МЕЛАНИЈА МИКЕШ

КОМПЈУТЕРСКА ОБРАДА ЛИНГВИСТИЧКИХ ПОДАТАКА — ДОСАДАШЊИ РЕЗУЛТАТИ, ПОТРЕБЕ И ПЕРСПЕКТИВЕ

Под овим називом одржан је 9. и 10. децембра 1977. године у Сарајеву научни скуп, који је организовао Институт за језик и књижевност у Сарајеву, под покровитељством Акамедије наука и умјетности Босне и Херцеговине. На скупу је било 65 заинтересованих лингвиста, инжињера и математичара из свих југословенских центара, поднесено је двадесет реферата и вођена је веома корисна дискусија о назначеној проблематици, а дата су и мишљења и сугестије за даље унапређење ове интердисциплинарне научне области.

Учесници овог скупа имали су прилике да се на основу поднесених реферата упознају са радом и резултатима до којих се дошло у појединим југословенским научноистраживачким центрима (1). У дискусији су дошли до изражаја неки основни и заједнички проблеми свих посленика у овој области (2). Радна и пријатељска атмосфера овог скупа омогућила је формирање неких заједничких мишљења и сугестија које ће бити упућене заинтересованим институцијама и стручним и научним асоцијацијама (3).

1. Преглед рада и резултата у појединим југословенским научноистраживачким центрима, који ћу покушати да дам на основу онога о чему сам се информисала на овом скупу, не претендује на целовитост. Оно што ћу изложити је само виђење једног лингвисте који нема никаква искуства у компјутерској лингвистици, али је своје присуство на овом скупу мотивисао, пре свега, својим дубоким уверењем да све послове у којима машина може да замени човека — и то боље и ефикасније — треба препустити машини, и тако ослободити знатну количину временског и људског потенцијала за креативни рад. Други мотив мог интересовања за проблеме компјутерске лингвистике проистиче из моје опредељености за примењену лингвистику, а то у овом конкретном случају значи да сматрам да резултати лингвистичких истраживања треба да допринесу усавршавању машина за што ефикаснију службу човеку.

1. 1. Прилазећи са тим истим мотивацијама и сређивању информација које сам добила на овом скупу, сасвим је разумљиво што се прво заустављам на раду и резултатима који су

постигнути у Загребу а одвијају се у оквиру успешне сарадње између Завода за лингвистику Филозофског факултета у Загребу и Електронског рачунског центра града Загреба, јер начела на којима се заснива сарадња између те две институције и начин на који се она одвија веома су блиски ономе што мотивише моје интересовање за компјутерску лингвистику. Наиме, на основу реферата које су поднели Рудолф Филиповић (»Језички корпус и његова компјуторска обрада у служби контрастивне анализе«), Маја Батанић-Чимбур (»Припрема текста за компјуторску обраду«), Милутин Цихлар (»О компјуторској обради конкорданција«) и Милан Могуш (»Употреба конкорданција при анализи текста«) стекла сам утисак да је иницијатива за сарадњу потекла од стране лингвиста, који су реално оценили предности компјутерске обраде језичког корпуса, усмерили поједине истраживаче у правцу специјализације за тај посао, оформили екипу и успешно укључили у решавање својих проблема специјалисте у области електронике. Они су у периоду од 1970. године остварили следеће пројекте: Контрастивне конкорданције српскохрватског и енглеског језика, Енглеско-хрватски лексикографски корпус, Компјуторска анализа текстова старије хрватске књижевности и Корпус савременог хрватског књижевног језика.

Рудолф Филиповић и Маја Батанић-Чимбур посветили су у својим рефератима доста пажње питањима припреме текста за компјутерску обраду. Откривање и исправљање грешака насталих приликом компјутеризације текста представља несразмерно дуготрајну и скупу фазу посла. Стога су искоришћене све могућности рационализације и скраћивања фазе припремања текста. Таквим размишљањима руководило се и приликом избора језичког корпуса за контрастивну анализу српскохрватског и енглеског језика, па је избор пао на корпус који се већ налазио на магнетској траци.

