

Raspon, opseg i sastav korpusa hrvatskoga suvremenog jezika

Marko Tadić

Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu

Osnovni je cilj ovoga priloga dati nacrtak korpusa velikog više desetaka milijuna pojavnica, korpusa referentnog za suvremeni hrvatski jezik.

Definicije i razjašnjenja temeljnih termina *zbirka tekstova* (ili *arhiv*), *korpus*, *računalni korpus* preuzimam iz konačnoga izvještaja projekta EU-a EAGLES¹ koji je, za sada, najpotpunije obuhvatio i obradio problematiku sastavljanja i računalne podrške korpusima te predložio standarde za njihovo kodiranje i obradbu.

Tamo se definiraju sljedeći termini:

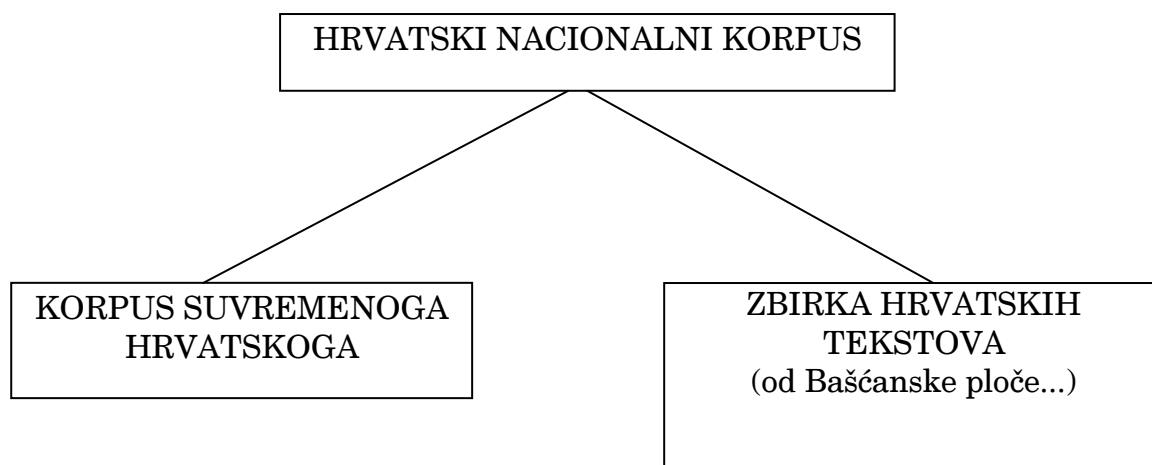
- *zbirka tekstova*: svaki skup tekstova koji je skupljen prema nekim kriterijima.
- *korpus*: zbirka jezičnih odsječaka koji su odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima upravo s ciljem da čine jezični uzorak.²
- *računalni korpus*: korpus koji je kodiran na standardan i dosljedan način s nakanom da bude otvoren za računalno pretraživanje

Usustavljen pregled kriterija (*vanjski*: tipovi tekstova, sudionici, prilike, socijalni okvir; *unutranji*: pojavljivanje posebnih jezičnih osobitosti unutar jezičnih odsječaka) za odabir jezičnih odsječaka može se pronaći u okviru projekta EAGLES: *Preliminary Recommendations on Text Typology*,³ a o njihovoj primjeni na hrvatski raspravlja se u Tadić (1996) gdje se predlaže struktura *Hrvatskoga nacionalnog korpusa*:

¹ <http://www.ilc.pi.cnr.it/EAGLES/home.html>.

² Valja svratiti pozornost na termin odsječak umjesto tekst jer u corpus ne moraju ulaziti čitavi tekstovi nego tek njihovi dijelovi koji su dovoljno veliki da čine korpusni uzorak. Nasuprot odsječcima, citati (potvrde) su premali jezični odsječci da bi činili korpusni uzorak.

³ Inačica izvješća iz lipnja 1996, str. 4.



Ovdje će se interes zadržati na lijevoj strani grafikona i pokušat će se predložiti moguću primjenu nekih od gore navedenih kriterija.

Raspon korpusa (dakako vremenski)

Vremenski raspon korpusa najlakše je definirati kao raspon između najstarijega i najmlađega teksta (jezičnoga odječka) koji je u korpusu. Dakako, korpus ne mora biti omeđen s oba kraja: on to može sezati do ili od neke točke.

