FILOLOGIJA 30–31, Zagreb 1998

UDK 808.62-323.1
Izvorni znanstveni članak
Primljen 6. V. 1998.
Prihvaćen za tisak 15. VI. 1998.

Marko Tadić
Zavod za lingvistiku Filozofskoga fakulteta
Ivana Lučića 3, HR-10000 Zagreb

RASPON, OPSEG I SASTAV KORPUSA HRVATSKOGA SUVREMENOG JEZIKA

Osnovni je cilj ovoga priloga dati nacrtak korpusa velikog više desetaka milijuna pojavnica, korpusa referentnog za suvremeni hrvatski jezik. Razmatraju se njegov vremenski raspon, opseg i sastav.

U ovome radu definicije i razjašnjenja temeljnih termina *zbirka tekstova* (ili *arhiv*), *korpus*, *računalni korpus* preuzimaju se iz konačnoga izvještaja projekta Europske unije EAGLES[1] 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:

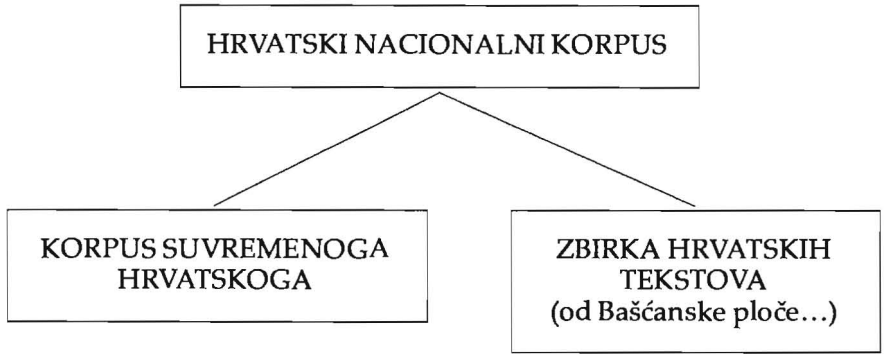
- *zbirka tekstova*: svaki skup tekstova koji je skupljen prema nekim kriterijima.
- *korpus*: zbirka jezičnih odsječaka koji su odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima upravo s ciljem da čine jezični uzorak.[2]
- *računalni korpus*: korpus koji je kodiran na standardan i dosljedan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

Ususlavljen pregled kriterija (*vanjski*: tipovi tekstova, sudionici, prilike, socijalni okvir; *unutarnji*: pojavljivanje posebnih jezičnih osobitosti unutar jezičnih odsječaka) za odabir jezičnih odsječaka može se pronaći u okviru pro-

[1] *http://www.ilc.pi.cnr.it/EAGLES/home.html.*

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

jekta 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*:



Ovdje će se interes zadržati na lijevoj strani grafikona i pokušat će se predložiti moguća primjena nekih od gorenavedenih kriterija.

Raspon korpusa (dakako vremenski)

Vremenski raspon korpusa najlakše je definirati kao raspon između najstarijega i najmlađega teksta (jezičnoga odsječka) koji je u korpusu. Dakako, korpus ne mora biti vremenski omeđen s oba kraja: on se može ograničiti samo jednom točkom u vremenu. U tom se slučaju uzimaju tekstovi do završne vremenske granice ili od te vremenske granice nadalje.

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 s leksičkim sastavima starijih korpusa (npr. *M-korpus* akademika Milana Moguša, koji obuhvaća razdoblje od 1935. do 1978.) moguće je tu intuitivnu spoznaju i potvrditi inventarski i frekvencijski. Kad se KSHJ ne bi započeo s 1990, onda bi takve mogućnosti za usporedbu nestalo.

Da ne bi bilo zabune: hrvatski je postojao kud i kamo prije te godine, ali

[3] Inačica izvješća iz lipnja 1996, str. 4.