Међутим, проблеми који искрсавају приликом припремања текста за компјутерску обраду (а који ће у великој мери бити отклоњени када се буду примењивали оптички читачи) не треба да обесхрабре лингвисте када се одлучују за машинску обраду језичких података, јер, према речима професора Филиповића, компјутерски обрађен материјал представља »драгоцени извор за сваког нашег анализатора, без обзира на то коју тему обрађује. Употреба материјала показала се не само врло корисна за обраду задатака унутар нашег пројекта већ и за анализу других питања у енглеском и хрватском или српском језику«. Он наводи неке конкретне примере, и каже: »Та документација из компјутерски обрађеног корпуса дала је потпуно нове вриједости анализама изнесеним у студијама (које су писане пре употребе овог корпуса)«.

Да би компјутерски обрађен језички материјал могао што шире да се примењује, па на тај начин оправда и време, и средства, и људски рад који је уложен у припремној фази, потребно је да се пре почетка преноса текста у тај медиј принципи коди-

рања текста разраде до те мере да он у свом коначном облику пружи могућност максималне и разнолике искористивости на свим лингвистичким нивоима, од фонолошког, преко морфолошког и синтаксичког, до семантичког и прагматског.

Милутин Цихлар, дипл. инг., приказао је у свом реферату повезаност у раду између човека и компјутера и показао где компјутер помаже човеку. Он је објаснио главне фазе рада приликом израде компјутерске конкорданције, а то су: утврђивање текста, исправак формалних грешака, тј. грешака које се могу отклонити помоћу компјутера, исправак логичких грешака, тј. грешака које се могу отклонити само визуелном контролом исписа добивених помоћу компјутера и основног текста, обликовање за штампање коначне верзије текста и обрада коначног текста у облик конкорданције. Свака од наведених фаза обухвата по неколико програма који се изводе на компјутеру, а уз то су нужне и људске интервенције.

Цихлар је изнео низ искустава у вези са конверзијом текста, дилему око избора улазног медија за обухват и унос података у компјутер која је условљена варијабилношћу текста, с обзиром на садржај и форму, и најзад је указао на све оне радње техничког карактера које приликом обраде текста врши компјутер уместо човека.

1. 2. У Љубљани се проблематиком компјутерске лингвистике интензивно баве у Институту »Јожеф Стефан«. О достигнућима компјутерске обраде и математичког моделирања словеначког језика, као и о неким дефиницијама и таксономијама компјутерске и математичке лингвистике, говорио је сарадник тог института Петер Танциг.

У реферату су наведена сва главна достигнућа из ове области у Словенији, при чему су истакнути радови у које је аутор био укључен. Демонстрирани су неки формални фазно-структурни модели главних синтаксичких структура у словеначком језику и њихово компјутерско генерисање. Поменута је и Данешова формализација словеначких реченица.

Танциг је, такође, рекао да је у Југославији мултидисциплинарно поље компјутерске лингвистике у почетној фази организованог развоја. Досад је главна особина тог развоја била неусклађеност напора на различитим подручјима компјутерске обраде природних језика, разбацаним и неповезаним по различитим крајевима земље. Он се залаже за такав развој компјутерске лингвистике у Југославији који би био прилагођен специфичним језичким условима у нашим срединама.

1. 3. Домаћин овог скупа, Институт за језик и књижевност у Сарајеву, који своју делатност у овој области остварује у сарадњи са »Енергоинвестом«, представио се са три реферата.

Марија Ковачић говорила је о могућности утврђивања ауторства литерарног дела помоћу компјутерске конкорданције. Служећи се искуствима загребачког пројекта »Компјуторска анализа текстова старе хрватске књижевности«, којим руководи

професор Могуш, учињен је покушај да се два текста која се припусују Петру Кочићу компјутерски обраде и да се њихове конкорданце упореде са конкорданцама језичког корпуса осталих Кочићевих дела. Иначе је компјутерска анализа Кочићевих дела предмет пројекта под називом »Језик Петра Кочића«, о којем је реферисао Милан Шипка, у вези са могућношћу коришћења обрнуте конкорданце с индексом ајтема у граматичкој анализи текста.

Према ономе што сам чула но овом скупу, лингвисти у Сарајеву су првенствено заинтересовани за компјутерску обраду књижевних текстова која може да послужи као основни материјал за даљу лингвистичку, лингвостилистичку и књижевно-стилистичку анализу. Чине се напори да се чак и стари црквенословенски текстови компјутерски обраде и тако учине доступнима за даља проучавања. О томе је говорила Херта Куна у реферату »Неки проблеми припреме текста за компјутерску конкорданцу босанских средњовјековних кодекса«. Њен реферат су са интересовањем пратили учесници скупа из редова математичара и инжењера, који су дали низ сугестија у вези са разрешавањем проблема које је она истакла.

