

Marko Tadić

Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu

Računalna obradba hrvatskoga i nacionalni korpus

Članak pokušava pokazati opravdanost dviju polaznih teza kojima se predlaže određenje statusa računalne obradbe hrvatskoga. Teze su da računalna obradba hrvatskoga mora imati status: 1) fundamentalnoga istraživanja u humanističko-društvenim znanostima; 2) strateškoga istraživanja za Republiku Hrvatsku.

Danas istraživati prirodni jezik bez pomoći računala nije samo mukotrпно i dugotrajno nego, uslijed ljudske nemogućnosti da se u obradi zamašne jezične građe održe kriteriji i(li) koncentracija, često i paraznanstveno. Računalna nam tehnologija u tome može u mnogome pomoći, no u ovome je trenutku primarni problem, barem kada je riječ o hrvatskome, kako se računala mogu (moraju) rabiti u istraživanjima jezika organiziranije i obuhvatnije nego do sada.

Na početku valja postaviti dvije teze koje će se u ovome radu pokušati obrazložiti. Naime, istraživanja s područja računalne obradbe hrvatskoga jezika (pri tome se misli na ono što se engleski zove *Natural Language Processing* u najširem smislu¹) moraju imati status:

1. fundamentalnoga istraživanja u humanističko-društvenim znanostima
2. strateškoga istraživanja za Republiku Hrvatsku.

1. Računalna obradba hrvatskoga kao fundamentalno istraživanje

Danas nikomu nije neobično da istraživanja subatomske čestice imaju status fundamentalnoga istraživanja. Tà spoznaje o strukturi atoma čine temelje za daljnja istraživanja u fizici, astronomiji, (bio)kemiji, medicini, tehnologiji itd. Također nikomu nije čudno što se do spoznaja o subatomske česticama može

1 Npr. u Boguraev & Briscoe 1989: 3.

doći isključivo uporabom visoke i skupe tehnologije. Ono što je možda začudno jest odsutnost takva istraživanja i takva pristupa u humanističko–društvenim znanostima. Tu ulogu, barem za one znanosti čiji se predmet istraživanja ostvaruje tekstem, može preuzeti upravo računalna obradba prirodnoga jezika.

Temelj za svako istraživanje teksta jest korpus bez obzira na to promatra li se kao jezična građa ili kao nešto drugo što se putem teksta/jezika tek ostvaruje. Za razliku od ostalih jezikoslovnih disciplina, korpusna lingvistika određena je ne toliko područjem istraživanja koliko metodološkom osnovicom na kojoj se temelji istraživanje. Stoga se korpusni pristup (ili korpusna metodologija) lako može primijeniti u različitim lingvističkim disciplinama: fonologiji, morfologiji, sintaksi, sociolingvistici, kognitivnoj lingvistici itd., i to najčešće u kombinaciji s drugim, tim disciplinama inherentnim, metodološkim postupcima. Današnji uvid u korpus ne može se ni zamisliti bez pomoći računala i svih mogućnosti koje ona pružaju pri pregledu i uređivanju građe. Stoga valja razlikovati *predračunalnu korpusnu lingvistiku*, koja svoje teorijske osnovice nalazi u američkih deskriptivista, od *računalne korpusne lingvistike*.² Tako postavljeno istraživanje temeljne jezične građe omogućuje potpuno nov uvid:

... I wish to argue that computer corpus linguistics (henceforth CCL) defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject. The computer, as a uniquely powerful technological tool, has made this new kind of linguistics possible. So technology here (as for centuries in natural science [sic!]) has taken a more important role than that of supporting and facilitating research: I see it as the essential means to a new kind of knowledge, and as »open sesame« to a new way of thinking about language.³

Računala se osim za usustavljanje i pretraživanje građe mogu koristiti i kao sredstvo za provjeru istraživačkih hipoteza. Stoga uporaba računala u lingvistici mora steći onakav status kakav ima uporaba računala u prirodnim znanostima, a to je uloga nezaobilaznoga alata za prikupljanje i obradu *istraživačkih podataka* kao i provjeru istraživačkih hipoteza računalnim *modeliranjem predmeta istraživanja* (njegove strukture i odnosa u kojima se njegove sastavnice nalaze). U slučaju lingvistike istraživački su podaci *jezična građa* (u obliku korpusa), a računalni modeli odgovaraju (strukturnim) modelima pojedinih (dijelova) *jezičnih podsustava* (u oblicima različitih modula za obradbu prirodnoga jezika npr. *taggeri*, *parseri*, *generatori*, itd.). No i tako konstruirani

