

# Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive\*

Marko Tadić (marko.tadic@ffzg.hr)

Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu

(<http://www.ffzg.hr/zzl/zzl-home.htm>)

Članak daje pregled obradbe hrvatskih korpusa. Navode se najznačajniji korpusni projekti od prvoga hrvatskoga računalno podržanoga korpusa — Gundulićev *Osman* (Bujas 1967), preko Moguševa milijunskoga korpusa do današnjih dana. Nadalje se rad usredotočuje na Hrvatski nacionalni korpus koji je središnji projekt na području korpusne lingvistike u Hrvatskoj danas. Hrvatski nacionalni korpus (Tadić 1996) čine dvije sastavnice: 1) reprezentativni 30-milijunski korpus suvremenoga hrvatskoga jezika (30M) i 2) Hrvatski elektronski tekstovni arhiv (HETA). U prvoj fazi sastavljanja Hrvatskoga nacionalnoga korpusa naglasak je na zaokruživanju 30-milijunskoga korpusa dok će se u drugoj fazi sav napor preusmjeriti na širenje obuhvata Hrvatskoga elektronskog tekstovnoga arhiva. U sadašnjem stanju rad na 30-milijunski korpusu, koji bi trebao biti završen 2000, u fazi je uznapredovala planiranja i javnoga testiranja probne inačice korpusa (7,68 milijuna pojava) putem WWW-a.

## 0. Uvod

Na samome bi početku valjalo omediti područje i pristup o kojem će ovdje biti riječi. Kad se ovdje govori o *hrvatskome korpusu* misli se na (računalno) podržane korpusne tekstova nastalih na hrvatskome jeziku.<sup>1</sup> Nadalje, valja imati na umu da je ovo *pogled lingvista* kojemu je do korpusa i njegove obradbe ponajprije stalo kao do sredstva za kvalitetniji jezični opis, a ne *pogled informatičara* kojemu su način, brzina, izvedba obradbe jezičnih podataka u prvome planu.

Kako bismo stekli uvid i razumjeli sadašnje stanje valja nam se vratiti unatrag. Ovdje će se dati pregled projekata u Republici Hrvatskoj kojima je osnovni cilj ili metodologija

---

\* Članak je neznatno doradena verzija izlaganja priređenog za XII. međunarodnom slavističkom kongresu u Krakovu, u rujnu 1998.

<sup>1</sup> U Hrvatskoj se, naime, obrađuju ili su se obrađivali i korpusi i na drugim jezicima: npr. specijalističkim se engleskim korpusima bave Boris Pritchard s Pomorskoga fakulteta Sveučilišta u Rijeci i Milica Gačić s Policijske akademije u Zagrebu.

bila ona iz domene korpusne lingvistike. Većina je tih projekata bila ili je još uvijek smještena u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

## 1. Jučer

Kako je hrvatski na teorijskoj razini doista rano stupio u dodir s računalnim pristupom prirodnome jeziku u nezaobilaznome radu »Broj u jeziku« što ga je napisao Bulcsú László,<sup>2</sup> bilo je za očekivati da će se obradba korpusa, kao jedno od područja na kojemu su računala ponajprije našla svoju primjenu u lingvistici, pojaviti istovremeno ili neposredno po tom. Međutim, pomalo je neobično što je prvi hrvatski korpus (veličine 100.000 pojavnica) sastavio i frekvencijski, premda ne i računalno, obradio psiholog Ivan Furlan u svojoj disertaciji *Raznolikost rječnika. Struktura govora*.<sup>3</sup>

Prvi računalno i iz lingvističkih pobuda obrađen korpus bio je barokni ep Ivana Gundulića *Osman* koji je priredio, frekvencijski obradio i konkordancijama popratio Željko Bujas 1967. za boravka na Sveučilištu u Austinu, u SAD.<sup>4</sup> Slijedio ga je ubrzo s istim autorstvom i tehnikom priređen *Povratak Filipa Latinovicza* Miroslava Krleže, a potom i Marulićeva *Suzana*.

U Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu 1968. pod voditeljstvom Rudolfa Filipovića započinje se veliki kontrastivni projekt pod nazivom *Yugoslav Serbo-Croatian -- English Contrastive Project* u kojem se po prvi puta u nas pokreće računalna obradba korpusa. Za nju je, zahvaljujući iskustvima stečenim u SAD, unutar projekta bio zadužen Željko Bujas. Pribavljen je *Brown korpus* na magnetnome mediju, prepolovljen na veličinu od 505.822 pojavnice uz očuvanje omjera svih 15 tekstovnih žanrova iz izvornoga korpusa.<sup>5</sup> Potom je morfosintaktički obilježen i preveden na ono što se tada zvalo tri standardne varijante hrvatskoga ili srpskoga. Time

---

<sup>2</sup> László Bulcsú (1959) »Broj u jeziku«, *Naše teme* 6/1959 i ispravljeni pretisak u *SOL* 10-11/1990.

<sup>3</sup> Furlan, Ivan (1961) *Raznolikost rječnika. Struktura govora*, neobjavljena disertacija, Filozofski fakultet Sveučilišta u Zagrebu, Zagreb.

<sup>4</sup> Dorađena, druga inačica konkordancije objavljena je u Bujas, Željko (1975a) *Ivan Gundulić »Osman«*. *Komputorska konkordancija*, Sveučilišna naklada Liber, Zagreb 1975.

<sup>5</sup> Bujas (1975b:49). Također i u Bujas (1969: 36)

su se po prvi puta dobili i paralelni korpusi. Na temelju tih prijevoda napravljena je konkordancija s morfosintaktičkim kategorijama kao stožernicama (točnije: 231 funkcionalna riječ i 257 morfosintaktičkih elemenata<sup>6</sup>) i dvojezična rečenična kartoteka s pomoću kojih se moglo pretraživati i engleski i prevedene korpusse. To je ujedno i prva uporaba računala u svjetskoj kontrastivnoj lingvistici.<sup>7</sup> Projekt je trajao do 1971. i osim izravnih svjetskih dostignuća u kontrastivistici, rezultirao je još za svoga trajanja i snažnim metodološkim utjecajem na druge projekte u Zavodu koji su počeli usvajati računalnu obradu korpusa kao standardnu istraživačku proceduru. Sama je obrada uvijek, ispravno, bila smatrana prije svega sredstvom obrade jezične građe, a ne ciljem.

U istoj je instituciji 1968. pod voditeljstvom Milana Moguša započet projekt pod nazivom *Jezik Marka Marulića*. Željko Bujas obradio je za taj projekt već spomenutu Marulićevu *Suzanu*. Godine 1970. projekt je proširen i preimenovan u *Kompjutorska analiza tekstova stare hrvatske književnosti*<sup>8</sup> i s njim je korpusna obrada hrvatskih tekstova dobila svoj pravi zamah. Do 1981. konkordirana su sva hrvatska Marulićeva djela, djela Barne Karnarutića, Zoranićeve *Planine*, Pelegrinovićeve *Jejupka* (3 verzije), djela Hanibala Lucića i Petra Hektorovića, Benetovićeve *Hvarkinja*, *Ranjinin zbornik*, Držićeve komedije, djela Ivana Bunića Vučića, Vitezovićeve djela, Kanižlićeva *Sveta Rožalija*, komedije Tituša Brezovačkoga i *Razvod istarski*<sup>9</sup> kao i, neovisno o tom projektu, *Balade Petrice Kerempuha* Miroslava Krležje. Tako dobiveni rezultati korišteni su ponajprije za nova kritička čitanja kao i potvrđivanje autorstva gdje je ono bilo dvojbeno (npr. Moguš (1976) o Maruliću<sup>10</sup>) ali i istraživanja na svim jezičnim razinama od leksičke do stilističke. Za tog je projekta hrvatska korpusna lingvistika bila u potpunosti u svjetskom trendu tada zvanom *Literary and linguistic computing*.