Ako se danas želi sastaviti korpus hrvatskoga suvremenog jezika onda se može poći od točke u vremenu koja je u mnogome prijelomna ne samo za hrvatski jezik već i za Hrvatsku kao državu i Hrvate kao narod. Riječ je, dakako, o godini 1990. Stoga bi se korpus suvremenoga hrvatskoga jezika (KSHJ nadalje) valjalo započeti s godinom 1990. Netko bi mogao prigovoriti da je takva odluka u potpunosti nelingvistička, gotovo politička. No čini se da se jezičnih argumenata za takvu odluku može naći jer svi, dakako intuitivno, osjećamo da smo od tada hrvatski mogli rabiti »slobodnije«, »spontanije« ili, gotovo pjesnički rečeno, mogli smo ga konačno »disati punim plućima«. Usporedbom sa leksičkim sastavima starijih korpusa (npr. *M-korpus* akademika Muguša koji obuhvaća razdoblje od 1935. do 1978.) moguće je tu intuitivnu spoznaju i potvrditi

inventarski i frekvencijski. Kad se KSHJ ne bi započeo sa 1990. onda bi takve mogućnosti za usporedbu nestalo.

Da ne bi bilo zabune: hrvatski je postojao kud i kamo prije te godine, ali korpusi, osobito suvremenoga jezika, moraju se ograničiti i započeti od neke točke u vremenu. Sve što je na hrvatskome nastalo prije 1990. može se uključiti u korpus koji se ne bi zvao korpus suvremenoga hrvatskoga jezika već bi pripadao u Zbirku hrvatskih tekstova tj. na desnu stranu prethodnoga grafikona.⁴

Opseg korpusa

Opseg korpusa, dakako, ovisi prije svega o njegovoj namjeni. Namjera je projekta *Računalna obradba hrvatskoga* u Zavodu za lingvistiku FF-a sastaviti korpus koji bi mogao ući u NERC tj. mrežu europskih referentnih korpusa. Kako je sastavljanje korpusa i suviše složen i skup pothvat da bi ga se moglo prepustiti pojedinačnim i *ad hoc* kasnijim uporabama, valja ga zamisliti i organizirati kao *višesvrhovito pomagalo*⁵ tj. kao jezični resurs ili izvor jezične građe koji će služiti većem broju istraživača. Ti istraživači mogu svoj predmet istraživanja promatrati na raznim jezičnim razinama i pristupati mu s različitih teorijskih osnovica. Dobro *neutralno sastavljen korpus* im to mora omogućiti. Korpus koji mora zadovoljiti više namjena ne može biti malen korpus jer se u tome slučaju ne može pojaviti statistički relevantna jezična raznolikost na svim onim jezičnim razinama na kojima se korpusu mora moći pristupiti.

Iskustva obradbe jedno- i višemilijunskih korpusa (Brown, LOB, ICAME) koji su osim istraživačke imali nakanu poslužiti i kao leksikografska građa, govore da su

⁴ Vidi u Tadić (1996) gdje se zastupa stav o potrebi stvaranja Hrvatskoga nacionalnog korpusa koji bi obuhvaćao, s jedne strane KSHJ sastavljen prema svim uzusima i regulama korpusne lingvistike s obzirom na zahtjev za reprezentativnošću takva korpusa (više uzoraka iste veličine, razna tematska područja, razni žanrovi, pisani i govoreni jezik itd.); i s druge strane Zbirke hrvatskih tekstova (Hrvatskog elektronskog arhiva) gdje bi se smještali tekstovi izvan KSHJ ali u njihovoj cjelini tj. u punome obimu nekoga djela. Takva bi zbirka tekstova isto tako bila obradiva svim računalnim alatima kao i sam KSHJ.

⁵ Vidi u Atkins-Zampolli 1994: 8 i 11. Također o međunarodnim standardima u sastavljanju jezičnih resursa, kojih su korpusi samo jedna vrsta, vidi završno izvješće EAGLES projekta: *Preliminary Recommendations on Corpus Typology* (svibanj 1996), *Preliminary Recommendations on Text Typology* (lipanj 1996), *Recommendations on Corpus Encoding* (listopad 1996).

nekolikomilijunski korpusi premali za rječničarsku svrhu. Stoga valja razmišljati o korpusu od 10 i više milijuna pojava tj. riječi tekućeg teksta.