1. 4. Реферати из Београда одликују се шароликошћу проблематике коју третирају и хетерогеношћу методологије коју примењују, а њихови аутори су махом научни посленици који припадају нелингвистичкој структури.

Инг. Томислав Томић и Петар Правица (Електротехнички факултет — Група за истраживање говора) у својим су рефератима — под насловом »Статистичка анализа српскохрватског текста помоћу рачунара« и »Акустичка анализа говора помоћу рачунара« — приказали она истраживања која у првом реду занимају електроничаре — нарочито у погледу разлагања и синтетизовања звучне компоненте људског говора — али чији резултати имају значаја и за лингвистичка истраживања. Посебно бих истакла настојања инг. Томића у статистичкој анализи српскохрватског језика, која се крећу у правцу израчунавања количине ентропије и информације у вербалном саопштењу.

Израчунавање ентропије помиње се и у реферату Првослава Плавшића, који је говорио о лексичкој и семантичкој анализи фреквенцијских речника ТВ-дневника и ТВ-драме. Међутим, како се из самог наслова реферата види, аутор се залаже за овакву анализу из много ширих побуда. Наиме, он се придружује мишљењу да је данас немогуће приступити изучавању језика без исцрпних података о дистрибуцији појединих величина и прегледа учесталости одређених језичких појава. Овакве статистике језика, помоћу компјутера, сматра аутор, погодне су за примену различитих лингвистичких теорија и суочавање са различитим моделима или методима приступа језичкој анализи.

1. 5. О статистичким испитивањима македонског језика помоћу компјутера обавестио је Драган Михајлов, са Математичког факултета у Скопју. На одабраним текстовима македонског

језика изведена су разна статистичка испитивања са циљем да се утврди статистичка структура језика. Испитивања су вршена на текстовима дневне штампе, средњошколским уџбеницима, прозе и поезије македонских писаца итд.

Научни посленици из Скопја представили су се и рефератом Олге Мишеске-Томић, са Филолошког факултета Универзитета «Кирил и Методиј« у Скопју. Она је у свом реферату дала преглед основних смерница већег броја пројеката у свету усмерених ка синтаксичко-семантичкој анализи, а у закључку је указала на кораке које би требало предузети да би »дубинска« анализа језика ушла у југословенске рачунске центре.

2. У свом излагању о оном што се догађа у југословенској компјутерској лингвистити указала сам само на оне моменте који, по мом мишљењу, имају значаја за ову научну област у нас, изостављајући при том све оно што сам сматрала маргиналним, или само добронамерним покушајем да се компјутерска лингвистика унапреди, али без озбиљнијих научних приступа тој проблематици.

Реферати и дискусија на овом скупу показали су да се у односу на формулисану тему скупа отишло дубље, а и сама тема је третирана у најширем смислу. Показало се и то да су заједнички проблеми свих наших истраживачких центара: стручно усавршавање лингвиста у компјутерској лингвистици, организовање припремне фазе обраде података и сачињавање парцијалних или интегралних лингвистичких образаца који би се могли успешно примењивати у програмима.

Размишљања која су изнета на овом скупу имају доста заједничких црта са размишљањима изнесеним у рефератима на XII међународном конгресу лингвиста који је одржан у Бечу крајем августа 1977. године, у секцији »Лингвистика и компјутер«. Стога бих се укратко осврнула на садржај тих реферата. Сматрам да ће на тај начин читаоци моћи боље да оцене напоре који се чине у нас у овој области, да констатују колико заостајемо у односу на достигнућа у свету и да сагледају ургентност истраживања у овој области, ако желимо да југословенска лингвистика, пре свега она коју називамо примењеном, задржи своје место у матици савремених лингвистичких токова.

Одмах бих констатовала да од 13 реферата у тој секцији 10 потичу из европских истраживачких центара, што сведочи о живом интересовању за ову проблематику на нашем континенту.