Također je u Zavodu za lingvistiku od 1972. do 1975. pod voditeljstvom Željka Bujasa trajao projekt *Englesko-hrvatski leksikografski korpus* čija je polazna ideja bila tretirati tekst dvojezičnoga rječnika kao korpus. To je dovelo do preokretanja i konkordiranja po

---

<sup>6</sup> Bujas (1975b: 53)

<sup>7</sup> cf. Bujas (1975b:44)

<sup>8</sup> Moguš (1975:66)

<sup>9</sup> Moguš (1975:67), Bratanić-Čimbur (1977:46-47), Bratanić-Čimbur(1979:40-41), Bratanić (1981:70-71)

<sup>10</sup> Moguš (1976)

hrvatskoj stožernici čitava Filipovićeva *Englesko-hrvatskoga rječnika*,<sup>11</sup> (veličine preko milijun pojava) što je za to vrijeme također bilo jedinstveno računalnolingvističko dostignuće u dvojezičnoj leksikografiji. Na tom je korpusu kasnije na Elektrotehničkom fakultetu radio Šandor Dembitz.

Kako su se obrade korpusâ starijih hrvatskih pisaca odvijale već uhodanim i usvojenom metodološkim obrascima ubrzo se pokazala potreba za korpusom kojim bi se moglo proučavati pojave prisutne u jezičnoj sinkroniji tj. korpusom reprezentativnim za suvremeni hrvatski jezik. Tako je godine 1976. pod vodstvom Milana Mogušâ pokrenut projekt *Korpus suvremenog hrvatskog književnog jezika* s primarnim ciljem sastavljanja jednomilijunskoga korpusa nazvanog Mogušev korpus. Stjecajem raznih, a to znači i ratnih, okolnosti taj je korpus čija se obrada sastojala od abecednih i frekvencijskih rječnika pojava, konkordancija, potom i strojno potpomognute lematizacije,<sup>12</sup> završen 1996. premda je njegova građa bila dostupna znatno ranije. Taj je korpus prvi pokušaj u hrvatskoj lingvistici da se na temeljima reprezentativnoga korpusa započne usustavljivanje i istraživanje jezične građe. Pokušaj je to koji stoji uz bok tadašnjim svjetskim trendovima u sastavljanju korpusa. Najznatnijim se rezultatom obrade Moguševa milijunskoga korpusa može držati prvi *Hrvatski čestotni rječnik* koji se upravo nalazi u tisku kod Školske knjige.<sup>13</sup>

Sam je korpus sastavljen prema sljedećim principima reprezentativnosti:

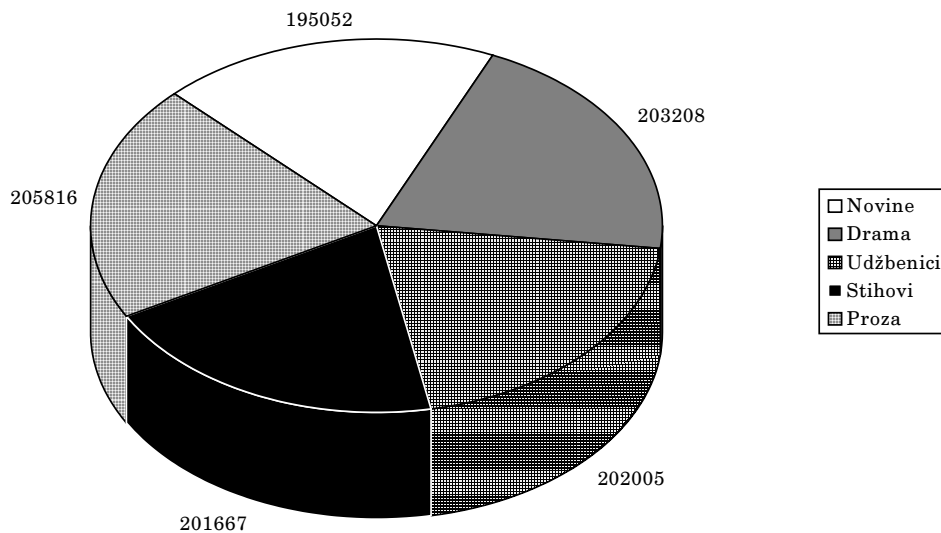
- 5 potkorpusa različitih tekstovnih žanrova:
  - Drama            20 uzoraka po 10.000 pojava
  - Novine            8 uzoraka po 25.000 pojava
  - Proza             20 uzoraka po 10.000 pojava
  - Stihovi           20 uzoraka po 10.000 pojava
  - Udžbenici       58 uzoraka po 3450 pojava