Prvi leksikografski projekt u potpunosti temeljen na korpusu (COBUILD, v. Sinclair 1987.) rezultirao je jednim od najboljih engleskih jednosvezačnih rječnika (Collins-COBUILD), a u njegovoj se srži nalazi korpus od 7,3 milijuna pojava koji je izabran iz većeg korpusa od 24 milijuna pojava. Danas COBUILD projekt obuhvaća desetak korpusa objedinjenih u *Bank of English* s preko 300 milijuna pojava.⁶ U Institut für Deutsche Sprache u Mannheimu sastavlja se korpus njemačkoga koji također obuhvaća preko 100 milijuna pojava, a *Trésor de la langue française* obuhvaćao je preko 200 milijuna pojava još početkom desetljeća.

Generacije korpusa:

- | | | |
|------|--------------------|--|
| I. | milijun pojava | (Brown) |
| II. | desetak milijuna | (Birmingham Collection of English Text 20 M) |
| III. | stotinjak milijuna | (Bank of English 320 M, IDS Mannheim preko 100M, TLF preko 200 M...) |

Kako u ovome trenutku i s prvim opsežnijim hrvatskim korpusom nije realno »zagristi« tako velik zalogaj, potrebno je ograničiti se na suvremeni jezik i sastaviti KSHJ ne veći od 30 milijuna pojava⁷ koji bi bio reprezentativan za suvremeni hrvatski i dovoljno velik da dà statistički relevantnu jezičnu raznolikost za npr. pisanje rječnika hrvatskoga jezika.

⁶ Vidi <http://clg1.bham.ac.uk/>.

⁷ Istraživanja na Moguševu M-korpusu (Moguš-Bratanić-Tadić 1998) pokazala su kakvi su omjeri porasta vokabulara s povećanjem veličine uzorka sa 5000 na 10000 ili 20000 pojava. Uzorak veličine 10000 pojava za M-korpus bio je najprihvatljiviji s toga gledišta. Za 30 milijunski korpus bilo bi potrebno 3000 takvih uzoraka što je s tehničke strane zahtjevno i resursno (ljudi/vrijeme) neisplativo. Valjalo bi obaviti istraživanje koje bi provjerilo što se s porastom vokabulara događa pri uvećanju uzorka sa 10000 na 50000 i 100000 pojava. Također bi za pripremu strojnih resursa valjalo provjeriti opseg takva korpusa. U Tadić (1992) pokazano je da milijun pojava hrvatskoga teksta zauzima oko 6 Mb memorije, dakle 30 milijuna pojava bi zauzelo oko 180 Mb memorije tj. opseg korpusa u Mb iznosio bi 180 što je za informatičare i te kako relevantan podatak.

