++, ali i mnogi drugi.[119]

Središnje je načelo DOM-a objektna orijentacija, gdje **objektni model** obuhvaća skup **objekata** od kojih svaki posjeduje određeno sučelje. Najviši objekt u hijerarhiji je **dokument-objekt**, i predstavlja korijenski element XML dokumenta. Primjer pristupanja korijenskom elementu korištenjem skriptnog jezika VB Script izgledao bi ovako:

(pr. 23)

```
root = source.documentElement
```

Varijabli *root* pridružen je korijenski element, u primjeru kodiranja posjetnice (primjer 14 str. 40) to je element *posjetnice*. Nakon pristupanja **dokument-objektu**, može se koristiti njegovo sučelje, odnosno sve njegove **metode** (*methods*) i **svojstva** (*properties*) u skladu s postavkama objektnog programiranja (*object-oriented programming*).[120] Kako je korijen element nadređen svima ostalim elementima u dokumentu, pri ispisu ove varijable:

(pr. 24)

```
document.write(root)
```

bio bi ispisan cijeli bi XML dokument.

Pristup ostalim elementima obavlja se preko **dokument-objekta** i pripadajućeg mu svojstva *childNodes*. Pri korištenju toga svojstva, svi elementi u dokumentu indeksirani su prema korijenu, a samo indeksiranje u ovom slučaju počinje od 0 (jer VB Script uvijek počinje indeksiranje od 0). Ukoliko se pristupa elementu *ime* (iz primjera 14 str. 40):

(pr. 25)

```
ime = root.childNodes.item(2)
```

[118] v. poglavlje 3.3.2.1.
[119] Spencer (1999):64
[120] Više na adresi: http://www.webopedia.com/TERM/o/object_oriented_programming_OOP.html

svojstvo *item* treba imati vrijednost 2, što je zapravo treći element u hijerarhiji.

XML DOM vrlo je efikasan, ali i kompleksan način manipuliranja XML dokumentima. Kombiniranje objektnog pristupa sa skriptnim jezicima, koji su također objektno orijentirani, daje mu više fleksibilnosti uz istodobnu preciznost u odnosu na manipulacije rađene s pomoću XSL-a.

3.3.2.9. Povezivanje (linking) u XML-u

Pri suvremenom kodiranju korpusa preporučuje se držanje svih ili većine oznaka u odvojenim dokumentima. Iz tih se odvojenih dokumenata **poveznicama** (*links*) referira na određeno mjesto u dokumentima u kojima se nalaze tekstovi na koje se oznake odnose.[121] Poveznice iz polaznoga dokumenta *jednoznačno* upućuju na specifično mjesto u dokumentu u kojem se nalazi tekst. Time se stvara hipertekstni oblik dokumenata, gdje poveznice imaju više semantičku nego uobičajenu navigacijsku ulogu. Tako se obavlja **udaljeno obilježavanje** (*stand-off mark-up, stand-off annotation*). Prednost je ovakvoga načina obilježavanja što izvorni tekst ne mora sadržati nikakve oznake (osim strukturnih oznaka), jer se one nalaze u odvojenom dokumentu s vezama prema izvornome tekstu. XML posjeduje mehanizme (omogućuju i udaljeno obilježavanje) koji su znatno napredniji u odnosu na mehanizme za povezivanje koje nudi SGML:

1. **XLink**: mehanizam za definiranje veza (jedno- ili višesmjernih) između dokumenata 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 dijelova.[122]

Na primjer, XPath izraz:

(pr. 26)

```
/div/p[2]/s[3]
```

[121] Ide (2000):28
[122] Ide (2000):28

određuje treću rečenicu <s> (koja je ujedno i element promatran u XML kontekstu) unutar drugoga odlomka <p> unutar svakoga elementa. Xpath omogućuje pristup i odsječcima teksta unutar elementa, što nije omogućeno u DOM-u. Izrazom

```
substring(p/s[2]/text(),4)
```

iz XML dokumenta

(pr. 27)

```
<p><s id="d3p8s4">Za tri dana trajanja pregovaranja
nije postignut sporazum o trgovini.</s><s
id="d3p8s5">Tim sporazumom trebalo je dogovoriti
razmjenu naftnih derivata.</s></p>
```

bio bi odabran niz koji počinje nakon četvrtoga pismena unutar druge rečenice, koja se nalazi unutar odlomka. Dakle, odabran bi bio niz: `"sporazumom trebalo je dogovoriti razmjenu naftnih derivata."`. Izraz

(pr. 28)

```
substring(p/s[2]/text(), 4, 22)
```

primijenjen na tekst iz primjera 27 odabire niz `"sporazumom trebalo"`.

Referiranje se obavlja navođenjem adrese elementa (bilo relativne ili apsolutne) koji je najbliži traženome nizu i brojevima koji označavaju početak odnosno završetak niza (u gornjem slučaju 4 i 22).

Xlink se može koristiti za povezivanje odgovarajućih odsječaka dvaju ili više izvornih tekstova, ili za povezivanje dokumenata koji sadrže oznake s pripadajućim dokumentima (izvornim tekstovima). U drugom slučaju, informacije o pojedinim pojavnicama (element <tok>), tj. njihove oznake povezuju se s nizovima pismena iz izvornoga teksta:

(pr. 29)

```
<tok  xlink:href =
"substring(p/s[2]/text(), 4, 14)">
```

Ovako stvorena poveznica uspostavila bi vezu s pojavnicom `"sporazumom"` iz primjera 27.

Kod udaljenog obilježavanja XML-om na obilježene se odsječke referira preko **URI-ja** (*Uniform Resource Identifier*)[123], **ciljanoga izvora** (*target resource*),

[123] Svrha URI-ja je osiguranje jedinstvenosti elemenata i otklanjanje potencijalne višeznačnosti kod različitih elemenata. U ovom slučaju URI je *http://www.hnk.ffzg.hr/dokument.xml#xptr*

proširenoga pokazivača (*extended pointer*) koji identificira element i tamo gdje je potrebno odabranog niza pismena iz sadržaja toga elementa na sljedeći način:

(pr. 30)

```
<tok  xlink:href =
“http://www.hnk.ffzg.hr/dokument.xml#xptr
(substring(p/s[2]/text(), 4, 22))”>
```

Obilježavanje koje je nastalo kao rezultat automatskog obrađivanja (npr. označavanje granica rečenica, pojavnica, veza između paralelnih tekstova itd.) često obuhvaća mnoštvo (ponekad i tisuće) veza prema istom vanjskom dokumentu. Ponavljanje imena dokumenta za svaki relevantni element značajno povećava količinu podataka i veličinu dokumenta, te se može smatrati redundantnim. XML za rješavanje tog problema koristi *xml:base* atribut koji se rabi za specificiranje nasljeđivanja nekog atributa. Na primjeru,

(pr. 31)

```
<chunk
     xml:base “http://www.hnk.ffzg.hr/dokument.xml#”>
<tok
     xlink:href =“xptr(substring(p/s[2]/text(), 4,
22))”/>
<tok
     xlink:href =“xptr(substring(p/s[2]/text(), 25,
32))”/>
</chunk>
```

dva <tok> elementa nasljeđuju vrijednost atributa *xml:base* specificiranog u <chunk> elementu. Ponavljanje nije potrebno jer su elementi <tok> potomci, tj. podređeni elementu <chunk>.

*3.3.2.10. XML/XSL obradnici (*processors*)*

Glavna je uloga XML/XSL obradnika primijeniti XSLT *stylesheet* na izvorni (*source*) XML dokument i prikazati rezultat kao novi dokument. Parser strukturu XML dokumenta "razbija" u strukturu stabla kojom je moguće manipulirati. Postoji nekoliko XSLT obradnika, a spomenut će se tri koja se najčešće koriste (svi se bez naknade mogu preuzeti putem WWW-a):

1.) Saxon: pisan je u Javi. Pokreće se iz "command prompta" (nije potreban ni poslužnik ni preglednik). Za Windowse (95/98/NT/2000) potrebno je preuzeti "Instant Saxon" paket ("Windows executable"). Nalazi se na adresi:
http://users.iclway.co.uk/mhkay/saxon/index.html

2.) xt: pisan je u Javi. Pokreće se iz "command prompta" (nije potreban ni poslužnik ni preglednik). Za Windowse (95/98/NT/2000) potrebno je preuzeti "xt" paket ("Windows executable"). Nalazi se na adresi:
http://www.jclark.com/xml/xt.html

3.) Microsoft MSXML processor: prikazuje rezultat u Internet Exploreru. Sadržan je u inačicama IE 5+. Može se koristiti i kao COM (najmanje IE 4) . Trenutna inačica obradnika je MSXML3. Nalazi se na adresi:
http://msdn.microsoft.com/xml

3.4. Standardi za kodiranje korpusa

U posljednjem se desetljeću naglo povećala produkcija velikih korpusa *(large-scale corpora*). Sve do tada većinu elektroničkih tekstova sastavljali su pojedini istraživači za svoje specifične potrebe ili istraživački instituti koji su sastavljali korpuse za vlastita proučavanja jezika. U načelu se svaki od sastavljača držao svoje strukture, načina kodiranja korpusa itd. S vremenom, elektronički su se tekstovi počeli sve više razmjenjivati među istraživačima. Postalo je jasno da sve postojeće sheme kodiranja imaju nedostataka. Neke su rađene za alate koje zahtijevaju specifičnu platformu, druge odražavaju poglede autora na svoje područje itd.[124] Zbog umnažanja napora pri svakoj pojedinačnoj izradi, lakše razmjene i mogućnosti višestrukoga

[124] Lawrer & Dry (1998):111

korištenja već gotovih jezičnih resursa, stvorila se potreba za ustanovljenjem jedinstvenog standarda za kodiranje korpusa. To je rezultiralo nastajanjem nekoliko standarda i preporuka za kodiranje. Dobro bi kodiran korpus trebao biti:[125]

- **višestruko uporabiv** (*reusable*),[126] potencijalno uporabiv u više istraživačkih projekata i za više namjena,
- **proširljiv** (*extensible*), u smislu mogućnosti daljnjega nadograđivanja postojećega korpusa.

3.4.1. TEI (*Text Encoding Initiative*)[127]

TEI je najveći međunarodni projekt u području definiranja standarda za elektroničku razmjenu ponajprije tekstovnih podataka pokrenut 1988. godine pod pokroviteljstvom *Association for Computers and the Humanities*, *Association for Computational Linguistics* i *Association for Literary and Linguistic Computing*.

Cilj je TEI-a bio napraviti smjernice za pripremu i razmjenu elektroničkih tekstova kako za znanstvena istraživanja, tako i za širok raspon uporaba za potrebe jezičnih tehnologija, pa i šire. Pored osnovne svrhe, razmjene tekstovnih informacija, obuhvaćeni su i drugi oblici informacija kao što su slika ili zvuk. Međutim, u lingvističkom smislu TEI ponajprije postavlja standarde za obilježavanje svih vrsta tekstova, a ne specijalizirano samo za jezične korpuse. TEI smjernice pokušavaju eksplicirati određene osobitosti tekstova kako bi se olakšalo njihovo obrađivanje različitim aplikacijama na različitim platformama.

Pri definiranju sheme obilježavanja u središtu su dvije težnje:

1. **koje** su osobine teksta koje trebaju biti kodirane (tj. učinjene eksplicitnima) u elektroničkom tekstu i
2. **kako** bi one trebale biti kodirane za platformski neovisnu razmjenu bez gubitka informacija.

Takve je zahtjeve 1988. godine najbolje zadovoljavao jezik za obilježavanje SGML (***S**tandard **G**eneralized **M**arkup **L**anguage*). TEI skup oznaka zasnovan je na SGML-u i definiran prema postojećoj praksi za kodiranje, a dizajniran je s ciljem da bude sveobuhvatan i ekstenzivan. Prva radna inačica pod imenom "*Guidelines for*

[125] Ide & Brew (2000):1
[126] v. poglavlje 4.1.
[127] TEI (1999)

Electronic Text Encoding and Interchange" (TEI dokument P1) izdana je 1990. godine. Aktualna inačica *TEI Guidelines* (TEI dokument P3)[128], kako se najčešće naziva ovaj dokument, pretrpjela je brojne nadopune, izmjene i revizije prvih dviju inačica, i datira iz travnja 1994. godine. Dokument objavljen u papirnatom i elektroničkom obliku vrlo je opsežan (1300 stranica). U potpunoj TEI shemi obilježavanja definirano je nekoliko stotina SGML elemenata. Stoga su za jednostavnije slučajeve uporabe razvijena dva podskupa TEI-a: *TEI-Lite* (TEI dokument U5)[129] i *TEI-Barebone* (TEI dokument U6).[130]

Ukoliko je dokument jedinstven (ne sastoji se od više tekstova), njegova se općenita struktura po TEI-ju prikazuje na sljedeći način:[131]

(pr. 32)

```
<tei.2>
<teiHeader>
  ...
  ...
</teiHeader>
  <text>
     <body>
        <!–tijelo teksta nalazi se u ovome dijelu. -->
     </body>
  </text>
 </tei.2>
```

Svi dokumenti obilježeni prema TEI standardu imaju isti korijenski element <*tei.2*>, unutar kojeg se nalaze svi elementi koji čine dokument. Oznakom <*text*> obilježava se bilo koja vrsta teksta koja je u dokumentu. U samom tijelu dokumenta (<*body*>) nalaze se nizovi jedinica (elementi) niže razine.

Svaki dokument koji je u skladu s TEI shemom mora sadržati TEI zaglavlje (<*teiHeader*>), koje se sastoji od četiri glavna dijela:

- <*fileDesc*>: odjeljak za opis datoteke. Sadrži potpuni bibliografski opis datoteke u kojoj se nalazi TEI dokument i njezinu bibliografsku referencu,
- <*encodingDesc*>: odjeljak za opis kodiranja. Njim se objašnjava odnos između elektroničkoga teksta i njegovoga izvornoga oblika,
- <*profileDesc*>: odjeljak za opis profila teksta. Sadrži klasifikacijske i ostale informacije o tekstu (vrsta teksta, uvjeti u kojima je sastavljan itd.),

[128] http://www.tei-c.org/Guidelines/index.htm
[129] http://www.tei-c.org/Lite/index.html v. str. 57
[130] http://www.hcu.ox.ac.uk/TEI/Vault/Bare/
[131] Sperberg-McQueen & Burnard (1990)

- *<revisionDesc>*: odjeljak bilježi sve promjene koje su izvedene do konačne inačice tog TEI dokumenta.

Nakon zaglavlja može slijediti oko šezdesetak unaprijed definiranih elemenata kao što su naslov, ime, kratice, citati i sl. koje zajedno s gore navedenima predstavljaju TEI osnovni skup oznaka.

Neka područja obilježavanja teksta (npr. jezični korpusi) zahtijevaju mnogo detaljniji opis od drugih. Iz tog razloga TEI definira osnovni skup elemenata, ali i dodatni skup oznaka koji se može po potrebi proširivati.

Zbog mogućnosti vrlo širokih strukturnih podjela (*division*) u samome tekstu, osnovni element za obilježavanje odsječka strukture je * * i * *. Taj se element koristi zajedno s atributom za vrstu (*type*) podjela (poglavlje, broj, dio, knjiga i slično). Na primjer:

(pr. 33)

```
<body>
 <div type='part' n='1'>
       <div type='chapter' n='1'>
       <!–tekst dijela 1, poglavlje 1 -->
       </div>
       <div type='chapter' n='2'>
       <!–tekst dijela 1, poglavlje 2 -->
       </div>
 </div>
 <div type='part' n='2'>
       <div type='chapter' n='1'>
       <!–tekst dijela 2, poglavlje 1 -->
       </div>
       <div type='chapter' n='2'>
       <!–tekst dijela 2, poglavlje 2 -->
       </div>
 </div>
 </body>
```

Nezaobilazan dio organizacijske strukture, barem kad je o proznim tekstovima riječ, je odlomak (paragraf) obilježen oznaka * * i * *:

(pr. 34)

```
<p id="d2p1">
   <seg id="d2p1seg1">Odoh ju&ccaron;er dolje u Pirej s Glaukonom
Aristonovim da se pomolim bo&zcaron;ici i s namjerom da ujedno vidim
kako &cacute;e prirediti svetkovinu, budu&cacute;i da su je sada prvi
put svetkovali.</seg>
</p>
```

TEI je nastojao definirati listu od preko 400 osobina teksta (predstavljenih elementima) koje bi jezikoslovac ili korisnik s područja humanističkih znanosti

mogao trebati. *TEI Guidelines* opisuju svaku od njih i daju primjere njihove uporabe. Kako nijedan popis nije sveobuhvatan, moguća su daljnja proširenja i promjene. Samo je manji dio oznaka obvezatan. Ostale oznake koriste se ovisno o potrebi onoga tko obilježava i onoga za koga je tekst obilježen. Proces kodiranja zamišljen je s otvorenom mogućnošću za dodavanje novih oznaka, na način da neki drugi istraživač može već obilježenom tekstu dodavati nove oznake prema svojim potrebama.

TEI smjernice izgrađene su po načelu prema kojem se svim tekstovima koji imaju neke zajedničke osobine mogu pridodati oznake za specifična područja ili tekstovne tipove/žanrove. Specijalizirani osnovni skupovi oznaka postoje za: poeziju, dramu, transkribirane govore, rječnike i terminološke podatke.

Suženi opseg TEI standarda objavljen je u dokumentu *TEI Lite*[132] s namjerom obuhvaćanja minimalnog broj elemenata (ti su elementi podskup potpune TEI sheme) koje bi većina korisnika trebala poznavati. Ciljevi definiranja TEI Lite su:[133]

- Treba uključiti većinu osnovnih oznaka TEI skupa, budući da one sadrže elemente relevantne za skoro sve tipove tekstova i oblike obrade teksta,
- treba biti kadar ravnati (*handle*) s dovoljno širokim rasponom različitih tekstova do razine detalja koji već postoje u praksi (npr. *Oxford Text Archive*),
- trebao bi biti koristan u izradi novih dokumenata kao i u kodiranju postojećih,
- trebao bi se moći koristiti širokim spektrom postojećih programa koji rabe SGML,
- trebao bi biti izveden iz potpunog TEI DTD-a poštujući ekstenzije koje su opisane u TEI smjernicama,
- trebao bi biti malen i jednostavan kako bi bio konzistentan.

Primjer teksta obilježena prema TEI standardu može se naći u dodatku A.

[132] Sperberg-McQueen & Burnard (1995)
[133] http://www.hcu.ox.ac.uk/TEI/Lite/U5-Intro.html

3.4.2. CES (*Corpus Encoding Standard*)[134]

CES je standard razvijen u suradnji europskih projekata MULTEXT[135] i EAGLES (*Expert Advisory Group on Language Engineering Standards*) s američkim partnerom Vassar College, te francuskim partnerom CNRS (*Centre National de la Recherche Scientifique*). EAGLES je projekt EU-a kojim je iskazana inicijativa Europske Komisije zadužene za ubrzano donošenje standarda za jezične resurse, sredstava za obradu znanja kroz formalizaciju jezika, jezike za obilježavanje, različite alate, te za vrednovanje resursa, alata i proizvoda. CES je sastavni dio *EAGLES smjernica*.

Namjena je CES-a da bude skup široko prihvaćenih standarda za kodiranje tekstova koji su standardi optimalni za korpusno zasnovane djelatnosti. Glavni je cilj CES-a određivanje minimalne razine kodiranja koju korpus mora zadovoljiti da bi se mogao smatrati standardiziranim u smislu deskriptivne reprezentacije (označavanje strukturalnih i tipografskih informacija), ali također i općenite arhitekture (ne bi li se postigla maksimalna prikladnost za uporabu u tekstovnim bazama podataka). CES također pokriva i transkribirane govorne tekstove.

Za razliku od *EAGLESove* definicije korpusa[136], CES definira korpus više u tehničkom smislu za potrebe lakšeg obrađivanja:

> Here, we use the term *corpus* to refer to any collection of linguistic data, whether or not it is selected or structured according to some design criteria. According to this definition, a corpus can potentially contain any text type, including not only prose, newspapers, as well as poetry, drama, etc., but also word lists, dictionaries, etc.[137]

CES razlikuje **primarne podatke** (*primary data*) u koje spadaju neobilježeni podaci u elektroničkom obliku i **jezikoslovno obilježavanje** (*linguistic annotation*) koje obuhvaća informacije pridodane primarnim podacima nastale kao rezultat neke jezikoslovne analize.

---

[134] CES (1996)

[135] *Multext* obuhvaća nekoliko projekata čiji su ciljevi razvoj standarda i specifikacija za kodiranje i obrađivanje jezičnih korpusa, razvoj alata, te samih korpusa i lingvističkih resursa koji uključuju te standarde.

[136] v. poglavlje 2.2.

[137] CES (1996):http://www.cs.vassar.edu/CES/CES1-0.html: Ovdje (op. u CES-u) se uporaba termina *korpus* odnosi se na bilo koji skup jezičnih podataka, bez obzira jesu li odabrani i strukturirani prema nekim kriterijima. Prema definiciji, korpus potencijalno može sadržavati bilo kakvu vrstu teksta, uključujući prozu, novine, dramu itd, pa čak i popise riječi ili rječnike. (prijevod moj)

CES omogućuje kodiranje ne-lingvističkih objekata u primarnim podacima kao što su:[138]

- jedinice diskursa, npr. odlomci, poglavlja itd. (zajedno s naslovima, fusnotama i sl.)
- elemente na nižoj razini od odlomka koji su zanimljivi za lingvističke analize bilo koje vrste, npr. rečenice, imena, kratice itd.

Pored toga CES određuje lingvističko obilježavanje teksta kao što je morfosintaktičko obilježavanje, sravnjivanje (*alignment*) paralelnih tekstova, fonetska transkripcija i sl.

Prvi princip razvijanja CES-a je smjer "odozdo-gore" (*bottom-up*), u smislu započinjanja obilježavanja s minimalnim jedinicama na koje se nadodaju veće. Drugi princip razvoja je da su sve ranije inačice CES DTD-ova u najvećoj mogućoj mjeri kompatibilne s novijim inačicama. Na taj se način postiže unatražna kompatibilnost, pa su ranije obilježeni tekstovi uskladivi s novijim DTD-ovima.

CES je primarno namijenjen razmjeni korpusnih podataka. Standard za razmjenu podataka nužno bi trebao biti neovisan od domene, aplikacije i platforme, te zbog toga u najvećoj mogućoj mjeri uopćen. U idealnom slučaju, trebao bi biti izražajan poput bilo kojeg lokalnog oblika (*format*) kako bi bez gubitka informacija omogućio lako prevođenje svih lokalnih oblika u oblik za razmjenu podataka.

CES je SGML aplikacija koja je u skladu s *TEI smjernicama*. Kako je TEI projekt koji se stalno razvija i nadopunjuje, a neki njegovi dijelovi još nisu dovršeni, CES je istodobno poslužio za nadopunjavanje TEI-a. Kao rezultat te činjenice pojavljuju se značajna područja u kodiranju korpusa prema CES-u koja nisu pokrivena TEI smjernicama. TEI su smjernice napravljene da budu uporabljive za što širi skup aplikacija i disciplina, te da budu što uopćenije i fleksibilnije pa zbog toga nisu dovoljno razradile specifično korpusno kodiranje. Većina aplikacija koristi TEI smjernice samo u onim dijelovima gdje one zadovoljavaju njihove potrebe. CES je upravo takva aplikacija i iskorištava one dijelove TEI-a koji su prikladni za kodiranje jezičnih korpusa. Određeni su dijelovi TEI smjernica, ovisno o potrebi, proširivani ili sužavani uvođenjem dodatnih ograničenja. Na primjer, dio koji se odnosi na sadržaj elementa u CES-u je pojednostavljen. TEI elementi nisu preimenovani tamo gdje

[138] CES (1996)

postoji mogućnost zabune, osim tri TEI-specifična elementa koja odražavaju svoju uporabu u CES-a:

- <tei.2> postaje <cesDoc>,
- <teiCorpus.2> postaje <cesCorpus>,
- <teiHeader> postaje <cesHeader>.

Tako bi elementi CES zaglavlja imali sljedeći raspored:[139]

(pr. 35)

```
<cesHeader>
     <fileDesc></fileDesc>
     <encodingDesc></encodingDesc>
     <profileDesc></profileDesc>
     <revisionDesc></revisionDesc>
</cesHeader>
```

CES preporučuje uporabu ISO 8859-X skupa pismena za arapska, ćirilićna, grčka, hebrejska i latinska pisma. Pismena koja se ne mogu dobiti iz preporučenoga skupa pismena rješavaju se pozivanjem na entitete (*entity references*). Preporučuje se uporaba ISO entiteta (*Public Entity Sets*).[140]

U današnje vrijeme mnogo korpusa nastaje iz elektroničkih podatka već kodiranih u nekom konvencionalnom formatu zapisa (npr. HTML, RTF itd.). Budući da se radi o velikim količinama podataka, transformacija u željeni oblik ne smije zahtijevati preveliku količinu ljudskog rada. CES je u tom smislu prvi počeo davati preporuke i postavljati standarde za iskorištavanje takvih tekstovnih izvora čime se izrada korpusa znatno ubrzala i olakšala.

[139] CES (1996)
[140] Više o SGML-u u poglavlju 3.3.1.

3.4.3. XCES[141]

XCES je nastao na osnovi iste arhitekture organizacije podataka kao i CES. Povod za izradu XCES-a bilo je osiguravanje najsuvremenijega oblika zapisa podataka za ANC-a (*American National Corpus*).[142] XCES je standard u nastajanju, te je prirodno da trpi brojne izmjene i nadopune. XCES, poput CES-a daje smjernice za obilježavanje različitih osobina u pisanome i govorenome tekstu, morfosintaktičko i sintaktičko obilježavanje, informacije o sravnjivanju, no velik je dio standarda još uvijek u izradi. Do sada postoje tri dovršena XCES DTD-a (*xcesDoc.dtd, xcesAna.dtd* i *xcesAlign.dtd*), i mogu se preuzeti s Web-adrese:

http://www.cs.vassar.edu/XCES

Iako je nastao iz SGML-a i njegov je podskup, XML (*e**X**tended **M**arkup **L**anguage*) ima brojne prednosti. U prvom se redu misli na XML-ova proširenja kao što je XSL (*e**X**tensible **S**tylesheet **L**anguage*), XSLT, XPath, Xpointer[143] i sl, ali i na prilagođenost XML-a za uporabu u WWW okružju. Zbog njihove sličnosti, konverzija CES DTD-a iz SGML-a u XML je relativno jednostavna, i obuhvaća samo nekoliko manjih sintaktičkih promjena koje ne zadiru u same sadržaje elemenata. Međutim, uz XML je razvijen i *XML shema jezik* (*XML Schema definition language*) koji omogućuje naprednije definiranje sadržaja u odnosu prema XML DTD-u. XML shema jezik uvodi tipove podataka (*datatypes*) za elemente i atribute kojima se mnogo preciznije može određivati sadržaj elemenata, kao i moguće zadane vrijednosti (*default values*) za atribute i elemente. Za sva tri do sada postojeća XCES DTD-a napravljene su pripadajuće XML sheme. Upravo su XML sheme snažno utjecale na razvoj XCES-a, ali i općenito na obilježavanje korpusa.

Primjer teksta obilježena prema XCES standardu može se naći u dodatku B.

[141] Ide, Bonhome, Romary (2000)
[142] Više o ANC-u na adresi: http://www.cs.vassar.edu/~ide/anc/
[143] Više o proširenjima XML-a u poglavljima 3.3.2.7. i 3.3.2.9.

4. Računalnojezikoslovni alati

S razvojem računalne lingvistike narasla je potreba za razvojem alata koji bi trebali obrađivati postojeće resurse, ali i omogućiti kreiranje novih. Većina je alata nastala u okviru znanstveno-istraživačkih institucija za istraživačke potrebe, ali se u zadnje vrijeme pojavljuju i komercijalni proizvodi namijenjeni većem broju korisnika.

Kako je uporaba računalnih korpusa sve više stjecala svoje mjesto nezaobilaznog sredstva za iscrpna jezična proučavanja, postajalo je sve jasnije da lingvisti i računalni stručnjaci trebaju surađivati kako bi razvili što kvalitetnije alate za obrade i analize korpusa.