---

<sup>11</sup> Filipović, Rudolf (1971) *Englesko-hrvatski rječnik*, Zora, Zagreb 1971.

<sup>12</sup> O postupcima pri obradbi Moguševa korpusa kao i lematizaciji vidi u Tadić (1992)

<sup>13</sup> Moguš-Bratanić-Tadić (1999)



- Vremenski raspon: 1935-1978.
- Veličina uzorka: testiranjem prirasta novih riječi između uzoraka veličine 5, 10 i 20 tisuća pojavnica odabran je uzorak veličine 10 tisuća čime je postignuta i bolja disperzija uzoraka prema različitim autorima.<sup>14</sup>

Mogušev je milijunski korpus, kad je zamišljen 1975/76, bio znatan već i time što je tada bio prvi milijunski korpus nekoga slavenskoga jezika, ali je, nažalost, zbog duljine projekta njegova veličina danas postala neadekvatna za narasle kako jezikoslovne potrebe (ponajprije istraživanja suvremena leksika) tako i uznapredovale tehnološke mogućnosti. To ni u kojem slučaju ne umanjuje njegov kapitalan i ishodišni položaj u hrvatskoj korpusnoj lingvistici prije svega zbog jasne i čvrsto definirane primjene kriterija reprezentativnosti pri sastavljanju. Nadalje, njegova je uporabivost pri istraživanjima na morfološkoj, sintaktičkoj pa i frazeološkoj razini neupitna, no najzanimljivijom se može smatrati mogućnost koju njegovo postojanje otvara, a to je da se npr. na planu leksika mogu usporediti podaci iz tekstova ostvarenih prije godine 1990. s onima ostvarenim poslije te godine. Kada Moguševa korpusa ne bi bilo, takve bi usporede bile teško ostvarive, a nikako ih ne bi bilo moguće egzaktno, frekvencijski iskazati.

Na Odsjeku za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu, 1976. frekvencijski je korpus tekstova Vjesnika i Večernjega lista (opsega 130.000 pojavnica)

<sup>14</sup> Detaljno o svim elementima korpusa vidi u predgovoru Moguš-Bratanić-Tadić (1999)

obradio Zorislav Šojat.<sup>15</sup> To je istraživanje prvo *naručeno* korpusno istraživanje tj. prvo istraživanje korpusa koje nije financirano iz sredstava Ministarstva znanosti.

Na Odsjeku za informacijske znanosti Filozofskoga fakulteta Sveučilišta u Zagrebu u prvoj je polovici osamdesetih sastavljen neuravnotežen korpus tekstova osnovnoškolskih udžbenika i novinskih tekstova (otprilike milijun pojavnica) pod vodstvom Damira Borasa uz sudjelovanje Miroslava Kržaka i suradnika. Na tom su korpusu početkom devedesetih rađena istraživanja probabilističkoga označivača (*taggera*) za hrvatski.<sup>16</sup>

Na poticaj EEZ-a 1988. bivša je Jugoslavija pozvana u projekt *Language Industries* pod pokroviteljstvom Vijeća Europe. Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu tada je bio koordinator između Ljubljane, Zagreba i Beograda i uspješno je 1989. u Dubrovniku organizirao konferenciju *Language Industries — Needs and Perspectives* na kojoj su se prvi put skupili stručnjaci iz zapadnoeuropskih i srednjo- i istočnoeuropskih zemalja. Tada je u području korpusne lingvistike uspostavljena znatnija suradnja Zavoda sa zapadnoeuropskim projektima i centrima (Pisa, Birmingham, Mannheim) prije svega u razmjeni stručnjaka i ideja.