2 Cf. Leech 1992: 106.

3 Leech 1992: 106. U mom prijevodu: Htio bih ustvrditi da računalna korpusna lingvistika (nadalje CCL) određuje ne samo novonastajuću metodologiju jezičnoga proučavanja nego i novi istraživački pogled i zapravo nov filozofski pristup predmetu. Računalo je, kao iznimno snažno tehnološko sredstvo, omogućilo tu novu vrstu lingvistike. Tako je tehnologija (kao što je to već stoljećima u prirodnim znanostima) dobila važniju ulogu od puke podrške i olakšavanja istraživanja. Vidim je kao esencijalno sredstvo za novu vrstu znanja i kao »Sezame, otvori se« novom načinu razmišljanja o jeziku. (isticanje moje).

jezični modeli mogu se dalje provjeravati u srazu s podacima skupljenim u korpusu.

Sve su filološke znanosti po definiciji usmjerene na istraživanje tekstova, ali od računalne obradbe i prezentacije jezične/tekstovne građe mogu i te kako koristiti imati znanost o književnosti⁴ (kad konzultira činjenice teksta) i povijest (kad zahtijeva uvid u dokumente što se inače smješta u pomoćne povijesne znanosti npr. arhivistiku), a dijelom i arheologija (kad zagléda u na(t)pise). Dakako, tu svoga interesa mogu pronaći i psiholozi i sociolozi.⁵

Ako je računalo potrebno pri skupljanju i obradbi jezične građe, a to je valjda danas već neupitno, onda se računalna obradba hrvatskoga u potpunosti uklapa u nacionalni istraživački prioritet istraživanja hrvatskoga jezika.⁶

2. Računalna obradba hrvatskoga kao strateško istraživanje za Republiku Hrvatsku

Sa sve većom uporabom računala u gotovo svim područjima ljudskoga života (ovdje posebno zanimanje mora izazvati jedna od najraširenijih uporaba računala, tj. za pisanje, »proizvodnju« teksta) i pojavom novih telekomunikacijskih sredstava, koja su gotovo u potpunosti upravljana računalima (npr. Internet i sl.), pojavljuje se potreba za *jezičnim alatima* (u okviru paradigme *jezičnih industrija*) kao pomagalima kako jezikoslovcima u istraživanju jezika i informatičarima u dizajniranju i programiranju sustava za obrade teksta, tako i svakodnevnim korisnicima računala pri manipulaciji tekstem.

The term 'language industries' was launched at the 1986 Tours Conference organized by Council of Europe (...) and covers both activities where computer assistance is being developed for the traditional applied linguistics professions (such as, for instance, lexicography, translation, and language teaching) and activities directed towards the development of new applications, made possible by new computational systems with NLP components. The latter include systems for natural language interfaces, speech analysis and synthesis, automatic indexing and abstracting, office

4 Sâm termin *literary and linguistic computing* koji pokriva područje jednostavnije strojne obrade tekstova (indeksi, konkordancije, čestotni i abecedni rječnici) za potrebe lingvista i stilista pojavljuje se prvi put 1966. cf. Atkins–Levin–Zampolli 1994: 22.

5 Valja naglasiti da se ovdje pokušava izbjeći ona metodološka zamka u koju su mnogi lingvisti djelomice upali (npr. Hjelmsov) tvrdnjom kako je upravo lingvistika ta, ili neki semiotičari tvrdnjom kako je upravo semiotika ta, koja se može smatrati epistemologijom društvenih znanosti. Računalna obradba teksta ne bi smjela pretendirati ni na kakvu epistemološku dimenziju. Dapače, njezina se svrha mora održati na razini istraživačkoga alata, instrumenta ili pomagala i ona je krajnje jednostavna: omogućiti znanstvenicima (i ne samo znanstvenicima) usustavljen i brz pristup velikim količinama teksta koji pristup omogućuje sagledavanje predmeta istraživanja iz novih i do sada vremenski i trudom neisplativih kutova.