> While it would be desirable if the linguist and the computer scientist were equally well versed in each other's disciplines, in most instances this is not case (…)[144]

Neka čvrsta ili sveobuhvatna podjela računalnolingvističkih alata danas još ne postoji, i bilo bi pretenciozno pokušati je obaviti u ovome radu. Postojeće podjele najčešće obuhvaćaju samo jedan podskup alata i ne mogu se smatrati sveobuhvatnima. S obzirom na *različitost operacijskih sustava*, samih *namjena* i brojnost računalnolingvističkih alata, njihova bi klasifikacija bila prilično kompleksan zadatak u kojem bi neizbježno dolazilo do miješanja kriterija klasifikacije. Stoga će u ovom radu biti izloženi samo mogući kriteriji klasifikacije.

Jedan od kriterija diobe mogao bi biti na alate koji *uvažavaju jezike za obilježavanje* i one koji ih *ne uvažavaju*. Treba napomenuti da bi se pod jezicima za obilježavanje uzimali u obzir samo oni alati koji uvažavaju standardne jezike za obilježavanje o kojima je bilo riječi (SGML, XML), a nikako brojni komercijalni HTML i slični alati namijenjeni objavljivanju na Internetu. Alati koji ne uvažavaju jezike za obilježavanje obrađuju podatke koji se nalaze u obliku običnoga teksta ili u nekom drugom formatu.

[144] Souter & Atwell (1993):25: Bilo bi poželjno da lingvist i informatičar budu podjednako upućeni u oba područja, što je u praksi rijedak slučaj(…) (prijevod moj)

Podjela bi se mogla zasnivati prema kriteriju *veličine tekstova* koje su alati u stanju obrađivati. Iako bi bilo teško odrediti koja je to veličina teksta danas koja bi razvrstavala alate u pripadajuće kategorije, treba istaknuti da su neki alati namijenjeni obrađivanju manjih tekstova (mjerenih u desetinama ili stotinama tisuća pojavnica), a drugi obrađuju 100-milijunske, pa i veće tekstove ili zbirke tekstova. Kako alati koji obrađuju manje tekstove obradu obavljaju uglavnom na lokalnom računalu, a ovi drugi na poslužniku, možda bi logičniji kriterij za podjelu mogao biti prema *vrsti računala koje obrađuje tekstove*. Alati koji obradu obavljaju na poslužniku u pravilu su znatno robusniji, složeniji, ali i fleksibilniji.

Jedna od važnijih podjela mogla bi biti na one alate koji obavljaju *prepoznavanje i obilježavanje* i na one koji obavljaju *pretraživanje i ekstrakciju* teksta. Kada je riječ o alatima koji obavljaju prepoznavanje i obilježavanje teksta, njihova bi daljnja podjela mogla biti prema *načinu* rada na one alate koji: [145]

1. uvažavaju kognitivnu vjerojatnost (*cognitively plausibile*),
2. ne uvažavaju kognitivnu vjerojatnost (*cognitively inplausibile*).

Prvi se povode za kognitivnim modelom, vodeći računa o tome kako čovjek obavlja neke zadatke. Takvi alati imaju "inteligenciju" za opisivanje ljudskog načina rješavanja zadataka i koriste je kao osnovu za strojno obavljanje "inteligentnih" zadataka. Ti alati rabe kompleksne skupove složenih pravila da bi implicitno znanje izrazili eksplicitno, najčešće u obliku baze znanja. Druga skupina alata koristi kvantitativne podatke kako bi generirali "inteligentno" ponašanje koje bi oponašalo ljudsko ali sasvim drukčijim putem od čovjeka. Kako je čovjek slab izvor kvantitativnih podataka, takvi se alati često koriste jezičnim resursima kako bi priskrbili dovoljnu količinu jezičnih podataka.

Korpusi, kao i ostali jezični resursi pridonose razvoju obiju navedenih skupina alata. Alatima koji uvažavaju kognitivnu vjerojatnost služe kao neiscrpan izvor za prikupljanje svih mogućih jezičnih pravila i situacija koje treba uzeti u obzir. U drugom slučaju, kad se žrtvuje kognitivna vjerojatnost u korist grube sile (*brute force*) matematičkog modeliranja, korpusi su *sine qua non* takvom pristupu, jer najveća snaga korpusa iskazuje se upravo u slučajevima kada je potrebna velika količina stvarnih podataka. Ipak, uporaba statističkog modeliranja ne isključuje automatski uvažavanje kognitivne vjerojatnosti. Žrtvovanje jednog pristupa u korist drugoga

[145] McEnery & Wilson (1996):118

svakako je pitanje stupnja, a ne apsolutnog odbacivanja jednoga pristupa. U realnosti postoji svega nekoliko sustava koji u potpunosti zanemaruju kognitivnu vjerojatnost, jer je doista teško napraviti sustav koji bi modelirao neki jezični podskup a da u njemu nema niti jednoga pravila.

Nadalje, postoje alati koji su *besplatni* (*public domain tools*) i oni koji su *komercijalni*. Toj bi podjeli svakako trebalo pridodati alate koji su besplatni samo za akademske, odnosno istraživačke potrebe (kao npr. CQP[146]).

Unatoč nepostojanju neke čvrste ili standardne podjele računalnolingvističkih alata danas, postoje neke čvrste činjenice koje se odnose na pojedine grupe alata:

- alati koji rade u UNIX/Linux okružju u pravilu istu količinu podataka obrađuju brže od onih koji se zasnivaju na Windows/MacOS okružju,
- alati koji uvažavaju jezike za obilježavanje znatno su uporabljiviji,
- alati koji uvažavaju suvremene standarde u velikoj su prednosti.

No, s obzirom na temu i problematiku ovoga rada, jedna bi od važnih podjela mogla biti na alate koji *ovise o specifičnom jeziku* kojega se obrađuje i na one koji *ne ovise o specifičnom jeziku*. Većina je alata iz prve skupine, pogotovo oni koji automatiziraju obilježavanje (morfosintaktičko, sintaktičko i sl.) napravljena za engleski jezik. Načelno u okviru Europe, što je manji broj govornika nekog jezika, manji je i broj alata specifičnih za jezik. Tako je za strojnu obradu hrvatskoga najveći nedostatak manjak alata upravo iz ove skupine.

Pri razvoju alata pojavljuju se slični problemi kao pri kodiranju korpusa: množenje napora različitih sastavljača, ili grupa sastavljača, mnoštvo inkompatibilnih standarda itd. Stoga se i na ovom području nastojalo pribjeći uspostavljanju međunarodnoga standarda.

[146] Koenig-Baumer (1999): CQP (*Corpus Query Processor*) je specijalizirani alat za pretraživanje korpusa napravljen u lingvističke svrhe. Dio je *IMS Corpus Workbench* skupa alata namijenjenih manipulaciji velikih obilježenih korpusa. CQP često koriste veliki korpusi nastali u okviru akademskih institucija.

4.1. Guidelines for Linguistic Software Development (GLOSIX)[147]

GLOSIX je nastao u sklopu EAGLES projekta i daje osnovne smjernice i preporuke standarda za sve aspekte razvoja jezikoslovnih programa (*software*), prikazivanja podataka (*data representation*), jezikoslovnog obilježavanja (*linguistic annotation*) itd.

Cilj je GLOSIX-a omogućiti razmjenu alata i podataka između različitih istraživača, kompatibilnost među alatima s potencijalno različitim namjenama, te pridonijeti razvoju pouzdanih i prenosivih visokokvalitetnih lingvističkih alata.

Jezikoslovni bi alat trebao posjedovati svojstvo **višestruke uporabljivosti** (*reusability*). Ona obuhvaća nekoliko aspekata koji su navedeni po redoslijedu kojim bi trebali biti primijenjeni, a svaki slijedeći ovisi i nadograđuje se na prethodni:[148]

1. **Uporabljivost** (*Usability*): zapreke koje utječu na smanjenu uporabljivosti mogu biti: slaba dokumentacija ili njezin nedostatak, nepouzdanost, itd.
2. **Prenosivost** *(Portability)*: pod tim se podrazumijeva da alati koji su napravljeni na jednom mjestu mogu odmah biti uporabljeni na drugom mjestu. Kao kratkoročni cilj ponajprije se misli na "slična" okružja (npr. različite inačice Unixa ili Windowsa). Idealna je prenosivost između različitih platformi (Unix, Windows, MacOs) ali to već pripada u dugoročne ciljeve.
3. **Kompatibilnost** (*Compatibility*): alati koji su razvijeni nezavisno trebali bi biti sposobni raditi u istom okružju u cilju izvođenja kompleksnih zadataka. To zahtjeva:
   - da alati mogu komunicirati, tj. rezultati obrade prvog alata mogu biti iskorišteni kao ulaz (*input*) drugome alatu,
   - da su njihove funkcionalnosti komplementarne i koherentne.

   Alati bi trebali biti kompatibilni s podacima i drugim resursima (npr. s leksikonima) koji su pohranjeni u zajedničkom formatu.
4. **Prilagodljivost** (*Extensibility*): podrazumijeva sposobnost alata da se mogu prilagoditi kako bi zadovoljili odgovarajuće potrebe, npr. da se na već postojeći alat može dodati novi dio, da se dio alata može zamijeniti drugim

[147] EAGLES(1996c)
[148] EAGLES(1996c): http://www.lpl.univ-aix.fr/projects/multext/LSD/LSD2.Oview.html#ToC8

dijelom itd. Idealna bi prilagodljivost za lingvističke svrhe bila sposobnost da se isti alat može rabiti za različite prirodne jezike.

4.2. Primjer računalnojezikoslovnoga alata: WordSmith alati *(tools*)

WordSmith tools 3.0 integrirani je i komercijalni paket lingvističkih alata za obradu elektroničkoga teksta. On na jednom mjestu obuhvaća skup alata koje korpusni lingvist u proučavanju jezika najčešće rabi, ali i alate koji su u stanju obavljati složenije obrade, pa i obilježavanje samoga korpusa. Prednost paketa (ili skupa) alata je što koriste iste ulazne podatke (tekstove) pohranjene samo na jednom mjestu, pa pojedini alati mogu rabiti rezultate obrade ostalih alata.

Program je izvorno namijenjen za rad pod *Windows 3.1x, 9.x* i *NT* operacijskim sustavima ali postoje emultori i za *Apple Mac* i *Unix* operacijske sustave. Minimalni zahtjevi računalnih resursa za korištenje programa svakako su vrlina:

- najmanje 4 MB RAM,
- najmanje 5 MB prostora na tvrdom disku,
- 386, ili noviji procesor.[149]

Program se jednostavno instalira pokretanjem samoraspakirajuće arhive. Već nakon tog koraka moguće je odabrati bilo koji alat – važno je to napomenuti kao pozitivnu osobinu – u grafičkom okružju. Svi su alati i pomagala (*utilities*) dohvatljivi iz glavnoga nadzornika alata (*Wordsmith Tools Controller*) što dodatno olakšava uporabu, osobito u slučajevima kad se istodobno obavlja nekoliko obrada različitim alatima nad istim tekstom. Kvaliteta ovoga skupa alata rezultat je činjenice što je autor Mike Scott izniman lingvistički ali i informatički stručnjak.

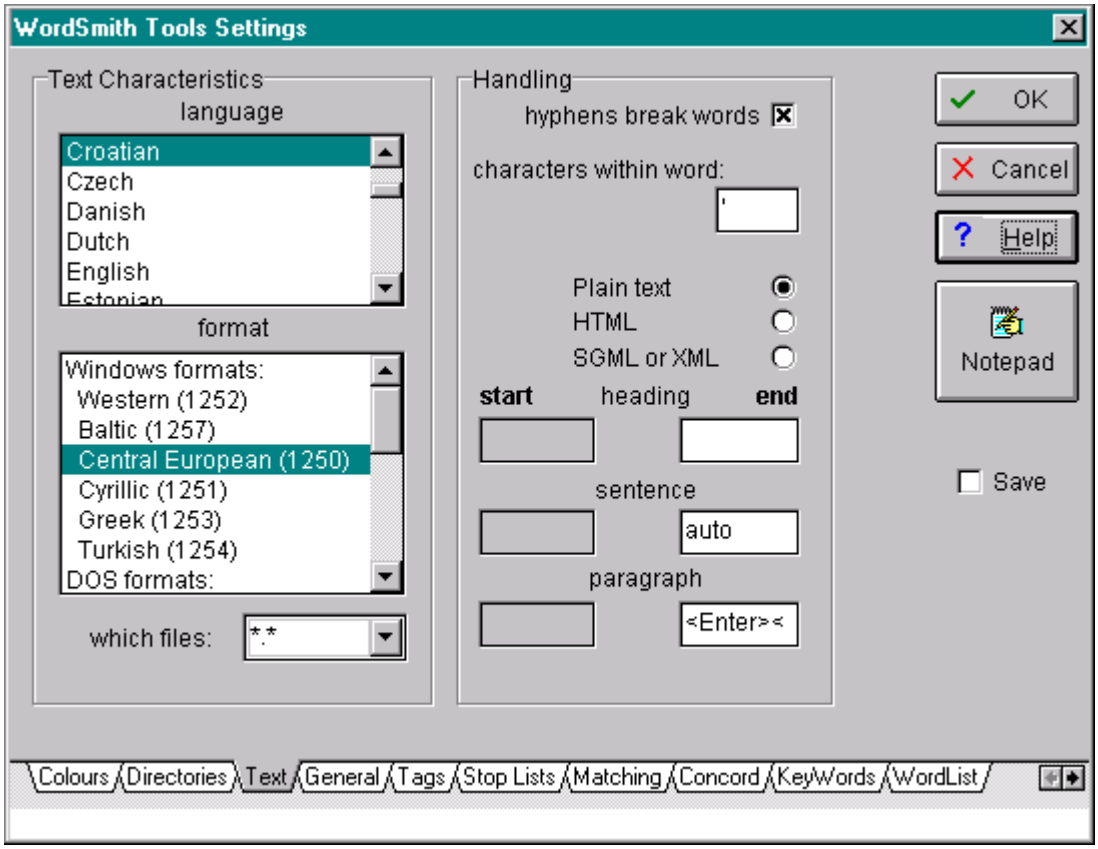
S obzirom da se funkcije i način rada pojedinih alata najzornije mogu prikazati na konkretnim primjerima, ogledni će primjerci pojedinih vrsta obrade biti prikazani na knjizi Viktora Žmegača *Bečka moderna*.[150]

---

[149] Scott (1999)

4.2.1. Glavni nadzornik (*the Controller*)

Iz glavnoga je nadzornika moguće nadzirati sve okupljene alate. Pored toga, u glavnom se nadzorniku odabiru tekstovne datoteke koje će se obrađivati, te se definiraju njihove postavke (*settings*).



Slika 4: Postavke ulaznoga teksta smještene u glavnom nadzorniku

Jedna od najvažnijih postavki da bi tekst bio ispravno obrađen je određivanje formata ulaznoga teksta. Ulazni tekstovi mogu se nalaziti u slijedećim formatima:

- običan tekst (*plain text*),
- HTML,
- SGML ili XML.

Format se ulaznoga teksta (tekstova) određuje u srednjem okviru. U gornjem slučaju, budući da se tekst nalazi u obliku običnoga teksta, odabran je *Plain text* ulazni format, te su postavljeni jezik i kodna stranica prikladni za hrvatski tekst. Ukoliko je tekst obilježen, oznake se mogu prilagoditi granicama zaglavlja, rečenice ili odlomka. O ostalim će postavkama biti riječi u poglavljima o alatima na koje se same postavke odnose.

[150] Žmegač (1998)

Nakon definiranja postavki slijedi odabir i spremanje (*storing*) ulaznog teksta (ili tekstova) koji će se obrađivati. Moguće je odabrati i spremiti jednu, dvije ili nekoliko ulaznih datoteka, ili pak datoteke iz cijelog direktorija (s mogućnošću odabira poddirektorija).

4.2.2. Popis pojavnica (*WordList*)

WordList je alat *WordSmith toolsa* koji generira popise pojavnica iz odabranih tekstnih datoteka. **Popis** je **pojavnica** lista u kojoj je svaka pojavnica popraćena podatkom o svojoj čestoti (frekvenciji).[151] Popis može sadržati i indeks gdje su uz svaku pojavnicu prisutne i reference na ona mjesta gdje se pojavljuju u tekstu. Popis pojavnica najčešće je razvrstan abecednim ili čestotnim redoslijedom. Pojavnice također mogu biti i odostražno abecedno razvrstane, tj. po svojim završecima. Iako postoji, rijedak je slučaj da je kriterij razvrstavanja pojavnica po broju slova pojavnice.

WordList - [new wordlist (F)]

File Settings Comparison Index Window Help

N	Word	Freq.	%	Lemmas
1	U	2,726	3.89	
2	I	2,500	3.57	
3	JE	2,093	2.99	
4	SE	1,129	1.61	
5	DA	933	1.33	
6	NA	885	1.26	
7	SU	639	0.91	
8	O	541	0.77	
9	S	442	0.63	
10	KOJI	440	0.63	
11	ZA	392	0.56	
12	NE	384	0.55	
13	OD	370	0.53	
14	A	345	0.49	
15	KAO	336	0.48	
16	ŠTO	331	0.47	
17	TO	281	0.40	
18	ILI	263	0.38	
19	NIJE	259	0.37	
20	KOJE	257	0.37	
21	IZ	251	0.36	
22	KOJA	251	0.36	
23	SAMO	219	0.31	
24	PREMA	203	0.29	

Slika 5: Popis pojavnica razvrstan po čestotnom redoslijedu

Svrha je popisa pojavnica višestruka:

- proučavanje vokabulara koji je korišten u tekstu,
- uspoređivanje čestote pojavnica u tekstovima različitih žanrova,
- identifikaciju zajedničkih grozdova (*cluster*) pojavnica i sl.

Specifičnost su *WordSmith* alata grozdovi (*clusters*). Oni predstavljaju n-arne nizove pojavnica koje slijede jedna drugu u tekstu osim onih razdvojenih znakovima interpunkcije. Moglo bi se reći da u usporedbi s kolokacijama[152] koje upućuju na lingvističku pozadinu, grozdovi predstavljaju puku informatičarsku sliku odnosa između pojavnica.

> The term phrase is not used here because it has technical senses in linguistics which would imply a grammatical relation between the words in it. [153]

Promotrimo grozdove na primjeru rečenice:

(pr. 36)

```
Tko rano rani, dvije sreće grabi.
```

Grozdovi veličine dviju pojavnica izgledali bi ovako:

```
Tko rano
rano rani
dvije sreće
sreće grabi
```

(“rani, dvije” nije grozd zbog zareza!)

Grozdovi ne prelaze znakove interpunkcije, a moguće je odabrati grozd dužine do 8 pojavnica.

U ovome se alatu također može obavljati usporedba dvaju ili više popisa pojavnica. Ta je procedura iznimno korisna u stilističkom proučavanju jezika. Usporedbom se mogu uočiti one pojavnice koje se znakovito češće pojavljuju u jednom tekstu nego u drugom.

Iz modula popisa pojavnica može se obavljati i lematizacija korpusa. Moguća su dva načina: ručni i automatski. Ručnom se lematizacijom pojavnice najprije ručno obilježavaju, a potom združuju pod istu lemu. Preporučljivo je pri obilježavanju koristiti abecedni popis jer se velik broj pojavnica koje pripadaju pod istu lemu nalazi

---

[151] Lawrer & Dry (1998):115

[152] Više o kolokacijama u poglavlju 4.2.4.

[153] Scott (1999): Termin fraza se ovdje ne rabi, jer ima u lingvistici više tehnički smisao koji bi upućivao na gramatički odnos između pojavnica. (prijevod moj)

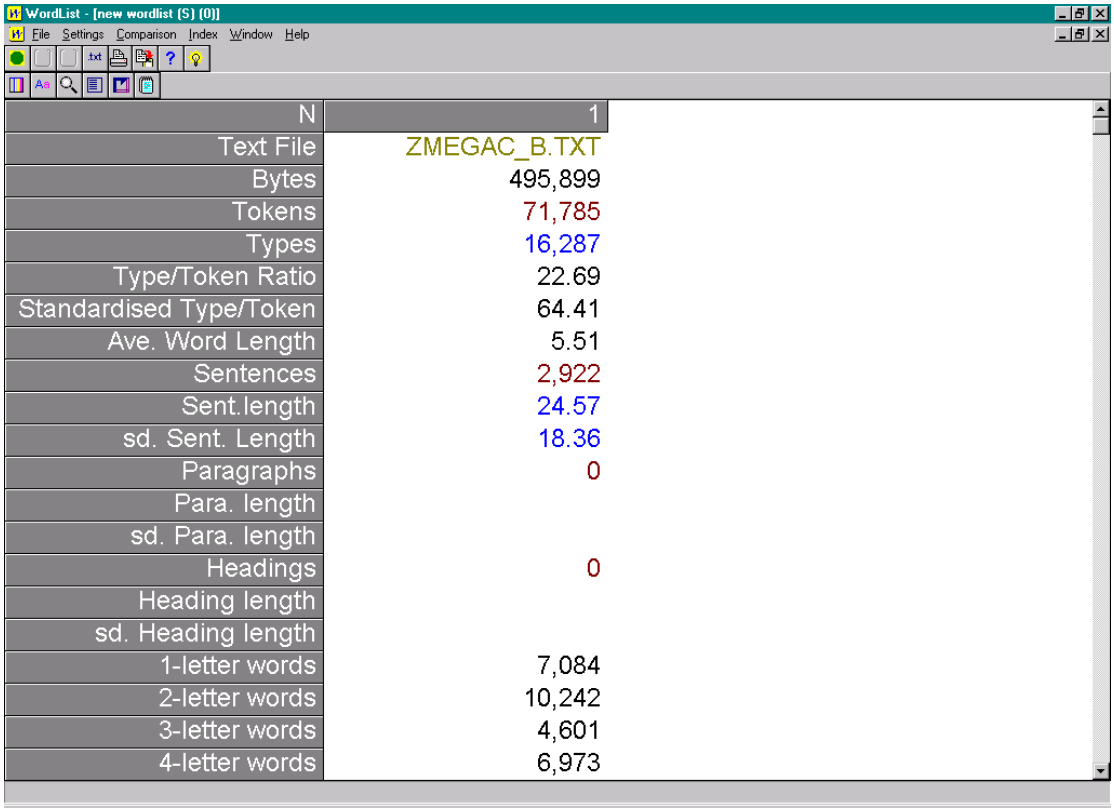
u bližoj okolini. Ukoliko se lematizacija obavlja automatski, ulazni parametar za lematizaciju tekstovna je datoteka koja sadrži popis lema u obliku:[154]

(pr. 37)

```
...
...
knjiga -> knjiga,knjige,knjizi,knjigu,knjigo,knjigom,knjigama
itd.
```

Program nakon pokretanja u popisu pojavnica uspoređuje pojavnice iz teksta s onima u datoteci, te ih smješta uz istu lemu.

Iz modula se popisa pojavnica generira i opća statistika teksta:



N	1
Text File	ZMEGAC_B.TXT
Bytes	495,899
Tokens	71,785
Types	16,287
Type/Token Ratio	22.69
Standardised Type/Token	64.41
Ave. Word Length	5.51
Sentences	2,922
Sent.length	24.57
sd. Sent. Length	18.36
Paragraphs	0
Para. length	
sd. Para. length	
Headings	0
Heading length	
sd. Heading length	
1-letter words	7,084
2-letter words	10,242
3-letter words	4,601
4-letter words	6,973

Slika 6: Statistički podaci o tekstu

Statistički podaci odnose se na kvantitativne informacije kao što su:

- **Tokens**: broj pojavnica u tekstu
- **Types**: broj različnica u tekstu
- **Type/Token Ratio**: omjer između broja različnica i pojavnica pomnožen sa 100. Što tekst sadrži "bogatiji" vokabular, ovaj omjer ima veću vrijednost, ali omjer različnica/pojavnica često oscilira i u ovisnosti od dužine teksta. Najvažnija je

[154] Scott (1999)

primjena kod usporedbe tekstova, ali zbog ovisnosti o dužini teksta omjer će biti relevantan samo ako su tekstovi približno jednake dužine.[155] Treba uzeti u obzir da je u našem slučaju taj omjer izračunat na visokoflektivnom jeziku kakav je hrvatski

- **Standardised Type/Token**: računa se omjer različnica/pojavnica svakih *n* pojavnica iz teksta (gdje je n moguće definirati; zadana postavka: n = 1000), te se zatim izračunava prosjek na taj način svih dobivenih intervala
- **Average Word Lenght**: broj je koji predstavlja prosječnu dužinu pojavnice mjerenu brojem pismena
- **Sentences**: broj rečenica izračunatih prema pismenima za ograničavanje rečenica
- **Paragraphs**, **Headings**: broj odlomaka i poglavlja je nula zato što ogledni uzorak teksta nema obilježene odlomke
- **1-letter words**: broj riječi dužine jednoga slova, n-letter words: broj riječi dužine n slova.

Iz ovoga se modula može napraviti indeksirana lista (*Index list*) svih pojavnica, tj. popis pojavnica s informacijom o mjestu pojavljivanja svake posebno. Ona se izrađuje relativno sporo u odnosu na ostale obrade, i izvodi se u tri faze:

1. odabiru se sve različnice iz liste pojavnica,
2. obrađuju se sve različnice s visokom čestotom koju postavlja sam korisnik u postavkama,
3. obrađuju se sve ostale različnice.

Indeksirani popis pojavnica znatno ubrzava obrađivanje konkordancija (jer se umjesto sekvencijalnog pretraživanja koristi indeksirano pretraživanje) ali omogućuje i lakše izračunavanje uzajamne obavijesnosti.

Uzajamna obavijesnost, UO (*Mutual Information, MI*) izračunava se uspoređivanjem vjerojatnosti supojavljivanja dviju pojavnica bilo koje jezične jedinice (ili bilo kojih vrijednosti gramatičkih kategorija u korpusu, npr. vrsta riječi) zajedno s vjerojatnošću da se pojave odvojeno.[156] Što je njihovo supojavljivanje češće to je veća i uzajamna obavijesnost. Moglo bi se reći da je UO okvirna mjera koja opisuje koliko jedna pojavnica govori o drugoj.[157] Na primjer, pojavnica *kupaći* se vrlo često nalazi u blizini pojavnice *kostim* jer su dio jedinice koje se sastoje od više

[155] Lawrer & Dry (1998):129
[156] McEnery & Wilson (1996):71
[157] Manning & Schütze (1999):178

riječi (*multi-word unit, MWU*). MWU su neprekinute kolokacije s visokom čestotom čiji je redoslijed jedinica najčešće fiksan, a ovise o domeni kojoj pripadaju.[158] Uzajamna se obavijesnost ne mora odnositi samo na one pojavnice koje se nalaze neposredno uz promatranu, već i na pojavnice koje mogu biti udaljene. S obzirom da je izračunavanje UO-a svih pojavnica po opsegu obrade zahtjevna operacija i zauzima mnogo vremena, najčešće je se izračunava samo za pojavnice s visokom čestotom, odnosno prema istraživačevoj potrebi.

Pojavnice koje iz određenog razloga nisu prikladne, ili ih istraživač ne želi uvrstiti u analizu upisuju se u zaustavan popis (*stop list*). **Zaustavan** je **popis** "negativan rječnik", popis riječi koje se neće uzimati u obzir pri obrađivanju.[159] On je obična tekstovna datoteka (s datotečnim nastavkom **.stp*) koja sadrži popis pojavnica što se isključuju iz obrade. Postavke za definiranje zaustavnoga popisa nalaze se u postavkama glavnoga nadzornika. Pojavnice se iz zaustavnoga popisa u datoteku unose velikim slovima, npr:

```
I,U,JE itd.
```

Popis pojavnica može biti iskorišten kao ulazni parametar za generiranje ključnih riječi (*keywords*). O ključnim će riječima bit više u idućem poglavlju.

4.2.3. Ključne riječi (*KeyWords*)

Ključne su riječi one pojavnice koje imaju neuobičajeno visoku čestotu u odnosu prema nekom normativu. Normativ je obično referentni korpus nekoga jezika (npr. za engleski bi to mogao biti BNC, za hrvatski HNK). Do ključnih se riječi dolazi usporedbom dva popisa pojavnica. Ključne su riječi vrlo koristan način za određivanje karakteristika ili žanra teksta. Promatranjem samog čestotnoga popisa pojavnica nekoga teksta nije moguće doći do ključnih riječi jer riječi s najvećom čestotom obično nose nisku obavijesnost (*i, u, je* itd).