Maja je Bratanić kao voditeljica hrvatskoga segmenta od 1990. do 1991. sudjelovala u međunarodnom projektu *Multilingual lexicography project* pod okriljem Vijeća Europe i vodstvom Johna Sinclaira. Sudjelovanje je bilo moguće ponajprije zahvaljujući postojećem milijunskom korpusu koji je dopuštao uvid u kontekst toliko potreban za pronalaženje prijevodnih ekvivalenata.

Na žalost, sva je suradnja zaustavljena 1991. srbijanskom agresijom na Republiku Hrvatsku. Zamrzavaju se sve aktivnosti od strane Europske unije tako da je odsutnost Hrvatske iz sudjelovanja u PHARE, TEMPUS, COPERNICUS programima, koji su formirali okvire za znanstveno-tehnološku suradnju, rezultirala stagnacijom i nemogućnošću praćenja razvitka korpusnolingvističkih istraživanja.

---

<sup>15</sup> Šojat (1976)

<sup>16</sup> Žubrinić (1995)

## 2. Danas

Lagano odmrzavanje tek 1995. ulaskom Zavoda za lingvistiku u međunarodni projekt koji promiče međueuropsku suradnju na području jezičnih resursa — TELRI. Međutim, ni tada, kao ni danas, Republika Hrvatska nema status punopravnoga člana projekta isključivo zbog političkih razloga.

Istodobno se, u Zavodu za lingvistiku koji je potvrdio svoj rodonačelni položaj na području korpusne lingvistike u Hrvatskoj obrađuju korpusi: Križanić (*Politika* 1988), Gundulić (ukupna djela 1989.), u suradnji sa Splitskim književnim krugom Marulić (*Judita, Dijaloška djela, Suzana, Vartal* 1990-1993), Mažuranić (*Smail-aga Čengić* 1994), *Trsatski statut, Vinodolski zakon* (1988), *Jačke* Mate Mešića Miloradića, *Škrinja* Luke Perkovića.

Stručnjaci Zavoda za lingvistiku od 1995. do 1997. sudjelovali su i u suradnji Hrvatske akademije znanosti i umjetnosti s Austrijskom akademijom znanosti na obradi korpusa Katančićeva prijevoda *Svetoga pisma* u okviru značajnoga prj. *Civilizacijska terminologija jugoistočne Europe*.<sup>17</sup>

U travnju 1996. pod vodstvom Vesne Muhvić-Dimanovski pri Ministarstvu znanosti i tehnologije prijavljen je projekt *Računalna obradba hrvatskoga jezika* kojem je jedan od primarnih zadataka sastavljanje više desetaka milijuna velikoga korpusa hrvatskoga jezika. U Tadić (1996) dana je struktura toga korpusa i predloženo je da ga se, po uzoru na češki i britanski, zove *Hrvatski nacionalni korpus* (odsada HNK).<sup>18</sup> Taj je projekt konačno u veljači 1998. dobio zasluženu pozornost Ministarstva u obliku financijske pomoći i računalne opreme.

Sam je HNK sastavljen od dvije sastavnice:<sup>19</sup>

---

<sup>17</sup> Moguš-Tadić (1997)

<sup>18</sup> Hrvatski se nacionalni korpus može konzultirati preko WWW-a na adresi:

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

<sup>19</sup> V. Tadić (1996) i Tadić (1998)

1. **30M:** reprezentativni 30 milijunski korpus suvremenoga hrvatskoga jezika (s tekstovima nastalim 1990. i nakon nje)
2. **HETA:** *Hrvatski elektronski tekstovni arhiv* čini neuravnotežena zbirka korpusa u koji se smještaju i obrađuju tekstovi ili stariji od 1990. ili tekstovi koji ne odgovaraju zahtjevima reprezentativnosti 30M a sami čine znatne korpuse (nekoliko milijuna pojavnica). Korpusi su u HETA obrađeni istom korpusnom metodologijom kao i u 30M.

