Српски језик у дигиталном добу -- The Serbian Language in the Digital Age

Duško Vitas, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, Mladen Stanojević

META NET

White Paper Series Серија белих књига

THE SERBIAN LANGUAGE IN THE DIGITAL AGE

СРПСКИ ЈЕЗИК У ДИГИТАЛНОМ ДОБУ

Duško Vitas
Ljubomir Popović
Cvetana Krstev
Ivan Obradović
Gordana Pavlović-Lažetić
Mladen Stanojević

Springer

White Paper Series Серија белих књига

THE SERBIAN LANGUAGE IN THE DIGITAL AGE

СРПСКИ ЈЕЗИК У ДИГИТАЛНОМ ДОБУ

Duško Vitas University of Belgrade
Ljubomir Popović University of Belgrade
Cvetana Krstev University of Belgrade
Ivan Obradović University of Belgrade
Gordana Pavlović-Lažetić University of Belgrade
Mladen Stanojević University of Belgrade

Georg Rehm, Hans Uszkoreit
(уредници, editors)

ПРЕДГОВОР PREFACE

Ова бела књига је део серије која промовише знање о језичким технологијама и њиховим могућностима. Намењена је наставницима језика, новинарима, политичарима, језичким заједницама и другима. Покривеност језичким технологијама и начин њихове употребе се у Европи разликују од језика до језика. Због тога се разликују и активности које је потребно спровести да би се подржала истраживања и развој, а неопходни кораци зависе од многих фактора, као што су сложеност језика или величина заједнице која га користи. Пројекат META-NET, мрежа изврсности коју финансира Европска комисија, спровео је анализу текућих језичких ресурса и технологија. Анализа је била усмерена на 23 званична европска језика, као и на друге значајне националне и регионалне језике у Европи. Резултати анализе сугеришу постојање многих значајних празнина у истраживањима за сваки језик. Детаљнија експертска анализа и процена текуће ситуације за сваки језик помоћи ће да се повећа утицај нових истраживања и умање могући ризици. Према стању из новембра 2011, META-NET повезује 54 истраживачка центра из 33 земље (стр. 81), који сарађују са заинтересованим странама из сфера предузетништва, државних институција, привреде, истраживачких организација, софтверских компанија, понуђача технологија и европских универзитета. Они заједно граде технолошку визију кроз развој стратешких истраживачких програма који показују како ће примене језичких технологија попунити постојеће празнине у истраживањима до 2020. године.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 84). The analysis focuses on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of future research.

As of November 2011, META-NET consists of 54 research centres in 33 European countries (p. 81). META-NET is working with stakeholders from economy (software companies, technology providers and users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Захваљујемо се ауторима беле књиге о немачком језику [1] што су дозволили да језички независне делове њиховог текста користимо у овом раду.

Израду ове беле књиге финансирали су Седми оквирни програм (FP7) и Програм подршке политици информационо-комуникационих технологија Европске комисије преко уговора T4ME (Уговор о финансирању 249 119), CESAR (Уговор о финансирању 271 022), METANET4U (Уговор о финансирању 270 893) и META-NORD (Уговор о финансирању 270 899).

The authors of this document are grateful to the authors of the White Paper on German for permission to re-use selected language-independent materials from their document [1].

The development of this White Paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

САДРЖАЈ CONTENTS

СРПСКИ ЈЕЗИК У ДИГИТАЛНОМ ДОБУ

THE SERBIAN LANGUAGE IN THE DIGITAL AGE

РЕЗИМЕ

У последњих 60 година Европа је постала јединствена политичка и економска структура, мада је културно и језички веома разноврсна. То значи да је, од португалског до пољског, од италијанског до исландског, свакодневна комуникација становника Европе, као и комуникација у сфери пословања и политике, нужно суочена са језичким препрекама. Институције Европске уније троше око милијарду евра годишње на одржавање своје политике вишејезичности, тј. на превођење текстова и говорне комуникације. Питање које се поставља јесте да ли је толико оптерећење неопходно. Модерне језичке технологије и лингвистичка истраживања могу значајно да допринесу брисању језичких граница. Језичке технологије у комбинацији са интелигентним уређајима и апликацијама могу у будућности да помогну Европљанима да се међусобно споразумевају једноставно и лако и да обављају послове и кад не говоре истим језиком.

Језичке технологије граде мостове за европску будућност.

Главни трговински партнери Србије су земље Европске уније, са уделом од преко 50% у укупној трговинској размени, при чему је извоз из Србије на тржиште ЕУ ослобођен царине у складу са Споразумом о стабилизацији и придруживању. Али језичке препреке могу да зауставе пословање, посебно за СМП (средња и мала предузећа) која немају финансијских средстава да их превазиђу. Једина (незамислива) алтернатива за вишејезичну Европу била би да један језик почне да доминира и на крају замени све остале језике.

Један традиционални начин превазилажења језичких баријера јесте учење страних језика. Међутим, без технолошке подршке, савладавање 23 званична језика земаља чланица Европске уније и око 60 других европских језика, за становнике Европе представља непремостиву препреку, баш као и за њену економију, политичке дебате и научни напредак.

Решење лежи у изградњи кључних потпорних технологија. Оне ће европским актерима понудити огромне предности, не само у оквиру заједничког европског тржишта већ и у трговинским односима са трећим земљама, посебно са привредама које се брзо развијају. Да би се постигао тај циљ и очувала европска културна и језичка разноврсност, неопходно је да се прво спроведе систематска анализа језичких специфичности свих европских језика, као и текућег стања њихове опремљености језичким технологијама. На тај начин ће језичке технологије послужити као јединствени мост међу европским језицима.

Језичке технологије као решење за будућност.

Алати за аутоматско превођење и обраду говора који се могу наћи на тржишту још увек не омогућавају остварење овог амбициозног циља. Главни актери на овом пољу су пре свега приватна профитна предузећа из Северне Америке. Још крајем 1970-их Европска

унија је препознала суштински значај језичких технологија као покретача европског јединства, и почела је са финансирањем првих истраживачких пројеката као што је био пројекат EUROTRA. У исто време, започели су национални пројекти, који су дали вредне резултате, али нису покренули и заједничку усклађену европску акцију. Насупрот овим појединачним и неповезаним напорима у финансирању, друга вишејезична друштва, као што су Индија (22 званична језика) и Јужна Африка (11 званичних језика) [2], недавно су почела дугорочне националне програме језичких истраживања и технолошког развоја.

Данас се главни актери на подручју језичких технологија ослањају на непрецизне статистичке приступе, који не користе дубље лингвистичке методе и знања. На пример, реченице се аутоматски преводе тако што се нове реченице пореде са хиљадама претходно „ручно" преведених реченица. Квалитет резултата у великој мери зависи од квантитета и квалитета расположивог корпуса узорака. Мада машинско превођење једноставних реченица може да пружи употребљиве резултате у језицима са довољном количином расположивог текстуелног материјала, ове плитке статистичке методе нужно доживљавају неуспех у случају језика са мањим обимом узорака или у случају реченица комплексне структуре.

Европска унија је због тога одлучила да финансира пројекте као што су EuroMatrix и EuroMatrixPlus (од 2006) и iTranslate4 (од 2010) који спроводе основна и примењена истраживања и стварају ресурсе за успостављање језичкотехнолошких решења високог квалитета за све европске језике. Анализа дубљих структурних својстава језика је једини начин да се изграде апликације које дају добре резултате на целом распону европских језика.

Европска истраживања у овој области већ су постигла бројне успехе. На пример, преводилачки сервиси Европске уније користе софтвер отвореног кода за машинско превођење MOSES, који је претежно развијен кроз европске истраживачке пројекте. Суштински пробој у области синтезе и препознавања говора на српском језику начинила је група са Факултета техничких наука Универзитета у Новом Саду. Развијен је низ апликација у области TTS и ASR на бази говорних и лексичких база података акцентованих облика речи. Препознавање и генерисање говора за српски комерцијализовала је фирма AlfaNum која је потекла са Универзитета у Новом Саду. AlfaNum има значајан број корисника међу српским фирмама. С друге стране, први корпус савременог српског језика, електронски морфолошки речник, паралелни француско-српски и енглеско-српски корпуси литерарних текстова, као и различити софтверски алати развијени су у оквиру заједничких пројеката Математичког факултета и Одсека за српски језик Филолошког факултета у Београду.

Језичке технологије
помажу уједињењу Европе.

Према увиду у досадашње стање, сви су изгледи да ће „хибридне" језичке технологије које комбинују дубинску обраду са статистичким методама бити у могућности да премосте јаз између свих европских језика, и шире. Како показује ова серија белих књига, постоји драматична разлика у степену припремљености када су у питању језичка решења и стање истраживања међу европским језицима. Српски језик је један од „мањих" европских језика и потребна су даља истраживања која ће омогућити да ефикасна решења која нуде језичке технологије уђу у свакодневну употребу.

Дугорочни циљ META-NET-a јесте да уведе језичке технологије високог квалитета за све језике, како би се постигло политичко и економско јединство кроз

културну разноврсност. Те технологије ће помоћи да се уклоне постојеће баријере и да се изграде мостови међу европским језицима. Ово захтева од свих заинтересованих учесника – у политици, истраживању, привреди и друштву – да уједине своје напоре за будућност. Ова серија белих књига допуњује друге стратешке активности које предузима META-NET (видети преглед у додатку). Ажурне информације као што су текућа верзија текста META-NET визије [2] или стратешки истраживачки план рада (Strategic Research Agenda, SRA) могу се наћи на META-NET веб локацији: http://www.meta-net.eu.

2

ОПАСНОСТ ПО НАШЕ ЈЕЗИКЕ И ИЗАЗОВИ ПРЕД ЈЕЗИЧКИМ ТЕХНОЛОГИЈАМА

Сведоци смо дигиталне револуције која драматично утиче на комуникацију и друштво. Најновија достигнућа на подручју дигиталних информационих и комуникационих технологија могу да се пореде са Гутенберговим изумом штампарске пресе. Шта ова аналогија може да нам каже о будућности европског информационог друштва и посебно наших језика?

Дигитална револуција се може упоредити са Гутенберговим изумом штампарске пресе.

После Гутенберговог изума, стварне продоре у комуникацији и размени знања остварила су дела као што је био Лутеров превод Библије. Током наредних векова у култури су развијене технике за бољу обраду језика и размену знања:

- правописна и граматичка стандардизација већих језика омогућила је брзо преношење нових научних и интелектуалних идеја;
- развој званичних језика омогућио је становницима да међусобно комуницирају унутар (често политичких) граница;
- подучавање и превођење језика омогућило је комуникацију која превазилази језичке границе;
- стварање уредничких и библиографских упутстава обезбедило је квалитет и расположивост штампаног материјала;
- појавом различитих медија, као што су новине, радио, телевизија, књиге и други облици, задовољене су различите потребе за комуникацијом.

У последњих двадесет година информационе технологије помогле су да се аутоматизују и поједноставе бројни процеси:

- софтвер за стоно издаваштво је заменио писаћу машину и слагање текста;
- Microsoft PowerPoint је заменио графоскопске фолије;
- документа се шаљу и примају електронском поштом често брже него факс машином;
- Skype се користи за јефтино телефонирање преко интернета и организовање виртуелних састанака;
- формати аудио и видео записа олакшавају размену мултимедијалних садржаја;
- претраживачке машине обезбеђују приступ веб странама преко кључних речи;
- *онлајн* услуге као што је Google Translate производе брз и приближни превод;
- друштвени медији као што су Facebook, Twitter, и Google+ поједостављују комуникацију, сарадњу и размену информација.

Мада су сви ови алати и апликације од велике помоћи, они још нису довољни да остваре одрживо ви-

шејезично европско друштво у коме је свакоме могућ слободан приступ информацијама и роби.