6 Ne valja brkati status fundamentalnoga istraživanja s nacionalnim istraživačkim prioritetima kao što su nacionalna povijest ili hrvatski jezik jer ti su prioriteti formirani na neepistemološkim temeljima. Vidi o nacionalnim istraživačkim prioritetima na WWW adresi Ministarstva znanosti i tehnologije Republike Hrvatske: <http://hvar.mzt.hr/>.

automation, machine translation, and, more generally, communications support.⁷

Sasvim je prirodno da se tako definirani jezični alati, unatoč mnogim metodološkim i tehničkim podudarnostima pri istraživanjima, moraju izraditi za svaki jezik odvojeno. Dakle, svaka jezična zajednica mora, ako želi održati korak s vremenom, oblikovati i pokrenuti projekte rezultat kojih moraju biti upotrebljivi jezični alati.⁸

Za razliku od »velikih« (tj. mnogoljudnih) jezika, situacija hrvatskoga nije jednostavna. Naime, dok se kod mnogoljudnih jezika za takve projekte relativno lako mogu naći privredni i neprivredni subjekti zbog zainteresiranosti velikoga tržišta, te su ti projekti većim dijelom financirani iz takvih izvora; za hrvatski se, kao jezik od svega pet milijuna govornika, u svjetlu imperativnosti tih istraživanja, nameće jedino rješenje u obliku pomoći jezične zajednice, tj. države koja bi morala pokrivati većinu troškova tih projekata.⁹ No, unatoč postojanju privrednih izvora financiranja ni u mnogoljudnih jezika ne izostaje podrška zajednice. Tako se u Cencioni & Klein (1994) navodi 28 projekata koje je financirala Komisija Europske zajednice od 1991. do 1994. sa zajedničkim proračunom od 69 milijuna Ecu (15 projekata s pojedinačnim proračunom većim od milijun Ecu).¹⁰

Tri su razloga zbog kojih Europska komisija potpomaže takve projekte i oni su za drugu tvrdnju postavljenu u uvodu još znakovitiji:

1. ... As perhaps the world's largest user of language technology, the Commission itself has a vested interest in encouraging this technology. Its multilingual documentation burden will only get heavier as the EU welcomes new members from Scandinavia, Central Europe, and Eastern Europe. (...) Success in the language engineering endeavour will be a strategic asset.
2. ... Commission's support for language processing is that language technology can play a vital role in maintaining the plurality of European culture, where no one language should be allowed to gain ascendancy.

7 Atkins & Zampolli 1994: 6. U mom prijevodu: Termin 'jezične industrije' lansiran je 1986. na skupu u Toursu koji je organiziralo Europsko vijeće (...) i pokriva područja u kojima se s jedne strane razvija računalna pomoć za tradicionalne primijenjenolingvističke profesije (kao što su npr. leksikografija, prevodenje, učenje jezika) i s druge strane područja usmjerena prema razvijanju novih aplikacija koje su omogućene novim računalnim sustavima sa sastavnicama koje obrađuju prirodnojezične podatke. Ova posljednja područja obuhvaćaju sustave za prirodnojezična sučelja, analizu i sintezu govora, strojno indeksiranje i pravljenje sažetaka, uredsku automatizaciju, strojno prevodenje i, općenito, podršku u telekomunikacijama.

8 Tome u tehničkim znanostima odgovara npr. istraživanje tvrdoće materijala kako bi se na temelju tih rezultata mogli izraditi alatni strojevi.

9 Kao što je to slučaj u npr. Finskoj koja ima otprilike istu populaciju kao i Hrvatska, strukturno složeniji jezik, ali znatno veća ulaganja upravo u njegovu računalnu obradu.

10 Cencioni & Klein 1994: 21–103.

3. ... Community supports language processing is because language has so much potential across the entire domain of telecommunications and information technologies, the so-called world of *telematics*. From education to consumer electronics, from office automation to special needs services, language therefore language processing plays a ubiquitous role.¹¹

Ne valja na ovome mjestu pomisliti kako su »europske integracije« jedini ili najvažniji razlog za potporu takvim projektima u Republici Hrvatskoj. Dapače, o svome jeziku, gledajući od najstarijih spomenika do danas, imamo premalo sustavno skupljene građe da bismo ga uopće mogli ozbiljno jezikoslovno sagledavati u cjelini i(li) sa svih onih strana koje nam istraživanja mogu nalagati. Nema je npr. u korpusnome obliku, dovoljno ni za kvalitetan jednosvežačni hrvatski rječnik, a kamoli za opsežnija djela i istraživanja na ostalim jezičnim razinama koja bi našoj djeci i unucima omogućila normalnu (dakle, »informatiziranu«) uporabu hrvatskoga u 21. stoljeću.