Također je moguće uspoređivati popis pojavnica velike količine novinskih tekstova (koja u ovom slučaju može poslužiti kao norma) s popisom pojavnica jednog novinskoga članka. One pojavnice iz članka koje imaju neuobičajeno visoku čestotu u

[158] Merkel & Andersson (2000):738
[159] Glossary (1999)

odnosu prema normativu predstavljaju ključne riječi promatranoga članka. Ukoliko se promatra više tekstova koji pripadaju istoj domeni (npr. uzorak od 500 tekstova poslovnih izvještaja) korisno je izraditi popis "ključnih-ključnih riječi" (*key-keywords*). On je popis ključnih riječi napravljen na osnovi svih uzoraka teksta. Ključne riječi napravljene na osnovi jednoga teksta odražavaju stanje samo jednog dokumenta (npr. učestali nazivi poduzeća samo u tom tekstu i njihovih proizvoda), a ne opći popis ključnih riječi svih 500 tekstova. Dakle, tim bi se načinom dobile stvarne ključne riječi kao npr. *prodaja, porast, zaposlenik* itd, a ne one specifične za samo jedan dokument.

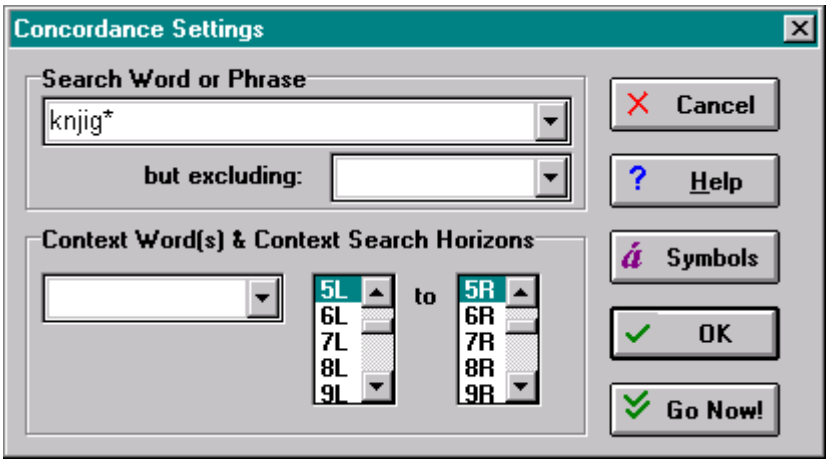
Potencijalna je uporaba i u ostalim jezikoslovnim granama kao što su učenje jezika, stilistika[160], forenzična lingvistika[161] itd.

4.2.4. Alat za konkordancije (*Concord tool*)

Današnji korpusi prevelikoga su opsega da bi se mogli pretraživali bez pomoći računala. Stoga se alat za konkordancije nalazi u samom središtu korpusne lingvistike i osnovni je alat korpusnoga jezikoslovca.[162] Osnovni je cilj alata za konkordancije omogućiti uvid u mnoštvo primjera određene pojavnice ili fraze u okolinama u kojima se one pojavljuju. Uvidom u okolinu često je moguće pronaći neko značenje ili način uporabe riječi koji nije opisan u rječniku ili uočiti pravilnost koja do tada nije primijećena.

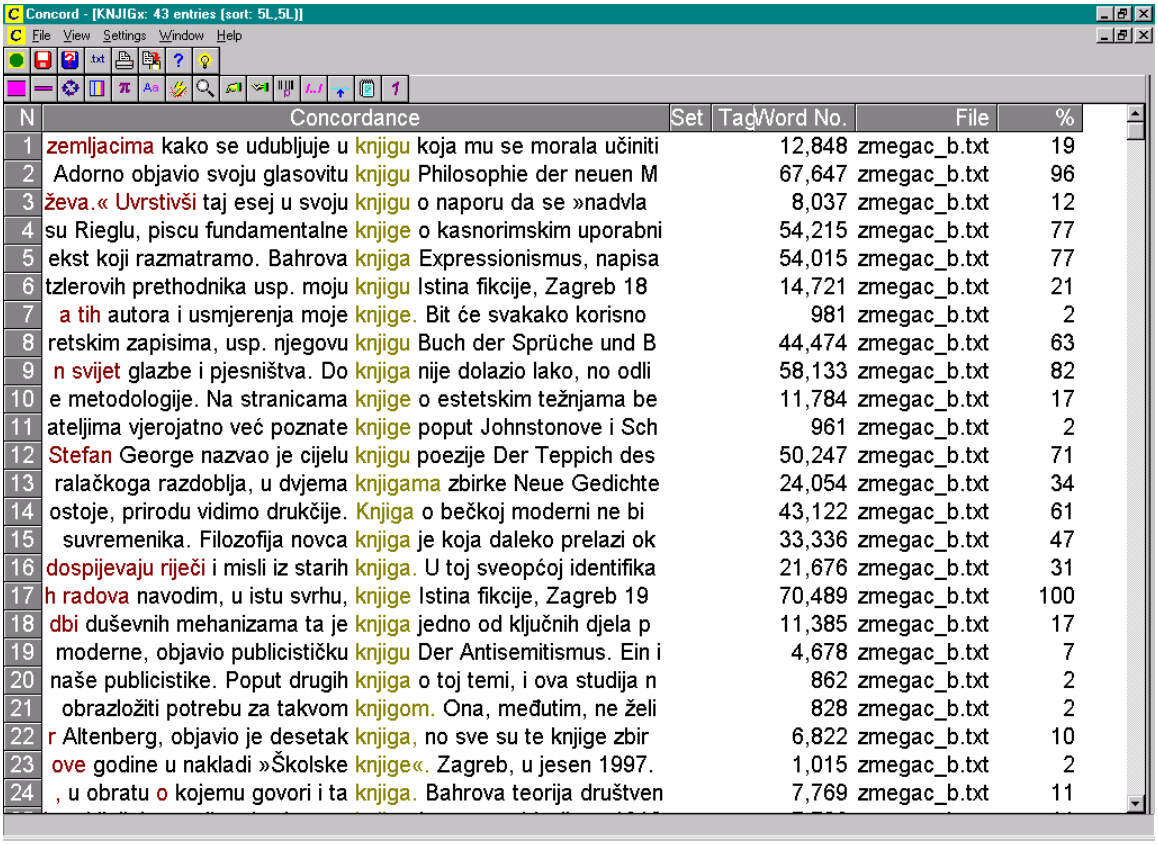
U okviru za postavke konkordancija unosi se pojavnica(e) za pretraživanje, a mogu se isključiti pojavnica(e) (korisno je ako se koriste zamjensko pisme, npr. isključi *knjigovodstvo*!) i odrediti pojavnica(e) koja mora biti sadržana u zadanom kontekstu:

[160] Scott (1999)
[161] cf. Gibbons (1994):27: Forensic Linguistics is the application of our linguistic and psychology knowledge of language to the legal and investigative domain. (Forenzična lingvistika primjenjuje jezično i psihologijsko znanje o jeziku na pravno i istražiteljsko područje. (prijevod moj))
[162] CCL (2000)



Slika 7: Okvir postavki konkordancija

Prikaz bi konkordancija pojavnice za pretraživanje (*search word*) *knjig** izgledao:



N	Concordance	Set	Tag	Word No.	File	%
1	zemljacima kako se udubljuje u knjigu koja mu se morala učiniti			12,848	zmegac_b.txt	19
2	Adorno objavio svoju glasovitu knjigu Philosophie der neuen M			67,647	zmegac_b.txt	96
3	ževa.« Uvrstivši taj esej u svoju knjigu o naporu da se »nadvla			8,037	zmegac_b.txt	12
4	su Rieglu, piscu fundamentalne knjige o kasnorimskim uporabni			54,215	zmegac_b.txt	77
5	ekst koji razmatramo. Bahrova knjiga Expressionismus, napisa			54,015	zmegac_b.txt	77
6	tzlerovih prethodnika usp. moju knjigu Istina fikcije, Zagreb 18			14,721	zmegac_b.txt	21
7	a tih autora i usmjerenja moje knjige. Bit će svakako korisno			981	zmegac_b.txt	2
8	retskim zapisima, usp. njegovu knjigu Buch der Sprüche und B			44,474	zmegac_b.txt	63
9	n svijet glazbe i pjesništva. Do knjiga nije dolazio lako, no odli			58,133	zmegac_b.txt	82
10	e metodologije. Na stranicama knjige o estetskim težnjama be			11,784	zmegac_b.txt	17
11	ateljima vjerojatno već poznate knjige poput Johnstonove i Sch			961	zmegac_b.txt	2
12	Stefan George nazvao je cijelu knjigu poezije Der Teppich des			50,247	zmegac_b.txt	71
13	ralačkoga razdoblja, u dvjema knjigama zbirke Neue Gedichte			24,054	zmegac_b.txt	34
14	ostoje, prirodu vidimo drukčije. Knjiga o bečkoj moderni ne bi			43,122	zmegac_b.txt	61
15	suvremenika. Filozofija novca knjiga je koja daleko prelazi ok			33,336	zmegac_b.txt	47
16	dospijevaju riječi i misli iz starih knjiga. U toj sveopćoj identifika			21,676	zmegac_b.txt	31
17	h radova navodim, u istu svrhu, knjige Istina fikcije, Zagreb 19			70,489	zmegac_b.txt	100
18	dbi duševnih mehanizama ta je knjiga jedno od ključnih djela p			11,385	zmegac_b.txt	17
19	moderne, objavio publicističku knjigu Der Antisemitismus. Ein i			4,678	zmegac_b.txt	7
20	naše publicistike. Poput drugih knjiga o toj temi, i ova studija n			862	zmegac_b.txt	2
21	obrazložiti potrebu za takvom knjigom. Ona, međutim, ne želi			828	zmegac_b.txt	2
22	r Altenberg, objavio je desetak knjiga, no sve su te knjige zbir			6,822	zmegac_b.txt	10
23	ove godine u nakladi »Školske knjige«. Zagreb, u jesen 1997.			1,015	zmegac_b.txt	2
24	, u obratu o kojemu govori i ta knjiga. Bahrova teorija društven			7,769	zmegac_b.txt	11

Slika 8: Ispis konkordancija pojavnice za pretraživanje *knjig**

Unesena pojavnica za pretraživanje nalazi se između lijeve i desne okoline, a ona se u kontekstu konkordancija naziva **stožernica** (*headword*). Uobičajeno je da su stožernice u konkordancijama razvrstane onim redoslijedom kojim se pojavljuju u tekstu, međutim moguće ih je razvrstati po brojnim parametrima. Desno-redana konkordancija je poredana abecednim redoslijedom stožernice i pojavnica koje slijede nakon nje. Takvo redanje može poslužiti za analizu onih slučajeva gdje stožernica

otvara neku frazu, ili je sadržana u njoj. Lijevo- redana konkordancija je poredana abecednim redoslijedom stožernice i pojavnice koja joj prethodi. Različitim se redanjem konkordancija omogućuje preslagivanje podataka, te tako olakšavaju brojne analize kao npr. pronalaženje karakterističnih leksičkih ili frazeoloških obrazaca.

WordSmith alat za konkordancije ima iznimno bogate mogućnosti odabira redanja. Moguće je postaviti i po nekoliko kriterija istovremeno po kojima će se ono provoditi.

Sintaksa za pretraživanje tj. postavljanje upita nad korpusom nije uvijek jednaka kod svih alata za konkordancije, ali najčešće se rabe slični simboli. Zadana postavka (*default setting*) za pretraživanje kao i kod većine alata postavljena je na neosjetljivost za mala i velika slova (*case-insensitive*). Sintaksa je za pretraživanje koju koristi ovaj alat sljedeća:

Simbol	Značenje	Primjer
*	bilo koji niz	knjig*
?	bilo koje pisme (uključujući interpunkciju)	knj?ga
^	bilo koje slovo abecede	^njiga
==	podešava osjetljivost za mala i velika slova	==Knjiga==
/	razdvaja alternativne pojavnice za pretraživanje	knjiga/knjigu

Tablica 4

Na konkretnim primjerima zadana pojavnica za pretraživanje dala bi sljedeće rezultate:

Pojavnica za pretraživanje	Pronalazi
knjiga	*Knjiga*, *knjiga*, *knjIGa* i sl.
knjig*	*knjiga*, *knjigama*, *knjigom* i sl.
k*ga	*koga*, *knjiga*, *kruga* i sl.
poput * knjiga	*poput drugih knjiga* i sl.
knjig?	*knjigu*, *knjige*, *knjiga* i sl.
knjig^	*knjigu*, *knjige*, *knjiga* i sl.
==knjiga==	*knjiga* (ali ne i *Knjiga* ili *KNJIGA*)
knjiga/djelo	*knjiga* ili *djelo*

Tablica 5

Pretraživanje je moguće ograničiti određivanjem onih pojavnica koje *moraju* ili *ne smiju* biti unutar zadane okoline stožernice. Na primjer, moguće je pretraživati pojavnicu *knjiga* uz uvjet da pojavnica *žalb** mora biti u definiranoj okolini. Za korištenje rezerviranih pismena u pretraživanju potrebno je rabiti znakove navoda.

Dobivene je konkordancije moguće pohraniti u internom formatu (*.*cnc*) ili u formatu tekstne datoteke (*.*txt*). Konkordancije se mogu ispisivati s kolokacijama, ali u tom se slučaju obrada usporava. **Kolokacije** su karakteristična supojavljivanja obrazaca pojavnica.[163] One su kombinacije stožernice i onih pojavnica koje se pojavljuju u njezinoj bližoj okolini. Njihova primjena i važnost značajna je u mnogim granama lingvistike, kao npr. leksikografiji, frazeologiji, učenju jezika, strojnoj obradi jezika (izradi vjerojatnosnog označivača), psiholingvistici (npr. za potkrjepu tvrdnje da se mentalni leksikon ne sastoji samo od pojedinih riječi, nego i od većih, frazeoloških jedinica) ili strojnom prevođenju itd. U postavkama glavnog nadzornika definira se vrijednost minimalne čestote da bi kolokacija bila relevantna. Ispisani broj kolokacija ovisit će i o postavkama lijeve i desne okoline tražene pojavnice koja će se uzimati u obzir za ulazak u kolokaciju. Zadana je postavka 5 pojavnica s obje strane (5,5), dok je maksimalna moguća postavka 25 pojavnica. Pored toga, omogućene su brojne postavke za kolokacije kao što su postavljanje minimalne čestote ili minimalnog broja pismena u pojavnici da bi bila uračunata kao kolokacija.

4.2.5. Pomagala (*Utilities*)

U ovom se modulu nalaze pomagala koja se često rabe u obradi ili pripremi obrade teksta. Modul sadrži nekoliko pomagala:

- **splitter**: pomagalo koje razdvaja velike datoteke u manje. Moguće je ubaciti simbol za kraj pojedinog odsječka teksta koji će program prepoznati i na tom će mjestu započeti nova tekstovna datoteka.
- **text converter**: višenamjensko pomagalo koje se koristi za nekoliko svrha: uređivanje (*edit*) tekstova, preimenovanje tekstovnih datoteka, prebacivanje datoteka u drugi direktorij ukoliko sadrže određenu pojavnicu ili frazu itd. Glavna

[163] McEnery & Wilson (1996):71

je svrha ovog pomagala zamjena nizova pismena u tekstovnoj datoteci (*search & replace*) i preoblikovanje tekstovnih datoteka u određeni oblik.

- **viewer**: pomagalo koje služi pregledavanju tekstova u različitim oblicima, a može se koristiti i za sravnjivanje tekstova.
- **aligner**: sravnjuje rečenice prevedenih tekstova iz dviju datoteka na razini rečenice.

5. Hrvatski računalni korpusi i računalnojezikoslovni alati

Premda je sastavljanje korpusa u Hrvatskoj započelo neznatno nakon pojave *Brown* korpusa, današnja je korpusna lingvistika u Hrvatskoj u zaostatku za suvremenim svjetskim dostignućima. Glavni je razlog tome zamrzavanje međunarodne znanstvene suradnje uzrokovane ratom od 1991. do 1997. godine, pa je odsutnost Hrvatske iz nekoliko međunarodnih projekata rezultirala stagnacijom i nemogućnošću praćenja razvitka korpusnolingvističkih istraživanja upravo u trenutku (početak 90-ih godina) kad je korpusna lingvistika doživjela svoj procvat.[164] Većina hrvatskih korpusa sastavljena je u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu, pa se ta ustanova može smatrati referentnom za hrvatsku korpusnu lingvistiku.

5.1. Kronološki pregled razvoja hrvatskih jezičnih korpusa i alata

Zanimanje za sastavljanje i istraživanje korpusa, u usporedbi s ostalim zemljama u Europi u Hrvatskoj se javilo prilično rano. Prvi je hrvatski korpus 1961. godine sastavio Ivan Furlan u svojoj disertaciji *Raznolikost rječnika i struktura govora*.[165] Korpus je obradio frekvencijski bez uporabe računala. Iako je opseg tog korpusa svega 100.000 pojavnica, važan je kao prvi korak u sastavljanju korpusa kod nas.

Prvi je računalno i korpusnom metodologijom obrađen tekst na hrvatskome jeziku ep Ivana Gundulića "Osman".[166] Obradio ga je Željko Bujas za boravka na

[164] Tadić (1997):393
[165] Moguš, Bratanić, Tadić (1999):5
[166] Tadić (1997):388

Sveučilištu u Austinu (SAD), 1967. godine. Taj je tekst popraćen frekvencijama i konkordancijama.[167]

Među važnije događaje na samim počecima korpusne lingvistike u Hrvatskoj svakako spadaju lingvistički projekti koji su se izvodili u Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu krajem šezdesetih godina prošloga stoljeća. Metodologijom rada, ti su projekti u potpunosti zasnovani na načelima korpusne lingvistike tog vremena. Pod vodstvom Rudolfa Filipovića, 1968. godine, u Zavodu započinje značajan projekt pod nazivom *Yugoslav Serbo-Croatian -- English Contrastive Project* u kojem se po prvi puta u nas pokreće opsežna računalna obradba korpusa. Kontrastivna istraživanja engleskoga i hrvatskoga zasnivala su se na analizi polovice *Brown* korpusa. Posebno je zanimljiva činjenica da je za potrebe istraživanja engleski korpus kodiran prema vrsti riječi i sintaktičkim funkcijama što je u to vrijeme bila još nepoznata praksa u svijetu.[168] Nakon što je korpus morfosintaktički obilježen i preveden na ondašnju inačicu hrvatskoga jezika, s engleskim je izvornikom (američka inačica) napravljen i prvi hrvatski paralelni korpus. Na temelju tih prijevoda napravljena je konkordancija s morfosintaktičkim kategorijama kao stožernicama i dvojezična rečenična kartoteka s pomoću koje se moglo pretraživati i engleski i hrvatski prijevod. To je ujedno bila i prva uporaba računala u kontrastivnoj lingvistici u svijetu.[169]

Ranih su sedamdesetih godina osim rada na dvojezičnom englesko-hrvatskom paralelnom korpusu izrađene i konkordancije djela nekoliko starijih hrvatskih pisaca.[170]

Prihvaćajući i nastavljajući Bujasov rad, pod vodstvom Milana Moguša pokrenut je dalekosežni projekt *Kompjutorska analiza tekstova starije hrvatske književnosti*. U Zavodu su od 1970. do 1981. godine napravljene konkordancije djela mnogih hrvatskih književnika. Tada je hrvatska korpusna lingvistika bila u potpunosti ravnopravna suvremenim svjetskim korpusnim dostignućima.[171]

Pod istim je vodstvom 1976. godine pokrenut projekt *Korpus suvremenog hrvatskog književnog jezika* s primarnim ciljem sastavljanja *Jednomilijunskoga korpusa hrvatskoga književnog jezika* znanog i pod nazivom *Mogušev korpus*. To je

[167] Bujas (1975)
[168] Bratanić (1991):149
[169] cf. Tadić (1997):389
[170] Bratanić (1991):149
[171] Tadić (1997):390

prvi korpus u hrvatskoj lingvistici sastavljan s namjerom da bude reprezentativan. Na osnovi tog korpusa izrađen je *Hrvatski čestotni rječnik*.[172]

Potkraj sedamdesetih godina pod vodstvom Dubravka Škiljana na Odsjeku za opću lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu provodila su se kvantitativna istraživanja jezika zagrebačkoga tiska.[173]

Korpus dnevnih novina opsega oko 130.000 pojavnica sedamdesetih je godina sastavio Zorislav Šojat,[174] s ciljem izrade čestotnoga rječnika.

Korpus sveukupnoga djela Ivana Gundulića dovršen je 1989. godine u Zavodu za lingvistiku, ali i na Odsjeku za informacijske znanosti Filozofskoga fakulteta Sveučilišta u Zagrebu pod vodstvom Damira Borasa. Taj je korpus djelomično obilježen.[175]

Sredinom devedesetih na Odsjeku za informacijske znanosti Filozofskoga fakulteta Sveučilišta u Zagrebu sastavljen je *Korpus tekstova udžbenika za osnovne i srednje škole u Republici Hrvatskoj* od oko milijun pojavnica.[176] Taj je rad značajan jer su na njemu rađena prva istraživanja za izradu vjerojatnosnog označivača, alata za lematizaciju (SOLAH, ***S**ustav za **o**značavanje, **l**ematizaciju i **a**nalizu tekstova **h**rvatskoga jezika*) i alata za segmentaciju rečenica[177] za hrvatski jezik.

5.2. Jednomilijunski korpus hrvatskoga književnoga jezika

Jednomilijunski se korpus hrvatskoga književnoga jezika počeo sastavljati 1976. u Zavodu za lingvistiku pod vodstvom Milana Moguša, pa je u javnosti poznat i pod nazivom *Mogušev korpus*. Sadrži 1.001.748 pojavnica tekstova objavljenih između 1935. i 1978. godine, a sastavljen je od pet potkorpusâ.[178] Iako je izrada korpusa počela znatno prije prihvaćanja SGML-a kao standarda za obilježavanje tekstova, bio je postavljen zahtjev za čuvanjem podatka iz kojeg dijela korpusa svaka pojavnica dolazi. U tu je svrhu korištena shema obilježavanja prema kojoj se obilježavao početak svakog uzorka u svakom od pet potkorpusâ, kao i relevantni

[172] Moguš, Bratanić, Tadić (1999):5
[173] Bratanić (1991):149
[174] Šojat (1976)
[175] Boras (1998):50
[176] Boras (1998):48
[177] Više o SOLAH-u u Boras (1998) i Žubrinić (1995)

dijelovi uzorka. Sličan je način obilježavanja korišten i u popularnom programu tog vremena COCOA.[179] Primjer sheme obilježavanja, objašnjenje njezinih dijelova kao i njezina primjena mogu se naći u dodatku C.

Glavni je cilj obrade korpusa izrada čestotnoga rječnika koji uključuje postupke prethodne izrade:

1. abecednoga rječnika,
2. čestotnoga rječnika pojavnica,
3. konkordancija pojavnica,
4. lematizacije pojavnica uz pomoć programa za konkordancije.

Ogledni rezultati postavljenih ciljeva pod 1.) i 2.) kao i ispis konkordancija, te izgled radnoga zaslona za lematizaciju također se nalaze u dodatku C.

Jedna je od bitnih kvaliteta ovoga korpusa njegova djelomična lematiziranost. Svaka je pojavnica lematizirana izravnim uvidom u okolinu u kojoj se u korpusu pojavila. Lematizacija je rađena na način da se svaka lema upisuje samo jedanput u lemarij[180] (popis lema), a zatim se uspostavlja odnos između već postojeće leme i pojavnice koja se ima lematizirati. Za samu lematizaciju izrađen je program[181] koji iz tekstova u bazi omogućuje trenutno dobivanje konkordancije za svaku pojavnicu pri čemu se neposrednim uvidom u okolinu uspostavlja odnos između leme i istopisne pojavnice.

Kao jedan od rezultata obrade ovoga korpusa objavljen je u tiskanom obliku *Hrvatski čestotni rječnik*, koji je zasigurno najiscrpniji i najopsežniji (ukupno 1224 stranice) rječnik ove vrste objavljen za hrvatski jezik. Natuknice u rječniku zapravo su leme. U njemu se nalaze sljedećim redoslijedom: [182]

1. čestotni rječnik (čestotni popis lema),
2. abecedni rječnik (abecedni popis lema uz oznaku čestote),
3. abecedni rječnik s pojavnicama (abecedni popis lema uz pripadajuće pojavnice s njihovim čestotama).

---

[178] Tadić (1991):170
[179] Tadić (1991):170
[180] Tadić (1991):175
[181] Tadić (1991):175 i Moguš, Bratanić, Tadić (1999):11
[182] Moguš, Bratanić, Tadić (1999):12

5.3. HNK (Hrvatski nacionalni korpus)

Računalna obradba hrvatskoga jezika projekt je MZT RH (*130718*) kojemu je jedan od osnovnih ciljeva sastavljanje najvećeg računalnoga korpusa u Hrvatskoj do sada. U sklopu projekta sastavlja se *Hrvatski nacionalni korpus* (HNK) koji je dobio pridjev *nacionalni* po uzoru na ostale nacionalne korpuse.[183] HNK se sastavlja u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu i sastoji se od dvije sastavnice:

1. **30-m**ilijunskoga korpusa suvremenoga hrvatskoga jezika (*30m*),
2. **H**rvatskoga **E**lektronskoga **T**ekstovnoga **A**rhiva (*HETA*).

U 30-milijunski korpus, koji se sastavlja prema principu reprezentativnosti ulaze tekstovi na hrvatskome jeziku od 1990. godine i mlađi prema određenim omjerima.[184] Trenutni opseg ove komponente HNK-a je 9.156.446 pojavnica.

HETA je zbirka tekstova koja se sastoji od tekstova starijih od 1990. godine i onih koji narušuju ravnotežu reprezentativnosti 30-milijunskoga korpusa. Stoga je planiranim opsegom znatno veća od 30-milijunskog dijela HNK-a i ima trend stalnoga rasta. HETA bi se po svojoj namjeni, ali i količini tekstova koji se tamo nalaze mogla smatrati svojevrsnim spremištem elektroničkih tekstova hrvatskoga jezika teoretski neograničena opsega.

Obje su sastavnice HNK-a s nezavisnim sučeljima pretražive putem Interneta.

5.3.1. Pretvaranje i priprema tekstova za unos u korpus

Većina tekstova koji ulaze u današnje korpuse pribavlja se u digitalnome obliku putem WWW-a ili iz DTP izvora. Prvi problem koji se pojavljuje s velikim brojem različitih izvora jest raznolikost formata u kojima tekstovi dolaze. S obzirom na dominaciju medija (Internet) ili aplikacije (npr. MS Word), najučestaliji su formati pribavljenih tekstova HTML, RTF ili DOC (*MS Word*) i PDF.

[183] Tadić (1996):610
[184] Omjeri su izneseni u Tadić (1996) i Tadić (1998)

Priprema i **pretvaranje** (*conversion*) tekstova koji ulaze u HNK-a zamišljena je u dvije faze:

1 . faza: izrađuje se neobilježena, ali putem Interneta slobodno pretraživa probna inačica korpusa. Cilj je pretvaranja u ovoj fazi dovesti tekstove u stanje običnoga teksta, dakle u ASCII oblik bez ili s malim brojem dodatnih oznaka.

S obzirom da prikupljeni tekstovi sadrže mnoštvo suvišnih oznaka potrebno ih je *filtrirati*, tj. ukloniti sve suvišne oznake. Za filtriranje ulaznih tekstova korišten je skup vlastitih programa (izrađenih u okviru projekta) i *Search&Replace V3.0* alat tvrtke *Funduc Software Ltd*.[185] Osim navedenih, postoji velik broj "traži i zamijeni" (*search & replace*) alata koji su ponekad dio većih aplikacija, ali postoje i specijalizirana samostalna rješenja u obliku posebnih programa ili skupova skripata. Svrha im je pronaći zadani niz pismena u dokumentu, te ga zamijeniti odgovarajućim korisnički definiranim nizom, ili ga pak brisati pretvaranjem u prazan niz. Kvalitetni "traži i zamijeni" alati omogućuju uporabu regularnih izraza, skripata i sl. Takvi su alati zbog mogućnosti preciznoga definiranja traženih nizova iznimno uporabljivi pri dovođenju različitih oblika teksta u stanje običnoga teksta, ali se mogu koristiti i za neke vrste obilježavanja. U drugom se slučaju koriste već postojeće oznake ulaznih oblika koje se zamjenjuju s odgovarajućim oznakama sheme obilježavanja.