2.1 ЈЕЗИЧКЕ ГРАНИЦЕ ПРЕДСТАВЉАЈУ СМЕТЊУ ЗА ЕВРОПСКО ИНФОРМАЦИОНО ДРУШТВО

Не можемо тачно да предвидимо како ће изгледати будуће информационо друштво. Али постоје велики изгледи да ће револуција у комуникационој технологији повезати на нове начине људе који говоре различитим језицима. То ствара притисак на појединце да уче нове језике и посебно на развојне тимове да стварају нове технолошке производе који ће обезбедити узајамно разумевање и приступ заједничком знању. У глобалном привредном и информационом простору, нови типови медија омогућавају бржу размену у којој учествују бројни језици, говорници и садржаји. Актуелна популарност друштвених медија као што је Википедија (Wikipedia), Фејсбук (Facebook), Твитер (Twitter), Јутјуб (YouTube) и, однедавно, Гугл+ (Google+) само је врх леденог брега.

Глобална економија и информациони простор суочавају нас са различитим језицима, говорницима и садржајима.

Данас можемо да допремимо гигабајте текста из целог света за свега неколико секунди, пре него што схватимо да је текст на језику који не разумемо. Према недавном извештају Европске комисије, 57% корисника интернета наручује робу и услуге на језицима који им нису матерњи. (Енглески је најчешћи страни језик а за њим следе француски, немачки и шпански.) 55% корисника чита садржаје на страном језику, док само 35% користи страни језик за писање електронских порука или коментара на мрежи [3]. До пре неколико година енглески језик је био lingua franca веба – огромна већина садржаја на вебу била је на енглеском језику. Ситуација је данас драстично промењена. Количина онлајн садржаја и на другим европским (као и азијским и средњоисточним) језицима доживела је праву експлозију.

Ова свеприсутна дигитална подељеност као последица језичких граница изазвала је изненађујуће мало пажње у јавности. Ипак, она поставља неодложно питање: „Који ће европски језици напредовати и опстати у умреженом друштву информација и знања, а који су осуђени да нестану?"

2.2 НАШИ ЈЕЗИЦИ СУ УГРОЖЕНИ

Проналазак штампарске пресе допринео је повећању обима размене информација у Европи, али је такође довео до изумирања многих европских језика. На регионалним језицима и језицима мањина ретко се штампало. Ово је довело до тога да су многи језици, као што су корнвалски или далматски, сведени на усмене облике преношења, што је довело до тога да су се све мање користили. Да ли ће интернет довести до истих последица када су наши језици у питању?

Разноврсност језика у Европи јесте једно од њених најдрагоценијих и најзначајнијих културних добара.

Око 80 језика Европе представља једно од њених најбогатијих и најважнијих културних добара и суштински чинилац њеног друштвеног модела [4]. Док ће језици као што су енглески или шпански вероватно преживети на дигиталном тржишту у настајању, многи би европски језици могли да постану не-

битни у умреженом друштву. Овакав развој ствари ослабио би позицију Европе у свету, а то би било и у супротности са стратешким циљем да се обезбеди подједнако учешће за сваког становника Европе без обзира на језик. Према УНЕСКО-вом извештају о вишејезичности, језици су суштински медијум за уживање основних људских права као што су изражавање политичких опредељења, образовање и учествовање у друштву [5].

2.3 ЈЕЗИЧКЕ ТЕХНОЛОГИЈЕ СУ КЉУЧНЕ ПОТПОРНЕ ТЕХНОЛОГИЈЕ

У прошлости, инвестиције у очување језика усмераване су на учење и превођење језика. Према неким проценама, европско тржиште превода, усменог превођења, локализације софтвера и глобализације страница на вебу 2008. год. је износило 8,4 милијарде евра, са очекиваним растом од 10% годишње [6]. Тај износ покрива само мањи део садашњих и будућих потреба у међујезичкој комуникацији. Најубедљивије решење које ће обезбедити дубину и ширину коришћења језика у Европи сутрашњице јесте коришћење одговарајуће технологије на исти начин на који се користе технологије за потребе транспорта, у енергетици или за особе са посебним потребама.
Дигиталне језичке технологије (којима је циљ да овладају свим облицима писаног и говорног језика) помажу људима да сарађују, послују, размењују знање и учествују у политичким и друштвеним дебатама без обзира на језичке баријере или њихове информатичке вештине. Технологије су често невидљиве у сложеним софтверским системима који нам помажу да:

- нађемо информацију помоћу претраживачких машина;
- користимо правописне и граматичке провере у програмима за обраду текста;
- разгледамо препоруке о производима у мрежним продавницама;
- саслушамо гласовна упутства у навигационим системима аутомобила;
- преводимо веб странице помоћу мрежних преводилаца.

Језичке технологије чине већи број основних апликација које омогућавају обраду језика у оквирима ширих програмских система. Сврха белих књига, састављених у оквиру META-NET-а, јесте да се опише степен развоја основних језичких технологија за сваки од европских језика.

> Европи су потребне робусне и приступачне језичке технологије за све европске језике.

Да би одржала свој положај у првим редовима светске иновативности, Европи ће бити потребне језичке технологије прилагођене свим европским језицима које ће бити робусне, свима приступачне и потпуно интегрисане у кључна софтверска решења. Без језичких технологија нећемо бити у стању да у скорој будућности остваримо одиста ефикасно интерактивно, мултимедијално и вишејезично корисничко искуство.

2.4 МОГУЋНОСТИ ЈЕЗИЧКИХ ТЕХНОЛОГИЈА

У свету штампане речи, брзо умножавање слике текста (стране) коришћењем штампарске пресе представљало је технолошки пробој. Људима је био препуштен тешки посао прегледања, читања, превођења и апстраховања знања. Требало је да дочекамо Едисона да бисмо снимили говорни језик, при чему је његова технологија производила само аналогне копије.

Дигиталне језичке технологије данас могу да аутоматизују сам процес превођења, генерисање садржаја и управљање знањем за све европске језике. С њима је могуће опремити кориснику блиске, текстуалне или говорне, сумеће (интерфејс) за кућне електричне апарате, машине, возила, рачунаре и роботе. Иако достигнућа истраживања и развоја омогућавају да се наслуте широке могућности, комерцијалне и индустријске примене су још у раној фази развоја. На пример, за многе европске језике машинско превођење достиже задовољавајући ниво тачности у оквиру специфичних домена, а експерименталне апликације обезбеђују вишејезичне информације, управљање знањем и генерисање садржаја.

Језичке технологије помажу да се превазиђе „хендикеп" језичке разноликости.

Као што је то случај са већином технологија, прве језичке апликације, као што су гласовне корисничке сумеће и дијалошки системи, развијене су за уско специјализоване домене, а њихова употребљивост је била често ограничена. Међутим, у образовној индустрији и индустрији забаве леже огромне тржишне могућности за интеграцију језичких технологија у игре, странице везане за културно наслеђе, производе за образовање кроз забаву, библиотеке, симулациона окружења или програме обуке. Мобилне информационе услуге, софтвер за рачунарски потпомогнуто учење језика, окружења за електронско учење, алати за самооцењивање и откривање плагијата само су неки од примера где језичке технологије могу да одиграју важну улогу. Популарност друштвених мрежа као што су Твитер (Twitter) и Фејсбук (Facebook) указује да постоје потребе за језичким технологијама које омогућавају надгледање поште, резимирање дискусија, детекцију трендова у испољеним ставовима, препознавање емотивних реакција, идентификацију повреда ауторских права или праћење злоупотреба.

Језичке технологије представљају огромну прилику за Европску унију. Оне могу да помогну у решавању комплексног питања вишејезичности у Европи – чињенице да различити језици природно коегзистирају у европском пословању, организацијама и школама. Али грађани желе да комуницирају изван језичких граница које још увек постоје на јединственом европском тржишту. Језичке технологије могу да помогну у превазилажењу ове последње препреке својом подршком слободном и отвореном коришћењу појединачних језика. Ако гледамо и корак даље, иновативне европске вишејезичне технологије представљаће узор за наше партнере по свету када они буду почели да обезбеђују ове технологије за своје вишејезичне заједнице. Језичке технологије могу се посматрати као облик „потпорних" технологија које помажу да се превазиђе „хендикеп" језичке разноликости и да језичке заједнице постану блискије. Најзад, једно активно истраживачко подручје јесте коришћење језичких технологија у операцијама спасавања у областима погођеним катастрофама, где успешно деловање може одлучивати о животу или смрти: будући интелигентни роботи са вишејезичним способностима у могућности су да спасу животе.

2.5 ИЗАЗОВИ ПРЕД ЈЕЗИЧКИМ ТЕХНОЛОГИЈАМА

Мада су језичке технологије последњих неколико година остварили значајан напредак, текући темпо технолошког напретка и иновације производа је сувише спор. Језичке технологије које су у широкој употреби, као што су граматичке и правописне провере, по правилу су једнојезичне, и постоје само за мали број језика. Онлајн услуге машинског превођења, мада су корисне да се брзо произведе прихватљи-

ва апроксимација садржаја документа, стварају пуно потешкоћа када је потребно да се произведу високо прецизни и потпуни преводи. Због комплексности природног језика, његово софтверско моделирање и тестирање у реалном свету је дуг и скуп посао, који захтева дугорочно финансирање. Европа мора да одржи своју пионирску улогу у суочавању са технолошким изазовима вишејезичне заједнице изналажењем нових метода којима ће убрзати развој на целој својој територији. У њих би спадале и иновације у области рачунарства и технике које користе потенцијале великог броја учесника (crowdsourcing).

Текући темпо технолошког напретка сувише је спор.

2.6 УСВАЈАЊЕ ЈЕЗИКА КОД ЉУДИ И МАШИНА

Да бисмо илустровали начин на који рачунари поступају са језиком и зашто их је тако тешко испрограмирати да употребљавају језик, бацимо поглед на начин како људи усвајају матерњи и страни језик, а онда погледајмо и како ради језичкотехнолошки систем.

Људи стичу језичке вештине на два различита начина. Бебе уче језик слушајући разговоре између родитеља, браће и сестара и других чланова породице. У узрасту од приближно две године деца изговарају своје прве речи или кратке фразе. То је могуће само захваљујући посебној генетској предиспозицији људи да опонашају, а потом и осмисле оно што чују.

Учење страног језика у старијем узрасту захтева више напора, углавном зато што дете не припада језичкој заједници оних којима је тај језик матерњи. У школи страни језици се обично усвајају учењем граматичких структура, речника и правописа кроз вежбања која описују језичко знање преко апстрактних правила, табела и примера. Учење страног језика са годинама постаје све теже.

Два главна типа система језичких технологија „усвајају“ језичке способности на сличан начин као људи. Статистички приступи (или „приступи вођени подацима“) стичу језичко знање из огромних колекција конкретних примера текстова. За обучавање система за проверу правописа, на пример, довољно је коришћење текстова на једном језику, али су за обучавање машинских преводилаца потребни тзв. паралелни текстови на два (или више) језика. Алгоритам машинског учења затим „учи“ обрасце превођења речи, кратких фраза и комплетних реченица.

Статистички приступи могу да захтевају милионе реченица јер квалитет резултата расте са порастом броја анализираних текстова. То је један од разлога што добављачи претраживачких машина жељно прикупљају што је могуће више писаног материјала. Исправка правописних грешака у програмима за обраду текста и сервиси као што су Google Search и Google Translate, ослањају се на статистички приступ. Велика предност статистике је у томе што машина учи брзо понављајући циклусе обуке, мада квалитет може да варира на непредвидљив начин.

Други приступ језичким технологијама, а посебно машинском превођењу, јесте изградња система заснованих на правилима. Експерти из лингвистике, рачунарске лингвистике и рачунарства морају најпре да граматичку анализу изразе кроз систем правила и да саставе листе речи (лексиконе). То је посао који захтева много времена и велики труд. Неки од водећих система машинског превођења заснованих на правилима у сталном су развоју већ више од двадесет година. Предност система заснованих на правилима је у томе што експерти могу детаљније да контро-

лишу обраду језика. То омогућује да се грешке у софтверу систематски поправљају, а кориснику пруже детаљне повратне информације, пре свега када се такви системи користе за учење језика. Због великих трошкова, језичке технологије засноване на правилима до сада су биле развијане само за велике језике.

Људи стичу језичке вештине на два различита начина: учењем примера и учењем језичких правила.