Imajući u vidu izneseno, sasvim je bjelodano da računalna obrada hrvatskoga *mora* imati status strateškoga istraživanja za Republiku Hrvatsku.

3. Nacionalni korpus kao ishodišno istraživanje s područja računalne obradbe hrvatskoga

Već je u prvome dijelu rečeno da je temelj za svako istraživanje teksta korpus. Sastavljanje korpusa hrvatskoga jezika mora biti primaran zadatak u ovoj fazi istraživanja hrvatskoga jezika uopće.¹² Jedan od razloga tome možemo naći pogleda li se smjer razvitka jezičnih teorija koje su, bile one generativne provenijencije ili ne, kad su izašle izvan okvira eksperimentalne hipoteze, dakle, kad ih se pokušalo provjeriti na stvarnom jezičnom materijalu, naišle na nedo-

11 Cencioni & Klein 1994: 8. U mom prijevodu: 1.... Kao vjerojatno najveći svjetski korisnik jezične tehnologije, Komisija sama ima interes za poticanje te tehnologije. Teret njezine višejezične dokumentacije još će se povećati kako EU bude primala nove zemlje-članice iz Skandinavije, središnje i istočne Europe. (...) Uspjeh u pokušajima jezičnoga inženjerstva bit će strateški dobitak. 2.... podrška Komisije [strojnoj] obradbi jezika u tome je što jezična tehnologija može igrati vitalnu ulogu u održanju pluralnosti europske kulture, gdje se ne smije dopustiti propadanje niti jednoga jezika. 3.... Zajednica potpomaže [strojnu] obradu jezika zbog toga što jezik posjeduje velik potencijal na čitavom području telekomunikacija i informacijskih tehnologija, na području tzv. *telematike*. Od prosvjete do kućanske elektronike, od uredske automatizacije do posebnih servisa, jezik, dakle, [strojna] obrada jezika ima sveobuhvatnu ulogu.

12 Od 28 projekata navedenih u Cencioni & Klein (1994) 10 ih se izravno bavi ili je izravno temeljeno na korpusima: ANTHEM (Advanced Natural Language Interface for Multilingual Text Generation in Health Care), COBALT (Construction, Augmentation and Use of Knowledge Bases from Natural Language Documents), DELIS (Descriptive Lexical Specifications and Tools for Corpus-based Lexicon Building), GIST (Generating Instructional Text), MULTEXT (Multilingual Text Tools and Corpora), RELATOR (European Network of Repositories for Linguistic Resources), RENOS (Reduction of Noise and Silence in Full Text Retrieval Systems for Legal Texts), SIFT (Selecting Information From Text), SISTA (Semi-automatic Indexing System for Technical Abstracts), TRANSLEARN (Interactive Corpus-based Translation Drafting Tool).

statak leksikonskih podataka. To je bilo osobito bolno za sustave koji su bili oblikovani kao računalni modeli:

Thus NLP systems face what has been called the 'lexical bottleneck' (...) limitations in system performance attributable to the need for larger lexicons. It has become of paramount importance to find sources of data (...) that allow the building of effective large lexicons (...). These needs sparked a renewed interest in corpus linguistics and in lexicography, as both those fields offered potential help in overcoming the lexical bottleneck.¹³

Ako se želi izbjeći ta situacija, korpus hrvatskoga mora biti sastavljen tako da se na njegovoj osnovi mogu izraditi veliki leksikoni, i to kako za ljude kao korisnike tako i za strojeve, tj. one sustave za računalnu obradbu hrvatskoga koji bi se takvim leksikonima morali služiti za obrade na ostalim jezičnim razinama.