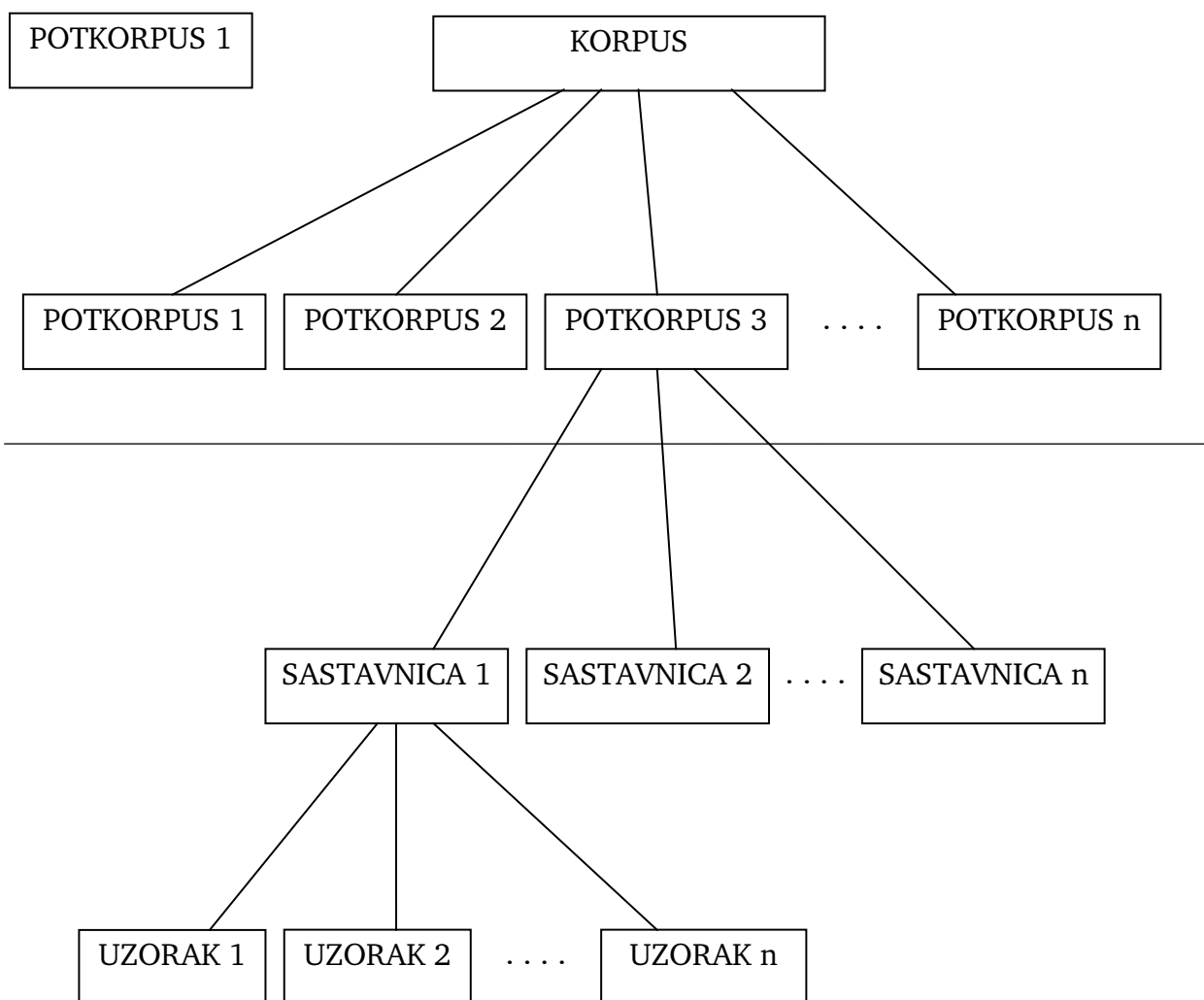
Sastav korpusa

Nije svaka zbirka tekstova korpus, a to pogotovo nije ako se želi reprezentativan korpus. Unatoč svim diskusijama o reprezentativnosti korpusa, koje ćemo ovdje preskočiti, i koliko god sastavljači željeli korpus učiniti što obuhvatnijim, pri njegovu sastavljanju valja zapravo krenuti od ograničenja. Neka će se od njih nametati sama od sebe (dostupnost elektronski pohranjenih tekstova, pristupačnost pojedinih tekstovnih žanrova itd.), a neka moraju postaviti sami sastavljači.

Jedno je sigurno, ako se želi korpus reprezentativan za neki jezik pri njegovu se sastavljanju mora paziti na:

- različita područja uporabe jezika (teme, discipline)
- tipove tekstova (knjige, novine, časopisi, brošure, prospekti, pisma itd.)
- dužina tekstova (knjige, pripovijetke, crtice, članci)
- žanrove (lijepa književnost, publicistika, znanost, udžbenici, novinski tekstovi itd.)
- medij ostvarivanja jezične poruke (pisani, govoreni jezik)
- autorove osobine (dob, spol)
- vrijeme nastanka teksta (u našem slučaju od 1990. do dana zaključenja korpusa)

Kako će se ti kriteriji primijeniti u samome korpusu tj. kako će se preslikati na opću strukturu korpusa rezultat je sastavljačeve diskrecione odluke. Idealna bi struktura korpusa morala izgledati kao na sljedećem grafikonu gdje isprekidana crta označuje granicu između dijelova koji su još uvijek korpusi i odsječaka tekstova koji više nisu korpusi.



Zastupljenost pojedinih kriterija i njihova primjena u dosadašnjim korpusima raznih jezika i raznih generacija može se vidjeti u sljedeće dvije tablice:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30					
LITERARY GENRE																																			
poetry			-																																
narrative				+																															
(auto)biography					+																														
novel/short story							+																												
historical																																			
sciencefiction																																			
humour																																			
theatre/drama			+																																
TOPIC																																			
topic			-				+					+																							
MEDIUM																																			
books																																			
letters/correspondence			+																																
newspapers			+																																
brochures/leaflets																																			
FICTION/NON-FICTION																																			
fiction			+																																
non-fiction																																			
STYLE																																			
distance																																			
popular/solemn			+																																
specialised/lay			+																																
(=technical)			+																																
OTHERS																																			
handbooks/textbooks			+																																
translations																																			