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 jezika* u Zavodu za lingvistiku Filozofskoga fakulteta sastaviti korpus koji bi mogao ući u NERC, tj. u mrežu europskih referentnih korpusa. Kako je sastavljanje korpusa isuviše 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* to im 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 jednomilijunskih i višemilijunskih korpusa (Brown, LOB, ICAME) koji su osim istraživačke imali nakanu poslužiti i kao leksikografska građa, govore da su nekolikomilijunski korpusi premali za rječničarsku svrhu. Stoga valja razmišljati o korpusu od deset i više milijuna pojavnica, tj. riječi tekućega 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 pojavnica koji je izabran iz većega korpusa od 24 milijuna pojavnica. Danas projekt COBUILD obuhvaća desetak korpusa objedinjenih u *Bank of*

[4] Vidi u Tadić 1996, gdje se zastupa stav o potrebi stvaranja Hrvatskoga nacionalnoga 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 i sto tako bila obradiva svim računalnim alatima kao i sam KSHJ.

[5] Vidi u Atkins–Zampoli 1994:8,11. Također o medunarodnim standardima u sastavljanju jezičnih resursa, kojih su korpusi samo jedna vrsta, vidi završno izvješće projekta EAGLES 1996a,b,c.

English s preko 300 milijuna pojavnica.[6] U Institut für deutsche Sprache u Mannheimu sastavlja se korpus njemačkoga koji također obuhvaća preko 100 milijuna pojavnica, a *Trésor de la langue française* obuhvaćao je preko 200 milijuna pojavnica još početkom desetljeća.

Generacije korpusa:

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

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 pojavnica[7] koji bi bio reprezentativan za suvremeni hrvatski i dovoljno velik da dâ statistički relevantnu jezičnu raznolikost, npr. za pisanje rječnika hrvatskoga jezika.

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.

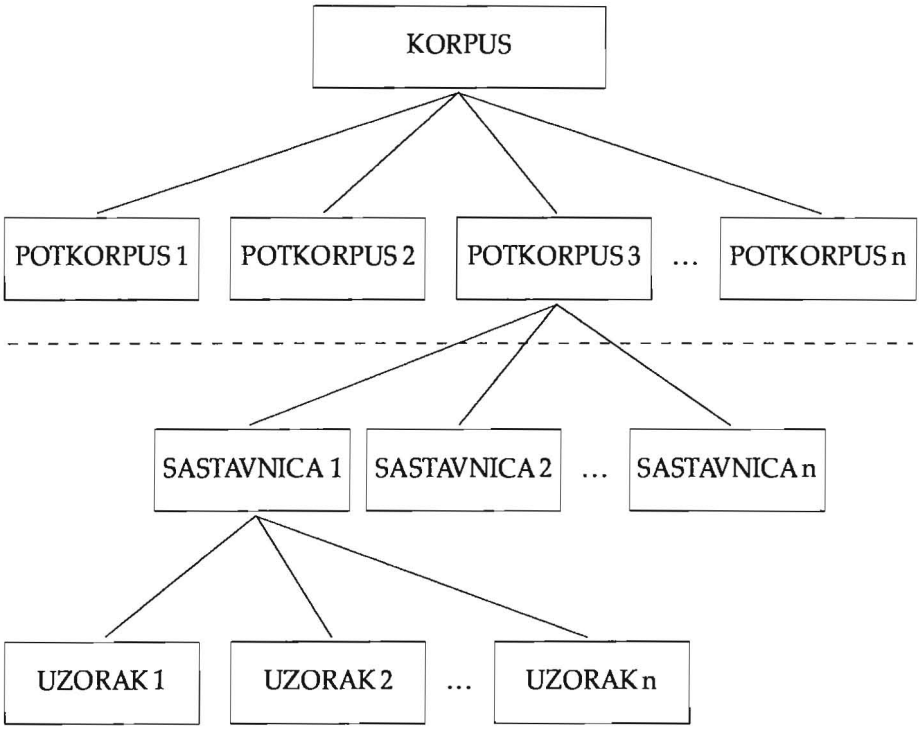
Jedno je sigurno, ako se želi korpus reprezentativan za neki jezik, pri njegovu se sastavljanju mora paziti na ovo:

[6] Vidi *http://clg1.bham.ac.uk/*.