Како предности и слабости статистичких система и система заснованих на правилима теже да се допуњују, текућа истраживања усмерена су на хибридне приступе који комбинују те две методологије. Па ипак, ови приступи су до сада били мање успешни у индустријским применама него у лабораторији.

Као што смо видели у овом одељку, многе апликације које су у широкој употреби у данашњем информационом друштву ослањају се у великој мери на језичке технологије. С обзиром на вишејезичност европске заједнице, ово се посебно односи на њен привредни и информациони простор. Мада су језичке технологије значајно напредовале последњих неколико година, још увек постоје огромне могућности за побољшање квалитета језичкотехнолошких система. У следећем одељку описаћемо улогу српског језика у европском информационом друштву и даћемо оцену текућег стања језичких технологија за српски језик.

3

СРПСКИ ЈЕЗИК У ЕВРОПСКОМ ИНФОРМАЦИОНОМ ДРУШТВУ

3.1 ОПШТИ ПОДАЦИ

Српски стандардни језик је национални стандардни језик Срба и званични језик у Републици Србији. Формиран је на основици млађих екавских и ијекавских штокавских јужнословенских дијалеката у форми коју му је одредио реформатор писаног језика код Срба Вук Караџић (1787–1864), који је истовремено реформисао и ћирилички алфабет и правопис. У 20. веку, у заједничкој држави Југославији тај језик је обухваћен називом српскохрватски који је имплицирао језичко заједништво са Хрватима (касније и другим народима чији је стандардни језик базиран на штокавским дијалектима). У последњој деценији 20. века уместо назива српскохрватски у Србији је у општој употреби назив српски језик. Устав Републике Србије из 2006. прописује: „Српски језик и ћирилично писмо биће у званичној употреби у Републици Србији" [7].

Према попису становништва из 2002, Србија има 7 498 001 становник [8], а српски је матерњи језик за 88,3% становништва [9]. Томе треба додати и становништво српске националности у другим крајевима бивше Југославије (чији број није лако одредити). Српска дијаспора, већином настала одласком на рад у иностранство и исељавањем због економских разлога, живи пре свега у појединим земљама централне и западне Европе, у САД, Канади и Аустралији (степен знања српског језика највише је условљен тиме о којој се генерацији исељеника ради). Према попису из 2002, већина Срба ван земље живи у Немачкој (102 799), затим у Аустрији (87 844) и Швајцарској (65 751).

> Стандардни српски језик је стандардни национални језик Срба и званични језик у Републици Србији.

Србија је вишејезична заједница. Према попису из 2002, националне мањине [10] су Мађари (3,91%), Бошњаци (2,1%), Роми (1,44%), Хрвати (0,94%), Црногорци (0,92%), Албанци (0,82%), Словаци (0,79%), Југословени (1,08%) као и друге мањине (Ашкалије, Бугари, Буњевци, Цинцари, Чеси, Горанци, Јевреји, Македонци, Немци, Муслимани, Румуни, Русини, Словенци, Турци, Украјинци и Власи, 2,45%). Структура мањинског становништва према језику је следећа: 3,8% мађарски, 1,8% бошњачки, 1,1% ромски, 0,8% албански, 0,8% словачки, 0,7% влашки, 0,5% румунски, 0,4% хрватски, 0,2% бугарски и 0,2% македонски. Остале језике говори 0,5% становника, док за 0,8% становника ови подаци нису познати. За неке мањинске језике у Србији постоји основно и средње образовање, и то за албански (55 основних/4 средње школе), мађарски (108/38), бугарски(26/-), румунски (27/2), русински (3/2), словачки (15/2), хрватски (7/1) [11]. Настава је праћена и издавањем уџбеника и лектире (нпр. у 2005. издато је укупно 526 уџбеника за основну и 283 за средњу школу) [9].

Службена употреба језика мањина је уређена законом о службеној употреби језика и писама [12], који обезбеђује да се закони и прописи објављују на језицима националних мањина. Ово укључује право обраћања републичким органима на свом језику и право да се добије одговор на том језику (у зависности од величине мањинске заједнице).

Превођење на српски или са српског је значајна активност. Током 2010. године је преведено 2549 дела (са енглеског 1438, са француског 215, са немачког 170, са италијанског 191, са шпанског 74, са мађарског 149). Део превода је са словенских језика (са руског 225, са чешког 4, са пољског 13, са словачког 21, са словеначког 19, са македонског 18, са бугарског 12). Што се тиче превода са српског на друге језике, у Србији је током 2010. објављен 591 наслов.

3.2 СПЕЦИФИЧНОСТИ СРПСКОГ ЈЕЗИКА

Српски језик има своје специфичности, које чине његову рачунарску обраду комплексним задатком.

3.2.1 Фонетика, фонологија, морфофонологија

Вокалски систем је једноставан (5 вокала), а консонантски релативно комплексан (25 консонаната). Вибрант *р* се у одређеним позицијама изговара као вокал и функционише као носилац слога (силабем), нпр. у речима *р̄рст̄* или *врст̄а*. У промени речи и творби речи постоји велики број фонемских алтернација (консонантских, вокалских и комбинованих) које се у неким случајевима комбинују на такав начин да два облика једне речи могу бити веома удаљена, нпр. номинатив сингулара именице „мисао“ је *мисао*, а инструментал сингулара *мишљу* (алтернације *а/ø, о/л, л+ј/љ/ с/ш*).

Акценатски систем од 4 акцента заснован је на два укрштена параметра: опозиција по дужини (кратки : дуги) и по тону (силазни : узлазни). Дистрибуција узлазних и силазних акцената је регулисана посебним правилима. У промени и у творби речи честе су акценатске алтернације. Пошто се акценатски знаци не бележе, у писаном тексту се јављају хомографи. На пример, значење речи *лук* се разликује према томе да ли је акценат краткосилазни или дугосилазни.

У доста речи и граматичких облика кодификована норма предвиђа изговор постакценатских дужина, али се оне у узусу све мање изговарају.

Скоро све речи су наглашене, али постоје и клитике: проклитике (већина везника и предлога, као и негација уз глагол) и енклитике (ненаглашени облици заменица и глагола и упитна партикула *ли*).

Изговор позајмљеница је фонетски прилагођен српском језику. Комбинације фонема (пре свега консонаната) у позајмљеницама често одступају од група које су типичне за изворне штокавске речи, као у примерима *софт̄вер*, *хардвер*, *инт̄ерфејс*. Има такође, нарочито у свакодневном узусу, одступања и од нормативне дистрибуције акцената.

Код једног броја лексема и облика постоје две варијанте изговора – екавска и ијекавска – етимолошки везане за некадашњи вокал звани *јат̄*, као што је показано у табели 1.

3.2.2 Морфологија

Постоји десет врста речи, са великим бројем подврста. Посебно су комплексни системи заменица и бројева. Не постоји члан.

Именице имају род као класификациону категорију (мушки, женски или средњи). Од значаја је и класификација према семантичком роду (мушки или женски). На пример, именица *ī̄азда* се мења као именица женског рода, али означава мушку особу.

		екавски	ијекавски
„цвет"	**сингулар**	*цвет* (дуго **е**)	*цвијет*
	плурал	*цветови* (кратко **е**)	*цвјетови*

1: Екавска и ијекавска варијанта изговора

Глаголи имају вид као класификациону категорију (свршени или несвршени). Известан број глагола има оба вида. Постоји више врста такозваних рефлексивних глагола.

Постоје три типа флексије: (а) деклинација (по броју и падежу код именица (видети табелу 2), по роду, броју, падежу и придевском виду код придева); (б) конјугација (веома комплексна); и (в) компарација (код градабилних придева и прилога). Све промене имају мањи или већи број ужих типова, као и известан број изузетака. Свуда постоје бројне фонемске и акценатске алтернације. Посебно треба истаћи велики број подударних облика, тј. облички синкретизам (морфолошку хомонимију). Последица флексије је да речнику од 120.000 лема одговара око 4,5 милиона флективних граматичких облика (ипак нема толико формалних речи јер су неки облици у појединим парадигмама истоветни).

Личне заменице (укључујући и рефлексивну заменицу) и помоћни, копулативни и егзистенцијални глагол „јесам" и помоћни глаголи „бити" и „хтети" имају и енклитичке облике, који се чешће користе од одговарајућих наглашених облика. На пример, датив једнине мушког и средњег рода личне заменице трећег лица гласи: *њему* (акцентовани облик) и *му* (енклитички облик).

Код именица, глагола и придева постоји веома развијена суфиксална творба речи. Код глагола је веома развијена и префиксација (добрим делом повезана и са аспекатским значењима). Композиција, у целини гледано, мање је развијена.

Постоји пуристички однос према калковима и кованицама, као и према тзв. есоцентричним именичким сложеницама, као нечем што не спада у аутентичну штокавску творбу. Овакав однос отежава лексичку и терминолошку елаборацију коришћењем творбе речи и један је од разлога веома великог броја позајмљеница.

Позајмљенице се већином уклапају у постојеће морфолошке и творбене типове, али од тога има одступања. На пример, неке стране речи се не мењају, као што су именице *Мери* и *скво* или придеви *фер* или *браон*.

Развијена творба речи (суфиксација, префиксација, у мањој мери композиција и разни комбиновани творбени начини) чине да се највећи број лексема може груписати у творбене породице односно лексикографска гнезда. Ту је посебно важно да један део творбених веза доводи до систематске (категоријалне) модификације значења основне речи, што знатно олакшава лексикографску обраду таквих случајева. На пример, за реч „глумац" твори се деминутив „глумчић" и аугментатив „глумчина", женски облик „глумица" и придеви „глумчев", „глумичин", „глумачки", итд.

Позајмљенице су у принципу фонолошки и морфолошки адаптиране, тј. прилагођене изговору и морфологији српског језика. И од њих се образују творбене породице.

	сингулар	паукал (2-4)	плурал
„прозор" (м. род)	*прозор*	*прозора*	*прозори*
„јаје" (с. род)	*јаје*	*јајета*	*јаја*
„жена" (ж. род)	*жена*	*жене*	
„вест" (ж. род)	*вест*	*вести*	

2: 4 типа именичке флексије

3.2.3 Лексика, фразеологија, терминологија, ономастика

Састав лексике одражава, с једне стране, штокавску основицу, и то не само у погледу оригиналног инвентара него и у погледу нових речи творених према новоштокавским творбеним моделима. С друге стране, фонд лексема одражава и језичку и културну историју српског народа, укључујући позајмљенице из црквенословенског, турског („мегдан"), руског („запета"), немачког („штрудла"), француског („руж") и, поготову у данашње време, енглеског („паркинг"). Томе треба додати, поготову у стручним терминологијама, интернационализме засноване на класичним језицима (грчком и латинском).

У области фразеологије посебно треба споменути идиоматске изразе, сликовита поређења, изреке и сл. који одражавају аутохтону имагинацију и језичку креативност. С друге стране, велики број лексикализованих израза је настао и настаје и даље калкирањем страних израза, данас пре свега енглеских.

Терминологија (и номенклатура) добрим делом се ослањала и ослања се и даље на поједине стране терминологије, путем превођења или позајмљивања (нарочито кад су у питању терминолошки интернационализми). Напори да се нађу изворна српска решења или да се постојећи термини посрбе имају одређене резултате, али не могу да иду у корак са све већим потребама у области терминологије и номенклатуре.

Ономастика представља важан део вокабулара српског језика, утолико више што се и овде стварају творбене породице речи.