У већини реферата нуде се решења за што успешније и рационалније коришћење компјутера у лингвистичке сврхе и говори се о финалним продуктима добивеним применом одређених поступака. »Многи лингвисти би могли добро да употребе компјутер за своја истраживања, али мало њих се користе овом могућношћу«, каже се у једном реферату (James L. Wyatt, A. pattern finding program for the linguist, Department of Modern Languages, Florida State Universitu). Стога аутор нуди један такав образац који, по његовом мишљењу, може помоћи

лингвистима да изнађу своје програме. Тај образац израђен је у виду конкретног програма писаног рачунарским језиком, који се може укључити у више компјутерских система. У једном другом реферату се каже да је најактуелнији проблем компјутерске лингвистике изналажење њене теоријске основе (R. Piotrowski, Computational linguistics and teht theoru, Herzen Pedagogical Institute, Leningrad). Аутор сматра да се питања компјутерске лингвистике не могу решавати искључиво у оквиру генеративне граматике и да се анализа природних језика мора више ослањати на теорију текста, па у складу са том теоријом треба примењивати најприкладнију технологију. Говорећи о предностима свог обрасца, један други аутор израчунава чак и у процентима економичност у односу на улазне и излазне податке (Friedrich H. Lang, Die maschinelle Generierung von Begriffssystemen, IBM Österreich, Wien).

Неки аутори су реферисали о томе како примењују поједине лингвистичке моделе у компјутерској лингвистици, на пример, модел теме и реме (Bátori István, Die kommunikativ motivierten Kategorien der Sprache und ihre Rolle im Verstehprozess, Tübingen), или модел дубинских падежа (Udo L. Figge, Die Semantik in der linguistischen Datenverarbeitung, Romanisches Seminar der Ruhr-Universität Bochum).

Међу рефератима који дају информације о финалним продуктима компјутерске обраде лингвистичких података истакла бих, пре свега, реферат који говори о компјутерском архиву за синтаксу финског језика (Osmo Ikola, Computer-Archiv für finnische Syntax, Universität Turku). Тај архив обухвата у свом најважнијем делу кодирани истраживачки програм помоћу којег се електронским путем могу добити подаци за било које питање синтаксе. Сем тога, треба поменути и реферат о компјутерској обради политичких неологизама (A. Petroff, Traitment informatique des types de formation des neologismes politiques, Centre d'Etude de la Néologie Lexicale, Université Paris X), као и реферат у којем се презентују стохастички модели и квантитативни аспекти гласовних образаца у говорном италијанском језику, а резултат су истраживања која се одвијају у Центру за фонетику Ц. Н. Р. и у Институту за примењену математику Универзитета у Падови.

3. Очигледно је да су учесници тог међународног скупа, више него учесници скупа у Сарајеву, посветили пажњу теоријским питањима компјутерске лингвистике, што је сасвим разумљиво ако се узме у обзир да је скуп у Сарајеву био први скуп овакве врсте у Југославији. Утолико је значајније што су учесници овог скупа изразили жељу да организатору, заинтересованим научним установама и стручној јавности упуте следећа своја мишљења и сугестије:

1 Потребно је регистровати све активности у вези с компјутерском обрадом језичких података у нашој земљи, и то:

а) на техничком нивоу (на техничким факултетима и у рачунарским центрима)

б) на лингвистичком нивоу (у лингвистичким институтима).

2 Такође је потребно организовати систематско прикупљање података о ауторима, насловима и садржајима објављених дела из ове области (књига, чланака и бележака), као и необјављених радова (докторских дисертација, магистарских и дипломских радова).

3. Прикупљени подаци уносили би се сваке године у Библиографију, која би се објављивала у једном лингвистичком (Савремена лингвистика, Загреб) и једном техничко-информативном часопису (Informatica, Ljubljana).

4. Договором заинтересованих научних институција и организација потребно је, у року од три године, организовати други скуп о овој проблематици. Организатору првог скупа даје се мандат за координацију свих послова у вези са утврђивањем места, термина и организатора новога скупа.

5. Ради популаризације компјутерске обраде језичких података, Савезу друштава за примењену лингвистику Југославије треба предложити да се у републичким и покрајинским друштвима оснују секције за примењену рачунарску лингвистику. Руководиоци тих секција чине координациони одбор, који ће, под руководством организатора претходног скупа, иницирати и водити даље акције — укључујући и организацију следећег скупа.