[7] Istraživanja na Moguševu M-korpusu (Moguš–Bratanić–Tadić) pokazala su kakvi su omjeri porasta vokabulara s povećanjem veličine uzorka s 5000 na 10000 ili 20000 pojavnica. Uzorak veličine 10000 pojavnica 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 i vrijeme) neisplativo. Valjalo bi obaviti istraživanje koje bi provjerilo što se s porastom vokabulara događa pri uvećanju uzorka s 10000 na 50000 i 100000 pojavnica. Također bi za pripremu strojnih resursa valjalo provjeriti opseg takva korpusa. U Tadić 1992 pokazano je da milijun pojavnica hrvatskoga teksta zauzimlje oko 6 MB memorije, dakle 30 milijuna pojavnica zauzelo bi oko 180 MB memorije, tj. ospeg korpusa u megabajtima iznosio bi oko 180, što je za informatičare i te kako relevantan podatak.

- različita područja uporabe jezika (teme, discipline)
- tipove tekstova (knjige, novine, časopisi, brošure, prospekti, pisma itd.)
- dužinu tekstova (knjige, pripovijetke, crtice, članci)
- žanrove (lijepa književnost, publicistika, znanost, udžbenici, novinski tekstovi itd.)
- medij ostvarivanja jezične poruke (pisani, govorni 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:

Tekstovna tipologija[8]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LITERARY GENRE															
poetry			−							+		−			
narrative				+								+	+	+	
(auto)biography					+							+		+	
novel/short story							+					+		+	
historical												+		+	
sciencefiction												+			
humour					+							+			
theatre/drama			+	+	−							−		+	
TOPIC															
topic			−				+					+	+		+
MEDIUM															
books										+		+			+
letters/correspondence			+									+			
newspapapers	+		+	+			+		+	+		+		+	+
brochures/leaets					+							+			+
FICTION/NON-F.															
fiction			+	+						+		+			+
non-fiction					+							+			+
STYLE															
distance			+					+	+			+	+		
popular/solemn			+	+	+				+						
specialised/lay (=technical)			+	+	+				+		+			+	
OTHERS															
handbooks/textbooks	+		+									+		+	+
translations		+													

1. Bou Written + Spoken
2. Hoz usnpec.
3. Svartik et al Written + Spoken
4. Juilland et al Written
5. Kucera et al Written
6. de Vriendt–de Man Spoken
7. Uit den Bogaart Written + Spoken
8. de Jong Spoken
9. Lara Written + Spoken
10. Allen Written
11. Altenberg Spoken
12. Birmingham Collection of English Texts Written
13. Birmingham Collection of English Texts Spoken
14. Gonzalez et al Written
15. Staphorsius Written
16. Feldweg Spoken

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
	+		+		+	+		+	+	+		+	+	+	newspapapers
			+			+		+		+		+			brochures/leaets
															FICTION/NON-F.
			+			+	+	+	+	+		+	+	+	fiction
										+		+	+	+	non-fiction
															STYLE
				+			+		+		+				distance
			+			+	+	+							popular/solemn
	+				+	+			+						specialised/lay (=technical)
															OTHERS
					+					+		+			handbooks/textbooks
			−					+		−					translations

17. Morales Written
18. Collins et al Written + Spoken
19. Summers (selective component) Written
20. Crowdy Spoken
21. Bindi et al Written
22. Martin et al Written + Spoken
23. Werkgroep Taalbank Written + Spoken
24. Atkins et al Written + Spoken
25. Biber Written + Spoken
26. Malaga Written
27. Malaga Spoken
28. Bank of English Written and Spoken
29. British National Corpus
30. Survey of English

Tematska tipologija[9]