- | | | |
|---|---|--|
| 1. Bou Written + Spoken | 3. Svartik et al Written + Spoken | 4. Juiliand et al Written |
| 5. Kucera et al Written | 7. Uit den Bogaart Written + Spoken | 8. de Jong Spoken |
| 9. Lara Written + Spoken | 11. Altenberg Spoken | 12. Birmingham Collection of English Texts Written |
| 13. Birmingham Collection of English Texts Spoken | 15. Staphorsius Written | 16. Feldweg Spoken |
| 17. Morales Written | 19. Summers (selective component) Written | 20. Crowley Spoken |
| 21. Bindi et al Written | 23. Werkgroep Taalbank Written + Spoken | 24. Atkins et al Written + Spoken |
| 25. Biber Written + Spoken | 27. Malaga Spoken | 28. Bank of English Written and Spoken |
| 29. British National Corpus | 30. Survey of English | |

⁸ EAGLES projekt, *Preliminary Recommendations on Text Typology*, inačica iz lipnja 1996, str. 32-33

Tematska tipologija⁹

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Religion	+			+	+		+	+		+		+	+	+	+	+	+	+		+
Technics/-ology			+		+	+		+	+	+	+	+		+	+	+		+		+
Law		+			+			+		+		+	+	+	+	+	+	+		+
Sports		+			+			+		+		+	+	+	+	+	+	+		+
Arts		+		+				+		+		+	+	+	+	+	+	+		+
Politics							+	+				+	+	+	+		+	+		+
History					+			+		+		+		+	+	+		+		+
Medicine				+				+		+		+		+	+	+		+		+
Philosophy					+			+		+		+		+	+	+		+		+
Economy					+			+				+		+	+	+		+		+
Education					+			+		+		+		+	+			+		+
Psychology					+			+		+		+		+	+			+		+
Science		+	+	+			+						+	+	+		+			+
Sociology					+			+		+		+		+	+	+				+
Leisure				+						+		+	+	+	+	+				+
Civilisation					+			+		+		+					+	+		+
Physics					+			+		+		+		+	+					+
Biology					+			+		+		+		+	+				+	+
Mathematics								+		+		+		+	+					+
Household								+		+		+				+		+		+
Travels					+			+				+				+		+		+
Anthropology								+				+		+	+			+		+
Military								+		+		+		+	+			+		+
Media/communication								+				+		+	+			+		+
Language								+				+		+	+	+				+
Literature										+		+		+	+					+
Architecture								+				+			+			+		+
Fashion/clothes								+				+		+				+		+
Computing								+				+				+		+		+
Agriculture										+				+	+			+		+
Geography										+				+	+	+		+		+
Ecology/environment								+						+				+		+
Trafc/transport												+		+		+				+
Chemistry												+		+	+					+
Finance												+		+		+				+

1. Bou
4. Kucera et al.
7. Altenberg
10. Staphorius
13. Crowdy
16. Werkgroep Taalbank
19. Bank of English

2. Svartik et al.
5. Uit den Bogaart
8. Birmingham Collection of English Texts
11. Morales
14. Bindi et al.
17. Biber
20. British National Corpus⁹

3. Juilland et al.
6. Lara
9. Gonzalez et al.
12. Summers (selective component)
15. Martin et al.
18. Malaga

Za KSHJ svakako bi valjalo uzeti u obzir sljedeća ograničenja:¹⁰

- pisani jezik (a ako to tehničke mogućnosti dopuste bar 10% govorena jezika)
- tekstovi izvornih hrvatskih govornika, ne prijevodi
- dugi i kratki tekstovi
- opća uporaba jezika, a prema potrebi specijalistička, tehnička
- razdoblje od 1990.
- tekst stvoren u stvarnoj komunikaciji, ne drama

⁹ EAGLES projekt, *Preliminary Recommendations on Text Typology*, inačica iz lipnja 1996, str. 34

¹⁰ V. slična ograničenja u Sinclair 1987:2.

- proza, ne poezija
- jezik odraslih tj. starijih od 16 godina
- standardni hrvatski, ne narječja

Također bi potpunost uzoraka trebalo planirati u skladu s rezultatima dvaju jednostavnih preliminarnih istraživanja:

- ispitati zastupljenost muških, ženskih i skupnih autora u novinama, časopisima i katalozima knjiga te prema tim rezultatima rasporediti autorsku zastupljenost u korpusu,
- ispitati čitanost pojedinih knjiga (popisi uspješnica postoje u knjižarama ili se poslužiti popisom najposuđivanijih knjiga u Gradskoj knjižnici) te tako vrednovati njihovu kandidaturu za ulazak u korpus jer je za očekivati da tekstovi koji su više »u prometu« imaju znatniji utjecaj na jezik.