Za 30M korpus iz hrvatskoga WWW prostora (.hr domena) skupljeno je oko tridesetak milijuna pojavnica koje, međutim, ne odgovaraju u potpunosti zahtjevima koje postavlja reprezentativni korpus. Također je napravljen poseban programski paket pod sustavom Windows NT za obradu, održavanje i pretraživanje korpusa. Taj paket omogućuje konvertiranje i pretprocesiranje tekstova dobivenih u HTML ili RTF formatu, dopušta pohranjivanje tekstova u HTML i SGML formatu s DTD-om prema minimalnoj PAROLE specifikaciji<sup>20</sup> za potrebe kasnije kompatibilnosti s drugim korpusima i alatima.

Pretraživanje trenutačne, neuravnotežene i nereprezentativne inačice korpusa (veličine 7,67 milijuna pojavnica) zasada je dostupno preko WWW-a u obliku jednostavne konkordancije uz odabir željenoga (pot)korpusa.

Sastav 30M korpusa koncipiran je prema EAGLES<sup>21</sup> preporukama o tekstovnoj i žanrovskoj tipologiji uz neprihvatanje tekstova prijevoda. Preostaje, uz podršku Ministarstva, očekivati i od nakladnika da se uključe u sastavljanje HNK ustupanjem svojih izdanja na elektronskome mediju.

Kako je jedan od osnovnih ciljeva HNK primjena u leksikografiji, sklopljen je ugovor o razmjeni podataka sa Institutom za hrvatski jezik i jezikoslovlje u kojem se sastavlja jednosvezačni hrvatski rječnik te je taj Institut tako postao najvećim korisnikom HNK-a.

---

<sup>20</sup> O PAROLE specifikaciji vidjeti na <http://svenska.gu.se/~ridings/textrep/textrep.html>

<sup>21</sup> O EAGLES projektu vidjeti na <http://www.ilc.pi.cnr.it/EAGLES/home.html>

### 3. Sutra

U godini 2000. *30M korpus* bi trebao biti završen, obrađen i objavljen, ako to tehnološke mogućnosti budu dopuštale, i na CD-ROM-u.

Nakon toga će se prići daljnjoj obradbi korpusa. Svaka će pojava biti popraćena morfosintaktičkim opisom u skladu s preporukama MulTextEast projekta<sup>22</sup> gdje se, kao i za slovenski ili češki, koriste gotove MTE etikete.

Valja se nadati da će do tada HNK steći status fundamentalnog istraživanja u humanističko-društvenim znanostima i, možda, strateškoga istraživanja za Republiku Hrvatsku.

Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu kani se s HNK uključiti u: NERC (Network of European Reference Corpora) kao i u međunarodne projekte i organizacije (TELRI II, ELSNET) u statusu punopravnoga člana.

Cilj je nakon završetka 30M korpusa dopunjavati HETA u koracima koji uključuju cijele korpuse bilo pojedinih autora ili više godišta pojedine publikacije kako bi jezikoslovnoj javnosti na raspolaganju mogla biti uvijek nova i do tada nedostupna jezična građa.

### Literatura

Bratanić-Čimbur, Maja (1977) »Kompjutorska analiza tekstova starije hrvatske književnosti«, Bilten Instituta za lingvistiku 2, Zagreb, 46.

Bratanić-Čimbur, Maja (1979) »Kompjutorska analiza tekstova starije hrvatske književnosti«, Bilten Instituta za lingvistiku 3, Zagreb, 40-41.

Bratanić, Maja (1981) »Kompjutorska analiza tekstova starije hrvatske književnosti«, Bilten Instituta za lingvistiku 4, Zagreb, 70-71.

Bratanić, Maja (1992) »Izgradnja hrvatske i višejezične baze podataka (I. faza)«, Bilten Zavoda za lingvistiku 6, Zagreb, 23-25.