Drugi je razlog za postavljanje korpusa u ishodište računalne obrade hrvatskoga zahtjevnost njegova sastavljanja u smislu ljudi i sredstava te iz toga proistječe potreba za njegovom *višestrukom iskoristivošću*, ali i *višesvrhovitošću*. Naime, pokazalo se da je sastavljanje višemilijunskih korpusa vrlo skup pot-hvat. Zato se moraju izbjegavati one situacije u kojima su se naponi znali udvajati i resursi rasipati kad bi svaki projekt konstruirao priručni korpus koji je, kad ga je trebalo nadopuniti, morao biti sastavljan praktično iznova. U tome smislu treba težiti za *višestrukom iskoristivošću*¹⁴ korpusa. Resursi su (ljudski i financijski) za hrvatski previše ograničeni da bi se moglo dopustiti sastavljanje većega broja *ad hoc* korpusa. Stoga bi trebalo preduhitriti takve sporadične pokušaje i pokrenuti sastavljanje obuhvatnoga korpusa hrvatskoga koji bi svojom veličinom i sastavom mogao biti reprezentativan za jezik u cjelini. Dapače, tome bi korpusu valjalo dati i pridjevnik *nacionalni* po uzoru na takve korpuse koji postoje za druge jezike.¹⁵

Za razliku od višestruke iskoristivosti, kojom se obuhvaća mogućnost uporabe korpusa od većega broja korisnika (ljudskih ili strojnih), *višesvrhovitost* korpusa vezana je uz zahtjev za teorijskom i uporabnom *neutralnošću* korpusa:

The (...) issue concerns the construction of new large-scale linguistic resources, explicitly designed to be multifunctional, that is, capable of serving, through appropriate interfaces, a wide variety of present and future research and applications. A crucial and controversial problem is whether

13 Atkins–Levin–Zampolli 1994: 21. U mom prijevodu: Tako su se sustavi za [strojnu] obradu prirodnoga jezika suočili s onim što se zove 'leksičko usko grlo' (...) ograničenjima u izvedbi sustava koja su ukazivala na potrebu za velikim leksikonom. Postalo je od presudne važnosti pronaći izvore podataka (...) koji bi omogućili stvaranje učinkovitih velikih leksikona (...). Te su potrebe pobudile obnovljeno zanimanje za korpusnu lingvistiku i leksikografiju jer su oba ova područja nudila moguću pomoć za prevladavanje 'leksičkoga uskog grla'.

14 Cf. Atkins–Zampolli 1994: 10.

15 Npr. British National Corpus cf. Atkins–Levin–Zampolli 1994: 29. Za trenutno dostupne podatke o BNC mnogo se može pronaći na WWW adresi <http://info.ox.ac.uk:80/bnc>.

it is possible (...) to make these linguistic resources 'polytheoretical', that is, usable in different theoretical frameworks.¹⁶

Prijepornost toga pitanja je li moguće sastaviti teorijski neutralan korpus valja ostaviti teorijskim raspravama. Sastavljačima korpusa to mora biti prezentna, ali ne i osnovna preokupacija koja više postoji kao teško dosegljiv ideal.

Kad se ima na umu sastav mogućega hrvatskog nacionalnog korpusa, mora se, ako se prihvati zahtjev za što je moguće većim obuhvatom građe, uočiti njezina raznorodnost koja se može svrstati u dvije kategorije. U prvoj bi se, sastavljenoj po načelima *elektronskoga tekstovnog arhiva*, našli stariji hrvatski tekstovi u cjelovitome obliku. Dakle, to bi trebao biti računalno podržan tekstovni arhiv koji bi, međutim, morao biti obilježen prema SGML standardu, tj. u skladu s *TEI smjernicama*¹⁷ i obrađen poput korpusa. U drugoj bi se kategoriji našao korpus suvremenoga hrvatskog jezika sastavljen prema kriterijima reprezentativnosti¹⁸ kojima bi se obuhvatio sinkronijski presjek¹⁹ hrvatskoga.

Dvokategoričnost odmah nameće pitanje o granici između ta dva skupa tekstova. Gdje bi morala biti razdjelnica starijih tekstova od suvremenoga hrvatskoga? U prvi, arhivski dio, svakako bi valjalo uključiti što je moguće više tekstova počevši od najstarijih na(t)pisa (Bašćanska ploča) do Ilirskoga preporoda ili druge polovice 19. stoljeća.²⁰ Druga, suvremena sastavnica mogla bi biti sastavljena prema metodološkom postupku spomenutom u bilješci 18, tj. bila bi starosti godine dana.