3.2.4 Синтакса, лингвистика текста

Што се тиче распореда реченичних конституената (субјекта, предиката, објекта итд.), српски језик спада у тзв. SVO језике са слободним редом речи (тачније речено: са слободним распоређивањем реченичних конституената). То значи да су у принципу све пермутације реченичних конституената дозвољене, а да је преферентни распоред: субјекат – предикат – објекат. Међутим, слободан не значи и анархичан; напротив, избор конкретног распореда је регулисан комбинацијама различитих синтаксичких, семантичких, прагматичких и стилских фактора, тј. ма колико разноврсни, распореди чине један веома комплексан функционални систем. Погледајмо реченицу на енглеском:

- *Mary gave John an apple. [Марија даде Јовану јабуку.]*

У српском се ова ситуација може изразити на 24 = 4! = 1*2*3*4 (број пермутација од четири речи) различитих начина:

- *Марија даде Јовану јабуку.*
- *Марија даде јабуку Јовану.*
- *Марија Јовану даде јабуку.*

	сингулар	паукал	плурал
Номинатив	*прозор*	*прозора*	*прозори*
Генитив	*прозора*		*прозора*
Датив	*прозору*		*прозорима*
Акузатив	*прозор*	*прозора*	*прозоре*
Вокатив	*прозоре*	*прозора*	*прозори*
Инструментал	*прозором*		*прозорима*
Локатив	*прозору*		*прозорима*

3: Пример именичке деклинације

- *Марија јабуку даде Јовану.*
- *Јовану даде Марија јабуку.*
- *Јовану Марија даде јабуку.*
- *Јабуку Марија даде Јовану.*
- *Јабуку Јовану даде Марија.*
- *Даде Марија јабуку Јовану.*
- *Даде Јовану јабуку Марија,* итд.

Поједини конституенти се исказују и енклитикама, које се распоређују на сасвим специфичан начин. Заменички субјекат се не мора исказати, него се може само подразумевати (тзв. нулти субјекат), као у примеру ***Ја*** *се зовем Марко* према *Зовем се Марко*. Значајан број реченичних образаца је формиран са разним типовима семантичких субјеката. Поред актива и пасива, постоји и специјалан начин формулисања реченице са неспецификованим хуманим агенсом, коришћењем облика повратног глагола. Негација се примењује и на глагол и на заменички конституент (тзв. двострука негација), нпр. *Овде* ***не*** *йознајем* ***никоī***. У српском постоји седам падежа: номинатив, генитив, датив, акузатив, вокатив, инструментал и локатив (видети табелу 3).

У српском језику постоји пет зависних падежа, који се сви комбинују и са предлозима. Сви ти падежи и предлошко-падежне комбинације су полисемични. И обрнуто, исто значење се у неким случајевима може исказати различитим падежима односно предлошко-падежним конструкцијама (падежна синонимија). Постоји такође и један број израза који имају функцију предлога, на пример, *йриликом* (+генитив).

У српском језику постоји развијен систем личних глаголских облика за исказивање временских и модалних значења (аспекат је класификациона категорија); сви ти облици су полисемични. Једна од специфичности глаголског система је да конструкција *да* + презент све више истискује инфинитив.

Конгруенција у роду, броју, падежу и лицу је један од битних аспеката синтаксе српског језика, а значајна је и за успостављање текстуалних веза. Категоризација контролора конгруенције (нарочито појединих типова именица, конструкција са бројевима и координативних израза), као и начини на које се та контрола испољава у разним конгруентним позицијама представља изузетно комплексно подручје.

Већина типова зависних реченица (нарочито односне, временске, условне и узрочне) имају више формалних и семантичких подтипова.

Код координативних реченица посебно је комплексан инвентар везника за копулативне и за адверзативне односе.

Везе међу исказима у тексту се успостављају текстуалним координаторима и текстуалним конекторима разних врста. Избор распореда реченичних конституената важан је за информативну кохеренцију и

прогресију, с једне, а за емфазу и истицање, с друге стране. Тзв. нулти субјекат и енклитички заменички облици су важна средства за контекстуализацију реченица.

3.2.5 Правопис

Традиционални српски алфабет је ћирилица, коју чини 30 графема. Данас се користи – све више – и латиница. Она такође има 30 графема (три од њих су диграми), који су у бијективном односу са ћириличким графемама. Међутим, званично писмо је само ћирилица (видети табелу 4). Што се тиче графије (односа графемског и фонемског система), графеме и фонеме стоје у бијективном односу.

На нивоу кодних схема, латинични диграфи *lj*, *nj*, *dž* могу бити кодирани као лигатуре или као диграфи. У првом случају, *Unicode* [13] обезбеђује, на пример, посебно кодове за лигатуре LJ, Lj и lj који су у случају диграфа представљени као комбинација два ASCII кода, нпр. L и J. Ово води у проблеме са транслитерацијом која се, у општем случају, може извршити аутоматски. На пример, сваки чланак на српској Википедији се може приказати и ћириличним и латиничним писмом.

Азбука у српском не предвиђа употребу латиничних карактера q, x, y, w нити латиничних карактера за записивање римских бројева, што може да доведе до деградације информације приликом транслитерације из латинице у ћирилицу. Тако, на пример, *www* може постати *њњњ*, а латинично *Petar II* може постати *Петар ИИ* уместо *Петар II*.

Оба алфабета се користе у савременој издавачкој продукцији. Према подацима из Народне библиотеке Србије, током 2010. објављене су укупно 12574 књиге. Од тог броја, 6459 је на ћирилици, 6050 на латиници, а 65 на другим алфабетима. Међу дневним листовима са широким кругом читалаца, Политика и Вечерње новости излазе на ћирилици, док је већина других листова (Блиц, Курир, Данас, итд.) на латиници.

Правопис је (квази)фонемског типа: са малим изузецима, реч се пише онако како се изговара (правило: „*Пиши као што говориш!*"), тачније речено, према свом фонемском саставу. Интерпункција је логичког, а не граматичког типа (слична француској и енглеској). Према правопису, стране речи се и ћирилицом и латиницом пишу онако како се изговарају, тј. ранскрибовано. И страна имена се такође транскрибују (нпр. уместо *Shakespeare* пише се, и изговара, *Шекспир* и *Šekspir*).

3.2.6 Српски и други стандардни језици штокавског порекла

Заједничка штокавска основица, међусобни утицаји и коегзистенција у оквиру исте државе и – концептуално – у оквиру заједничког српскохрватског језика чине да за рачунарску обраду других језика штокавске провенијенције (хрватског, бошњачког, црногорског) треба разрешити сличне проблеме. То отвара велике могућности за синергију или бар за продуктивну сарадњу, као и за рационалан и економичан приступ решавању заједничких проблема. Томе доприноси и постојање знатних језичких ресурса за некадашњи заједнички српскохрватски језик (граматике и речници), у којима, истина, није поклањана дужна пажња диференцијацијама унутар штокавског стандарднојезичког простора. У ствари, овде се не ради о превођењу текстова с једног страног језика на други, него о *адаптирању* текстова састављених на језицима са истом дијалекатском основицом и са тесно повезаним развојем.

Стандардни језици штокавског порекла морају да реше сличне проблеме. То отвара велике могућности за продуктивну сарадњу.

ћирилица	А	Б	В	Г	Д	**Ђ**	Е	**Ж**	З	И	Ј	К	Л	**Љ**	М
	а	б	в	г	д	**ђ**	е	**ж**	з	и	ј	к	л	**љ**	м
латиница	A	B	V	G	D	**Đ**	E	**Ž**	Z	I	J	K	L	**Lj**	M
	a	b	v	g	d	**đ**	e	**ž**	z	i	j	k	l	**lj**	m
ћирилица	Н	**Њ**	О	П	Р	С	Т	**Ћ**	У	Ф	Х	Ц	**Ч**	**Џ**	**Ш**
	н	**њ**	о	п	р	с	т	**ћ**	у	ф	х	ц	**ч**	**џ**	**ш**
латиница	N	**Nj**	O	P	R	S	T	**Ć**	U	F	H	C	**Č**	**Dž**	**Š**
	n	**nj**	o	p	r	s	t	**ć**	u	f	h	c	**č**	**dž**	**š**

4: Српска слова

Главни проблеми се, у ствари, тичу појава везаних за елаборацију штокавског језгра и, посебно, за терминологију.

3.3 САВРЕМЕНИ РАЗВОЈ

Промене крајем двадесетог и почетком двадесет првог века обухватају следеће:

- Уместо заједничког српскохрватског стандардног језика, сада званично постоје четири национална стандардна језика. Конкретно, у Србији је сада званични језик српски, а не више српскохрватски. Због недавних сеоба изазваних ратним збивањима делимично је промењена дијалекатска слика у Хрватској и Босни и Херцеговини (у подручјима захваћеним ратним збивањима).
- Уочавају се све веће промене у лексици и фразеологији и у терминологији, везане за политичке, друштвене и економске промене у Србији и отварање према свету, али и за усклађивање законодавства, стандарда и терминологије са законодавством, стандардима и терминологијом који важе у Европској унији. Посебно се уочава утицај енглеског језика, и то не само због културолошких и економских момената који важе и за друге европске земље него и зато што се за усклађивање са Европском унијом као изворници узимају текстови/верзије на енглеском језику.
- Латиница се све више употребљава (сем у званичним текстовима).
- Текстови на српском језику се све више реализују у дигиталном облику (употреба рачунара, електронско издаваштво, интернет, SMS-поруке).

3.4 НЕГОВАЊЕ ЈЕЗИКА У СРБИЈИ

3.4.1 Рад на нормирању и неговању језика

Овде се може навести следеће:

- Године 1997. створено је међуакадемијско и међууниверзитетско тело под називом *Одбор за стандардизацију српског језика* [14], у коме су пред-

ставници одговарајућих институција из Србије, Црне Горе и Републике Српске (у БиХ).

- Уместо раније опште српскохрватске норме, сада се спецификује норма српског језика.
- Нема пуризма у односу на кроатизме (речи преузете из хрватског).
- Израђен је правопис српског језика.
- Подржава се употреба ћирилице, која се сматра угроженом све већом употребом латинице, нарочито код млађих генерација.
- Наставни програми и уџбеници у основној и средњој школи усклађени су са новом стандарднојезичком ситуацијом.

Стандардизација српског језика институционализована је кроз Одбор за стандардизацију српског језика, међуакадемијско и међууниверзитетско тело.

3.4.2 Осавремењивање норме

Одбор за стандардизацију српског језика је организовао израду серије описно-нормативних монографија које треба да прикажу савремено стање језика и понуде нормативна решења (досада су обрађене: творба речи, синтакса и фонологија). Донет је већи број нормативних препорука. Званични правопис је два пута осавремењиван.

3.4.3 Неговање језичког узуса

Одбор за стандардизацију српског језика (својим препорукама), Друштво за српски језик и књижевност (публикацијама и организовањем такмичења из српског језика и језичке културе за ученике основних и средњих школа), Матица српска (организовањем рада на изради правописа, својим публикацијама и организовањем саветовања о језику), Вукова задужбина (својим публикацијама и организовањем трибина и саветовања о језику) и разне друге институције, поједине издавачке куће и редакције дневних листова и редакције радио и ТВ програма, као и језички стручњаци и љубитељи матерњег језика труде се да дају свој допринос чувању правилности и чистоте српског језика у писаној и усменој употреби.

3.4.4 Одговор на све већи утицај енглеског језика

Истиче се потреба за замењивањем енглеских речи и израза српским, као и за замењивањем калкираних преведеница са енглеског (аутентичним) српским речима и изразима. (Шире узев, овде спада и отпор све већој употреби латинице.)

3.4.5 Побољшање стања у области лексикографије

Поклања се све већа пажња лексикографији, једнојезичној и двојезичној. Израђен је велики једнотомни речник савременог српског језика, за којим се осећала велика потреба. Модернизује се рад на изради великог академијског речника српског језика. Преводе се закони и прописи који важе у Европској унији [15], као и међународни стандарди [16], укључујући терминолошке стандарде.

3.5 ЈЕЗИК И ОБРАЗОВАЊЕ

Предмет *Срйски језик и књижевносш* је један од битних предмета у основној и средњој школи. Међутим, настава је концентрисана на правилно писање и говор, знање о језику (о граматици и лексици), знање о историји књижевних (писаних) језика код Срба и о постанку српског стандардног језика. На оваквој настави су базирана и масовна такмичења из матерњег језика (почев од виших разреда основне школе). Недовољно пажње се поклања практичној употреби

језика и функционалној писмености. Жеља да се настава по својим циљевима и стандардима приближи настави у Европској унији, као и незадовољавајући успеси ученика на PISA тестирању представљају подстицаје за модернизацију наставе језика и за инсистирање на функционалној писмености и комуникационим способностима. То се одражава и на текућу реформу школства (циљеви наставе језика, стандарди постигнућа, силабуси), као и на побољшање квалитета уџбеника. На факултетима углавном недостају курсеви из српског језика који би систематски оспособљавали будуће стручњаке за успешну професионалну комуникацију и одговарајућу функционалну писменост. Примена језичких технологија свакако може допринети модернизовању наставе, нпр. применом система за рачунарски потпомогнуто учење језика (CALL).