Za postavljanje strukture korpusa najlakši bi kriterij na prvoj razini mogao biti medij ostvarivanja, potom tip teksta, a ostale bi razine slijedili ostali kriteriji. Tako bi KSHJ obuhvaćao:¹¹

- pisani tekst
 - Knjige
 - Proza (romani (povijesni, krimići...), pripovijetke, crtice, dnevni, eseji)
 - Publicistika (knjige, članci, kronike)
 - Znanost (knjige, rasprave, članci raznih struka)
 - Udžbenici (srednjoškolski i sveučilišni udžbenici raznih struka)
 - Priručnici (tehnički, kulinarski, odgoj djece, domaćinski...)
 - Zakoni (Narodne novine, zakonski tekstovi, pravnički časopisi)
 - Novine
 - dnevni
 - nadregionalni
 - regionalni

¹¹ V. Sinclair 1987:23 za izvrstan i u mnogo čemu inspirativan pregled sastava COBUILD korpusa.

- tjednici
- dvotjednici
- Časopisi
 - tjednici
 - dvotjednici
 - mjesečnici
 - višemjesečnici
- Brošure, prospekti
- Korespondencija
 - privatna
 - službena
- govoreni tekst
 - formalni/pripremljeni (predavanja, izlaganja, nastupi)
 - neformalni/ad hoc (dijalozi npr. RTV i druge javne diskusije itd.)

Dakako da predložena struktura daje samo grub nacrt koji će u mnogome, kad se u samo sastavljanje korpusa krene, trebati razraditi i tako strogo odrediti te, prema zacrtanim kriterijima, u KSHJ uvrstiti svaki pojedini tekst.

Mnoštvo je materijala već dostupno u elektronskome obliku bilo u samim izdavačkim kućama (računalna priprema za tisak i/li elektronsko izdavaštvo), bilo preko CARNET-a.¹² Nadalje, tu su i HINA-in servis kao i vijesti HRT-a koji su također dostupni preko Interneta.

Literatura

Andrijašević, Marin; Vrhovac, Yvonne (ur.) (1990) *Informatička tehnologija u*

primijenjenoj lingvistici, Hrvatsko društvo za primijenjenu lingvistiku, Zagreb.

Atkins, B. T. S. –Zampolli, A. (1994) *Computational Approaches to the Lexicon*, Oxford University Press.

EAGLES projekt (1996a) *Preliminary Recommendations on Corpus Typology*, svibanj 1996, <http://www.ilc.pi.cnr.it/EAGLES/home.html>.

¹² Npr. Narodne novine na <http://www.nn.hr> gdje je moguće pristupiti tekstovima svih trenutačno važećih zakona u Republici Hrvatskoj.

- EAGLES projekt (1996b) *Preliminary Recommendations on Text Typology*, lipanj 1996,
<http://www.ilc.pi.cnr.it/EAGLES/home.html>.
- EAGLES projekt (1996c) *Recommendations on Corpus Encoding*, listopad 1996,
<http://www.ilc.pi.cnr.it/EAGLES/home.html>.
- Engwall, Gunnel »Not Chance but Choice: Criteria in Corpus Creation« u: Atkins–
Zampolli (1994), str. 49-82.
- Ide, Nancy (1995) *Encoding Standards for Linguistic Corpora* u: Rettig-Pajzs-Kiss
(1995), str. 65-78.
- Johansson, Stig (1994) »Encoding a Corpus in Machine-Readable Form: The Approach
of the Text Encoding Initiative« u: Atkins–Zampolli (1994), str. 83-102.
- Moguš, Milan; Bratanić Maja; Tadić, Marko (1998) *Hrvatski čestotni rječnik*, Školska
knjiga-Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. (u tisku)
- Rettig, Heike; Pajzs, Júlia; Kiss, Gábor (ur.) (1995) *TELRI, Proceedings of the First
European Seminar »Language Resources for Language Technology« in Tihany*.
- Sinclair, John (ur.) (1987) *Looking Up. An account of the COBUILD Project in lexical
computing*, Collins, London-Glasgow.
- Tadić, Marko (1990) *Zašto nam je potreban višemilijunski referentni korpus?* u:
Andrijašević & Vrhovac (1990), str. 95-98.
- Tadić, Marko (1992) *Od korpusa do čestotnoga rječnika hrvatskoga književnog jezika*,
Radovi Zavoda za slavensku filologiju, 27, str. 169-178.
- Tadić, Marko (1996) *Računalna obradba hrvatskoga i nacionalni korpus*, Suvremena
lingvistika, 41-42, str. 603-611.