2 . faza: izrađuje se obilježeni korpus u XML formatu prikladan za složenije pretrage. U prvom se redu obilježava struktura dokumenta kao što su naslovi, odlomci, rečenice i sl. Dovođenje tekstova korpusa u takav oblik zahtijeva više koraka pretvaranja i obrade, osobito iz razloga što ulazni tekstovi nisu obilježeni po planiranoj shemi. Ovisno o vrsti ulaznoga teksta, pretvaranje se odvija dvama postupcima:

a) iskorištavaju se postojeći kodovi ulaznih tekstova i zamjenjuju odgovarajućim kodovima planirane sheme obilježavanja. Postojeći kodovi uvelike olakšavaju pretvaranje jer je struktura dokumenta već djelomično ili potpuno obilježena.

b) ukoliko nije moguće koristiti postojeće kodove ulaznih tekstova, oni se brišu te se naknadno posebnim alatom ubacuju oznake planirane sheme obilježavanja. Ovaj je postupak znatno složeniji i često zahtjeva dodatnu ljudsku intervenciju.

[185] *Search&Replace V3.0* alat može se naći na adresi: http://www.funduc.com/

Pismena ulaznih tekstova kodirana su različitim skupovima za kodiranje pismena (CP-1250, CP-852, CROSCII itd.). S obzirom da XML podržava UNICODE skup pismena, u ovoj je fazi potrebno obaviti i pretvaranje pismena. Postoji zaista velik broj, najčešće besplatnih (*freeware*) alata za pretvaranje pismena, a kod pretvaranja za drugu fazu HNK-a rabljen je besplatni alat CCP.[186]

Pretvaranje mnoštva ulaznih formata tekstova u jedinstveni XML format zasigurno je najsloženiji problem druge faze. Kako je s razvojem Interneta broj datoteka u HTML obliku postao dominantan, najviše će pozornosti biti usmjereno na nj. Sam koncept HTML-a dopušta više slobode nego u XML-a u sintaktičkom smislu, kao što je npr. opcionalno zatvaranje oznaka, relativno proizvoljan redoslijed oznaka, pridodavanje različitih naziva istoj funkciji (naglašeni tekst često se označava s <emph>, ili <b>) itd. Iz tog je razloga čest slučaj da svaki izdavač HTML dokumenata rabi vlastitu shemu obilježavanja. Sintaksa XML-a u tom smislu mnogo je stroža, pa zbog fleksibilnosti HTML-a nije moguće jednim "korakom" pretvoriti HTML datoteke različitih shema obilježavanja u jednoobrazne XML dokumente. Stoga se pretvaranje zasniva na korisničkim definiranim skriptama (*user-defined scripts*) gdje korisnik sam određuje u koje će XML oznake biti pretvorene određene HTML oznake i pod koji uvjetima.

Dakle, pretvaranje se tekstova obavlja posebnim **alatima za pretvaranje** (*conversion tools*). Ono se odnosi na oblike datoteka, ali i na sama pismena, pogotovo kada je riječ o tekstovima na jeziku sa specifičnim pismenima kakav je hrvatski. Iako postoji velik broj alata za pretvaranje, u nekim je slučajevima potrebno razvijati vlastite alate, što je bio slučaj u obje faze razvoja HNK.

5.3.2. 2XML: alat za pretvaranje HTML i RTF oblika u XML

2XML je alat koji je u suradnji s tvrtkom *Softlex d.o.o.* razvijen za potrebe korpusnih projekata Zavoda za lingvistiku. Namijenjen je pretvaranju HTML i RTF datoteka u XML oblik i tokenizaciji teksta. Navedena dva ulazna formata datoteka odabrana su iz razloga što se danas daleko najveći broj tekstovnih datoteka nalazi ili se lako pohranjuje upravo u te oblike. Sam se alat sastoji od dva nezavisna modula:

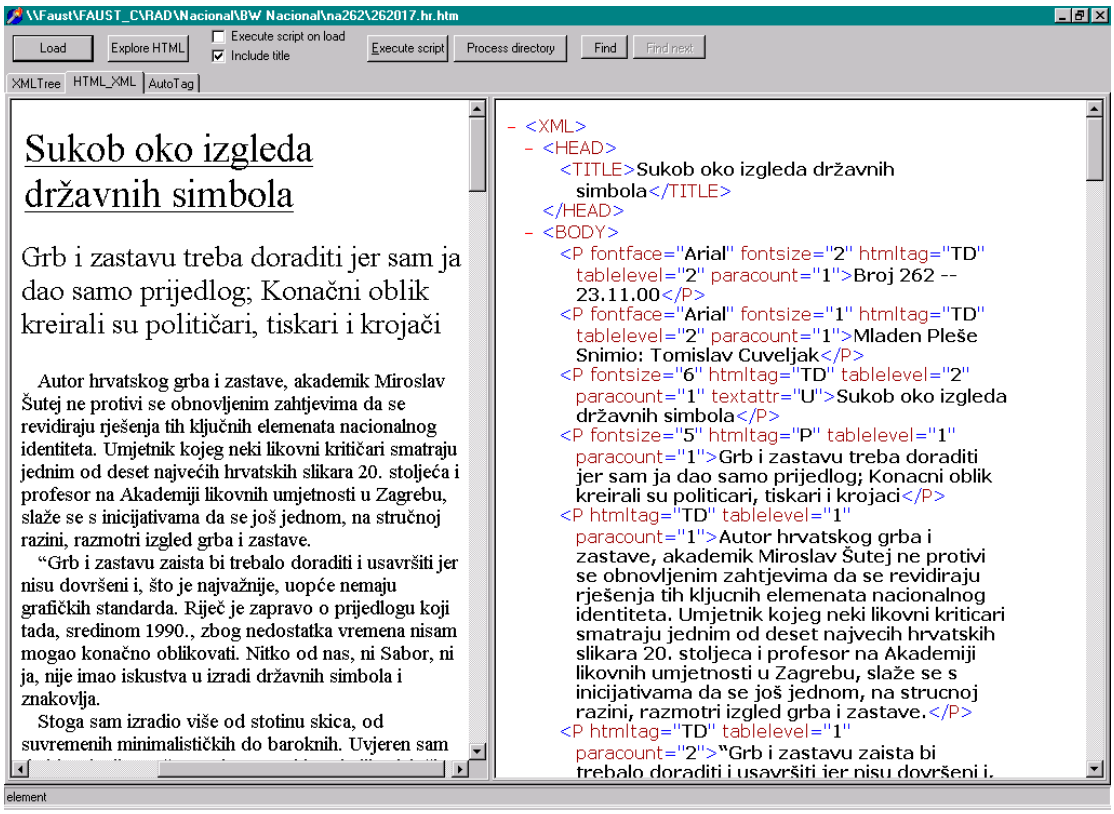
[186] CCP se može naći na Bulaja (1999)

1. prvi modul obavlja pretvaranje HTM(L) ili RTF oblika u XML,
2. drugi modul obavlja tokenizaciju teksta.

5.3.2.1. Pretvaranje HTM(L) ili RTF oblika u XML

Pretvaranje se izvodi u dva koraka:[187]

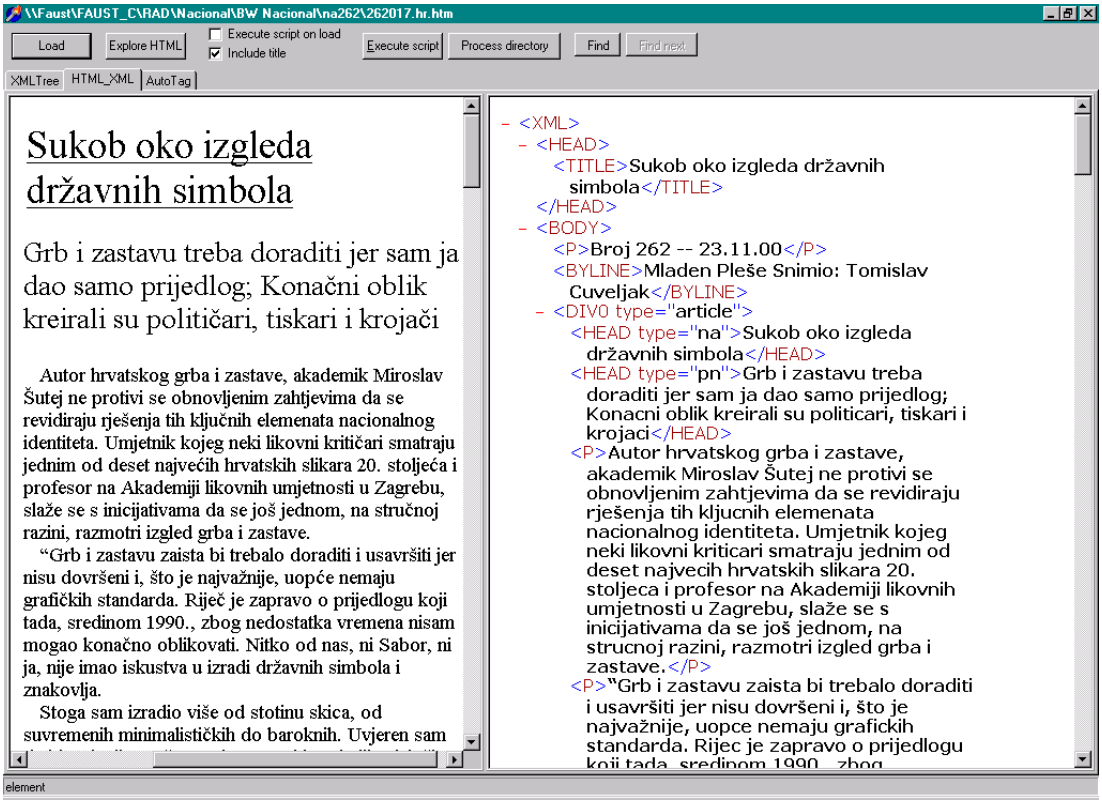
1. program stvara "nečisti" XML obilježen samo razinama odlomka <P>, gdje su sačuvane neke vrijednosti HTML atributa (slika 9).



Slika 9: Prvi korak pretvaranja HTML-a u XML

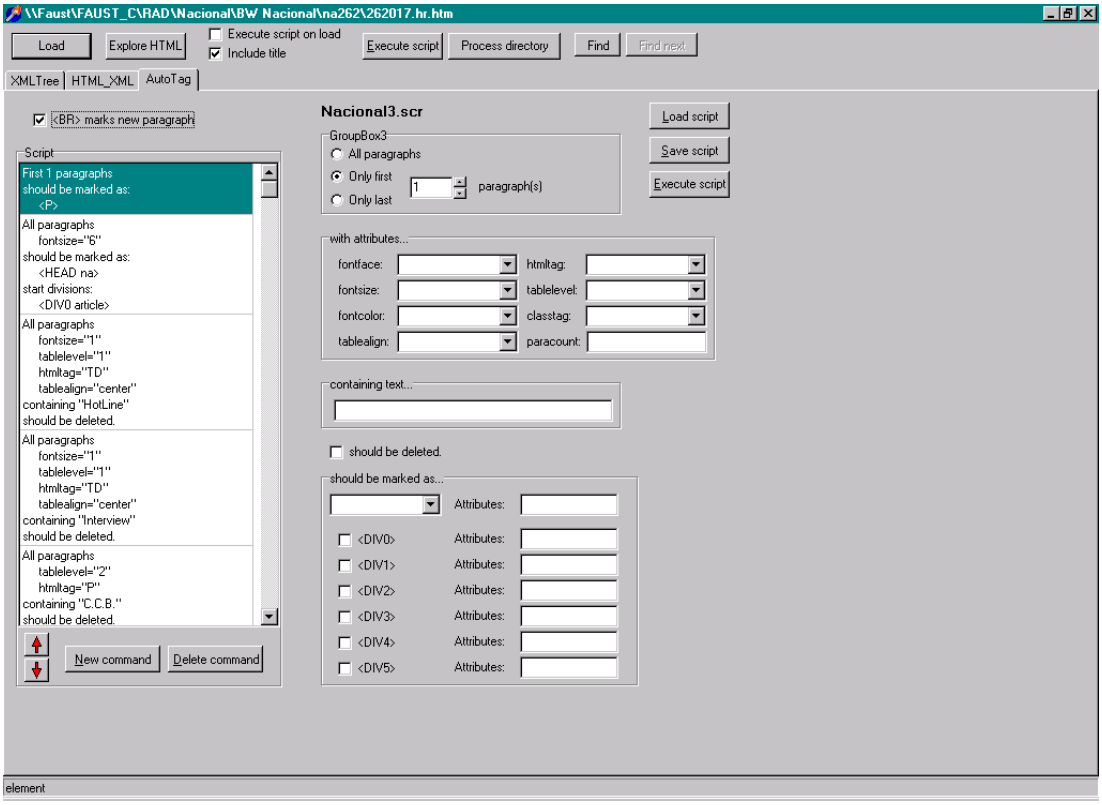
2. pokretanjem skripte, datoteka iz prvog koraka pretvara se u "čisti" XML, gdje se sačuvane vrijednosti atributa pretvaraju u korisnički definirane XML oznake i atribute, a sve otvorne oznake dobivaju pripadajuće zatvorne parnjake (slika 10).

[187] Tadić (2000c):525



Slika 10: Drugi korak pretvaranja HTML-a u XML

Skripte se sastavljaju posebnim sučeljem:



Slika 11: Sučelje za pisanje skripata u programu 2XML

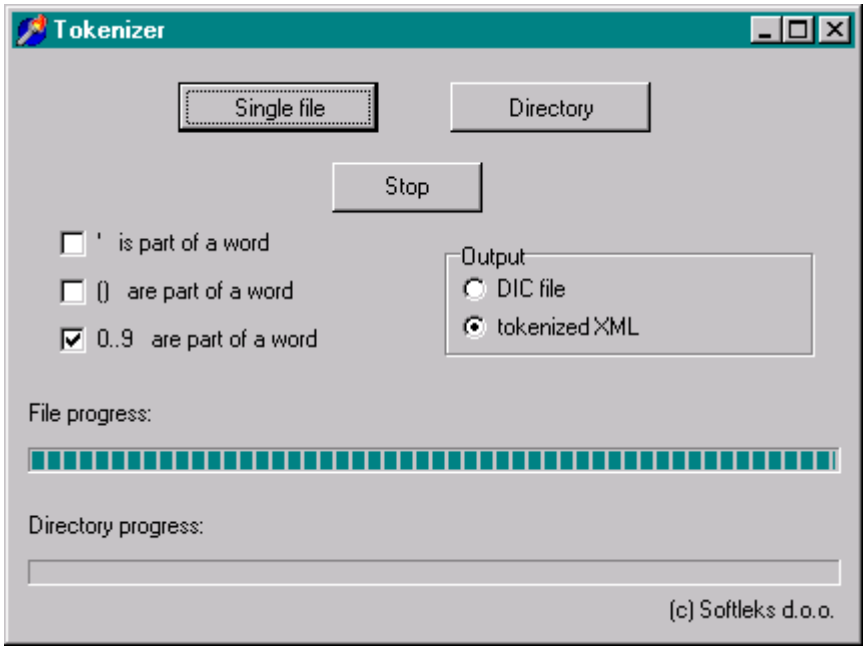
Za sastavljanje skripata nije nužno znanje programiranja već se ono svodi na odabiranje ponuđenih opcija i upisivanje parametara za određene elemente. Skripte se sastoje od niza naredbi koje se nalaze u lijevom okviru. Nove se naredbe dodaju uputom "*New command*". Nakon toga se specificiraju odlomci na koje se naredba odnosi, te HTML elementi, vrijednosti atributa i ostali parametri koje HTML elementi mogu imati kako bi bili odabrani za obrađivanje tom naredbom. Kao parametar za prepoznavanje HTML elementa moguće je unijeti tekst koji se u tom elementu mora pojaviti. Nakon specifikacije parametara po kojima će HTML elementi biti odabrani potrebno je definirati XML elemente u koje će biti pretvoreni. Izlazni oblik XML elemenata (ime i atributi elementa) definira se u okviru "*should be marked as...*", a sadržaj im se preslikava iz izvorne HTML datoteke. Druga je mogućnost brisanje čitava HTML elementa označavanjem "*should be deleted*".

Pretvaranje više istovrsnih[188] datoteka omogućeno je njihovim smještanjem u isti direktorij, nakon čega se pokreće naredba "*Process directory*". Jedina radnja koju je potrebno obaviti za unos tekstova u korpus je naknadno ručno pridodavanje zaglavlja.

5.3.2.2. Tokenizacija

U drugom se modulu obavlja tokenizacija tekstova koji se već nalaze u XML obliku. Ulazni dokument može biti samo jedna datoteka ili pak cijeli direktorij dokumenata. U sučelju se odabirom ponuđenih opcija određuje status apostrofa i brojeva, tj. hoće li oni biti obrađivani kao sastavni dio pojavnice ili ne.

[188] Pod istovrsnom se datotekom smatra ista shema HTML ili RTF obilježavanja



Slika 12: Sučelje modula za tokenizaciju

Rezultat obrade može biti pohranjen u dva oblika:

1) u obliku *dic* datoteke:

(pr. 38)

```
<XML>       262017.hr   1     X
<HEAD>      262017.hr   6     X
<TITLE>     262017.hr   12    X
Sukob       262017.hr   19    R
oko         262017.hr   25    R
izgleda     262017.hr   29    R
državnih    262017.hr   37    R
simbola     262017.hr   47    R
</TITLE>    262017.hr   54    X
</HEAD>     262017.hr   62    X
<BODY>      262017.hr   69    X
<P>         262017.hr   75    X
Broj        262017.hr   78    R
262         262017.hr   83    B
--          262017.hr   87    I
23.11.00    262017.hr   91    B
</P>        262017.hr   99    X
<BYLINE>    262017.hr   103   X
Mladen      262017.hr   111   R
Pleše       262017.hr   118   R
Snimio      262017.hr   125   R
:           262017.hr   131   I
```

oznaka/pojavnica

izvor/dokument

udaljenost od početka teksta

vrsta oznake/pojavnice

```
Tomislav    262017.hr   133   R
Cuveljak    262017.hr   142   R
</BYLINE>   262017.hr   150   X
<DIV0 type="article">   262017.hr   159   X
<HEAD type="na">  262017.hr   180   X
Sukob       262017.hr   196   R
oko         262017.hr   202   R
izgleda     262017.hr   206   R
...
```

U prvom se stupcu nalaze oznake ili pojavnice iz teksta. Obavijest o izvoru (drugi stupac) važna je kod pronalaženja i identifikacije dokumenata iz datoteka u kojima su smješteni. U trećem su stupcu brojčano izražene udaljenosti pojavnice (ili oznake) od početka teksta, tj. *byte-offset*. Oznake u posljednjem stupcu mogu imati četiri vrijednosti koje određuju vrstu pojavnice:

```
R    riječ,
B    broj,
I    interpukcija,
X    oznaka.
```

2) u obliku tokenizirane XML datoteke:

(pr. 39)

```
<XML><HEAD><TITLE><W type="R">Sukob</W> <W type="R">oko</W> <W
type="R">izgleda</W> <W type="R">državnih</W> <W
type="R">simbola</W></TITLE></HEAD><BODY><P><W
type="R">Broj</W> <W type="B">262</W> <W type="I">--</W>  <W
type="B">23.11.00</W></P><BYLINE><W type="R">Mladen</W> <W
type="R">Pleše</W> <W type="R">Snimio</W><W type="I">:</W> <W
type="R">Tomislav</W> <W type="R">Cuveljak</W></BYLINE><DIV0
type="article"><HEAD type="na"><W type="R">Sukob</W> <W
type="R">oko</W> <W type="R">izgleda</W> <W
type="R">državnih</W> <W type="R">simbola</W></HEAD><HEAD
type="pn"><W type="R">Grb</W> <W type="R">i</W> <W
type="R">zastavu</W> <W type="R">treba</W> <W
type="R">doraditi</W> <W type="R">jer</W> <W type="R">sam</W>
<W type="R">ja</W> <W type="R">dao</W> ... </XML>
```

Kod tokeniziranih datoteka svaka je pojavnica eksplicitno obilježena oznakom W uz koju stoji atribut koji može imati vrijednosti R, B i I.

Rezultati se obrade za oba oblika automatski pohranjuju u direktorij u kojemu se nalazi polazna datoteka (ili datoteke).

5.3.3. Pohranjivanje tekstova u bazu i povezivanje sa sučeljem

Tekstovi koji ulaze u korpus mogu biti pohranjeni u bazu podataka i/ili kao tekst-datoteke na tvrdom disku računala U drugom slučaju pretraživanje podataka znatno je sporije, ali je mogućnost ažuriranja tekstova jednostavnija i lakša. Ukoliko je korpus većega opsega ili zahtijeva brže obrađivanje i pretraživanje, nužno ga je pohraniti u bazu podataka. Time se s jedne strane postižu višestruka povećanja brzine pretraživanja, ali s druge strane sama obrada postaje složenija, zahtjevnija s aspekta obrazovanih ljudskih resursa i utrošenih strojnih resursa.

Da bi se pretraživao korpus veličine HNK-a nužno ga je pohraniti u bazu podataka. Za pohranu u bazu neophodno je razbijanje tekućega teksta u pojavnice, tj. tokenizacija. Tek nakon tokenizacije, uz očuvanje obavijesti o izvoru i udaljenosti svake pojavnice od početka izvornoga teksta moguće je unositi pojavnice u bazu podataka. Iz tog su razloga svi izvorni tekstovi u prvom koraku pohranjeni u internom *.dic* formatu:

(pr. 40)

```
Dok delorko 301
sam delorko 305
šetao delorko 309
nekim delorko 315
drvoredima delorko 321
zagrebačkih delorko 332
parkova delorko 344
pa delorko 353
i delorko 356
onima delorko 358
na delorko 364
Zrinjevcu delorko 367
nekoliko delorko 378
puta delorko 387
me delorko 392
je delorko 395
...
```

udaljenost od početka teksta

izvor

pojavnica

Podaci se u navedenom obliku unose u bazu čime je omogućeno indeksirano pretraživanje pojavnica uz očuvanje obavijesti o izvorima u kojima se one nalaze. Nakon što su pri pretraživanju korpusa pojavnice pronađene, njihova se konkordancija generira, tj. lijeva se i desna okolina ispisuju iz *.hnk* datoteka. To su datoteke u internom formatu koje imaju isti naziv kao izvori u drugom stupcu, a sadrže tekstove u "pseudohtml" obliku prikladnome za prikazivanje u WEB-pregledniku. Datoteke u *hnk* formatu izgledaju:

(pr. 41)

```
<HTML> <HEAD> <META HTTP-EQUIV="Content-Type" CONTENT="text/html;
charset=windows-1250"> <META HTTP-EQUIV="Author" CONTENT="Marko
Tadić, Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u
Zagrebu"> <Delorko, Olinko: Dnevnik bez nadnevaka, MH, Zagreb 1996.>
</HEAD> <BODY> <NASLOV>I.</NASLOV> Dok sam šetao nekim drvoredima
zagrebačkih parkova (pa i onima na Zrinjevcu) nekoliko puta me je
bila iznenadila kiša svojim neočekivanim pljuskovima. Kiša s jednog
neba na kojemu nije bilo toliko oblaka da bi ...
...
```

U početnoj fazi testiranja, dok je korpus bio veličine do 2 milijuna pojavnica, tekstovi su bili pohranjeni u bazu podataka *Access 97*. Osnovni je problem te baze što ona nije namijenjena velikoj količini podataka, a pristup je za vrijeme pretraživanja dopušten samo jednom korisniku. Kako je rasla veličina korpusa, ali i broj korisnika koji ga pretražuje, korpus je u sljedećoj fazi pohranjen u *Microsoft SQL Server 6.5*. bazu podataka. Taj je sustav održavanja baza podataka mnogo robusniji, namijenjen većim količinama podataka i dopušta istovremeno pretraživanje većem broju korisnika ovisno o broju licencija.

Sama baza i Web-sučelje povezani su s pomoću *Microsoft ASP (Active Server Pages)* tehnologije. ASP je Microsoftova tehnologija koja se koristi za pisanje skripta smještenih na poslužniku s ciljem kreiranja dinamičnih i interaktivnih WEB-aplikacija.[189] U načelu, ASP datoteka nalik je HTML datoteci, ali sadrži skripte koje se izvode prije nego što se podaci šalju prema korisničkom pregledniku (*user's browser*). Velika je prednost ASP-a mogućnost kombiniranja s drugim tehnologijama kao što su: XML, COM (*Component Object Model*) ili HTML.[190] Kod veoma zahtjevnih upita za pretraživanje vrijeme pretrage same baze može zauzeti iznimno

[189] Microsoft (2000b)

mnogo vremena (pogotovo u slučaju uporabe zamjenskoga pisma na početku upita). Zbog tehničkih ograničenja maksimalno dozvoljeno vrijeme za pretraživanje baze HNK-a po jednom upitu postavljeno je na 900 sekundi. ASP stranice na poslužniku HNK-a povezuju sučelje s bazom preko ODBC-a (*Open Database Connectivity*). ODBC je *Microsoftova* tehnologija koja omogućuje pristup bilo kojim podacima iz bilo koje aplikacije bez obzira u kojoj se bazi podataka podaci nalaze.[191]

5.3.4. Pretraživanje probne inačice HNK-a

Jedan je od ciljeva HNK-a dostupnost putem WWW-a. Početna stranica HNK-a nalazi se na Web-adresi

http://www.hnk.ffzg.hr/

i za sada je slobodno pretraživa njegova probna inačica.

Pretraživanje se korpusa zasniva na *client-server* tehnologiji, gdje korisnik postavlja upit iz preglednika (*browser*). Nakon obrade upita na poslužniku (*server*) korisniku se vraćaju obrađeni podaci. Poslužnik je povezan na Internet vezom od 100 Mbita/s, koja u potpunosti udovoljava današnjim zahtjevima protočnosti i brzine. Za pretraživanje HNK-a moguće je koristiti bilo koji preglednik, uz napomenu da će rezultat biti kvalitetnije oblikovan novijim inačicama. Stoga se preporučuju inačice *Netscape* 3.+ ili *Internet Explorer* 3.+.

Prva je specifičnost HNK-a u odnosu prema drugim nacionalnim korpusima njegova slobodna i besplatna pretraživost putem WWW-a. Ostali su nacionalni korpusi, barem za sada, ograničili slobodno pretraživanje, ili je pretraživanje slobodno i besplatno samo na manjem uzorku (npr. BNC, CNC itd.).