3.6 МЕЂУНАРОДНИ АСПЕКТИ

Званична употреба и настава српског језика у државама у којима живе делови српског народа регулисана је законодавством тих држава. Нестанак заједничког српскохрватског језика и званично постојање посебних језика штокавске провенијенције одразило се на организацију наставе некадашњег српскохрватског језика на иностраним универзитетима, као и на називе факултетских одсека на којима се држала та настава: сада за те језике, дакле и за српски језик (и књижевност), постоје посебни програми и дипломе, са већим или мањим комбиновањем предмета, а одсеци имају збирне називе. У Србији се наставља пракса организовања летњих школа за странце, али сада за српски, а не српскохрватски језик. Такође се шаљу домаћи наставници да раде као лектори на катедрама у иностранству. За децу српског порекла организује се у појединим земљама додатна настава из матерњег језика. Потреба усклађивања законодавства и терминологије са оним у Европској унији, утицај англо-америчке културе у области забаве и медија и ефекти глобализације све јаче доводе српски језик у везу са другим језицима, нарочито енглеским, и дају преводилаштву све већи подстицај и значај.

3.7 СРПСКИ ЈЕЗИК НА ИНТЕРНЕТУ

Анкета [17] извршена 2010. године говори да 50,8% становништва редовно користи рачунар и интернет, а 43,7% становништва никада није користило рачунар. Према другом извору [18], чак 55,9% популације користи интернет, при чему је стопа раста 926,8% у периоду 2000–2010. Према истом извору, у Србији на дан 31. августа 2010. било је 2 237 680 корисника Фејсбука, што представља 30,5% укупне популације. Електронске услуге јавне администрације (e-government) користи свега 13,2% становништва, док 38,5% не би никада користило ове услуге. Трговину преко интернета користило је свега 13% становништва. Према Републичком заводу за статистику Републике Србије [19], коришћење информационо-комуникационе опреме показује раст.

Према истом извору, 96,8% фирми користило је интернет 2010. године у поређењу са 90,2% у 2006, док је 67,5% фирми имало своју веб локацију 2010. године у поређењу са 52,9% у 2006. У 2010. години 70,6% користило је услуге јавне администрације.

Подаци Републичког завода за статистику из истраживања из 2010. на узорку од 2.400 домаћинстава и исто толико појединаца старости од 16 до 74 године показују да интернет прикључак има 39 одсто анкетираних, највише у Београду – 51 одсто [20]. Да приступ глобалној мрежи не зависи само од техничких могућности него и од зараде, види се из податка да 83% домаћинстава са месечним приходима

вишим од 600 евра има интернет, док га код оних са примањима нижим од 300 евра има 29% домаћинстава. Највише људи, 91%, светској мрежи приступа са десктоп рачунара, петина са мобилног телефона, а нешто мање од тога са лаптопа.

Кад је реч о типу везе, скоро половина домаћинстава у Србији која користе интернет има ADSL прикључак, четвртина кабловски интернет, а мобилне уређаје за повезивање користи 29% испитаника. Најчешће се приступа од куће (84%), затим с посла, од куће друге особе, у школи и на факултету, а тек 3,8% из интернет кафеа. Најзаступљенија категорија на мрежи су студенти, чак 95%. Ако није реч о пословним обавезама, интернет се највише користи за електронску пошту – 78%, затим за забаву (игре, филмови, музика) – 55%, за читање штампе – 41% и за учење – 23%. Најпопуларније српске веб странице су портали са вестима (Блиц [21], Б92 [22], Наслови [23] и РТС [24]). Најпосећенији домаћи портал је Krstarica [25], која укључује претраживачку машину, ажурне дневне вести из Србије, каталог локалних страница груписаних по тематици и разноврсне друге садржаје.

Експеримент започет 2005. увођењем локалне претраживачке машине *Pogodak*, која је претрагу прилагођавала морфологији српског, окончан је 2010. као непрофитабилан.

Википедија на српском представља извор разноврсних језичких података. Она садржи око 142.000 чланака и налази се на 28. месту [26] у свету у погледу броја објављених чланака. Википедија на српскохрватском [27] је мања и има око 40.000 чланака. Слободан приступ језичким подацима је могућ и преко портала *Растко* [28], *Антологија српске књижевности* [29] и *Транспоетика* [30], који садрже углавном књижевне текстове.

Видљивост појединих страна са садржајем на српском је привремено драматично пала током 2010. као последица преласка са топ-домена yu на rs.

Најчешће коришћена веб апликација је претрага веба. Она укључује аутоматску обраду језика на више нивоа, што ће бити детаљније описано у другом делу овог текста. Оваква обрада укључује префињене језичке технологије које се разликују за сваки језик. За српски, како је већ поменуто, проблеми настају због односа између латиничног и ћириличног писма, екавских и ијекавских варијација, графемских варијација у облику леме, као и морфолошког богатства. Користи које корисници интернета и добављачи садржаја на вебу могу да имају од језичких технологија можда су мање очигледне, на пример, у аутоматском превођењу веб садржаја са једног језика на други. Упркос високој цени ручног превођења, релативно мало језичких технологија је развијено и примењено у односу на уочене потребе. Разлог за то може бити у сложености српског језика и бројним технологијама које је потребно упослити за развој типичне језичке апликације. У следећем одељку представићемо преглед језичких технологија и основне области примене, као и оцену текуће ситуације у подршци језичким технологијама српског језика.

4

ЈЕЗИЧКЕ ТЕХНОЛОГИЈЕ ЗА СРПСКИ ЈЕЗИК

Језичке технологије су софтверски системи пројектовани за рад са природним језицима. Због тога се ове технологије често подводе под термин „технологија природних језика“. Природни језици се јављају у говорном и писаном облику. Иако је говор најстарији и са становишта човекове еволуције најприроднији начин језичке комуникације, комплексне информације и свеобухватно људско знање се бележе и преносе у писаном облику.

Говорне и текстуалне технологије обрађују и производе језик у ова два облика и оба користе речнике и граматичка и семантичка правила. То значи да језичке технологије повезују језик са различитим облицима знања независно од медија (говорних или текстуалних) којима су представљена. Слика 5 илуструје пејзаж језичких технологија.

Када комуницирамо, ми комбинујемо језик са другим начинима комуникације и другим информационим медијима. Говор се, на пример, комбинује са гестикулацијом и мимиком. Дигитални текстови се повезују са сликама и звуком. Филмови могу да садрже језик и у говорном и у писаном облику. Према томе, говорне и текстуалне технологије се преклапају и кооперирају са многим другим технологијама које олакшавају обраду мултимодалне комуникације и мултимедијалних докумената.

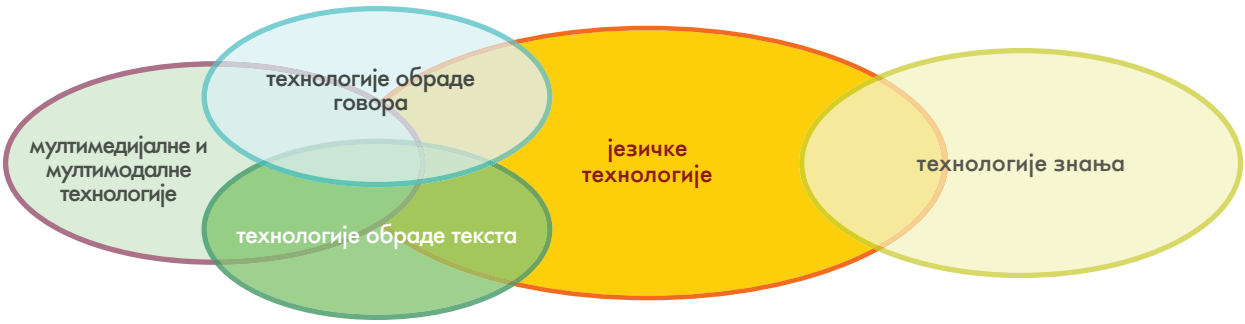
У тексту који следи размотрићемо главне области примене језичких технологија, а то су језичке провере, претраживање веба, технологију говора и машинско превођење. Ово укључује апликације и основне технологије као што су:

- исправљање правописних грешака;
- подршка састављању текста;
- рачунарски потпомогнуто учење језика;
- претраживање информација;
- екстракција информација;
- одговори на питања;
- резимирање текста;
- препознавање говора, и
- синтеза говора.

Језичке технологије представљају добро дефинисану истраживачку област са обимном општом литературом. Заинтересовани читаоци се упућују на следеће референце: [31, 32, 33, 34].
Пре него што размотримо наведене области примене, укратко ћемо објаснити архитектуру типичног језичкотехнолошког система.

4.1 АРХИТЕКТУРЕ АПЛИКАЦИЈА

Типичне софтверске апликације за обраду језика састоје се од неколико компонената, које одражавају различите аспекте језика. Слика 6 приказује веома поједностављену архитектуру на коју се може наићи у типичном систему за обраду текста. Прва три модула обрађују структуру и значење улазног текста:



5: Контекст језичких технологија

1. Припремна обрада: чишћење података, анализирање или уклањање форматирања и откривање улазног језика. У српском језику овај модул може да помогне у разрешавању ћириличко-латиничког двојства, као и екавско-ијекавског двојства.
2. Граматичка анализа: проналажење глагола и његових објеката, модификатора и осталих конституената, као и откривање структуре реченице.
3. Семантичка анализа: разрешавање вишезначности (тј. утврђивање одговарајућег значења речи у датом контексту); разрешавање анафора (тј. на шта се односе заменице) и референци у изразима; и представљање значења у машински читљивом облику.

Након анализе текста, модули посвећени специфичним задацима обављају многе различите операције, као што су аутоматско резимирање и прегледање база података. Овај поједностављен и идеализован опис архитектуре апликација илуструје сложеност апликација језичких технологија.

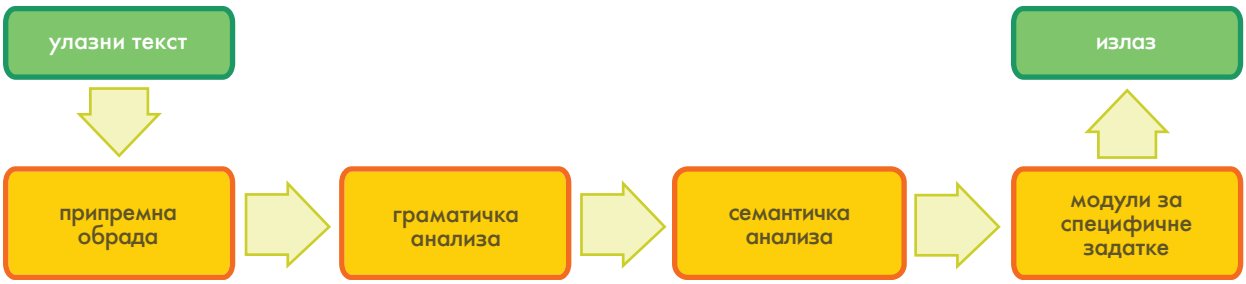
Пошто уведемо основна поља примене, даћемо кратак преглед стања у истраживању и образовању за језичке технологије, а закључићемо прегледом прошлих и текућих истраживачких програма. На крају овог одељка представићемо како по проценама стручњака изгледа позиција основних језичких алата и ресурса у простору чије димензије мере доступност, зрелост, квалитет и слично. Општа ситуација језичких технологија за српски језик резимирана је табелом 12.

4.2 ОСНОВНА ПОЉА ПРИМЕНЕ

У овом одељку посветићемо пажњу најважнијим језичкотехнолошким алатима и ресурсима и даћемо преглед активности на подручју језичких технологија у Србији.