	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		+	+		+			+		+		+		+	+			+		
Sciences				+			+						+	+			+			+
Sociology					+			+		+		+		+	+	+				+
Leisure				+						+		+	+	+	+	+				+
Civilisation					+			+		+		+					+	+		
Physics					+			+		+		+		+	+					
Biology					+			+				+		+	+			+		
Mathematics								+		+		+		+	+					
Household								+		+		+				+		+		
Travels					+			+				+				+		+		
Anthropology								+				+		+	+			+		
Military								+		+		+		+				+		
Media/communications								+				+		+	+			+		
Language								+				+		+	+	+				
Literature										+		+		+	+					
Architecture								+				+			+			+		
Fashion/clothes								+				+		+				+		
Computing								+				+				+		+		
Agriculture										+				+	+			+		
Geography										+					+	+		+		
Ecology/environment								+						+				+		
Trafic/transort												+		+		+				
Chemistry												+		+	+					
Finance												+		+		+				

1. Bou
2. Svartik et al.
3. Juilland et al.
4. Kucera et al.
5. Uit den Bogaart
6. Lara
7. Altenberg
8. Birmingham Coll. of Eng. Texts
9. Gonzalez et al.
10. Staphorius
11. Morales
12. Summers (selective component)
13. Crowdy
14. Bindi et al.
15. Martin et al.
16. Werkgroep Taalbank
17. Biber
18. Malaga
19. Bank of English
20. British National Corpus

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

Također bi popunjenost 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 pojednih knjiga (popisi uspješnica postoje u knjižarama i novinama ili se poslužiti popisom najposuđivanjih knjiga u gradskim knjižnicama) te tako vrednovati njihovu kandidaturu za ulazak u korpus jer je 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, dnevnici, 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
    dnevnici
    nadregionalni

[8] EAGLES, inačica iz lipnja 1996, 32–33.
[9] EAGLES, inačica iz lipnja 1996, 34.
[10] V. slična ograničenja u Sinclair 1987:2.
[11] Vidi Sinclair 1987:23 za izvrstan i u mnogo čemu inspirativan pregled sastava korpusa COBUILD.

regionalni
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/ili 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, Yvonne Vrhovac (ur.). 1990. *Informatička tehnologija u primijenjenoj lingvistici*. Zagreb : Hrvatsko društvo za primijenjenu lingvistiku.

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

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

EAGLES, projekt. 1996b. *Preliminary Recommendations on Corpus Typology*, lipanj 1996, *http://www.ilc.pi.cnr.it/EAGLES/home.html*.

EAGLES, projekt. 1996c. *Preliminary Recommendations on Corpus Typology*, listopad 1996, *http://www.ilc.pi.cnr.it/EAGLES/home.html*.

Engwall, Gunnel. Not Chance but Choice: Criteria in Corpus Creation. U knj. Atkins–Zampolli 1994, 49–82.

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

Ide, Nancy. 1995. Encoding Standards for Linguistic Corpora. U knj. Rettig––Pajzs–Kiss 1995, 65–78.

Johansson, Stig. 1994. Encoding a Corpus in Machine-Readable Form: The Approach of the Text Encoding Initiative. U knj. Atkins–Zampoli 1994, 83–102.

Moguš, Milan, Maja Bratanić, Marko Tadić. *U tisku. Hrvatski čestotni rječnik.* Zagreb : Školska knjiga i Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Rettig, Heike, Júlia Pajzs, Gábor Kiss (ur.). 1995. *TELRI, Proceedings of the First European Seminar »Language Resources for Language Techonology«* in Tihany.

Sinclair, John (ur.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing.* London–Glasgow : Collins.

Tadić, Marko. 1990. Zašto nam je potreban višemilijunski referentni korpus? U knj. Andrijašević–Vrhovac 1990, 95–98.

Tadić, Marko. 1992. Od korpusa do čestotnoga rječnika hrvatskoga književnog jezika. *Radovi Zavoda za slavensku filologiju* 27, 169–178.

Tadić, Marko. 1996. Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika* 41–42, 603–611.

Time-span, size and composition of the corpus of the Croatian contemporary language

Summary

The aim of this contribution is to give a general sketch of the reference corpus for contemporary Croatian. Its time-span, size (several dozens of millions running-words) and structure of subcorpora as well as constituents and samples typology are being discussed.

Ključne riječi: korpus, računalni pristup, rječnik, hrvatski jezik
Key words: corpus, computational approach, dictionary, Croatian