Kolikog bi opsega morao biti hrvatski nacionalni korpus? Za arhivski dio ne bi se smjela postavljati ograničenja. Namjena je te sastavnice ionako inventarizacijska, a ne reprezentacijska. Za drugu bi se sastavnicu morali postaviti uvjeti opsega, no pri tome ne treba iznova otkrivati što je već za druge jezike poznato. Ako se pogledaju uvodna objašnjenja u spomenutom BNC-u, može se

16 Atkins–Zampolli 1994: 11. U mom prijevodu: Radi se o sastavljanju novih jezičnih resursa velikih razmjera, izričito sklopljenih s nakanom da budu multifunkcionalni, što znači, kadri služiti, putem odgovarajućih sučelja, širokomu rasponu sadašnjih i budućih istraživanja i aplikacija. Ključni je i prijeporni problem u tom slučaju je li moguće napraviti takve jezične resurse 'politeorijskima', što znači, uporabljivima u različitim teorijskim okvirima.

17 Vidi o tome поближе u: Sperberg–McQueen & Burnard 1990; Spaeth 1991; Burnard 1990; Johansson 1994.

18 Problem sastava korpusa i njegove reprezentativnosti poznata je tema korpusne lingvistike i ovaj članak ni u kojem slučaju ne kani ulaziti u tu raspravu. Vidi o tome u Engwall 1994.

19 Vidi npr. korpusna istraživanja na Sveučilištu u Birminghamu (<http://clq1.bham.ac.uk>) gdje postoje dva osnovna korpusa tzv. COBUILD i Bank of English (sada već preko 300 milijuna pojavnica). Sinkronijski se presjek obuhvaća korpusom od oko 200 milijuna pojavnica starosti godine dana. To je metodološki iznimno zanimljiv postupak jer protokom vremena tekstovi koji postaju stariji od godinu dana ispadaju iz korpusa, a njihova se mjesta popunjavaju novima. Takav pak korpus ostavlja za sobom u vremenu trag jer se sve nove potvrde i sva nova značenja bilježe u leksičku bazu podataka. Za trenutačne podatke o COBUILD-u <http://titanica.cobuild.collins.co.uk>.

20 Vidi o povijesti hrvatskoga u Mogaš 1994.

uočiti nakana naručitelja²¹ kao i izravna uporabna vrijednost rezultata takva projekta:

The British National Corpus is a very large (over 100 million words) corpus of modern English, both spoken and written (...) The Corpus is designed to represent as wide a range of modern British English as possible. The written part (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) includes a large amount of unscripted informal conversation... (...) This generality, and the use of internationally agreed standards for its encoding, encourage us to believe that the Corpus will be useful for a very wide variety of research purposes, in fields as distinct as lexicography, artificial intelligence, speech recognition and synthesis, literary studies, and all varieties of linguistics.²²

Uz komponentu sinkroničnosti ovako zamišljen korpus hrvatskoga mogao bi zadovoljiti mnoge istraživačke,²³ a i potrebe komercijalnih korisnika ukoliko bi rezultati takva projekta mogli biti dostupni javno, npr. preko CARNET-a.²⁴

Literatura

- Allen, James (1987) *Natural Language Understanding*, Benjamin/Cummings, Menlo Park.
- Andersen, Poul (1995) *Cooperation with Central and Eastern Europe in Language Engineering* u: TELRI, str. 9–20.
- Andrijašević, Marin; Vrhovac, Yvonne (ur.) (1990) *Informatička tehnologija u primijenjenoj lingvistici*, Hrvatsko društvo za primijenjenu lingvistiku, Zagreb.
- Årsmelding/Annual Report (1995) Humanistisk Datasenter/Norwegian Computing Centre for the Humanities, Bergen
- Atkins, B. T. S.; Levin, B.; Zampolli, A. (1994) *Computational Approaches to the Lexicon: An Overview* u: Atkins & Zampolli (1994), str. 18–45.
- Atkins, B. T. S.; Zampolli, A. (ur.) (1994) *Computational Approaches to the Lexicon*, Oxford University Press, Oxford 1994.