4.2.1 Провера језика

Свако ко користи алат за обраду речи какав је Microsoft Word наишао је на компоненту за проверу, која указује на грешке у правопису и нуди исправке. Први програми за исправку правописних грешака поредили су листу речи извађених из текста са речником правилно исписаних речи. Данас су ти програми постали веома напредни. Коришћењем језички зависних алгоритама за **граматичку анализу** могу да се препознају грешке везане за морфологију (нпр. облици множине), синтаксу, као што је одсуство глагола или неслагање глагола са субјектом у лицу, броју или роду, на пример у '**Они је йисало йисмо.*' Па ипак, већина програма за проверу правописа



6: Типична архитектура система за обраду текста

неће пронаћи грешке у следећем тексту [35]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Да би се могле уочити овакве грешке, у многим случајевима је потребна и анализа контекста. На пример: да ли реч треба да буде написана великим словом у српском језику или не:

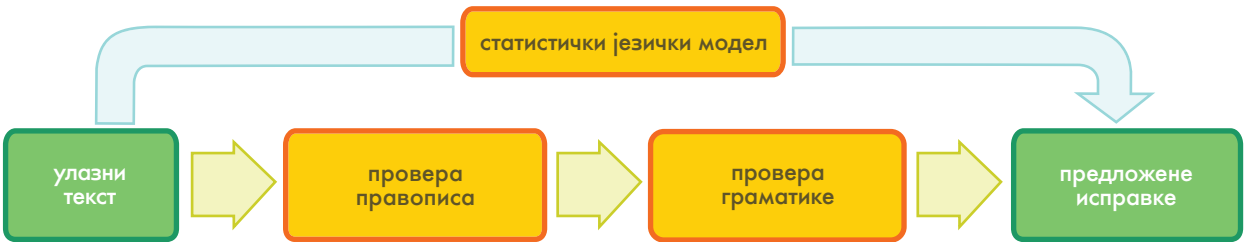
- Дивио се *Ружи*.
- Дивио се *ружи*.

Да би се ово постигло, могу се користити **граматике** специфичне за дати језик, што захтева много рада врхунских стручњака да би се оне уградиле у софтвер, или се могу користити такозвани статистички језички модели. Такви модели се заснивају на израчунавању вероватноће да се одређена реч појави у специфичном окружењу (нпр., испред или иза одређених речи). На пример, секвенција речи *ӣлава лаӣуна* много је вероватнија од секвенције *ӣлава Лаӣуна* (*Лаӣуна* је издавач). Статистички језички модели се могу аутоматски извести из велике количине (исправних) језичких података (који се зову **текстуални корпуси**). До сада су ови приступи коришћени и процењивани за податке на енглеском језику. Они се, међутим, не могу увек директно применити на српски језик имајући у виду његов слободан ред речи и богату флексију.

Провера језика се не користи само у алатима за обраду текста; она се примењује и у системима за подршку писању.

Први покушаји да се развије софтвер за проверу правописа за српски језик учињени су још крајем 1970-их [36] и били су мотивисани проблемима на које су наилазиле велике издавачке куће. Данас је слободан модул за проверу правописа за српски језик доступан за OpenOffice [37] на различитим оперативним системима, а постоји и занатски израђени производ, пакет RAS [38], који је развила компанија Srbosof и који се мора засебно инсталирати за сваког корисника.

Провера језика се не користи само у алатима за обраду речи; она се примењује и у „системима за подршку писању текста", тј. софтверским окружењима у којима се пишу приручници и друга документација за сложене производе информационих технологија, здравствене заштите, инжењерства и др. Плашећи се жалби купаца због погрешног коришћења и захтева за одштетом до којих би могло доћи јер су инструкције за употребу биле лоше или их они нису добро разумели, компаније су почеле све више пажње да посвећују техничкој документацији усредсређујући



7: Провера језика (статистичка; заснована на правилима)

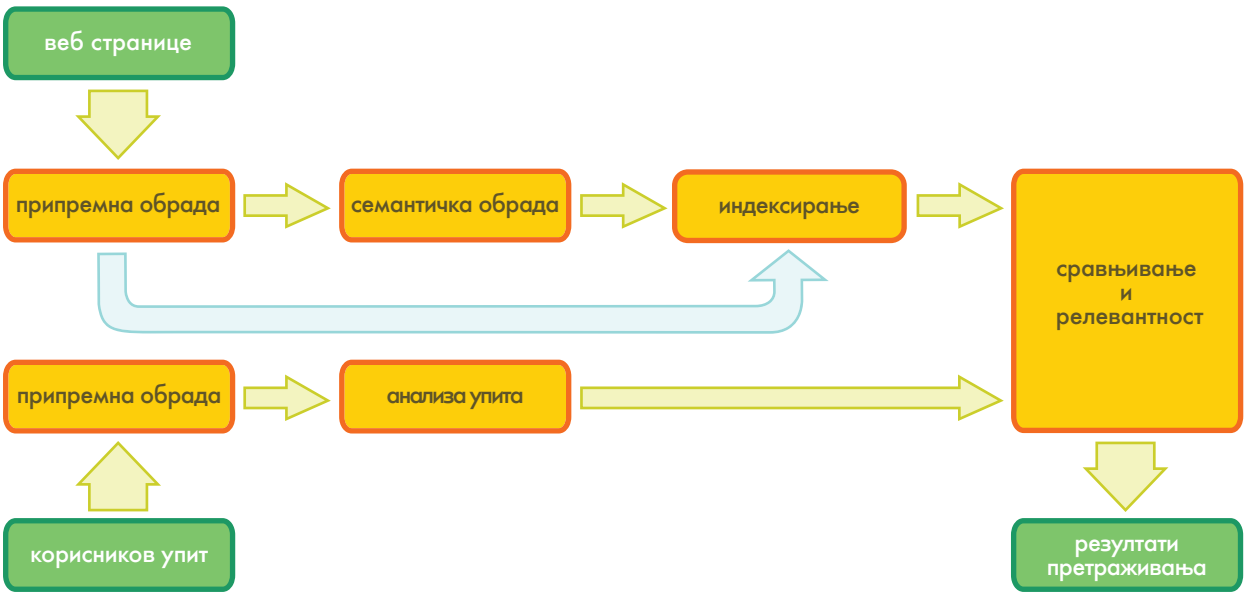
се истовремено на међународно тржиште (кроз превод или локализацију). Напредак у обради природних језика довео је до стварања софтвера за подршку писању текста који помаже ауторима техничке документације да користе речник и реченичне структуре усклађене са правилима струке и да поштују терминолошка ограничења која њихова компанија намеће. Провера језика није потребна само у системима за проверу правописа и за подршку писању текста већ је важна и за рачунарски потпомогнуто учење језика, а примењује се и за аутоматску корекцију упита који се постављају машинама за претраживање веба, као што су Гуглови предлози типа 'Да ли сте мислили на...'

4.2.2 Претраживање веба

Данас је претраживање веба, интранета и дигиталних библиотека вероватно најраспрострањеније коришћење језичких технологија, које је ипак недовољно развијено. Машина за претраживање Гугл (Google), која је отпочела са радом 1998, данас се користи за око 80% свих упита на вебу широм света [39]. Глаголи „гуглати/изгуглати" су у редовној употреби у српском језику. Ни сумећа (интерфејс) за претраживање ни приказ пронађених резултата нису се значајно променили од прве верзије. У текућој верзији Гугл нуди могућност исправке погрешно написаних речи, а такође је уградио и основне могућности за семантичку претрагу које могу да побољшају тачност претраге анализирањем значења упитних термина у контексту [40]. Успех Гугла показује да уз велику количину расположивих података и уз коришћење ефикасних техника за индексирање тих података, приступ који се заснива углавном на статистици може да доведе до задовољавајућих резултата.

Па ипак, за озбиљније захтеве за информацијама неопходно је укључивање и дубљег лингвистичког знања за **семантичку анализу**. Експерименти са коришћењем **лексичких ресурса** као што су **тезауруси** у машински читљивом облику и **онтолошки језички ресурси** (нпр. WordNet за енглески или СрпНет за српски), доводили су до побољшања у проналажењу страница коришћењем синонимних термина, нпр. *атомска енергија* и *нуклеарна енергија*, или претраживањем преко још слабије повезаних термина какви су *бели лук* и *чешњак*.

Следећа генерација машина за претраживање ће морати да укључи још много напредније језичке технологије, посебно да би могле да се изборе са упитима који се састоје од питања или неке друге врсте реченице уместо од листе кључних речи. На пример, за корисников упит *Дај ми листу компанија које су преузеле од стране других компанија у последњих пет година*, језичкотехнолошки систем мора да анализира реченицу на синтаксичком и семантичком нивоу, као и да обезбеди индекс који омогућава брзо проналажење релевантних докумената. За добијање задовољавајућег одговора треба да се примени синтаксичко парсирање да би се анализирала граматичка струк-



8: Претраживање веба

тура реченице и да би се утврдило да се траже компаније које су преузете, а не оне које су преузеле друге компаније. Такође, израз *у последњих пет година* треба да се обради да би се утврдило на које се године односи. Коначно, упит који се обрађује треба да се сравни са огромном количином неструктурираних података да би се пронашли делићи информација које корисник тражи. Ово се обично назива „проналажење информација", што укључује претраживање и рангирање релевантнијих докумената. Осим тога, да би се генерисала листа компанија, потребно је да се из документа изваде информације да се одређена ниска речи односи на име компаније. Овај процес зове се „препознавање именованих ентитета".

Још захтевнији су покушаји сравњивања упита са документима који су записани на различитим језицима. За вишејезично проналажење информација потребно је да се аутоматски преведе упит на све могуће изворне језике, а затим да се пронађена информација преведе на циљни језик.

Следећа генерација претраживачких машина мораће да укључи много напредније језичку технологију.

Све већи проценат података је доступан у формату који није текстуалан, што повећава захтеве за сервисима који омогућавају мултимедијално проналажење информација, нпр. проналажење информација у сликама и аудио и видео подацима. За аудио и видео датотеке то укључује модул за препознавање говора, да би се конвертовао говорни садржај у текст или фонетску репрезентацију са којом се корисников упит може сравњивати.

Популарне локације у Србији које нуде могућности претраживања, као што су B92 и Крстарица, ослањају се углавном на сервисе Гугла [41]. Покушај да се уведе машина за претраживање која би обављала искључиво претрагу надоле домена .rs и која би била делимично прилагођена специфичним својствима српског језика напуштен је 2010. године као непрофитабилан. Одређен број малих и средњих предузећа

ради на проширивању претраживачких сервиса, али углавном за стране партнере и за енглески језик.

У истраживачком окружењу су обављени експерименти са системима за проширивање упита који су машинама за претраживање слали упите проширивране морфолошким речницима и вишејезичним семантичким мрежама. Ови експерименти су дали занимљиве и корисне резултате у разноврсним доменима.

4.2.3 Говорна интеракција

Говорна интеракција је једно од многих подручја примене која зависе од говорне технологије, тј. технологија за обраду говорног језика. Технологија говорне интеракције је основа за израду сумећа које дозвољавају кориснику да комуницира са машинама користећи говорни језик уместо графичког дисплеја, тастатуре или миша. Данас се такве гласовне корисничке сумеће (voice user interfaces – VUIs) обично користе за потпуно или делимично аутоматизоване сервисе које компаније преко телефона нуде корисницима, запосленима или партнерима. Пословни домени који се у великој мери ослањају на гласовну корисничку сумећу јесу банкарство, логистика, јавни превоз и телекомуникације. Технологија за говорну интеракцију се осим тога користи и за сумеће са одређеним уређајима, нпр. у навигационим системима у колима, и за коришћење говора као алтернативе графичким или осетљивим на додир корисничким сумећама, нпр. у паметним телефонима.

Говорна интеракција састоји се од четири технологије:

1. Аутоматско **препознавање говора** (Automatic speech recognition – ASR) је задужено за утврђивање које речи су стварно изговорене када је дата секвенција звукова коју је произвео корисник.
2. Разумевање природног језика подразумева анализу синтаксичке структуре корисниковог исказа и његову интерпретацију у складу са наменом одређеног система.
3. Управљање дијалогом одређује коју акцију треба предузети за дати корисников улаз и дате функционалности система.
4. **Синтеза говора** (текст у говор, или Text-to-Speech, TTS) се користи за трансформацију одговора система у звукове које ће корисник примити као излаз.