Druga je specifičnost HNK-a u odnosu prema većim nacionalnim korpusima da se obrada izvodi u *Windows NT* okružju. To je rezultiralo nekim prednostima, ali i istodobnim poteškoćama u realizaciji. Osnovna je poteškoća bila nedostatak programa za pohranu i pretraživanje tako velikih korpusa u Windows okružju. Iz tog su razloga jednim dijelom iskorištene postojeće baze, a drugim je dijelom program razvijan u okviru samog Zavoda za lingvistiku. Sama teorijska

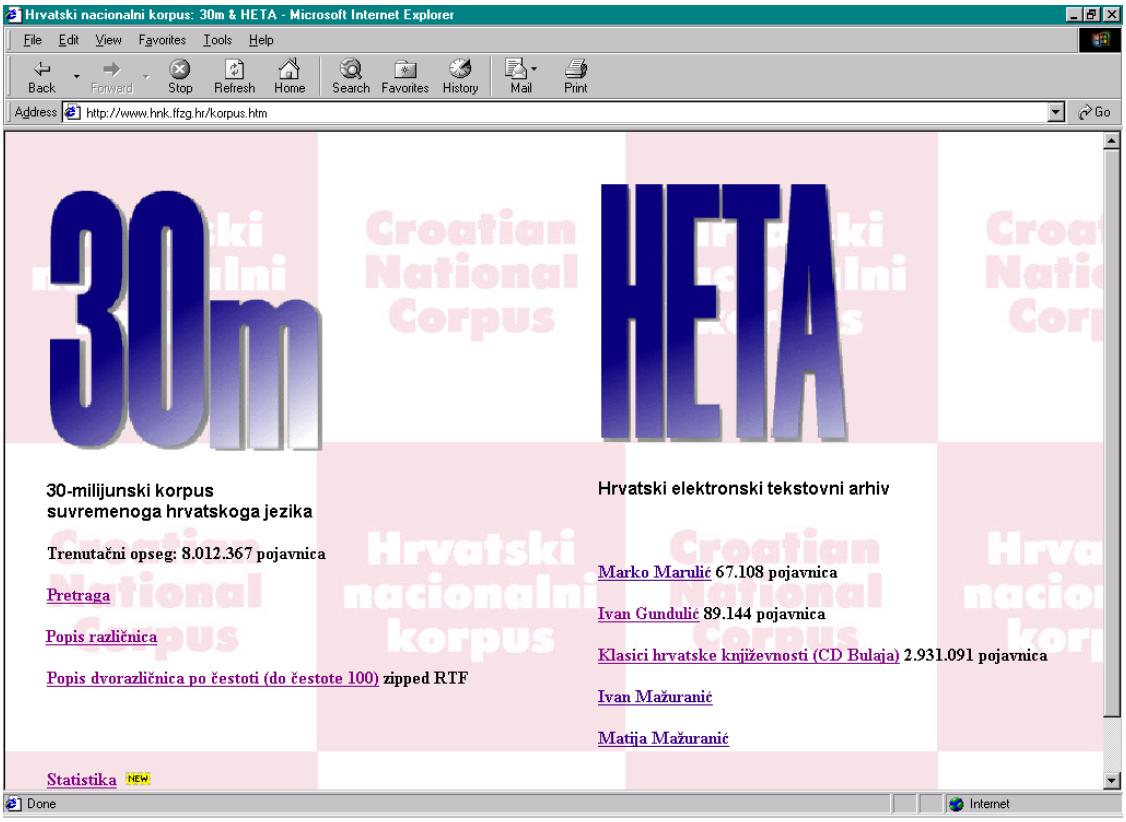
---

[190] Microsoft (2000b)
[191] Webopedia (2001): http://www.webopedia.com/TERM/O/ODBC.html

osnova i struktura HNK-a iznesene su u Tadić (1996) i Tadić (1998). Konkretna se realizacija dviju sastavnica HNK-a i mogućnosti njihova odabira mogu naći na adresi:

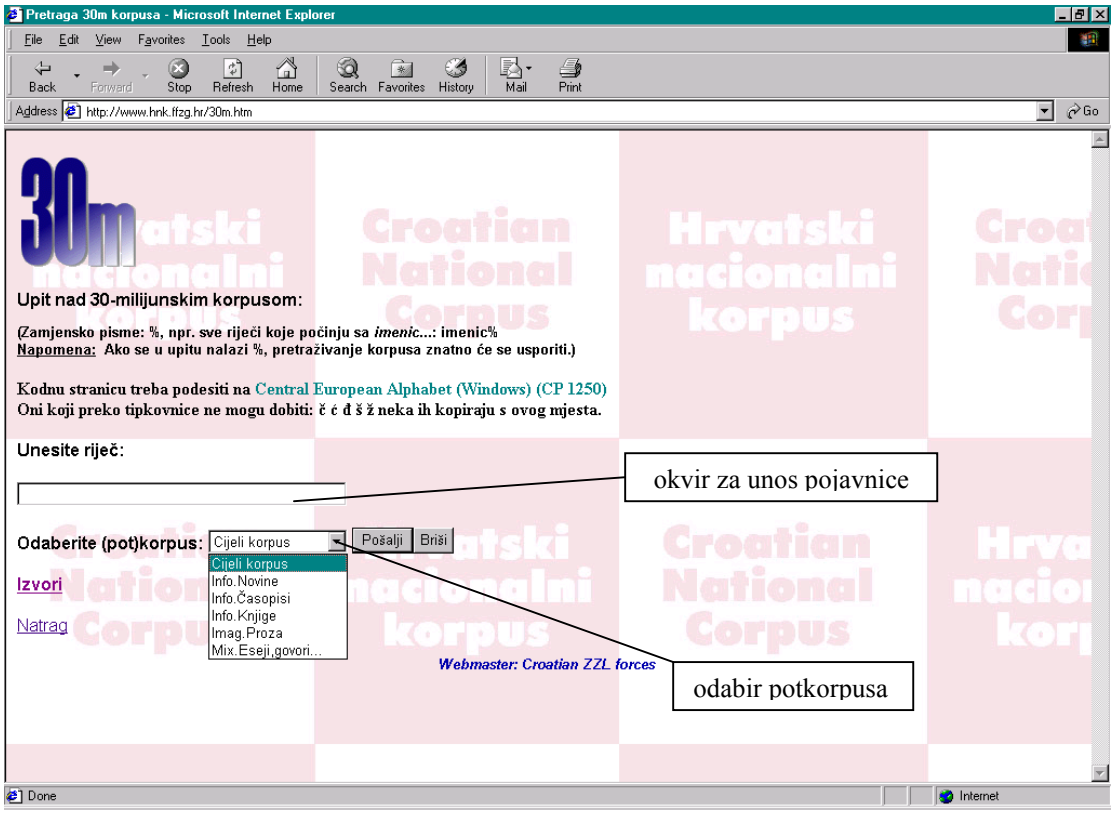
http://www.hnk.ffzg.hr/korpus.htm

Na istoj se adresi nalaze i poveznice prema popisu različnica i dvorazličnica 30m korpusa (slika 13):

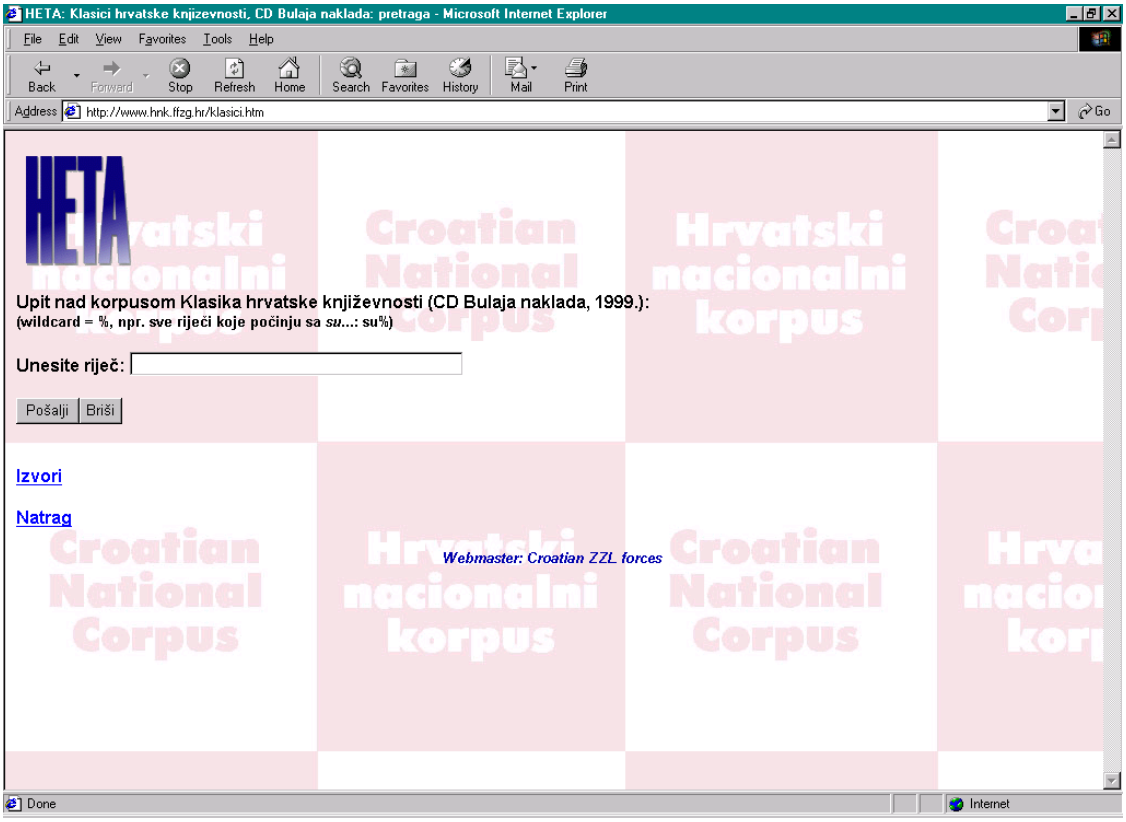


Slika 13: Stranica za odabir sastavnica HNK-a: 30m i HETA

Odabranim poveznicama moguće je pristupiti sučeljima za pretragu korpusa. S obzirom da 30m obuhvaća nekoliko potkorpusa, njegovo sučelje (slika 14) nudi više mogućnosti u odnosu na ostala sučelja za pretaživanje e-tekstova (slika 15):



Slika 14: Sučelje za pretragu 30-milijunskoga korpusa



Slika 15: Sučelje za pretragu korpusa klasika hrvatske književnosti

Sučelje prema Internet-korisniku izvedeno je u HTML-u. Glavno načelo prema kojem je izrađeno sučelje jednostavnost je uporabe uz istovremenu fleksibilnost za korisnika (slika 14). Pri pretraživanju moguće je odabrati cijeli korpus ili pak jedan od nekoliko potkorpusa (potkorpus novina, časopisa, knjiga, proze ili ostalo, kojim pripadaju najrazličitiji žanrovi kao npr. eseji, govori i sl.). Pojavnica za pretraživanje unosi se u okvir za unos. Postavljanje upita neosjetljivo je na mala i velika slova (*case-insensitive*). Tako bi za upit:

(pr. 42)

```
Brod, brod, BROD ili bRod,
```

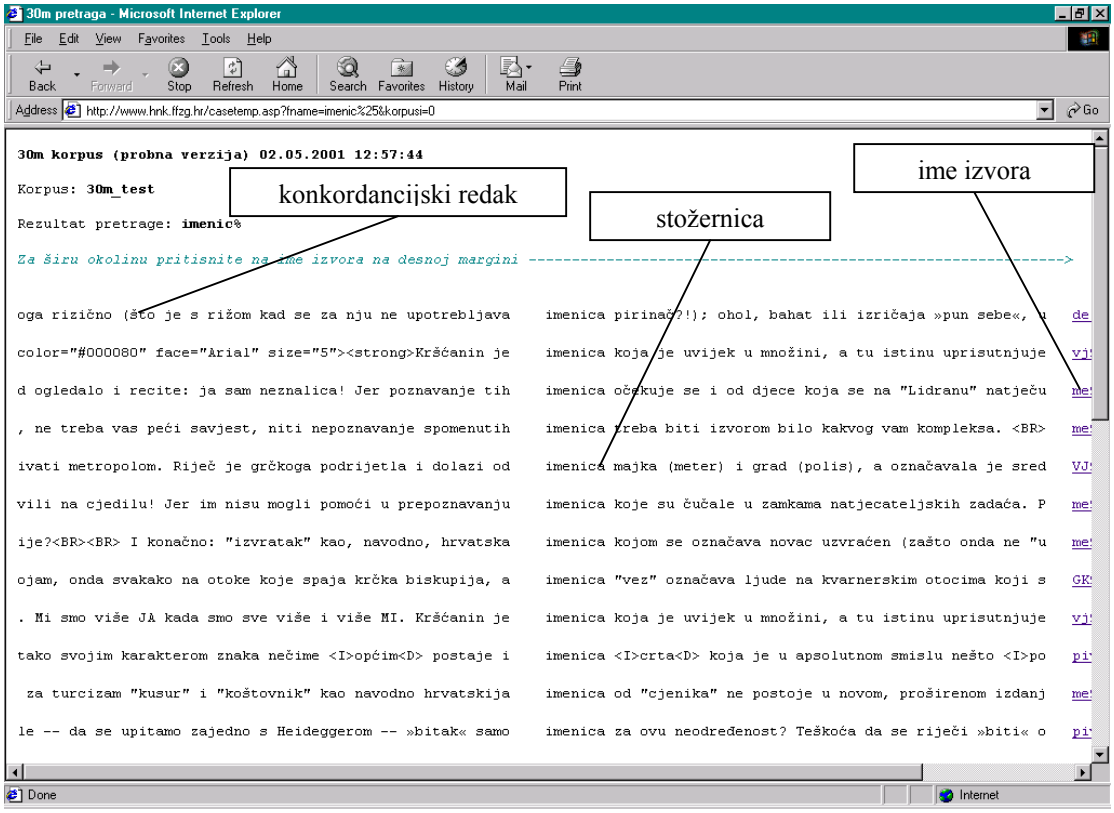
rezultat obrade bio isti, tj. sve pojavnice u kojima se pojavljuje "brod" u bilo kojoj konordanciji velikih i malih slova. Za postavljanje upita moguće je koristiti **zamjensko pisme** (*wildcard*). Zamjensko pisme zamjenjuje jedno ili više pismena. Ukoliko je unos za pretragu:

(pr. 43)

```
imenic%
```

rezultat su obrade svi oblici pojavnica koji se nalaze u korpusu a započinju sa zadanim nizom (npr. *imenica, imenice, imenicom itd.*). Zahtjevniji upiti zauzimaju više vremena za obradu na poslužniku, ali i za prikaz rezultata. Ukoliko se zamjensko pisme nalazi na početku pojavnice koja se pretražuje, pretraživanje će biti znatno usporeno ili u iznimno zahtjevnim slučajevima nemoguće. Razlog je tomu što je baza podataka indeksirana po počecima pojavnica, pa umjesto indeksiranog pretraživanja započinje sekvencijalno pretraživanje baze.

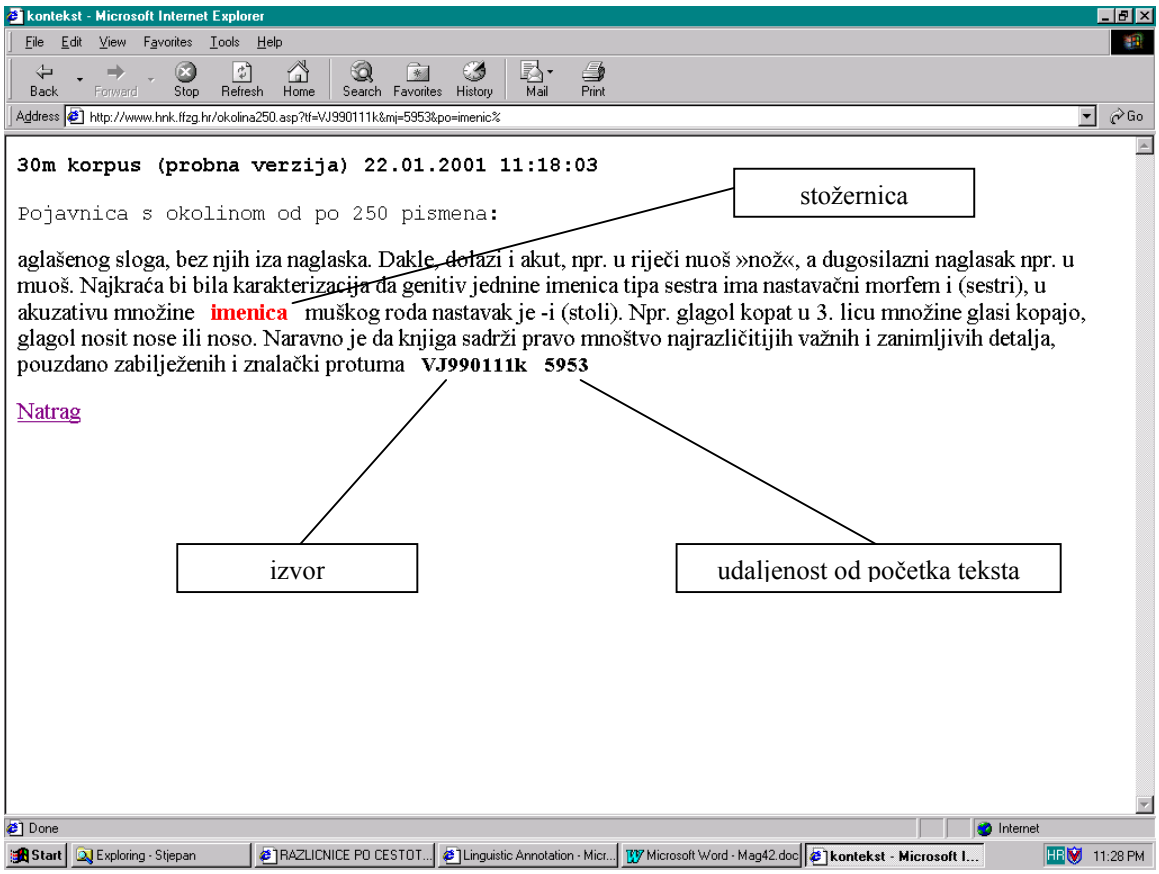
Provjera (*validation*) ulaznih podataka obavlja se već u sučelju, a za nju je uporabljen skriptni jezik *Microsoft VBScript*. Jedan je od glavnih ciljeva provjere rasterećenje poslužnika od prezahtjevnih, nemogućih, neispravnih ili zlonamjernih upita. To se ponajprije odnosi na nepropuštanje pojavnica iz zaustavne liste (u slučaju HNK-a čine je pojavnice s vrlo visokom čestotom kao npr. "i", "u", "je" itd.). Maksimalan broj pismena za unos ograničen je na 25. Unos brojčanih znakova nije dopušten, kao ni znakova "+" i "&".



Slika 16: Ispis konkordancija pojavnice za pretraživanje: imenic%

Rezultat je obrade upita KWIC (***Key-Word In Context***) [192] konkordancija u HTML formatu koja se izručuje u korisnikov preglednik (slika 16). Konkordancijski reci ispisani su istim redoslijedom kojim su se pojavili u korpusu, gdje uz svaki konkordancijski redak slijedi ime izvora uz koji se nalazi i redni broj retka u toj konkordanciji. Broj između ta dva podatka govori koliko je pismena tražena pojavnica udaljena od početka teksta u kojemu se nalazi. Na dnu je stranice ukupan broj pronađenih pojavnica, kao i relativna čestota (broj pronađenih pojavnica/broj svih pojavnica). Neposredan uvid u lijevu i desnu okolinu stožernice ograničen je na 60 pismena. Pritiskom na izvor konkordancije moguće je dobiti uvid u maksimalnu okolinu KWAL (***Key-Word And Line***)[193] koja je 250 pismena s lijeve i desne strane (slika 17):

[192] KWIC oblik konkordancije gdje se stožernice nalaze unutar unaprijed definirane lijeve i desne okoline.
[193] KWAL je oblik konkordancije koja dopušta nekoliko redaka konteksta s lijeve i desne strane okoline.



Slika 17: Maksimalna dužina konkordancije (250 pismena s lijeve i desne strane stožernice)

5.3.5. Rezultati obrade probne inačice 30m korpusa

Kvantitativna obrada jezičnog inventara probne inačice 30m korpusa rađena je u nekoliko smjerova. Kako su najčešći postupci u kvantitativnoj obradi brojanje, razvrstavanje i računanje korištena je baza *Access 97*. Obrada je rađena na probnoj inačici korpusa opsega 7.6 Mw.[194]

Za probnu inačicu 30m korpusa napravljeni su potpuni čestotni, abecedni i odostražni popisi pojavnica. Rezultati su pohranjeni u bazu, te je njima moguće

[194] *Mw* je kratica za broj pojavnica u milijunima

jednostavno manipulirati i izvoditi daljnje kvantitativne obrade. Navedeni popisi pojavnica mogu se naći u dodacima D1, D2 i D3.

Osim popisa pojavnica napravljen je i čestotni popis dvopojavnica (*bigrams*). U njemu se nalaze neposredni parovi pojavnica razvrstani po padajućoj čestoti. Popis dvopojavnica može se naći u dodatku D4. Na osnovi tog popisa izračunate su uzajamne obavijesnosti svih 3.414.743 parova dvopojavnica.

5.4. Hrvatsko-engleski paralelni korpus

Sastavljanje hrvatsko-engleskoga paralelnog korpusa započelo je koncem 1999. godine u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu s primarnim ciljem izrade paralelnog korpusa, ali i testiranja programa za sravnjivanje (*alignera*) za buduće paralelne korpuse.[195] Korpus sadrži 100 brojeva časopisa *Croatia Weekly* na obama jezicima, što otprilike iznosi 1.6 Mw na hrvatskome i 1.9 Mw na engleskome jeziku. Paralelni je korpus obilježen XML-om i automatskim postupkom sravnjen (*aligned*) na razini rečenice.

Tekstovi su pribavljeni u dva oblika (hrvatski u ASCII obliku bez ikakvih oznaka i engleski u *QuarkXPress 3.32* formatu) koje je trebalo u nekoliko koraka pretvaranja dovesti u oblik XML dokumenta, a samo pretvaranje u XML rađeno je s pomoću 2XML alata.

5.4.1. Segmentacija i sravnjivanje rečenica paralelnoga korpusa

Umetanje graničnih elementa rečenice (<S>) obavljeno je uporabom *Search & Replace V3.0*[196] alata tvrtke *Funduc Software Ltd.* s pomoću posebno programirane skripte. Tipično je prepoznavanje rečenice u tekstu obavljano traženjem niza:

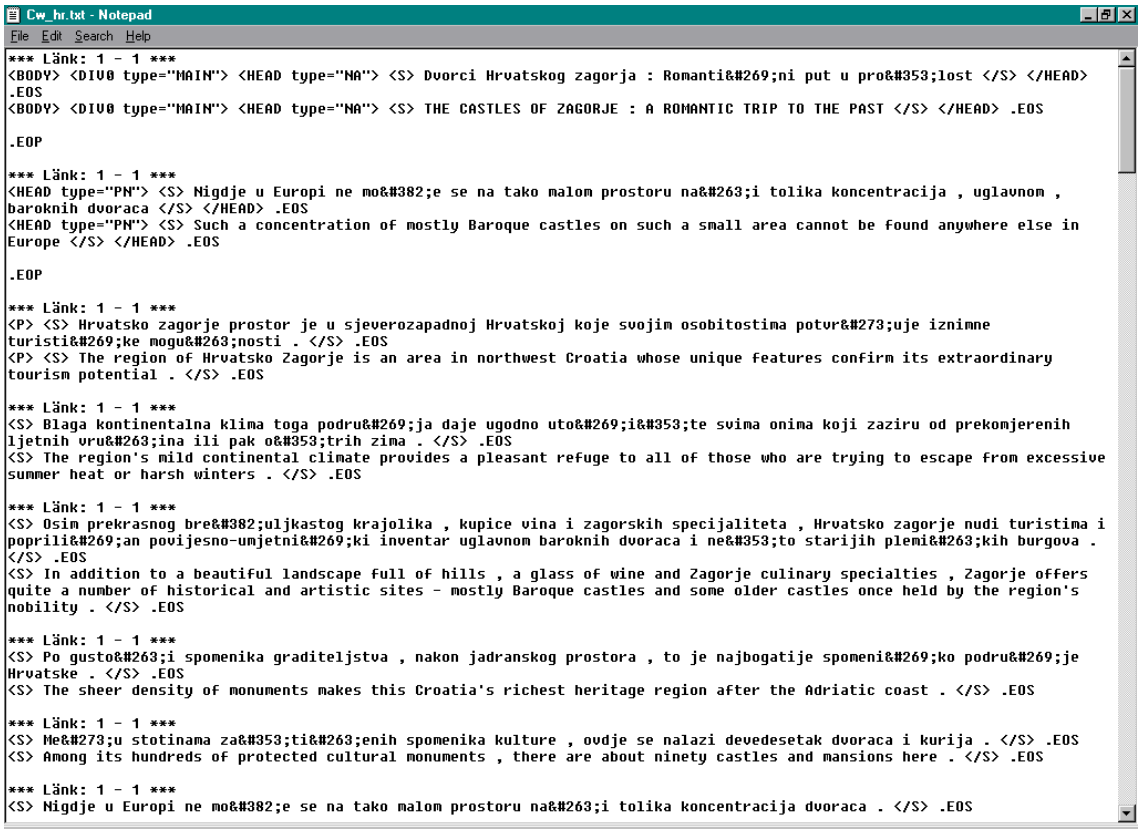
(pr. 44)

```
Interpunkcija (., ?, !) + razmak (space) + veliko slovo
```

[195] Tadić (2000c):527
[196] Više u poglavlju 5.3.2.

U drugom se koraku eliminiraju iznimke kao npr. *dr., prof., mr., ms., miss., ing., st., sv.*, i sl. Veliki problem automatskom prepoznavanju rečenice predstavljaju navodnici, koji se često rabe nekonzistentno i proizvoljno. Također, prepoznavanje granica rečenica u hrvatskim tekstovima znatno otežavaju i redni brojevi. Otprilike 28 % rednih brojeva pisanih znamenkama u hrvatskim tekstovima ujedno su i završeci rečenica, pa se taj dio treba obraditi uvidom u svaki pojedinačni slučaj.[197]

Sravnjivanjem se definiraju eksplicitne veza među odsječcima tekstova paralelnoga korpusa. U biti, **sravnjivanje** (*alignment*) je povezivanje elemenata (rečenica, fraza ili riječi) koje su uzajamni prijevodi.[198] Danas se sravnjivanje rečenica može izvesti automatski s visokim stupnjem točnosti. Alat koji obavlja sravnjivanje zove se **program za sravnjivanje** (*aligner*), a za sravnjivanje hrvatskih i engleskih rečenica korišten je program za sravnjivanje *Vanilla aligner*, autora Pernille Danielsson i Daniela Ridingsa.[199]



Slika 18: Neposredni rezultat sravnjivanja rečenica *Vanilla* programom za sravnjivanje

[197] Tadić (2000c):527
[198] CCL (2000)
[199] Tadić (2000c):526

Uporabljena inačica programa za sravnjivanje radi pod DOS operacijskim sustavom što donekle otežava rukovanje programom, ali su rezultati obrade znatno precizniji u usporedbi s nekim drugim testiranim programima s grafičkim sučeljima.

Jedan od problema s izlaznim rezultatima je što su ispisani i elementi koji su nadređeni rečeničnom elementu. Taj je problem riješen filtriranjem suvišnih oznaka. Mnogo više vremena utrošeno je na neposredan uvid i naknadno ručno obrađivanje pogrešno sravnjenih rečeničnih ekvivalenata.