21 Akademsko-industrijski konzorcij u sastavu: Oxford University Press (voditelj), Longman, Chambers Harrap, Oxford University Computing Services, Lancaster University's Unit for Computer Research on the English Language i British Library.

22 BNC: <http://info.ox.ac.uk:80/bnc/whatbnc.html>. u mom prijevodu: BNC je vrlo velik korpus (preko 100 milijuna pojava) suvremenoga engleskoga (govorenoga i pisanoga) (...) Zamišljen je s ciljem da predstavlja britanski engleski što je obuhvatnije moguće. Pisani dio (90%) uključuje između ostalih vrsta tekstova npr. izvratke iz regionalnih i nacionalnih novina; stručnih periodičnih časopisa za sve dobi i zanimanja; akademske knjige i lijepu književnost; objavljena i neobjavljena pisma i memorandumne; školske i sveučilišne eseje. Govoreni dio (10%) uključuje veliku količinu nepisanih nejavni razgovora... (...) Ova općenitost (korpusa) kao i uporaba međunarodno dogovorenih standarda za njegovo kodiranje, potiče našu nadu da će Korpus biti koristan cijelome nizu istraživačkih pristupa, i to u toliko različitim područjima kao što su leksikografija, umjetna inteligencija, prepoznavanje i stvaranje govora, znanost o književnosti i sve vrste jezikoslovnih disciplina.

- Boguraev, Bran; Briscoe, Ted (ur.) (1989) *Computational Lexicography for Natural Language Processing*, Longman–John Wiley & Sons, London–New York.
- Burnard, Lou (1990) *The Text Encoding Initiative: A Progress Report*, Humanistiske Data 3/1990, str. 52–58.
- Cencioni, Roberto; Klein, Ewan (ur.) (1994) *Linguistic Research & Engineering (LRE). An Overview*, Directorate–General XIII, Commission of the European Communities.
- Engwall, Gunnel (1994) *Not Chance but Choice: Criteria in Corpus Creation u: Atkins & Zampolli (1994)*, str. 49–82.
- Johansson, Stig (1994) *Encoding a Corpus in Machine–Readable Form: The Approach of the Text Encoding Initiative u: Atkins & Zampolli (1994)*, str. 83–102.
- Leech, Geoffrey (1992) *Corpora and theories of linguistic performance u: Svartvik (1992)*, str. 105–122.
- Mackie, Andrew W. (1992) *Preliminaries to a Parsed Corpus: Format and Usage u: Mackie–McAuley–Simmons (1992)*, str. 243–260.
- Mackie, Andrew W.; McAuley, Tatyana K.; Simmons, Cynthia (1992) *For Henry Kučera*, Studies in Slavic Philology and Computational Linguistics, Michigan Slavic Publications, Ann Arbor.
- Moguš, Milan (1995) *Povijest hrvatskoga književnoga jezika*, Nakladni zavod Globus, Zagreb.
- Sinclair, John (ur.) (1987) *Looking Up. An account of the COBUILD Project in lexical computing*, Collins, London–Glasgow.
- Spaeth, Donald A. (1991) *Living with Guidelines. The First TEI European Workshop*, Humanistiske Data 2/1991, str. 81–85.
- Sperberg–McQueen, C. M.; Burnard, Lou (1990) *Guidelines for the Encoding and Interchange of Machine–Readable Texts*, Text Encoding Initiative, Chicago–Oxford.
- Svartvik, Jan (ed.) (1992) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Mouton, Berlin.
- 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.
- TELRI (1995) *Language Resources for Language Technology*. Proceedings of the First European Seminar, Tihany.
- Winograd, Terry (1983) *Language as a Cognitive Process*. Addison–Wesley, Reading.
- Žubrinić, Tomislava (1995) *Mogućnosti strojnoga označivanja i lematiziranja korpusa tekstova hrvatskoga jezika*, magistarski rad, Filozofski fakultet Sveučilišta u Zagrebu, Zagreb.

Natural Language Processing of Croatian and the National Corpus

The article aims to show the applicability of two initial arguments by which the status of the NLP of Croatian is trying to be defined. These two arguments are: 1) NLP of Croatian must have a status of fundamental research in the field of humanities; 2) NLP of Croatian must have a status of strategic research of the Republic of Croatia.