Главни изазов је поседовање система за аутоматско препознавање говора који препознаје речи које је корисник изговорио што је прецизније могуће. Ово захтева или да се ограничи опсег могућих корисникових исказа на ограничен скуп кључних речи или да се ручно изграде језички модели који покривају широки опсег корисникових исказа на природном језику. Коришћењем техника машинског учења, језички модели могу да се изграђују аутоматски из **говорних корпуса**, тј. великих колекција говорних аудио датотека и њихових текстуалних транскрипција. Ограничавање исказа даје као резултат прилично ригидну и нефлексибилну гласовну корисничку сумећу коју корисници невољно прихватају. С друге стране, креирање, подешавање и одржавање богатих језичких модела може значајно да увећа трошкове. Па ипак, гласовне корисничке сумеће које користе језичке моделе на почетку дозвољавају кориснику да слободно изразе своје намере – подстичући га, на пример, питањем *Како Вам могу помоћи?* – показују већи степен аутоматизованости и прихватања.

> Технологија говора је основа за изградњу сумећа које омогућују кориснику да комуницира говорним језиком.

За генерисање излазног дела гласовне корисничке сумеће компаније теже коришћењу унапред снимљених исказа професионалних говорника. За статичке

9: Дијалошки систем заснован на говору

исказе код којих коришћене речи не зависе од конкретног контекста у коме се користе нити од личних података датог корисника, резултат може бити за корисника сасвим задовољавајући. Међутим, што је садржај исказа динамичнији, утолико више може да расте корисниково незадовољство због лоше прозодије до које долази због спајања појединачних аудио-датотека. Насупрот томе, данашњи системи за трансформацију текста у говор су супериорнији у погледу прозодијске природности динамичких исказа, иако их је још потребно оптимизовати.

Сумеће на тржишту говорне интеракције су значајно стандардизоване у току последње деценије када су у питању њихове различите технолошке компоненте. Дошло је и до велике консолидације тржишта, посебно у домену система за аутоматско препознавање говора и за претварање текста у говор. На овом пољу, националним тржиштима земаља G20 – што значи економски јаких земаља са значајном популацијом – доминира свега 5 актера из целог света, при чему су Nuance (САД) и Loquendo (Италија) најприсутнији у Европи. У 2011. год. Nuance је објавио преузимање Loquendo, што представља даљи корак у консолидацији тржишта.

Методе за препознавање и синтезу говора су у Србији, као и на ширем простору бивше Југославије, развијане углавном у електроинжењерском окружењу уз сарадњу стручњака за фонетику. Први напори су били усмерени на препознавање изолованих фонема. Значајан помак је у овом домену учинила група са Техничког факултета Универзитета у Новом Саду када је израдила, поред говорних база података, лексичку базу од преко 4 милиона акцентованих облика речи српског језика и више од 3 милиона облика речи хрватског језика. Коришћењем ових ресурса развијене су различите апликације из домена аутоматског препознавања говора и претварања текста у говор. Препознавање и синтеза говора за српски су ушли у комерцијалну употребу кроз фирму AlfaNum, која је потекла са Универзитета у Новом Саду. Ова компанија успешно послује и у другим државама које су настале на простору бивше Југославије – у Хрватској, Македонији, Босни и Херцеговини и Црној Гори. Компанија AlfaNum има значајан број корисника међу српским компанијама.

Када преводи на српски, Гуглов преводилац такође нуди основне могућности претварања текста у говор за резултате превођења, али без уграђених акцената.

Гледајући даље од данашњег технолошког стања, може се рећи да ће доћи до значајних промена захваљујући ширењу паметних телефона као нове платформе за управљање корисничким односима, која ће се користити поред већ постојећих канала – телефона, интернета и електронске поште. Ова тенденција ће утицати и на коришћење технологије за говорну интеракцију. С једне стране, на дуже стазе

ће опадати потражња за гласовном корисничком сумећом за телефонске услуге. С друге стране, коришћење говорних могућности као приступ паметним телефонима добиће на значају. Ову тенденцију подржава напредак који се већ може уочити у тачности препознавања говора независних говорника помоћу говорних сервиса за диктирање који се већ нуде као централизоване услуге корисницима паметних телефона.

4.2.4 Машинско превођење

Идеја да би се дигитални рачунари могли користити за превођење природних језика настала је 1946, после чега је уследило значајно финансирање истраживања у овој области педесетих година и потом осамдесетих година прошлог века. И поред свега, **машинско превођење** (Machine Translation – MT) и даље не успева да испуни велика очекивања која је подстакло у тим раним данима.

На основном нивоу, машинско превођење једноставно замењује речи из једног природног језика речима из неког другог. Ово може да буде корисно у неким предметним доменима који користе веома ограничен формализован језик, као што је језик временских прогноза. Међутим, за добар превод текстова који нису толико стандардизовани, треба сравнити веће текстуалне јединице (фразе, реченице или целе пасусе) са најближим паром у циљном језику.

На основном нивоу, машинско превођење једноставно замењује речи једног природног језика речима другог језика.

Овде највећа потешкоћа лежи у томе што су природни језици вишезначни, што ствара изазове на различитим нивоима, јер треба, на пример, отклонити вишезначност речи на лексичком нивоу („јагуар" може да буде назив животиње и аутомобила) или утврдити повезаност предлошких фраза на синтаксичком нивоу, као у:

- *Полицајац је усūео да ūримеūи човека без двоūледа.*
- *Полицајац је усūео да ūримеūи човека без револвера.*

Један начин да се изгради систем машинског превођења заснива се на лингвистичким правилима. За превођење између сродних језика могуће је и директно превођење у случајевима који наликују наведеним примерима. Ипак, системи засновани на правилима (или на знању) анализирају улазни текст и креирају посредну симболичку интерпретацију, из које се потом генерише текст на циљном језику. Успех ових метода веома зависи од постојања исцрпних лексикона са морфолошким, синтаксичким и семантичким информацијама, и великих скупова граматичких правила које су пажљиво израдили искусни лингвисти. Ово је веома дуг и скуп процес.

Како је крајем осамдесетих година прошлог века снага рачунара порасла и појефтинила, дошло је до већег интересовања за статистичке методе у машинском превођењу. Статистички модели се изводе из анализе двојезичних текстуалних корпуса, какав је, на пример, **паралелни корпус** Europarl, који садржи текстове Европског парламента на двадесетједном језику.

Под условом да постоји довољно података, статистичко машинско превођење може да изведе довољно добро приближно значење текста на страном језику, тако што обрађује паралелне верзије и проналази прихватљиве речи. Међутим, за разлику од система заснованих на знању, статистичко машинско превођење (или превођење засновано на подацима) често генерише неграматички излаз. С друге стране, осим што захтева мање људског напора за писање граматика, превођење засновано на подацима може да покрије специфичности језика које измичу

10: Машинско превођење (статистичко; засновано на правилима)

системима заснованим на знању, као што су идиоматски изрази.

Пошто се јаке и слабе стране машинских система заснованих на знању, односно подацима, допуњују, истраживачи данас једногласно теже хибридним приступима који комбинују обе методологије. То се може урадити на више начина. Један начин је да се користе и системи засновани на знању и системи засновани на подацима, а да засебан модул за селекцију одлучи шта је најбољи излаз за сваку реченицу. Међутим, за дугачке реченице, дуже од, рецимо, дванаест речи, код оваквог приступа ни један резултат неће бити савршен.

Боље је решење које комбинује најбоље делове сваке реченице добијене из различитих извора, што може бити доста сложено јер није увек очигледно шта су одговарајући делови код вишеструких могућности и јер их, осим тога, треба и поравнати.

Машинско превођење представља посебан изазов за српски језик.

Што се тиче везе српског и страних језика, проблеми зависе од природе специфичног језика (да ли има развијену морфологију, да ли има слободну или фиксирану дистрибуцију реченичних конституената, да ли користи чланове, да ли је записан ћириличним или латиничним писмом, да ли користи логичку или граматичку интерпункцију итд.). Међутим, овде се не ради само о томе шта су проблеми већ и о могућности да се сарађује на решавању сличних проблема. У том смислу би сарадња са пројектима везаним за рачунарску обраду других словенских језика била посебно корисна. Овде су такође важне лексичке и терминолошке везе, наиме у којој мери је неки страни језик утицао на развој српског. У овом подручју би требало тражити сарадњу са пројектима чији је циљ рачунарска обрада оних језика који су служили и још увек служе као кичма развоја српског, а то су, пре свега, енглески, француски, немачки и руски.

Требало би додати да се одвијају и контрастивна истраживања српског и неких страних језика. Нажалост, има недовољно сарадње између лингвиста који се баве српским као матерњим језиком и оних лингвиста који се као стручњаци за стране језике укључују у контрастивна истраживања. Други проблем је недовољан број великих двојезичних речника.

Највећа потреба за језичким технологијама у Србији је на пољу превођења. Постоје нека специјализована друштва (Друштво књижевних преводилаца Србије, Друштво научних и стручних преводилаца Србије), нека локална мала и средња предузећа (нпр. Elitence и Proverbum) и неке стране компаније (нпр. WorldLingo) које нуде професионалне преводилачке услуге или слободан машински превод заснован на фразама

(нпр. Google Translate, WorldLingo). Неке од њих користе власничке електронске речнике за свој рад, а WorldLingo нуди и шире услуге машинског превођења (веб локације, текст, документа, електронске поруке, API итд.).

Осим добро познатог и слободно доступног Гуглов ог статистичког система за превођење, који укључује и српски, ниједан други систем за машинско превођење за српски није произведен, осим неких почетних радова (нпр. у оквиру пројекта SEE-ERA) и малих експерименталних система.

Међутим, генерички статистички системи за машинско превођење какав је Google Translate подржавају српски у значајној мери, посебно за превођење на енглески и са енглеског. Ипак, за друге језичке парове перформансе су слабе, а добијени превод је често неразумљив, а понекад и смешан. То је резултат недовољне величине паралелних корпуса који је за те језичке парове коришћен за обуку система за статистичко машинско превођење.

Још увек се сматра да се много може урадити на побољшању квалитета система за машинско превођење. Изазови обухватају прилагођавање језичких ресурса датом предметном или корисничком домену и укључивање терминолошких база и преводилачких меморија у постојеће радне процесе.

Акције за процењивање омогућавају да се пореде квалитет система за машинско превођење, различити приступи, као и статус система за машинско превођење за различите језичке парове. Следећа табела 11 (стр. 30), представљена у оквиру пројекта Европске комисије Euromatrix+, приказује перформансе по паровима за 22 од 23 службена европска језика (недостаје ирски). Резултати су рангирани према BLEU процени, која даје више оцене за боље преводе [43]. Човек-преводилац постиже резултат од око 80 поена.

Најбољи резултати (приказани зеленом и плавом бојом) постигнути су за језике који имају користи од значајних истраживачких напора у оквиру сарадничких програма и за које постоје многи паралелни корпуси (нпр., енглески, француски, холандски, шпански и немачки), а најлошији (приказани црвеном бојом) за језике који нису могли да користе сличне претходне напоре или који су веома различити од других језика (нпр., мађарски, малтешки, фински).

4.3 ДРУГЕ ОБЛАСТИ ПРИМЕНЕ

Изградња апликација заснованих на језичким технологијама укључује опсег подзадатака који се не виде увек на нивоу интеракције са корисником, али који обезбеђују значајне функционалности система „испод хаубе". Сваки од њих представља важан истраживачки задатак који се развио у засебну поддисциплину у оквиру рачунарске лингвистике.