5.4.2. Kodiranje paralelnoga hrvatsko-engleskoga korpusa

Postoji nekoliko načina kodiranja paralelnih korpusa, a oni bi se prema osnovnom ustrojstvu mogla podijeliti u dvije grupe:

1. pohranjivanjem pokazivača u izdvojeni dokument (rezultira uporabom brojnih identifikatora za element rečenice <S>),
2. kodiranje nadahnuto prevodilačkim memorijama (TMX).[200]

Slijedeći preporuke udaljenog obilježavanja (*stand-off annotation*)[201] kao optimalnoga načina kodiranja korpusa, sravnjeni paralelni korpus ima oblik:[202]

(pr. 45)

DOKUMENT 1:

```
<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
provo&#273;enje mirne …</S>
<S id="CW010199803190201hr.S3">Stoga, treba izbje&#263;i svaki
&#269;in koji …</S>
</HEAD>
<P>
<S id="CW010199803190201hr.S4">Vlada Republike Hrvatske obvezala je
…</S>
 …
</P>
 …
</DIV0>
```

[200] Tadić (2000c):528
[201] Ide (2000):28
[202] cf. Tadić (2000c):528

DOKUMENT 2:

```
<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: (Povezivanje sravnjenih rečenica):

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

Ovaj način kodiranja korpusa u skladu je s XCES standardom. Uz svaku rečenicu (dokumenti 1 i 2) umetnut je jedinstveni identifikator (id) gdje prvi dio zapisa čuva obavijest o izvoru, a zadnja dva pismena ispred točke označavaju jezik kojem pripada rečenica. Podatak iza točke redni je broj rečenice u dotičnom članku. Tako obilježene rečenice sparuju se iz trećega dokumenta referiranjem na identifikatore s pomoću *Xlink* mehanizma koji je detaljnije opisan u poglavlju 3.3.2.9.

5.5. Probna inačica pretraživanja XML-om obilježenih tekstova

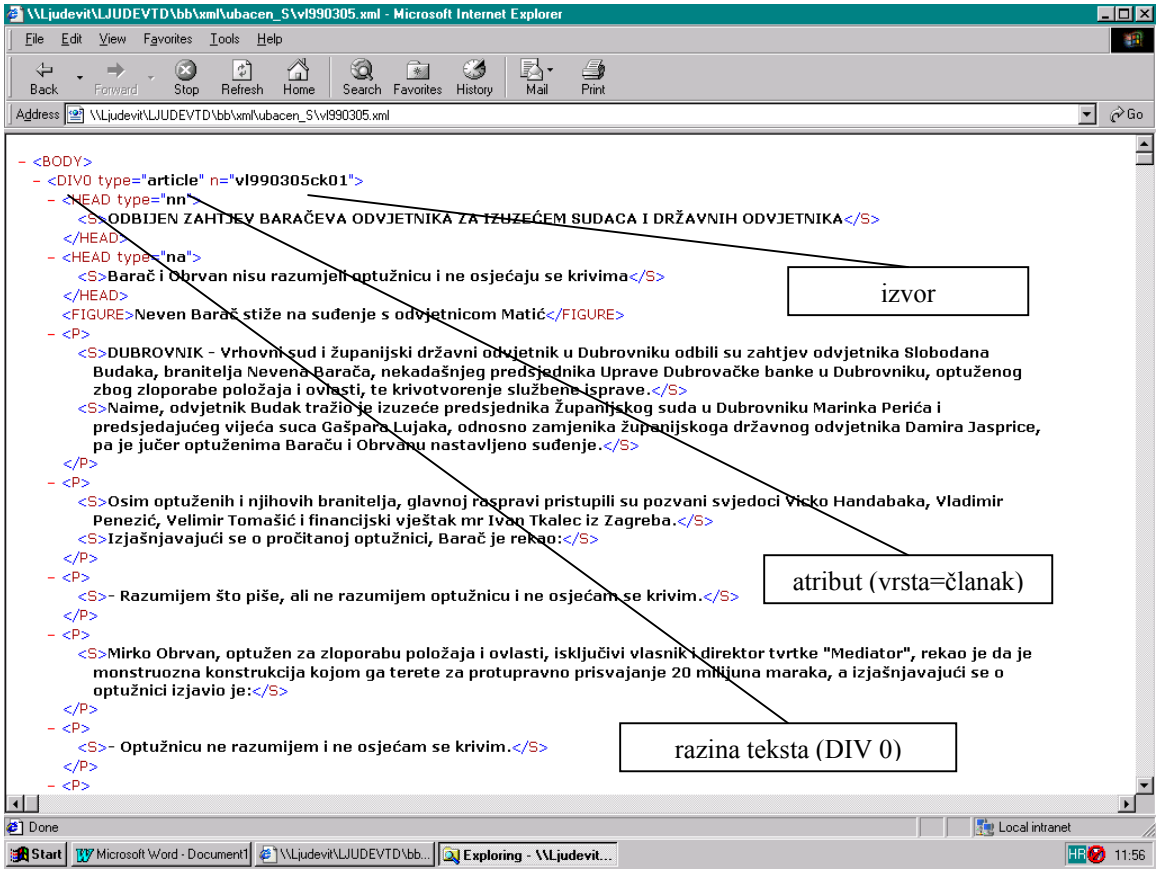
Dio tekstova HNK-a (oko 250.000 pojavnica) probno je obilježen XML-om prema XCES standardu. Odluka o kodiranju XML-om donesena još početkom 1999. godine pokazala se dalekovidnom, osobito nakon kasnijeg prihvaćanja XML-a kao međunarodnog standarda za obilježavanje korpusa. U prvom se koraku obilježavanja unose podaci o izvoru tekstova, te se obilježava fizička struktura npr. naslovi, podnaslovi, potpisi pod sliku itd. U sljedećem koraku obilježavaju se odlomci, a nakon toga rečenice.

Razine su teksta prema XCES-u čvrsto definirane, a oznake najčešće imaju ovaj redoslijed:

1. razina: najviša je razina koja obuhvaća cijeli tekst. Korijenski element <BODY> ima pripadajuću zatvornu oznaku (</BODY>) na samom kraju teksta.
2. razina: razina je razdjela teksta (*division*), i označena je elementom <DIV0>. On sadrži dva atributa: prvi (*type*) koji govori o vrsti teksta u odsječku (članak, tekst u okviru uz članak i sl.) i drugi atribut (*n*) koji jedinstveno označava izvor odsječka (ime, datum izdanja, rubriku i broj članka, u slučaju kodiranja novinskih članaka).[203] Odsječci mogu imati podređene odsječke, pa se oni označavaju prema hijerarhiji na DIV1, DIV2, itd.
3. razina: razina je naslova, potpisa pod sliku ili odlomka, ovisno o tome što se u tekstu pojavljuje:
   - <HEAD> je oznaka za naslov. Sadrži atribut (*type*) koji eksplicira vrstu naslova, npr. na = naslov, nn = nadnaslov, pn = podnaslov i sl.),
   - <FIGURE> označava tekst koji ide uz sliku (potpis pod sliku),
   - <P> je oznaka za odlomak (paragraf).
4. razina: razina je rečenice. Označena je s <S> i pojavljuje se unutar elemenata <P> i <HEAD>.
5. razina: najniža je razina i označava pojavnice u tekstu. Pojavljuje se uz element <S>. Oznaka za pojavnicu je <W>, a sadrži atribut (*type*) koji govori je li označena riječ (R), interpunkcija (I) ili broj (B).

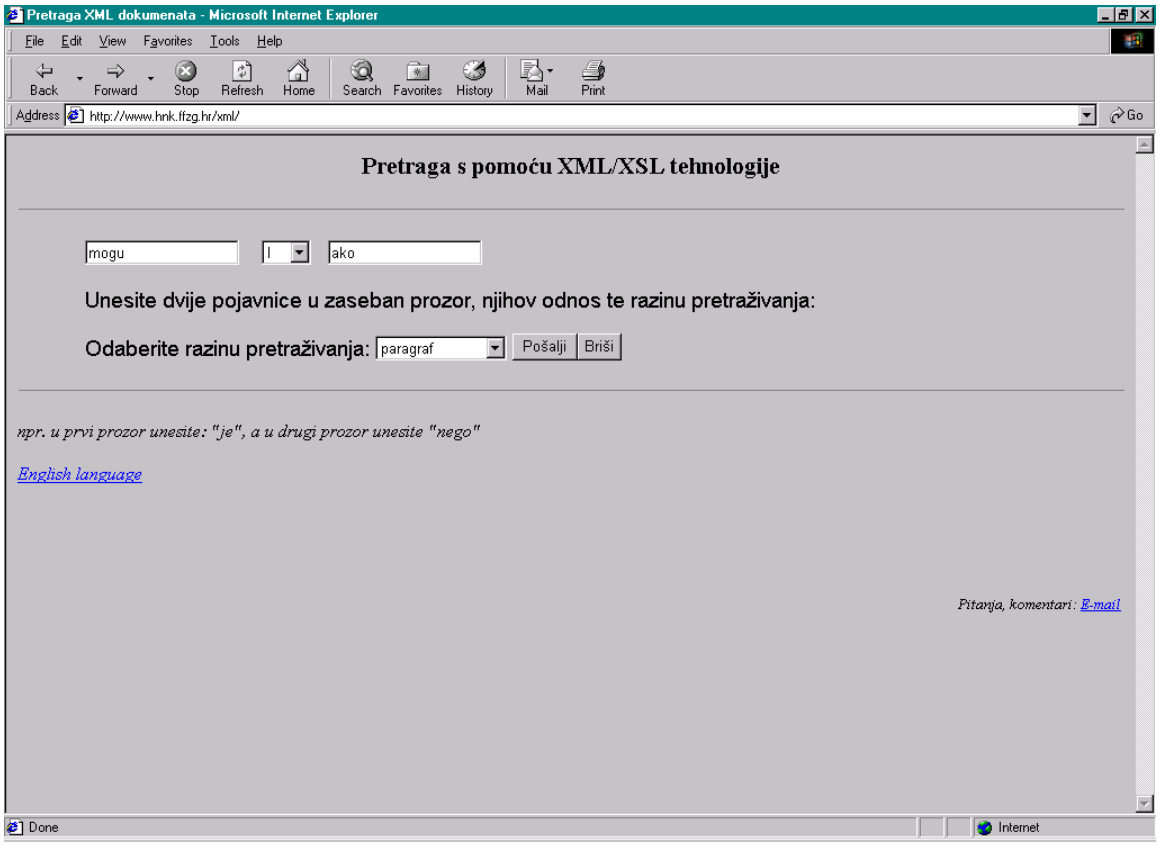
Primjer teksta obilježenog na svim ovim razinama nalazi se u dodatku B, a do razine rečenice obilježeni tekst u IE 5.5 izgleda (slika 19):

[203] Npr. ime izora *vl990305ck01* sadrži obavijest o:
vl = Večernji list, ime izvora
99 = 1999, godina izdanja,
03 = ožujak, mjesec izdanja,
05 = 5, dan u mjesecu izdanja,
ck = crna kronika, ime rubrike,
01 = broj članka u rubrici



Slika 19: XML-om obilježen tekst do razine rečenice

WWW-sučelje prema korisniku izvedeno je u HTML-u, ali nudi znatno bogatiji izbor mogućnosti pretraživanja u odnosu prema pretraživanju neobilježenoga HNK korpusa. Budući da je označena fizička struktura, omogućeno je pretraživanje prema razinama teksta, npr. razini rečenice ili odlomka. Uz odabir razine pretraživanja, moguće je istovremeno s pomoću Booleovih izraza (I, ILI i NE) definirati odnose između dviju pojavnica (slika 20). Dakle, moguće je postaviti slijedeći zahtjev: "Pronađi sve odlomke koji sadrže pojavnicu *mogu* i pojavnicu *ako*!". U prvi se okvir za unos unosi prva pojavnica za pretraživanje, odabire se Booleov izraz I (AND), u drugi se okvir za unos unosi druga pojavnice za pretraživanje, te se zatim postavlja razina pretraživanja na razinu odlomka:



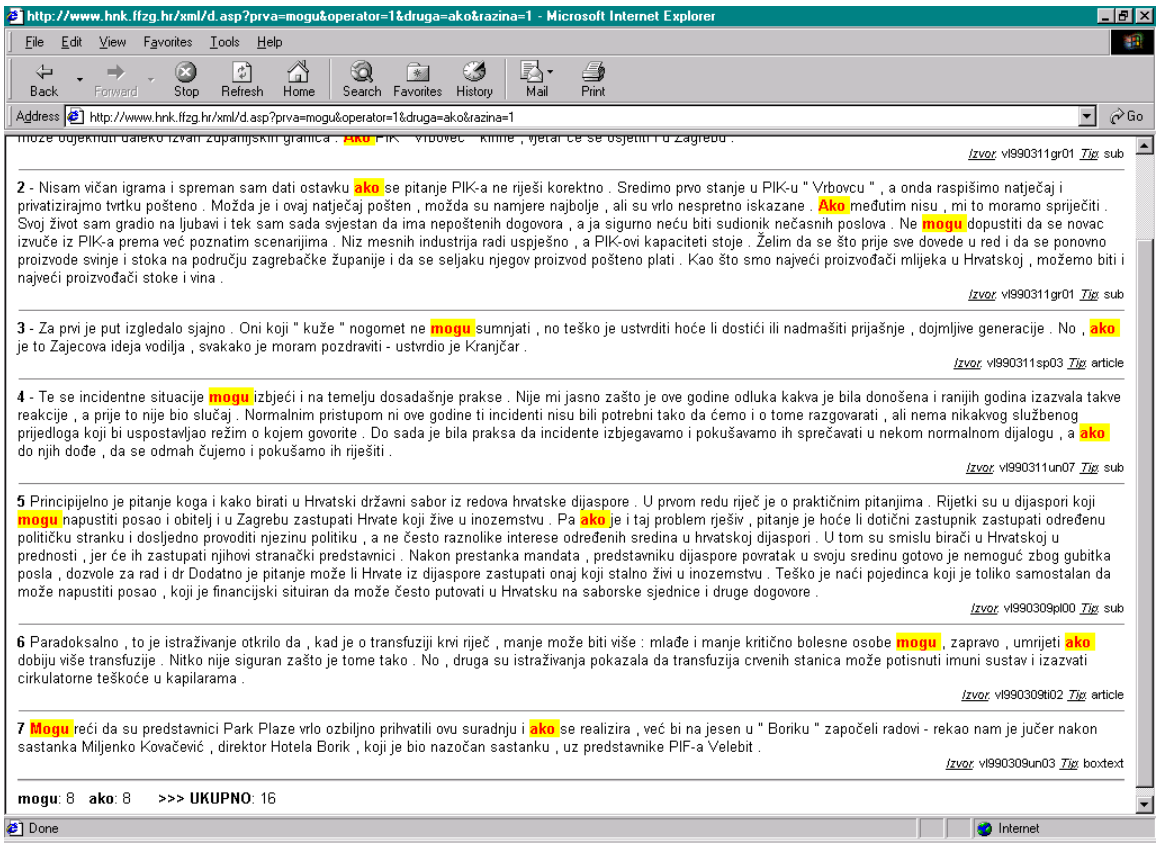
Slika 20: Sučelje za pretragu korpusa s pomoću XML/XSL tehnologije

Rezultat su obrade konkordancije pronađenih pojavnica u HTML formatu, popraćene informacijama o rednom broju, elementu odabrane razine pretraživanja, te izvoru i vrsti teksta u kojemu se nalazi tražena pojavnica (slika 21). Na dnu stranice rezultata nalazi se broj pronađenih pojavnica, te njihov zbroj (ukoliko su u ulaznim podacima unesene dvije pojavnice).

Za sada je veliki nedostatak što se tekstovi nalaze u XML datoteci koja nije smještena u relacijsku bazu podataka, već se kao XML dokument nalazi na poslužniku. Taj oblik pohrane znatno usporava pretragu, pogotovo ukoliko se veličina tekstova poveća. XML parser još uvijek je prespor za obrađivanje većih količina podataka kakve danas nameću suvremeni standardi korpusnih opsega. No, pohranjivanje XML datoteka u bazu podataka tehnički je tek nedavno riješeno[204], pa u probnoj inačici ovog oblika pretraživanja još nije primijenjeno. To je cilj sljedeće faze, čime bi se obrada korpusa znatno ubrzala.

[204] MS SQL Server 2000, više na http://www.microsoft.com/sql/

Program za pretraživanje ujedinjuje nekoliko tehnologija: HTML, ASP, XML/XSL i CSS. Kako je XML/XSL relativno mlada tehnologija, još nije riješeno prosljeđivanje vrijednosti varijabli iz tekstnog okvira HTML datoteke (tj. samih pojavnica za pretraživanje) prema XSL datoteci. Stoga je bilo nužno pronaći alternativno rješenje koje se pokazalo kao znatno složeno, ali uporabljivo. Napravljen je takav program koji ne prosljeđuje direktno vrijednost varijable (tj. unesene pojavnice za pretraživanje) prema XSL datoteci, već se XSL datoteka dinamički generira zajedno s unesenom vrijednošću varijable. Kôd XSL datoteke može se naći u dodatku E.



Slika 21: Rezultat obrade korpusa s pomoću XML/XSL tehnologije

5.6. Planovi i budući koraci razvoja HNK-a

U prethodnome je poglavlju spomenut kao jedan od važnijih ciljeva pohranjivanje XML-om obilježenih tekstova u relacijsku bazu podataka. Time bi se znatno ubrzalo pretraživanje, a veličina korpusa ne bi predstavljala tehničko ograničenje.

Kad je riječ o samim tekstovima korpusa, planira se POS i MSD obilježavanje tekstova prema *Multext-East*[205] specifikaciji, pa bi i samo pretraživanje bilo omogućeno prema tim kategorijama. Specifikacija morfosintaktičkoga opisa za hrvatski već je izrađena.[206] Nužan je preduvjet za takvo obilježavanje izrada morfološkog leksikona za hrvatski jezik. Tako su već u fazi obilježavanja morfološki uzorci promjene natukničke liste prvoga izdanja Anićeva rječnika, iz čega će se generirati *Hrvatski morfološki leksikon* s pripadajućim POS i MSD kategorijama. Obilježavanje se radi prema metodologiji razrađenoj u Tadić (1994). Na taj će se način stvoriti uvjeti za izradu automatskoga POS označivača za hrvatski, što je jedan od važnijih ciljeva, ali ujedno i najsloženiji dio treće faze obrade HNK-a. S obzirom na bogatstvo morfologije hrvatskoga, taj će označivač u odnosu prema označivačima za druge jezike imati veći broj oznaka, pa je za pretpostaviti i manju točnost.[207]

Dakle, u trećoj se fazi planira morfosintaktički obilježen *30m* korpus pohranjen u relacijsku bazu, pretraživ po svim obilježjima i pojavnicama korpusa.

[205] Multext-East (1996): MULTEXT-EAST je projekt Europske unije čija je namjera proširiti djelovanje MULTEXT projekta na srednje i istočnoeuropske jezike. Ciljevi su Multext-Easta: testiranje i prihvaćanje jezičnih standarda, razvoj obilježenih višejezičnih korpusa, razvoj morfo-leksičkih resursa i prihvaćanje korpusnih alata MULTEXT-a.

[206] Nalazi se zajedno sa specifikacijom za još sedam jezika na: http://nl.ijs.si/ME/V2/msd/html

6. Zaključak

Korpusna je lingvistika u posljednja tri desetljeća doživjela svoj nagli uspon. Danas, gotovo da nema značajnijeg jezika za koji nije sastavljen, ili pokušao biti sastavljen barem milijunski korpus. Za većinu važnijih jezika sastavljeni su i nacionalni korpusi ili se ulažu napori u njihovu sastavljanju. Iako je računalna obrada korpusa započela u Americi (*Brown corpus*), u Europi se posljednjih tridesetak godina barem jednako brzo razvijala. Hrvatska je korpusna lingvistika držala korak sa svjetskom do početka osamdesetih godina prošloga stoljeća, a danas nastoji uhvatiti priključak primjenom suvremenih tehnologija i standarda u sastavljanju korpusa.

Zbog brojnih razloga navedenih u ovome radu, suvremeni bi jezični korpusi trebali biti standardizirani i međusobno kompatibilni. Za sastavljanje dobrog korpusa usklađenost sa svjetskim standardima za obilježavanje u ovome je trenutku gotovo važnija od ostalih faktora, kao npr. brzine pretraživanja. Stoga je u ovome radu posvećena znatna pozornost standardima i jezicima za obilježavanje. *Hrvatski nacionalni korpus*, kao posljednji sastavljani hrvatski korpus u potpunosti je u skladu s tim načelom, te slijedi najnovije standarde i preporuke po kojima se sastavljaju suvremeni nacionalni korpusi (XCES, UNICODE).

S obzirom na uniforman oblik i strukturu ulaznih podataka korpusa, mnogi već postojeći alati koji nisu ovisni o specifičnome jeziku mogu se primjenjivati na sve korpuse koji su u skladu sa standardom. Stoga danas nije prioritetno razvijati posebne alate koji bi bili korišteni za pojedine korpuse već je mnogo korisnije većinu napora uložiti u dovođenje korpusa u standardan oblik koji se može obrađivati postojećim alatima. Razvoj posebnih alata potreban je u slučajevima kada ih je zbog specifičnosti jezika nužno razvijati, kada postojeći alati ne zadovoljavaju uvjete u kojima se sastavlja korpus, ili ne zadovoljavaju potrebe obrade korpusnih podataka.

U prvom je slučaju najveći nedostatak u hrvatskoj korpusnoj lingvistici nepostojanje POS označivača. Preduvjet za njegovu izradu sastavljanje je

[207] Erjavec (1999)

elektroničkoga morfološkoga leksikona koji se trenutno sastavlja u Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu. Drugi je preduvjet razvoja POS označivača pribavljanje velike količine morfosintaktički obilježene jezične građe, pa će HNK biti iskorišten kao prikladan kvalitativan i kvantitativan izvor za tu vrstu obrade prirodnoga jezika.

S obzirom na okružja obrade, postojeći alati za obrade velikih korpusa najčešće su namijenjeni *Unix* okružju. Kako se HNK obrađuje na *Windows NT* platformi kojoj nedostaju alati za pohranu i pretraživanje velikih korpusa, bilo je nužno razvijati alate za pretraživanje i povezivanje korpusa sa WWW-om koji bi zadovoljili postavljene ciljeve. Unatoč okružju, rezultati bi obrade trebali uvijek biti jednaki, bez obzira kojim se alatom sama obrada odvija, a jedina razlika može biti u upotrijebljenim računalnim resursima.

Dodatak A

Elektroničko izdanje Platonove *Države* obilježeno SGML-om na razini rečenice prema TEI standardu.[208]

```
<tei.2>
<teiHeader type="text" date.created="03 Nov 1997">
<fileDesc>
<titleStmt>
      <title type="main">Dr&zcaron;ave</title>
      <title type="gmd">Hrvatska elektronska verzija</title>
      <author>Platon</author>
      <respStmt>
        <resp>sastavio</resp>
        <name teiform="name">Marko Tadi&cacute;</name></respStmt>
</titleStmt>
<publicationStmt>
<distributor>Zavod za lingvistiku Filozofskoga fakulteta
Sveu&ccaron;ili&scaron;ta u Zagrebu</distributor>
</publicationStmt>
<sourceDesc>
<biblStruct lang="HR">
<monogr>
<author>Platon</author>
<title lang="HR" type="main">Dr&zcaron;ava</title>
<imprint>
<pubPlace>Zagreb</pubPlace>
<publisher>Fakultet politi&ccaron;kih nauka Sveu&ccaron;ili&scaron;ta
u Zagrebu, Sveu&ccaron;ili&scaron;na naklada Liber</publisher>
<date value="1977">1977</date></imprint></monogr>
</biblStruct></sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc>
<p>Tekst je utipkan u okviru projekta Ra&ccaron;unalna obradba
hrvatskoga knji&zcaron;evnog jezika.</p></projectDesc>
</encodingDesc>
<profileDesc>
      <langUsage>
      <language id="HR">hrvatski</language></langUsage>
      <textDesc>
      <channel mode="w"></channel>
      <constitution type="single"></constitution>
      <derivation type="translation">Preveo s gr&ccaron;koga Martin
Kuzm&cacute; 1905, preradio Zlatko Ga&scaron;parovi&cacute; 1942,
preradio i usporedio s gr&ccaron;kim izvornikom Damir Salopek
1976</derivation>
      <domain></domain>
      <factuality></factuality>
      <interaction></interaction>
      <preparedness></preparedness>
      <purpose></purpose>
      </textDesc>
    </profileDesc>
    <revisionDesc>
      <change>
      <date value="03-11-1997">03-11-1997</date>
```

[208] cf. Erjavec, Lawson, Romary (1998)

```
      <respstmt teiform="respStmt">
        <name>Marko Tadic</name>
        <resp>Full TEI Encoding</resp>
      </respstmt>
      <item teiform="item">Delivary of final encoded version</item>
      </change>
    </revisionDesc>
</teiHeader>
<text>
<body id="b1">
<div type="book" id="d1">
<head id="h1">I</head>
<div id="d2">
<milestone unit="folio" n="327">
<p id="d2p1">
<seg id="d2p1seg1">Odoh ju&ccaron;er dolje u Pirej s Glaukonom
Aristonovim da se pomolim bo&zcaron;ici i s namjerom da ujedno vidim
kako &cacute;e prirediti svetkovinu, budu&cacute;i da su je sada prvi
put svetkovali. </seg>
<seg id="d2p1seg2">Lijepim mi se u&ccaron;inio i doma&cacute;i ophod,
ali se isto tako pristalim &ccaron;inio i ophod u kojem su
i&scaron;li Tra&ccaron;ani. </seg>
<seg id="d2p1seg3">Nakon
<milestone unit="folio" n="b">&scaron;to se pomolismo i nagledasmo,
htjedosmo oti&cacute;i u grad.
</seg>
<seg id="d2p1seg4">Kad smo tako krenuli ku&cacute;i, ugleda nas
izdaleka Polemarh Kefalov i nalo&zcaron;i sluzi da potr&ccaron;i te
nam re&ccaron;e da ga pri&ccaron;ekamo. </seg>
<seg id="d2p1seg5">Sluga me uhvati otraga za odijelo i re&ccaron;e:
</seg></p>
<p id="d2p2">
<seg id="d2p2seg1">- Veli vam Polemarh da pri&ccaron;ekate.
</seg></p>
<p id="d2p3">
<seg id="d2p3seg1">Ja se okrenem i upitam gdje je on. </seg></p>
<p id="d2p4">
<seg id="d2p4seg1">- Evo dolazi za vama; nego pri&ccaron;ekajte.
</seg></p>
<p id="d2p5">
<seg id="d2p5seg1">- Pa pri&ccaron;ekat &cacute;emo - re&ccaron;e
Glaukon. </seg></p>
<milestone unit="folio" n="c">
<p id="d2p6">
<seg id="d2p6seg1">I malo poslije do&dstrok;o&scaron;e Polemarh, i
Adimant, brat Glaukonov, i Nikerat Nikijin i neki drugi, koji su,
&ccaron;ini se, sudjelovali pri sve&ccaron;anosti. </seg>
<seg id="d2p6seg2">Tada re&ccaron;e Polemarh: </seg></p>
<p id="d2p7">
<seg id="d2p7seg1">- Sokrate, u grad ste, mislim, krenuli? </seg></p>
<p id="d2p8">
<seg id="d2p8seg1">- Ne misli&scaron; krivo - rekoh ja. </seg></p>
<p id="d2p9">
<seg id="d2p9seg1">- Vidi&scaron; li - re&ccaron;e on - koliko nas
je? </seg></p>
...
</div></div>
</body></text>
</tei.2>
```

Dodatak B

Elektroničko izdanje *Večernjeg lista* tokenizirano uporabom 2XML alata i obilježeno XML-om prema XCES standardu:

```
<?xml version="1.0"?>
  <!DOCTYPE cesDoc PUBLIC "-//CES//DTD XML cesDoc//EN"
                                  "xcesDoc.dtd" [
        ]>
<cesDoc version="3.19">
  <cesHeader type="text"  version="3.19">
    <fileDesc>
      <titleStmt>
        <h.title>Electronic version of Vecernji list,
vl990311</h.title>
        <respStmt>
          <respType>TEI markup prepared by</respType>
          <respName>Marko Tadic</respName>
        </respStmt>
      </titleStmt>
      <extent>
          <wordCount>114776</wordCount>
          <byteCount units="bytes">1320303</byteCount>
      </extent>
      <publicationStmt>
        <distributor>ELAN ZAG and project MZT RH 130729</distributor>
          <pubAddress>Institute of linguistics, Ivana Lucica 3, 10
000 Zagreb, Croatia</pubAddress>
          <telephone>+38516120142</telephone>
          <fax>+38516156879</fax>
          <eAddress>zzl@ffzg.hr</eAddress>
          <idno>76676665676</idno>
        <availability status="restricted" region="Croatia">
        </availability>
        <pubDate>1999-12-20</pubDate>
      </publicationStmt>
      <sourceDesc>
        <biblStruct>
          <monogr>
            <h.title>Vecernji list</h.title>
            <h.author>Du. Tadic</h.author>
            <imprint>
              <pubPlace>Zagreb</pubPlace>
              <publisher>Vecernji list</publisher>
              <pubDate>1999-03-11</pubDate>
            </imprint>
          </monogr>
        </biblStruct>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <projectDesc>ZAG contribution to ELAN project consists of 2
million corpus of written Croatian language compiled of texts from
leading Croatian daily newspaper "Vecernji list". The corpus has been
collected in the Institute of linguistics, Faculty of Philosophy,
University of Zagreb in the frame of the project "Croatian
contribution to the ELAN project" granted by the Ministry of Science
and Technology of Republic of Croatia under No. 130729 and
ELAN</projectDesc>
```

```
    </encodingDesc>
    <profileDesc>
      <langUsage>
        <language id="hr" iso639="hr">Croatian</language>
      </langUsage>
      <textClass>
      </textClass>
    </profileDesc>
</cesHeader>
<text lang="HR">
<BODY>
<DIV0 type="article" n="vl990311ck01">
<HEAD type="nn">
      <tok type="R"><orth>TRAGEDIJA</orth></tok>
      <tok type="R"><orth>U</orth></tok>
      <tok type="R"><orth>VOLTINOM</orth></tok>
      <tok type="R"><orth>NASELJU</orth></tok>
      <tok type="R"><orth>U</orth></tok>
      <tok type="R"><orth>ZAGREBU</orth></tok>
</head>
  <head type="na">
    <s>
      <tok type="R"><orth>Elektri&#269;ara</orth></tok>
      <tok type="R"><orth>pod</orth></tok>
      <tok type="R"><orth>tu&#353;em</orth></tok>
      <tok type="R"><orth>ubila</orth></tok>
      <tok type="R"><orth>struja</orth></tok>
   </s>
 </head>
  <head type="pn">
    <s>
      <tok type="I"><orth>*</orth></tok>
      <tok type="R"><orth>Tragedija</orth></tok>
      <tok type="R"><orth>u</orth></tok>
      <tok type="R"><orth>ku&#263;ici</orth></tok>
      <tok type="R"><orth>odigravala</orth></tok>
      <tok type="R"><orth>se</orth></tok>
      <tok type="R"><orth>pred</orth></tok>
      <tok type="R"><orth>o&#269;ima</orth></tok>
      <tok type="R"><orth>sedmogodi&#353;njeg</orth></tok>
      <tok type="R"><orth>sina</orth></tok>
      <tok type="I"><orth>,</orth></tok>
      <tok type="R"><orth>koji</orth></tok>
      <tok type="R"><orth>je</orth></tok>
      <tok type="R"><orth>neprestano</orth></tok>
      <tok type="R"><orth>ponavljao</orth></tok>
      <tok type="I"><orth>:</orth></tok>
</s>
...
...
</DIV0>
</BODY>
</text>
</cesDoc>
```

Dodatak C

C1:

Shema obilježavanja *Jednomilijunkoga korpusa hrvatskoga književnoga jezika*, objašnjenje njezinih dijelova i njezina primjena:[209]

```
$#pxxxggs$
      p = potkorpus (Drama, Novine, Proza, Stihovi, Udžbenici)
      xxx = broj uzorka u potkorpusu
      gg = godina izdanja djela (?? za nepoznate)
      s = sastavnica tekstovne strukture (G glavni naslov, I
      nadnaslov, N naslov, P podnaslov, T tijelo teksta, S sažetak, O
      potpis, K kazalo/sadržaj, M motto, posveta)
      npr.