---

<sup>22</sup> Erjavec–Lawson–Romary (1998)

- Bujas, Željko (1967) »Concordancing as a Method in Contrastive Analysis«, SRAZ 23, 49-62.
- Bujas, Željko (1969) »Computers in the Yugoslav Serbo-Croatian/English Contrastive Analysis Project«, ITL Review for Applied Linguistics, Spring 1969, 35-42.
- Bujas, Željko (1975a) *Ivan Gundulić »Osman«*. *Komputorska konkordancija*, Sveučilišna naklada Liber, Zagreb 1975.
- Bujas, Željko (1975b) »Computers in the Yugoslav Serbo-Croatian : English Contrastive Project«, Bilten Instituta za lingvistiku 1, Zagreb, 44-58.
- Erjavec, Tomaž–Lawson, Ann–Romary, Laurent: *East meets West — A Compendium of Multilingual Resources*, TELRI-IDS, Mannheim, 1998.
- Filipović, Rudolf (1971) *Englesko-hrvatski rječnik*, Zora, Zagreb 1971.
- Furlan, Ivan (1961) *Raznolikost rječnika. Struktura govora*, neobjavljena disertacija, Filozofski fakultet Sveučilišta u Zagrebu, Zagreb.
- László Bulcsú (1959) »Broj u jeziku«, *Naše teme* 6/1959 i ispravljani pretisak u *SOL* 10-11/1990, str. 121-154.
- Moguš (1975) »Kako su se Marulićeva djela našla u kompjuteru«, Bilten Instituta za lingvistiku 1, Zagreb, 65-68.
- Moguš (1976) »Je li Marulić autor Firentinskog zbornika«, *Radovi Zavoda za slavensku filologiju* 14, Zagreb, 44-60.
- Moguš, Milan–Tadić, Marko (1997) *Katančićev prijevod Svetoga pisma: računalna obrada, abecedni, čestotni rječnik i konkordancija*, Austrijska akademija znanosti-HAZU, Beč-Zagreb 1997.
- Moguš, Milan–Bratanić, Maja–Tadić, Marko (1999) *Hrvatski čestotni rječnik*, Zavod za lingvistiku-Školska knjiga, Zagreb.
- Šojat, Zorislav (1976) *Čestotni rječnik Vjesnika i Večernjeg lista*, Zagreb.
- Tadić, Marko (1992) »Od korpusa do čestotnog rječnika hrvatskoga književnog jezika«, *Radovi zavoda za slavensku filologiju* 27, Zagreb, 169-179.
- Tadić, Marko (1996) »Računalna obradba hrvatskoga i nacionalni korpus«, *Suvremena lingvistika* 41-42, 603-612.
- Tadić, Marko (1998) »Raspon, opseg i sastav korpusa suvremenog hrvatskoga jezika« *Filologija* 30-31, 337-347.
- Žubrinić, Tomislava (1995) *Mogućnosti strojnoga označavanja i lematiziranja korpusa tekstova hrvatskoga jezika*, magistarski rad, Filozofski fakultet Sveučilišta u Zagrebu.

# **Croatian Corpus Processing: History, State-of-art and Perspectives**

*dr. Marko Tadić* (marko.tadic@ffzg.hr)

Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu

(<http://www.ffzg.hr/zzl/zzl-home.htm>)

This article gives a survey of Croatian corpus processing. It lists the most important projects since the first Croatian computer corpus (Gundulić's *Osman*) up to the present time. The article focuses on the Croatian National Corpus which is the central project in the field of corpus linguistics in Croatia today. The Croatian National Corpus consists of two parts: 1) representative 30-million Corpus of Contemporary Croatian Language and 2) Croatian Electronic Text Archive. The 30-million Corpus covers the first phase of the Croatian National Corpus while the effort in the second phase will be concentrated on the widening of the contents of the Croatian Electronic Text Archive. The 30-million Corpus, which is now at the stage of advanced planning and software and pilot corpus (7,67 million of running words) testing, should to be finished in the year 2000.