$#P03050G$
      Luka Perković: Škrinja
      $#P03050N$
      Djed
      $#P03050T$
      Bile su u djeda Vuje dvije krave. Treba odmah reći: nisu to
bile samo krave, jer su osim vimenom služile gospodaru i jarmom.
Njihov rog nije bio oštar da ubode, kad mu priđeš, već blag i mek, da
i glavu nasloniš na nj, ako ti je umorna. Njihova noga nije bila
opaka da udari, kad s loncem pod vime dođeš, već pitoma i poslušna,
da ti i zadnju kap dopusti izmusti. A kad su u jarmu bile, eh zažmiri
tada, da ti milota suzu ne izmami: njihov hod nije bio ohol da se
svidi, niti mogao da zaplaši, već ustrajan i jednakomjeran da ti usta
sama od sebe šapću: stanite malo, krave, da vas poljubim u to vaše
ćelo pametno.
      ...
      $#EOD$
```

Napomena: $ na početku i kraju naredbe odgovara početku i završetku oznake, što danas odgovara <oznaka> i </oznaka>.

[209] cf. Tadić (1991):171

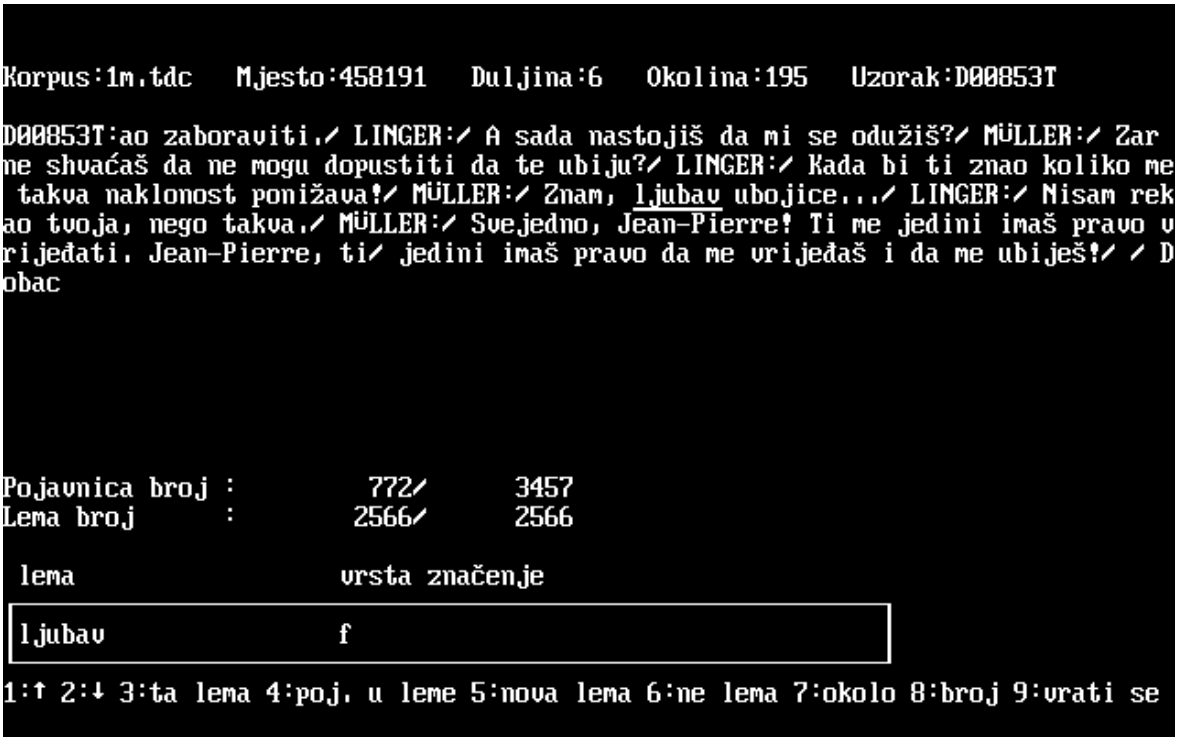
C2:

Abecedni i čestotni popis pojavnica *Jednomilijunskoga korpusa hrvatskoga književnoga jezika*:

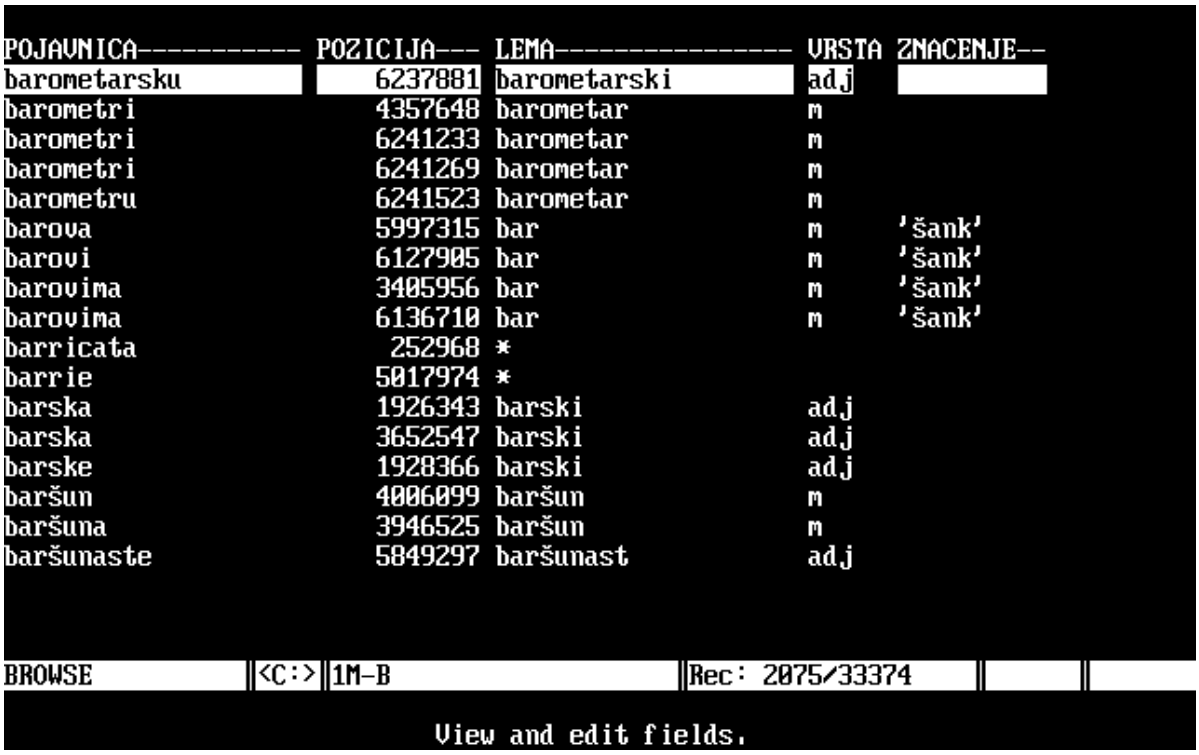
abecedni popis pojavnica		čestotni popis pojavnica	
a	7852	i	41923
a-l	2	u	27093
aa	1	je	24298
a-a-a	2	se	23630
aaa	4	da	15872
aaaa	2	na	13436
aaaaa	1	za	8320
aaaaaa	3	a	7852
aaaan	4	ne	7852
aaah	1	su	7596
aaan	1	od	6319
aaann	1	to	6019
aarhusa	1	što	5951
abada	1	s	5623
abadi	1	kao	4778
abatjouri	1	sam	4777
abažurom	2	o	3665
abbot	1	ja	3647
abdeselam	1	će	3433
abdominalno	1	bi	3399
abdu	1	ti	3371
abdula	2	koji	3286
abdulah	1	iz	3268
abdul-hamida	1	sve	3246
abdurahmana	1	nije	2996
abe	1	mi	2870
abeba	2	ili	2760
abebi	2	kako	2732
abeceda	2	ali	2709
abecede	1	samo	2642
abecedi	1	kad	2579
abecedno	1	tako	2525
abecednom	2	pa	2464
abecedu	3	po	2436
abesiniji	1	još	2301
abesiniju	1	te	2290
abesinku	1	do	2282
abiogeneza	1	ga	2167
abnormalni	1	ni	2157
abnormalnim	1	sa	2002
abonent	1	koje	1930
abonman	1	me	1921
abrazije	1	li	1867
...		...	

C3:

Radni zaslon programa za lematizaciju:[210]



Baza nakon lematizacije:[211]



[210] cf. Tadić (1991):175
[211] cf. Tadić (1991):176

C4:

KWIC oblik s naznakom uzorka s lijeve strane, razvrstan prema stožernici te desnoj okolini[212]

```
S01669T ljubav razilazi se s maglom /Ah ta    ljubav sva već istrošena / /Konaćno
D00853T e sam sažeo svu svoju neizživljenu    ljubav, sve ono lijepo što je /čovj
P01567T ve ode ... fuć ... ljubav ili nije    ljubav, sve vam dode isto. Decki /s
D004??T  tvoj prljavi novac, /prodao svoju    ljubav, svoga boga! Ona je moj rasp
S014??T jigu rastvorenu /okom u oko /čitaj    ljubav svoju. / /Ne treba više noć
S00763T gu više spavati, ni sniti, /a nije    ljubav što mi srce kida, /ni strah
S01877T  srce tvoje,  /sva snaga vatrena. /   ljubav, /što se u ljubav pretvorila
S00647T maš dvadeset godina /i djetinjastu    ljubav, što tepa noćnoj kiši? /Obla
S014??T /Znam vaš san, sastanke ilegalne, /   ljubav što u jedno srce ne može da
S00878T o gdje su krila koja donose veče, /   ljubav što u ledu izgara? / /Tražim
S01262T e onda i kroz žile stare; /duša je    ljubav što u sreći guče; /ti prsa h
U01477T ke /ljudske strasti — ljubomora i     ljubav, te zakon i ljubav prema dom
D020??T ijek, /I neka se urote vrazi svi, /   Ljubav tu imam za lijek. /RATKO I J
S003??T s koji sjaji iz bjeline /rasuta je    ljubav tvoga višeg sjaja /U njemu s
S014??T je bit će pjesme moje, /svrstat ću    ljubav u bunu stihova /kad se u pje
D00853T no, jedini od svih ljudi, /iskazao    ljubav u kojoj nije bilo ni trunka
S01262T mnogo čega posrće u hodu;  /tako i    ljubav — u ljudskome rodu — /sve
D01263T tovana /supruga njeguje evandeosku    ljubav u najbanalnijoj psećoj gužvi
S02078T /Anđeli, vi ste ruke, /koje bacaju    ljubav u nepovrat, /vi ste oči /koj
D004??T je patnje, moja krv, moja /vjera u    ljubav, u sreću, u čovjeka vape za
S00231T nuci — to su slavoluci /Kroz koje     ljubav u trijumfu ide! / / /HIMNA Z
D00853T aklonost ponižava! /MÜLLER: /Znam,    ljubav ubojice... /LINGER: /Nisam
```

[212] cf. Tadić (1991):174

Dodatak D

D1:

Abecedni popis pojavnica 30m korpusa:

	POJAVNICA	FRQ
1.	a	51365
2.	á	1
3.	Â	1
4.	AA	11
5.	A-a	1
6.	aÂ	1
7.	aaa	5
8.	Aaaa	4
9.	AAA-AAA	1
10.	AAAAAh	1
11.	Aaan	1
12.	Aachenskim	1
13.	Aachenu	3
14.	Aadenin	1
15.	AAG	2
16.	Aalders	1
17.	Aalena	2
18.	Aalenu	1
19.	Aarauu	1
20.	Aarhusa	1
21.	Aarne	2
22.	Aarne-Thompson	15
23.	Aarne-Thompsonovim	1
24.	Aarne-Thompsonovu	5
25.	Aaron	1
26.	Aarona	1
27.	Aaronu	1
28.	Aartsbischopejlik	1
29.	Aartsen	1
30.	AAS	1
31.	aau	1
32.	aaup	1
33.	AB	9
34.	Aba	1
35.	ababc	1
36.	Abache	4
37.	Abachea	2
38.	Abacheom	2
39.	Abacheovo	1
40.	Abacus	1
41.	Abadukonderu	2
42.	Abadukondre	3
43.	Abadukondrea	1
44.	Abadukondreu	2
45.	abadžija	1
46.	Abakus	1
47.	Abana	1
48.	Abanaca	1
49.	Abasa	1
50.	abažura	1
51.	ABB	67
52.	ABBA	1
53.	Abbada	1
54.	Abbadessa	1
55.	Abba-Oče	1
56.	Abbarra	1
57.	Abbasa	1
58.	Abbasu	1
59.	Abbe	3
60.	Abbé	1
61.	Abbea	1
62.	Abbey	1
63.	ABC	22
64.	Abdala	1
65.	Abdallaha	1
66.	Abdel	3
67.	Abderićanima	1
68.	abdicirali	1
69.	Abdić	18
70.	Abdića	28
71.	Abdićem	7
72.	Abdićev	1
73.	Abdićevih	2
74.	Abdićevim	1
75.	Abdićevo	7
76.	Abdićevoj	1
77.	Abdićevom	1
78.	Abdiću	3
79.	abdikacija	1
80.	abdikacije	1
81.	abdikacijom	2
82.	abdikaciju	1
83.	abdomen	8
84.	abdomena	31
85.	abdomenu	15
86.	abdominalna	4
87.	abdominalne	5
88.	abdominalni	1
89.	abdominalnih	4
90.	abdominalno	1
91.	abdominalnom	3
92.	abdominalnoperinealne	1
93.	abdominalnoperinealnim	1

94. abdominalnu	3
95. abdukcijom	1
96. Abdula	1
97. Abdulah	3
98. Abdulaha	8
99. Abdulahu	1
100. Abdul-Jabbar	1
101. Abdullah	7
102. Abdullaha	13
103. Abdullahaa	1
104. Abdullahu	1
105. Abdulsalam	1
106. Abdus	1
107. Abdus-selam	1
108. Abdykalydova	1
109. ABE	4
110. Abeceda	2
111. abecedariju	2
112. abecedi	2
113. abecedni	1
114. abecednim	5
115. abecedno	1
116. abecednom	2
117. abecednome	1
118. abecedu	2
119. Abel	4
120. Abela	3
121. Abelarda	1
122. Abele	2
123. Abelom	1
124. Abelu	1
125. Abendroth	1
126. Abe-Novaka	1
127. Abenteurer	1
128. ABEPERIŠA	2
129. aber	2
130. aberacija	2
131. aberacije	7
132. Aberdeana	1
133. Aberdeenu	1
134. Aberdine	1
135. Aberta	1
136. Abesinija	1
137. A-Bi	1
138. ABIĆ	14
139. abidćevštine	1
140. ABiH	4
141. Abijatara	1
142. abîme	3
143. Abimelek	1
144. abiotički	1
145. Abitanti	2
146. ablacija	4
147. ablacije	1
148. ablacijsku	1
149. ablativni	1
150. Ablondi	5
151. Ablondiju	1
152. ABN-AMRO	1
153. Abner	1
154. abnormalne	2
155. abnormalnih	2
156. abnormalnim	1
157. Abnormalno	4
158. abnormalnog	2
159. abnormalnost	1
160. abnormalnosti	1
161. abolicijom	1
162. abolicionističkih	1
163. abolira	1
164. abolirale	1
165. aboliran	3
166. A-bomba	1
167. A-bombama	1
168. A-bombe	7
169. A-bombu	1
170. abonenti	1
171. abonentsku	1
172. abonomana	1
173. abonomanom	1
174. aboridžina	5
175. aboridžinima	3
176. aboridžinske	2
177. Aboridžinskog	1
178. Aboriđinke	1
179. aboriginâ	1
180. Aborigini	1
181. aboriginima	1
182. abortion	5
183. abortira	1
184. abortirale	1
185. abortirane	1
186. abortirati	2
187. abortivno	1
188. abortus	21
189. abortusa	12
190. abortusima	1
191. abortusom	4
192. abortus-pilula	1
193. abortusu	3
194. Abraham	8
195. Abrahama	3
196. Abrahamom	2
197. Abrahamov	2
198. Abrahamova	1
199. Abrahamove	1
200. Abrahamovo	2

...

D2:

Odostražni popis pojavnica 30m korpusa:

POJAVNICA	FRQ
1. a	51365
2. á	1
3. Â	1
4. AA	11
5. A-a	1
6. aÂ	1
7. aaa	5
8. Aaaa	4
9. AAA-AAA	1
10. Čudaaa	1
11. Lejla-Lejlaaa	2
12. Paaa	1
13. baą	1
14. CAA	4
15. NFCA-a	1
16. giricaa	1
17. SDA-a	1
18. događaa	1
19. EAA	1
20. udrugaa	1
21. Abdullahaa	1
22. HIAA	1
23. poskupljenjaa	1
24. moojaa	1
25. chardonnayjaa	1
26. kakaa	1
27. maa	2
28. Yamaa	1
29. FIMA-a	1
30. SANAA	1
31. Spaa	1
32. Râa	1
33. kontejneraa	1
34. Peteraa	1
35. IRA-a	1
36. smatraa	1
37. ETA-a	1
38. dva-a	1
39. Vjesnikovaa	1
40. gospodarstvaÂ	1
41. Ba	5
42. B-a	4
43. Aba	1
44. baba	16
45. Agbaba	1
46. Ali-Baba	1
47. visibaba	2
48. Ježibaba	1
49. skladaba	2
50. zamjedaba	2
51. primjedaba	19
52. naredaba	2
53. odredaba	42
54. priredaba	15
55. usporedaba	1
56. uredaba	4
57. izvedaba	5
58. praizvedaba	1
59. plovidaba	1
60. prosudaba	2
61. faba	2
62. GABA	1
63. Ahaba	1
64. Diaba	1
65. Pundjaba	1
66. Punjaba	1
67. razglaba	1
68. Microlaba	1
69. urlaba	1
70. slaba	108
71. preslaba	5
72. Barnaba	4
73. Raba	22
74. Baraba	2
75. Taraba	1
76. graba	5
77. zloraba	1
78. zloporaba	33
79. uporaba	68
80. zlouporaba	21
81. upopraba	1
82. svraba	3
83. Saba	4
84. SAB-a	9
85. kasaba	1
86. štaba	60
87. generalštaba	21
88. Švaba	1
89. Szába	1
90. žaba	10
91. predodžaba	7
92. narudžaba	2
93. čovjek-žaba	2
94. izložaba	48
95. pritužaba	1
96. optužaba	11
97. ABBA	1
98. Barabba	1
99. BBB-a	8

100. Obba	1
101. curosbobba	1
102. ACCBA	1
103. ECB-a	1
104. lučba	1
105. skladba	39
106. uskladba	1
107. sladba	1
108. obradba	5
109. Preradba	4
110. gradba	1
111. izradba	5
112. Svadba	23
113. sljedba	7
114. sljedbâ	1
115. isljedba	1
116. primjedba	37
117. naredba	16
118. Odredba	75
119. priredba	33
120. poredba	1
121. rasporedba	1
122. usporedba	106
123. Uredba	9
124. Provedba	50
125. izvedba	62
126. praizvedba	21
127. Gojidba	3
128. gnojidba	90
129. selidba	8
130. Hranidba	8
131. plijenidba	1
132. Ženidba	12
133. krunidba	1
134. Kosidba	3
135. prosidba	1
136. plovidba	13
137. rezidba	1
138. godba	1
139. prilagodba	13
140. nagodba	6
141. pogodba	4
142. Kodba	2
143. SDB-a	4
144. Udba	26
145. UDB-a	8
146. uljudba	4
147. neuljudba	1
148. sudba	3
149. posudba	2
150. prosudba	22
151. beba	25
152. Marija-Beba	1
153. Feba	1
154. Gottlieba	1
155. jeba	4
156. zajeba	2
157. ždrijeba	2
158. primjeba	1
159. Makeba	2
160. Galeba	9
161. Kaleba	1
162. Gleba	1
163. koleba	2
164. pokoleba	3
165. MEBA	3
166. neba	140
167. doneba	1
168. Magreba	3
169. Zagreba	1464
170. Foto-Zagreba	1
171. pogreba	16
172. Maghreba	1
173. Škreba	35
174. preba	1
175. treba	7106
176. trěba	2
177. zatreba	20
178. potreba	523
179. upotreba	145
180. zloupotreba	27
181. jastreba	7
182. Dedakovića-Jas...	1
183. ustreba	5
184. naštreba	1
185. vreba	12
186. poseba	1
187. Bat-Šeba	1
188. teba	1
189. potrteba	1
190. cveba	2
191. Weba	22
192. web-a	2
193. Zeba	3
194. DŽEBA	1
195. DGB-a	1
196. KGB-a	5
197. HRHB-a	3
198. HZHB-a	1
199. IB-a	2
200. Habiba	1
201. Hebiba	1
202. Ćiba	1
203. FIBA	1
204. giba	6
205. nagiba	24

...

D3:

Čestotni popis pojavnica 30m korpusa:

POJAVNICA	FRQ
1. i	287853
2. u	252533
3. je	241718
4. se	143249
5. da	118606
6. na	113855
7. su	78945
8. za	75651
9. s	51981
10. a	51365
11. od	50944
12. ne	46232
13. to	41924
14. koji	40958
15. o	40860
16. što	40859
17. kao	33338
18. iz	30621
19. će	29584
20. bi	28607
21. sam	26298
22. nije	26128
23. Kako	24925
24. te	23734
25. ili	21255
26. ali	20019
27. do	19133
28. koje	18542
29. sve	18158
30. samo	17109
31. koja	15752
32. jer	14914
33. po	14513
34. tako	14454
35. biti	14190
36. više	13925
37. bio	13812
38. još	13746
39. pa	13254
40. godine	13168
41. već	12968
42. bilo	12888
43. kad	12769
44. može	12673
45. Mi	12265
46. smo	12139
47. sa	11859
48. Prema	11674
49. ni	11460
50. Hrvatske	11271
51. nakon	11268
52. zbog	10993
53. li	10869
54. ga	10775
55. No	9886
56. Ako	9781
57. nego	9586
58. on	9253
59. uz	9143
60. toga	8987
61. prije	8437
62. bez	8411
63. ima	8407
64. bila	8053
65. svoje	7938
66. Hrvatskoj	7767
67. ih	7601
68. dana	7593
69. godina	7486
70. tome	7466
71. taj	7413
72. mu	7383
73. Nisu	7305
74. treba	7106
75. vrlo	7019
76. jedan	6902
77. oko	6897
78. bih	6760
79. dok	6747
80. mogu	6727
81. vrijeme	6717
82. Ja	6648
83. ljudi	6629
84. kada	6485
85. oni	6374
86. danas	6369
87. sada	6235
88. bili	6010
89. nas	5995
90. tu	5952
91. nema	5802
92. između	5725
93. ona	5565
94. kod	5542
95. ipak	5530
96. gdje	5527
97. kojima	5518
98. me	5482
99. način	5463

100. Hrvatska	5447
101. koju	5410
102. tom	5275
103. među	5273
104. pod	5241
105. zato	5202
106. nam	5181
107. nekoliko	5148
108. Broj	5128
109. uvijek	5090
110. dr	4937
111. svojim	4860
112. svi	4857
113. Nacional	4834
114. čak	4831
115. Dobro	4805
116. onda	4715
117. ono	4709
118. nešto	4699
119. dva	4670
120. također	4605
121. rekao	4579
122. predsjednik	4575
123. protiv	4566
124. Naime	4530
125. im	4526
126. kojoj	4507
127. upravo	4496
128. ta	4457
129. prvi	4425
130. tri	4331
131. svoju	4329
132. iako	4308
133. dio	4290
134. ste	4251
135. mora	4239
136. tek	4195
137. svoj	4157
138. kojem	4144
139. odnosno	4134
140. posto	4017
141. tada	4000
142. svih	3985
143. Međutim	3967
144. neki	3926
145. drugi	3884
146. Pri	3881
147. pred	3853
148. koliko	3832
149. tko	3819
150. Zagrebu	3805
151. Zagreb	3802
152. neće	3796
153. hrvatski	3763
154. njih	3759
155. Riječ	3753
156. osim	3744
157. Dakle	3725
158. života	3723
159. pitanje	3713
160. radi	3697
161. njegova	3662
162. put	3631
163. nisam	3627
164. tim	3627
165. mnogo	3586
166. druge	3571
167. dvije	3547
168. prava	3533
169. možda	3509
170. život	3480
171. gotovo	3454
172. kaže	3442
173. pak	3418
174. ništa	3382
175. jednom	3366
176. ove	3357
177. čovjek	3353
178. niti	3333
179. mogao	3313
180. zapravo	3304
181. ti	3248
182. sv	3244
183. hrvatskog	3190
184. kuna	3190
185. strane	3183
186. one	3173
187. znači	3167
188. svim	3155
189. Stoga	3136
190. Hrvatsku	3118
191. vam	3089
192. imaju	3080
193. vlasti	3051
194. predsjednika	3046
195. zašto	3045
196. posebno	3039
197. poslije	3027
198. stranke	3019
199. njega	2985
200. mjesto	2980
201. neke	2951
202. reći	2915
203. malo	2831
204. njima	2829
205. često	2809

...

D4:

Čestotni popis dvopojavnica 30m korpusa:

POJAVNICA12	POJAVNICA1	POJAVNICA2	FRQ
1. da je	da	je	20795
2. da se	da	se	17464
3. je u	je	u	11796
4. što je	što	je	10400
5. koji je	koji	je	8999
6. koji su	koji	su	7726
7. i u	i	u	7547
8. se u	se	u	7188
9. da su	da	su	6704
10. je i	je	i	6599
11. su se	su	se	6545
12. to je	to	je	6495
13. da će	da	će	6185
14. je to	je	to	5913
15. će se	će	se	5877
16. bi se	bi	se	5848
17. da bi	da	bi	5782
18. je da	je	da	5514
19. u Hrvatskoj	u	Hrvatskoj	5330
20. Kako je	Kako	je	5076
21. što se	što	se	4863
22. koja je	koja	je	4805
23. kao i	kao	i	4492
24. koji se	koji	se	4405
25. i na	i	na	4373
26. je na	je	na	4021
27. i to	i	to	3699
28. se na	se	na	3583

29. se i	se	i	3538
30. ne može	ne	može	3537
31. u Zagrebu	u	Zagrebu	3509
32. i da	i	da	3506
33. koje su	koje	su	3411
34. kao što	kao	što	3335
35. su u	su	u	3328
36. mi je	mi	je	3207
37. se da	se	da	3176
38. se ne	se	ne	3158
39. što su	što	su	3157
40. je bio	je	bio	3143
41. kad je	kad	je	3132
42. koje je	koje	je	3098
43. Kako bi	Kako	bi	3080
44. ne bi	ne	bi	2983
45. o tome	o	tome	2854
46. ono što	ono	što	2781
47. bio je	bio	je	2697
48. u tom	u	tom	2679
49. i za	i	za	2658
50. kako se	kako	se	2583
51. u kojoj	u	kojoj	2570
52. jer je	jer	je	2510
53. ne samo	ne	samo	2454
54. u kojem	u	kojem	2429
55. ali i	ali	i	2427
56. a ne	a	ne	2395
57. je za	je	za	2384
58. više od	više	od	2382
59. a u	a	u	2310
60. koja se	koja	se	2253
61. mu je	mu	je	2200

…

D5:

Popis dvopojavnica hrvatsko-engleskog paralelnog korpusa poredan prema padajućoj vrijednosti uzajamne obavijesnosti (UO):

POJAVNICA 1	FRQ 1	POJAVNICA 2	FRQ 2	POJAVNICA 12	FRQ 12	UO
1.etničko	10	čišćenje	12	etničko čišćenje	10	13,9419263968436
2.Sveti	13	Otac	13	Sveti Otac	11	13,5854410799199
3.Svete	16	Stolice	11	Svete Stolice	11	13,5268888975648
4.Madeleine	10	Albright	17	Madeleine Albright	10	13,4394260563144
5.štednih	12	uloga	18	štednih uloga	11	13,2314330140386
6.Crnoj	16	Gori	19	Crnoj Gori	14	13,0863163061788
7.Europskoj	19	uniji	15	Europskoj uniji	13	13,0725105066537
8.Europsku	22	uniju	12	Europsku uniju	12	13,0674572789275
9.kardinala	21	Alojzija	11	kardinala Alojzija	10	12,9970679510361
10. međunarodnu	17	zajednicu	14	međunarodnu zajednicu	10	12,9539992291442
11. Crna	19	Gora	15	Crna Gora	11	12,83150240715
12. pomorstva	13	prometa	26	pomorstva prometa	13	12,8264491794237
13. Alojzija	11	Stepinca	27	Alojzija Stepinca	11	12,7720013954013
14. Crne	16	Gore	25	Crne Gore	14	12,6903876298476
15. Nacionalnom	17	parku	19	Nacionalnom parku	11	12,6509301615081
16. Dinka	17	Šakića	26	Dinka Šakića	15	12,6458769337818
17. Carlos	13	Westendorp	30	Carlos Westendorp	13	12,6199983019562
18. Pavao	16	II	25	Pavao II	13	12,5834724259311
19. istočnoj	30	Slavoniji	18	istočnoj Slavoniji	17	12,5375361417643
20. istočne	22	Slavonije	19	istočne Slavonije	13	12,519969483625
21. međunarodnoj	30	zajednici	14	međunarodnoj zajednici	11	12,2720749985359
22. B	39	rith	14	B rith	14	12,2414866787025
23. B	39	nai	15	B nai	15	12,2414866787025
24. nai	15	B	39	nai B	15	12,2414866787025
25. turističke	33	sezone	18	turističke sezone	14	12,1199246988216
26. kosovskih	24	Albanaca	39	kosovskih Albanaca	22	12,1159557966187

27.	Republiku	27	Srpsku	24	Republiku Srpsku	15	12,0939294902887
28.	Republici	27	Srpskoj	16	Republici Srpskoj	10	12,0939294902887
29.	svemu	26	sudeći	17	svemu sudeći	10	12,0609144330607
30.	Slobodan	17	Milošević	33	Slobodan Milošević	12	11,9799944376771
31.	logora	27	Jasenovac	18	logora Jasenovac	10	11,9240044888463
32.	ovim	28	prostorima	21	ovim prostorima	12	11,9121790534495
33.	Zlatko	20	Mateša	35	Zlatko Mateša	12	11,6606402864536
34.	helsinškog	11	odbora	59	helsinškog odbora	11	11,6442458482029
35.	Haaški	15	sud	44	Haaški sud	11	11,6199983019562
36.	oružanih	19	snaga	49	oružanih snaga	15	11,5711421356145
37.	državni	32	sabor	20	državni sabor	10	11,5268888975648
38.	visoki	25	predstavnik	26	visoki predstavnik	10	11,5045210845363
39.	posebnim	16	odnosima	45	posebnim odnosima	11	11,4944674198724
40.	visokog	19	predstavnika	35	visokog predstavnika	10	11,4716064620636
41.	nautičkog	10	turizma	67	nautičkog turizma	10	11,460799707107
42.	Privredna	11	banka	71	Privredna banka	11	11,3771417780601
43.	Franje	24	Tuđmana	43	Franje Tuđmana	14	11,3230165641991
44.	Novom	29	groblju	28	Novom groblju	11	11,3209845990169
45.	državnog	50	sabora	37	državnog sabora	24	11,2585418428822
46.	ratne	47	zločine	54	ratne zločine	32	11,2174125437237
47.	prvom	35	redu	23	prvom redu	10	11,1959720194501
48.	Ivan	71	Pavao	16	Ivan Pavao	14	11,1844967001177
49.	Ante	40	Jelavić	23	Ante Jelavić	11	11,1408304652577
50.	Međunarodna	63	zajednica	60	Međunarodna zajednica	45	11,134571474786

...

Dodatak E

Jedan od oblika XSL datoteke[213] probne inačice pretraživanja s posoću XML/XSL tehnologije:

```
<?xml version="1.0" encoding="windows-1252" ?>
       <xsl:stylesheet xmlns:xsl="http://www.w3.org/TR/WD-xsl"
xmlns="http://www.w3.org/TR/REC-html40" result-ns="">
      <xsl:script language="JavaScript">
var a=0;var b=0;var c=0;
 function brojic()
 { c++; }
 function brojib()
 { b++; }
 function brojia()
 { a++; }
      </xsl:script>
<xsl:template match="/">
    <HTML>
    <BODY>
    <DIV Style="font-family: Arial, sans-serif; font-size: 10pt;
text-align: left">
      <xsl:apply-templates select="BODY" />
    <B>
    <HR />ti
    </B>:
      <xsl:eval>a;</xsl:eval>
    <B>takvi</B>:
      <xsl:eval>b;</xsl:eval>
    <B>>>> UKUPNO</B>:
      <xsl:eval>a+b;</xsl:eval>
    </DIV>
    </BODY>
    </HTML>
      </xsl:template>
<xsl:template match="*">
  <xsl:apply-templates />
</xsl:template>
<xsl:template match="*/S[$any$ W $ieq$ 'ti' and $any$ .//W/text()
$ieq$ 'takvi']">
  <HR size="1" />
  <DIV Style="font-family: Arial, sans-serif; font-size: 10pt; text-
align: left">
  <B>
  <xsl:eval>brojic();</xsl:eval>
  <xsl:eval>c;</xsl:eval>
  </B>
<xsl:for-each select=".//W">
  <xsl:choose>
    <xsl:when test=".[. $ieq$ 'ti' or . $ieq$ 'takvi']">
  <SPAN STYLE="color:red;background-color:yellow; font-weight:bold">
     <xsl:value-of />
  </SPAN>
```

[213] XSL datoteka pri svakom se upitu nanovo dinamički generira

```
  </xsl:when>
   <xsl:otherwise>
     <xsl:value-of />
   </xsl:otherwise>
  </xsl:choose>
 </xsl:for-each>
<xsl:for-each select=".//W">
   <xsl:choose>
    <xsl:when test=".[. $ieq$ 'ti']">
      <xsl:eval>brojia();</xsl:eval>
    </xsl:when>
 <xsl:when test=".[. $ieq$ 'takvi']">
   <xsl:eval>brojib();</xsl:eval>
 </xsl:when>
   </xsl:choose>
</xsl:for-each>
<DIV Style="font-family: Arial, sans-serif; font-size: 8pt; text-
align: right">
  <u>
  <i>Izvor</i>
  </u>:
    <xsl:value-of select="ancestor(*[@n])/@n" />
  <u>
  <i>Tip</i>
  </u>:
    <xsl:value-of select="ancestor(*[@type])/(@type)" />
  </DIV>
  </DIV>
   </xsl:template>
  </xsl:stylesheet>
```

Dodatak F

F1:

Dio opće statistike pristupa poslužniku na adresi http://www.hnk.ffzg.hr/ na kojem se nalazi probna inačica HNK-a:

Item	Value
Hits	261182
Total Data Transferred	7.28 gigabytes
Total Visiting Users	28871
Time Period	November 27, 1998, 08:43 AM to December 31, 2000, 11:47 PM
Average Hits per User	9.05
Average Users per Day	37.69
Average Data Transferred per Day	9.74 megabytes
Hits cached by Client	67983 (26.03%)
Report generated on	January 11, 2001 at 11:44 AM
Incomplete downloads/file requests	3037 (1.16%)
Log spans a period of	766 days
Total failed requests	16574 (6.35%)
Unique IP Addresses	9480
Average Data Transferred per User	264.50 kilobytes
Average Hits per Day	340.97
Average Data Transferred per Hit	29.24 kilobytes
Each user has visited approximately	3.05 times
Hits on Pages	123105
Hits on Files	18620
Hits on Images	102883

F2:

Države, odnosno domene s kojih je pristupano probnoj inačici HNK-a do 1. siječnja 2001. godine, te broj pristupa i njihov razmjer:

Domain Name	Hits	Percentage
Croatia/Hrvatska (.hr)	119542	63.01%
Germany (.de)	18312	9.65%
Austria (.at)	2848	1.50%
Slovenia (.si)	1378	0.73%
Australia (.au)	870	0.46%

Italy (.it)	770	0.41%
Yugoslavia (.yu)	698	0.37%
Sweden (.se)	669	0.35%
Bosnia and Herzegowina (.ba)	502	0.26%
United Kingdon (.uk)	396	0.21%
Switzerland (.ch)	326	0.17%
Denmark (.dk)	276	0.15%
Hungary (.hu)	205	0.11%
Israel (.il)	166	0.09%
Greece (.gr)	145	0.08%
Macedonia (.mk)	118	0.06%
Finland (.fi)	96	0.05%
Portugal (.pt)	71	0.04%
Ukraine (.ua)	65	0.03%
Bulgaria (.bg)	64	0.03%
Estonia (.ee)	57	0.03%
Old ARPA-net (.arpa)	47	0.02%
United Arab Emirates (.ae)	40	0.02%
Argentina (.ar)	32	0.02%
Chile (.cl)	21	0.01%
Mexico (.mx)	17	0.01%
Oman (.om)	15	0.01%
Turkey (.tr)	10	0.01%
Hong Kong (.hk)	7	0.00%
India (.in)	5	0.00%
.int	5	0.00%
Lithuania (.lt)	3	0.00%
Singapore (.sg)	2	0.00%
Kyrgzstan (.kg)	2	0.00%
Thailand (.th)	1	0.00%
Dominican Republic (.do)	1	0.00%
Guadeloupe (.gp)	1	0.00%

Literatura:

1. Ball, N. Catherine (1996), *Concordances and Corpora*, Corpus linguistics course, Department of Linguistics, Georgetown University, Washington DC, preuzeto 22. studenog 2000, sa WWW: http://www.georgetown.edu/cball/corpora/tutorial.html
2. Bank of English (2000), *Bank of English*, Collins *COBUILD*, preuzeto 23. ožujka 2001, sa WWW: http://titania.cobuild.collins.co.uk/boe_info.html
3. Bekavac, Božo (2000), *XML Workshop*, Croinfo 2000, Dubrovnik, radionica održana 17. listopata 2000, PPT prezentacija, nalazi se na WWW: http://www.hnk.ffzg.hr/xml-ws01
4. Boras, Damir (1998), *Teorija i pravila segmentacije teksta na hrvatskom jeziku*, doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu
5. Bratanić, Maja (1991), *Korpusna lingvistika ili sretan susret*, Radovi Zavoda za slavensku filologiju 27, Zagreb, str. 145-160
6. Bray, T., Paoli, J., Sperberg-McQueen, C. M. (1998), *Extensible Markup Language (XML) Version 1.0. W3C Recommendation*, preuzeto 11. studenog 2000, sa WWW: http://www.w3.org/TR
7. Brew, Chris (2000), *XML and Linguistic Annotation*, Radionica, PPT prezentacija, preuzeto 2. prosinca 2000, sa WWW: http://www.ling.ohio-state.edu/~cbrew
8. British National Corpus (BNC) (1997), *What is the BNC?*, Oxford University Computing Services, Oxford, posljednja promjena: 23. siječnja 2001, preuzeto 18. veljače 2001, sa WWW: http://info.ox.ac.uk/bnc/what/index.html
9. Bujas, Željko (1975), *Ivan Gundulić "Osman", Kompjutorska konkordancija*, Sveučilišna naklada Liber, Zagreb
10. Bulaja, Zvonimir (1999), *Klasici hrvatske književnosti*, CD-Rom, Bulaja naklada, Zagreb
11. Burnard, Lou & Sperberg-McQueen, C. M. (1995), *TEI Lite: An Introduction to Text Encoding for Interchange*, broj dokumenta: TEI U 5, Chicago-Oxford, preuzeto 21. veljače 2001, sa WWW: http://www-tei.uic.edu/orgs/tei/intros/teiu5.html
12. Burnard, Lou (1991), *What is SGML?*, preuzeto 11. veljače 2001, sa WWW: http://euclid.ii.metu.edu.tr/~corpus/sgml/article.html
13. Center of Computational Linguistics (CCL) (2000), *Systematic Dictionary of Corpus Linguistics*, Vytautas Magnus University, Kaunas, Litva, preuzeto 18. veljače 2001, sa WWW: http://donelaitis.vdu.lt/publikacijos/SDoCL.htm
14. Chomsky, Noam (1991), *Jezik i problemi znanja*, SOL, Zagreb
15. Clark, James & DeRose, Steve (1999) *XML Path Language (XPath),* Version 1.0. W3C Recommendation, preuzeto 11. studenog 2000, sa WWW: http://www.w3.org/TR/xpath
16. Clark, James (1999), *XSL Transformations (XSLT),* Version 1.0. W3C Recommendation, preuzeto 11. studenog 2000, sa WWW: http://www.w3.org/TR/xslt
17. Corpus Encoding Standard (1996), *Document CES 1*, Version 1.4, listopad 1996, preuzeto 28. studenog 2000, sa WWW http://www.cs.vassar.edu/CES/

18. Delač, Damir (1999), *Hrvatski računalni rječnik*, Infocentar 1999, preuzeto 15. veljače 2001, sa WWW: http://www.infocentar.hr/hrr

19. EAGLES (1996a), *Preliminary Recommendations on Corpus Typology*, svibanj 1996, preuzeto 23. studenog 2000, sa WWW: http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html

20. EAGLES (1996b), *Preliminary Recommendations on Text Typology*, lipanj 1996, preuzeto 22. studenog 2000, sa WWW: http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html

21. EAGLES (1996c), *Guidelines for Linguistic Software Development*, GLOSIX verzija 0, travanj 1996, preuzeto 22. studenog 2000, sa WWW: http://www.lpl.univ-aix.fr/projects/multext/LSD/LSD2.html

22. Erjavec, Tomaž & Lawson, Ann & Romary, Laurent (1998), *East meets West – A compendium of Multilingual Resources*, TELRI CD-ROM, Disk 1

23. Erjavec, Tomaž (1997), *Introduction to SGML: Standard Generalized Markup Language*, EUROLAN 1997, Summer School in Corpus Linguistics, Tusnad (Rumunjska)

24. Erjavec, Tomaž (1998), *The MULTEXT-East Slovene Lexicon*, Proceedings of the 7th Electrotechnical Conference ERK '98, Portorož, Slovenija, Vol. B, str. 189-192, preuzeto 28. ožujka 2000, sa WWW: http://nl.ijs.si/et/Bib/ERK98/erk/

25. Erjavec, Tomaž (1999), *Tagging Slavic Corpora*, pozvano predavanje na Sveučilištu u Tübingenu 15. prosinca 1999, preuzeto 25. studenog 2000, sa WWW: http://nl.ijs.si/et/talks/SFB441/tue-slides/

26. Gibbons, John (1994), *Language and the Law*, London, Longman

27. Glossary (1999), *Institut für deutsche Sprache, Abteilung Lexik, Multilinguale Forschung*, posljednja promjena: 3. studenoga 1999, preuzeto 29. studenog 2000, sa WWW: http://solaris3.ids-mannheim.de/mlfglossar.html

28. Godfrey, John J. & Zampolli, Antonio (1996), *Language Resources*, u Survey of the State of the Art in Human Language Technology, Cambridge University Press, ISBN 0-521-59277-1

29. Grishman, Ralph & Calzolari, Nicoletta (1996), *Lexicons*, u *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, ISBN 0-521-59277-1

30. Grover, Claire & Matheson, Colin & Mikheev, Andrei (2000), *TTT: Text Tokenisation Tool*, Language Technology Group, Human Communication Research Centre, University of Edinburgh, Edinburgh, preuzeto 4. studenog 2000, sa WWW: http://www.ltg.ed.ac.uk/software/ttt/tttdoc.html

31. Henderson, John Charles (1999), *Exploiting diversity for natural language parsing*, doktorska disertacija, *The Johns Hopkins University*, kolovoz 1999, Baltimore, Maryland, preuzeto 15. veljače 2001, sa WWW: http://nlp.cs.jhu.edu/~jhndrsn/thesis.pdf

32. Hrvatski nacionalni korpus (HNK) (1999a), *Što je HNK?*, Zavod za lingvistiku Filozofskoga fakulteta u Zagrebu, Zagreb, posljednja promjena: 02. veljače 2001, preuzeto 18. veljače 2001, sa WWW: http://www.hnk.ffzg.hr/cilj.htm

33. Hrvatski nacionalni korpus (HNK) (1999b), *Pretraga 30m korpusa*, Zavod za lingvistiku Filozofskoga fakulteta u Zagrebu, Zagreb, posljednja promjena 6. veljače 2001, preuzimano od 20. veljače do 29. ožujka 2001, sa WWW: http://www.hnk.ffzg.hr/30m.htm

34. Hrvatski nacionalni korpus (HNK) (1999c), *Pretraga s pomoću XML/XSL tehnologije*, Zavod za lingvistiku Filozofskoga fakulteta u Zagrebu, Zagreb, preuzeto 23. veljače 2001, sa WWW: http://www.hnk.ffzg.hr/xml/

35. 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

36. Ide, Nancy (2000), *The XML Framework and Its Implications for Corpus Access and Use*, Data Architectures and Software Support for Large Corpora, LREC2000 Workshop Proceedings, ELRA, Paris-Athens, str. 28-32

37. Joscelyne, Andrew (1991), *The bigger the better? Corpora (co-) builder John Sinclair*, Language Industry Monitor, preuzeto 18. veljače 2001, sa WWW: http://www.lim.nl/monitor/sinclair.html

38. Koenig-Baumer, Esther (1999): *Corpus Query Processor (CQP), User's Manual*, posljednja promjena: 16. kolovoza 1999, preuzeto 8. veljače 2001, sa WWW: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/node1.html

39. Kučera, Henry & Francis, W. Nelson (1967a), *Brown Corpus manual*, Providence, Rhode Island, Department of Linguistics, Brown University, prošireno i dopunjeno izdanje, preuzeto 28. studenog 2000, sa WWW: http://helmer.hit.uib.no/icame/brown/bcm.html

40. Kučera, Henry & Francis, W. Nelson (1967b), *Computational Analysis of Present-Day American English*, Brown University Press, Rhode Island

41. Kuikka, Eila & Nikunen, Erja (1998), *Survey of software for structured text*, posljednja promjena: 1. siječnja 1998, preuzeto 22. studenog 2000, sa WWW: http://www.cs.uku.fi/~kuikka/systems.html

42. Lager, Torbjörn (1995), *A logical approach to computational corpus linguistics*, doktorska disertacija, Department of linguistics, Göteborg University

43. László, Bulcsú (1993), *Pabirci redničnoga i obavjestničkoga pojmovlja oko razumnih sustava*, u Obrada jezika i prikaz znanja, Zavod za informacijske studije, Zagreb, str. 11-73

44. Lawrer, M. John & Dry, Helen Aristar (1998), *Using Computers in Linguistics*, Routledge, New York

45. Leech, G., Garside, R., Bryant, M. (1994): *CLAWS4: The tagging of the British National Corpus*, u Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), Kyoto, Japan, str. 622-628, članak se može naći i na: http://www.comp.lancs.ac.uk/computing/research/ucrel/papers/coling.html

46. Leech, Geoffrey (1991), *The state of the art in corpus linguistics* u English corpus linguistics: studies in honour of Jan Svartvik, ur. K. Aijmer & B. Altenberg, Longman, London and New York, str. 8-29

47. Leech, Geoffrey (1993), *Corpus annotation schemes*, Literary and Linguistic Computing, Oxford University Press

48. Linguistic Data Consortium (LDC) (1999), *About Linguistic Data Consortium*, University of Pennsylvania, preuzeto 19. studenog 2000, sa WWW: http://www.ldc.upenn.edu/About/

49. LREC (2000) zbornik, Atena, 31. svibnja-2. lipnja 2000, ELRA, Pariz-Atena 2000, Vol. I, Vol. II i Vol III

50. Manning, Christopher D. & Schütze, Hinrich (1999), *Foundations of Statistical Natural Language Processing*, MIT Press Cambridge, Massachusets London

51. McEnery, Tony & Wilson, Andrew (1996), *Corpus Linguistics*, Edinburgh University Press

52. McEnery, Tony & Wilson, Andrew (1996a), dodatak knjizi *Corpus Linguistics*, Edinburgh University Press 1996, preuzeto 13. ožujka 2001, sa WWW: http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

53. Merkel, Magnus & Andersson, Mikael (2000), *Knowledge-lite extraction of multi-word units with language filters and entropy thresholds,*u Proceedings of RIAO'2000, Collége de France, Paris, France, 12-14 travnja 2000, Vol. I, str. 737-746, preuzeto 18. veljače 2001, sa WWW: http://stp.ling.uu.se/~corpora/plug/paper/merkel-andersson-RIAO-2000.pdf

54. Microsoft (2000a), *Windows Scripting Technologies*, Microsoft Corporation, Redmond, preuzeto 13. studenog 2000, sa WWW: http://msdn.microsoft.com/scripting/

55. Microsoft (2000b), *ASP Tutorial*, Microsoft Corporation, Redmond, preuzeto 13. studenog 2000, sa WWW: http://msdn.microsoft.com/workshop/c-frame.htm#/workshop/server/Default.asp

56. Microsoft XML 3.0 SDK (2000), Microsoft Corporation, Redmond, preuzeto 11. studenog 2000, sa WWW: http://msdn.microsoft.com/library/default.asp?URL=/library/psdk/xmlsdk/xmls6g53.htm

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

58. MULTEXT (1996), *Multext - Document MUL1*, Version 0.1, posljednja promjena 22. travnja 1996, preuzeto 29. studenog 2000, sa WWW: http://www.lpl.univ-aix.fr/projects/multext/

59. Multext-East (1996), *Multilingual Text Tools and Corpora for Central and Eastern European Languages*, projekt COPERNICUS 106, Naslovnica, preuzeto 28. studenog 2000, sa WWW: http://www.lpl.univ-aix.fr/projects/multext-east/

60. Nancy, Ide & Bonhomme, Patricie & Romary, Laurent (2000): *XCES: An XML-based Encoding Standard for Linguistic Corpora*, LREC (2000) zbornik, Atena, 31. svibnja-2. lipnja 2000, ELRA, Pariz-Atena 2000, Vol. II

61. Oxford Text Archive (OTA) (2001), *What Is The Oxford Text Archive*, University of Oxford, Oxford, posljednja promjena: 12. veljače 2001, preuzeto 18. veljače 2001, sa WWW: http://ota.ahds.ac.uk/

62. Palmer, David D. (1994), *SATZ - An Adaptive Sentence Segmentation System*, Computer Science Division, University of California at Berkeley, preuzeto 12. ožujka 2001, sa WWW: http://sunsite.berkeley.edu/TR/UCB:CSD-94-846

63. Scott, Mike (1999), *WordSmith Tools 3.0 (demo version) Help*

64. Souter, Clive & Atwell, Eric (1993), *Corpus-based computational linguistics*, GA, Amsterdam-Atlanta

65. Spencer, Paul (1999), *XML Design and Implementation,* Wrox Press, Birmingham

66. Sperberg-McQueen, C. M. & Burnard, L. (1990) *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, Text Encoding Initiative, Chicago-Oxford

67. Šojat, Zorislav (1976), *Čestotni rječnik Vjesnika i Večernjeg lista*, Zagreb

68. Tadić, Marko (1991), *Od korpusa do čestotnoga rječnika hrvatskoga književnog jezika*, Radovi Zavoda za slavensku filologiju 27, Zagreb, str. 161-168.

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

70. Tadić, Marko (1996), *Računalna obradba hrvatskoga i nacionalni korpus*, Suvremena lingvistika 41-42, Zagreb, str. 603-612

71. Tadić, Marko (1997), *Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive*, Suvremena lingvistika 43-44, Zagreb, str. 387-394

72. Tadić, Marko (1998), *Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika*, Filologija 30-31, Zagreb, str. 337-348

73. Tadić, Marko (1999), *Hrvatski čestotni rječnik*, pozvano predavanje, Filozofski fakultet, Sveučilište u Ljubljani, 1999-10-26, preuzeto 22. studenog 2000, sa WWW: http://www.hnk.ffzg.hr/txts/mt4frq_ljubljana/index.htm

74. Tadić, Marko (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, str. 132-137

75. Tadić, Marko (2000b), *Information Retrieval Meets Human Language Technology*, CUC2000 Zbornik, CD-ROM, Zagreb, 24-26. rujna 2000, CARNet, Zagreb

76. Tadić, Marko (2000c), *Building the Croatian-English Parallel Corpus*, LREC 2000 zbornik, Atena, 31. svibnja-2. lipnja 2000, ELRA, Pariz-Atena 2000, Vol. I, str. 523-530

77. TEI (1999), *TEI Home Page*, posljednja promjena 7. srpnja 2000, preuzeto 28. studenog 2000, sa WWW: http://www.tei-c.org/

78. Unicode Consortium (2000), *The Unicode Standard Version 3.0*, Addison-Wesley Longman, Inc, Reading, Massachussets

79. Unicode Consortium (2001), *What is Unicode?*, posljednja promjena: 02. veljače 2001, preuzeto 18. veljače 2001, sa WWW: http://www.unicode.org/unicode/standard/WhatIsUnicode.html

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

81. Webopedia (2001), *On-line rječnik i pretraživač informatičke tehnologije*, preuzimano od rujna 2000. do travanja 2001, sa WWW: http://www.webopedia.com/
82. Žmegač, Viktor (1998), *Bečka moderna*, MH, Zagreb
83. Žubrinić, Tomislava (1995), *Mogućnosti strojnoga označavanja i lematiziranja korpusa tekstova hrvatskoga jezika*, Magistarski rad, Filozofski fakultet Sveučilišta u Zagrebu