На пример, одговарање на питања је постало активно истраживачко подручје, за које су изграђени анотирани корпуси и отпочела су научна такмичења. Идеја је да се крене даље од претраживања заснованог на кључним речима (на које машине одговарају целом колекцијом релевантних одговора) ка ситуацији у којој корисник поставља конкретно питање, а систем пружа један одговор. На пример:

> *Пӣтање: Са колико го̄дина је Нил Армсӣронӣ крочио на Месец?*
>
> *Одго̄вор: 38.*

Иако је ово очигледно повезано са већ поменутим основним претраживањем веба, одговарање на питања је данас пре свега заједнички термин за различите истраживачке теме као што су: које типове питања треба разликовати и како треба с њима поступати, како треба анализирати и поредити документа

	Циљни језик — Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

11: Машинско превођење између 22 EU-језика – Machine translation between 22 EU-languages [42]

која потенцијално садрже одговоре (да ли она садрже супротстављене одговоре?), и како се специфична информација – заправо одговор – може поуздано извући из документа, не запостављајући при томе контекст у коме се налази.

Ова област је повезана са задатком екстракције информација, облашћу која је била изузетно популарна и утицајна у време „статистичког заокрета" у рачунарској лингвистици почетком деведесетих година. Циљ екстракције информација је да се идентификују специфични делићи информација у специфичним класама докумената; то може да буде, на пример, откривање кључних актера у преузимању компанија на основу извештавања у новинама. Други сценарио на коме се радило били су извештаји о терористичким инцидентима, где је проблем био да се текст преслика у шаблон у коме су спецификовани извршилац, мета, време и место инцидента, и шта је њиме постигнуто. Централна карактеристика екстракције информација је попуњавање шаблона специфичног за неки домен, због чега је то још један пример технологије 'иза сцене', која представља јасно разграничено истраживачко подручје, али која из практичних разлога мора да се угради у одговарајуће окружење апликације.

Два „гранична" подручја, која понекад имају улогу самосталне апликације, а понекад помоћне компоненте („испод хаубе"), јесу резимирање текста и **генерисање текста**. Резимирање се, очигледно, односи на задатак скраћивања дугачког текста, и њега као функцију нуди MS Word. Оно ради углавном на статистичким основама, тако што прво идентификује „важне" речи у тексту (на пример, речи које се у конкретном тексту често јављају, док се у текстовима у начелу јављају много ређе), а затим утврђује у којим се реченицама јавља пуно важних речи. Ове рече-

нице се затим издвајају из текста и из њих се саставља сажетак. У овом сценарију, који је комерцијално уобичајен, резимирање се своди на просту екстракцију реченица, а текст на подскуп својих реченица. Алтернативни приступ, коме се такође посвећују неки истраживачки напори, састоји се у генерисању потпуно нових реченица које не постоје у изворном тексту. Ово захтева дубље разумевање текста, што значи да је тај приступ (за сада) знатно мање робустан. Све у свему, генерисање текста у већини случајева није самостална апликација, већ је уграђено у шире софтверско окружење, као што је клинички информациони систем у коме се подаци о пацијентима скупљају, складиште и обрађују. Генерисање извештаја је само једна од многих примена резимирања текста.

Унутар ових поменутих подручја се, кад је реч о српском, спроводе врло успешни експерименти везани за препознавање именованих ентитета, као дела проблема екстракције информација. Очекује се убрзани развој система за екстракцију информација и одговарање на питања, имајући у виду опсег изграђених морфолошких речника и локалних граматика.

Постоје и друга подручја на којима се примењују језичке технологије. Једно од њих је откривање плагијаторства, које користи језички независне технологије, али се може проширити претрагом за једноставним парафразама текста. Истраживање које иде у овом правцу за научне чланке у Србији је реализовала компанија CEON [44].

4.4 ОБРАЗОВНИ ПРОГРАМИ

Језичке технологије су интердисциплинарно подручје које захтева знања многих стручњака, лингвиста, стручњака за рачунарство, математичара, филозофа, психолингвиста и неуролога, да поменемо само неке. Као такво, оно још није добило сталну позицију у високом образовању у Србији и углавном је ограничено на појединачне курсеве у оквиру општијих постдипломских студијских програма. Парадоксално, упркос оваквом стању, у оквиру истраживачке станице Петница [45] се сваке године организују мали истраживачки семинари за средњошколце са темама из рачунарске лингвистике.

На нивоу универзитетских студија, теме из области рачунарске лингвистике су присутне на студијама из рачунарства, електронике, библиотекарства, лингвистике и психологије, и то на универзитетима у Београду и Новом Саду. Предмети који су понуђени студентима дају основне појмове о процесу обраде природних језика, али су у функцији формирања студената за друкчије профиле. На Математичком факултету у Београду, на редовним студијама су присутни курсеви из лексичке анализе и истраживања података (енгл. data mining), поред курсева који обрађују фундаментална математичка знања потребна у обради природних језика (посебно статистика, алгебра и логика), док на докторским студијама постоји већи избор предмета из области технологија природних језика. Најтемељније образовање на овом подручју стичу студенти Групе за библиотекарство и информатику на Филолошком факултету у Београду, док на другим групама тог факултета постоји највише један уводни курс. У оквиру студија српског језика није предвиђено образовање на подручју обраде природних језика. На Филозофским факултетима у Београду и Новом Саду, на групама за психологију постоје курсеви из психолингвистике на којима се студенти упознају са статистичким методама обраде језика. На техничким факултетима се изучавају методе од значаја за обраду говора. Курикулум који даје специјалност у домену рачунарске лингвистике или језичких технологија не постоји ни на једном од факултета.

4.5 НАЦИОНАЛНИ ПРОЈЕКТИ И ИНИЦИЈАТИВЕ

Индустрија језичких технологија је у Србији релативно неразвијена у поређењу са водећим економијама земаља Европске уније, и то из више разлога. Главна покретачка снага иза развоја језичких технологија у Србији су углавном домаћа мала и средња предузећа, али и неке стране компаније, које понекада обезбеђују подршку за српски језик у разноврсним апликацијама које траже подршку језичких технологија. Пошто не постоји национални програм подршке развоју језичких технологија, њихов развој и примена се одвијају често на некоординиран начин. Постоје бар три правца којима се језичке технологије уводе у Србију: (а) кроз државне научне и развојне пројекте, (б) преко (првенствено) страних фирми које уз рачунарску опрему пружају и одређени облик језичке подршке и (в) кроз интерни развој у оквиру домаћих организација какве су, нпр., издавачке куће или преводилачке агенције. Активности у ова три правца се одвијају, осим изузетно, независно једне од других.

С друге стране, рачунарски писмено становништво у Србији је навикло да користи графичку корисничку сумећу (интерфејс) на енглеском језику, иако неки од њих можда и не знају енглески. Локализоване верзије њима понекад изгледају чудне и непрецизне и нису вољни да их користе. Једине апликације које у великом броју користе графичку корисничку сумећу, на српском су различите пословне, финансијске и рачуноводствене апликације, укључујући и SAP ERP систем. Ипак, има примера локализоване графичке корисничке сумеће познатих софтверских продаваца као што је Microsoft (нпр. Windows, Office), Google или Oracle (локализација Open Office, финансирана у периоду од 2008. до 2011. од стране Министарства за телекомуникације и информационо друштво кроз пројекат на Математичком факултету [46]).

Научни пројекти које финансира Министарство за образовање и науку тек у најновијем циклусу научних пројеката (период 2011–2014) препознају интердисциплинарност. До 2010. године научни пројекти (па тиме и критеријуми за њихову евалуацију) били су оштро раздвојени на подручја математике (коме је подређено рачунарство), језика и технолошких дисциплина. У таквом амбијенту је било тешко реализовати природни спој дисциплина које су у основи развоја језичких технологија. У оваквом контексту било је неопходно успоставити везе између истраживања на подручју српског језика и информатике.

Први такав пројекат, под називом „Интеракције текста и речника“, формиран је 2002. године као заједнички пројекат катедара за српски језик Филолошког факултета у Београду и Филозофског факултета у Новом Саду и Математичког факултета у Београду. У оквиру овог пројекта је формиран први корпус савременог српског језика [47] доступан преко веба, а који данас има преко 300 корисника са различитих универзитета и института у земљи и иностранству. У оквиру овог пројекта је започета и конструкција електронског морфолошког речника српског језика према тзв. LADL формату [48]. Овај пројекат је настављен као заједнички пројекат Катедре за српски језик Филолошког факултета у Београду и Математичког факултета у периоду од 2006. до 2010. под називом „Теоријско-методолошки оквир за модернизацију описа српског језика” и од 2011. до 2014. као „Српски језик и његови ресурси: теорија, опис и примене”. Кроз ове пројекте је довршена конструкција електронског речника простих речи и започет рад на конструкцији речника сложених речи, развијени су паралелни француско-српски и енглеско-српски корпус литерарних текстова, описане су локалне граматике за поједине сегменте српског (посебно за именоване ентитете), као и различити соф-

тверски алати, од којих посебан значај има радна станица LeXimir, која омогућава интеграцију и трансформацију хетерогених лексичких ресурса.

Упоредо са овим истраживањима у области језика, у области друштвених наука је финансиран пројекат „Фундаментални когнитивни процеси и функције", који је реализован на Катедри за психологију Филозофског факултета у Београду. Овај пројекат је, поред осталог, имао за циљ да испита могућност аутоматске анотације текста полазећи од анотираног корпуса [49], развијеног још током педесетих година, а деведесетих преведеног у електронски облик.

Синтеза и препознавање говора на Техничком факултету Универзитета у Новом Саду се реализује кроз пројекте технолошког развоја почев од 2005. године, и то „Развој говорних технологија на српском језику и њихова примена у 'Телекому Србија'" (2005–2007), „Говорна комуникација човек-машина" (2008–2010), „Развој дијалошких система за српски и друге јужнословенске језике" (2011–2014). Они пружају подршку различитим апликацијама и сервисима за претварање текста у говор и аутоматско препознавање говора, који укључују системе за интерактивне гласовне одговоре (IVR), пословне телефонске системе, позивне центре, пријављивање гласом, праћење реклама, уочавање речи, и др.

У оквиру других области науке развијани су појединачни ресурси од значаја за језичке технологије, али без непосредне интеракције са већ наведеним пројектима. Поменимо као примере геолошки српско-енглески тезаурус [50] и фолклористичку базу ДАБИ Балканолошког института САНУ [51].

Упоредо са националним пројектима, српске научне институције су биле укључене и у различите међународне пројекте везане за подручје језичких технологија. Током периода санкција Уједињених нација, одржавање одређеног нивоа активности је било могуће захваљујући учешћу у пројектима TELRI I и II [52]. Иако српске истраживачке групе у то време нису могле да учествују на пројекту MULTEXT-East [53], оне су ипак произвеле корисне ресурсе у формату који је тај пројекат дефинисао: морфосинтаксички опис српског језика, поравнату верзију српског превода романа „*1984*" Џорџа Орвела, његову лематизирану и морфосинтаксички етикетирану верзију и исцрпан речник који покрива комплетну лексику романа „*1984*".

Ситуација у разним доменима обраде српског језика је различита, али значајан напредак постоји у развоју корпуса, морфолошкој анализи, електронским речницима, као и у екстракцији именованих ентитета.

По укидању санкција, посебно је значајан био пројекат BalkaNet [54], који је омогућио развој семантичке мреже типа WordNet за српски. Кроз билатералну сарадњу са Француском је развијен српски део вишејезичне лексичке базе властитих имена Prolex [55], а у оквиру пројекта Intera једномилионски паралелизовани енглеско-српски корпус, који је лематизиран и морфолошки анотиран. Овај корпус је послужио за обучавање тагера и за експерименте у поравнавању на нивоу речи и у аутоматском превођењу.

Српски учесници су били укључени у два регионална пројекта. Један од њих, SEE-ERA.NET – Building Language Resources and Translation Models for Machine Translation (Изградња језичких ресурса и преводилачких модела за машинско превођење), био је усмерен на јужнословенске и балканске језике (ICT 10503 RP, 2007–2008). Његов главни допринос био је развој једносмерних преводилачких модела који се ослањају на вишејезичне ресурсе великих димензија, у ствари на корпус *Acquis Communautaire*. Међутим, пошто документа која улазе у овај ресурс у то

време још нису била преведена на српски, преводилачки модел није био произведен за српски. Превод законске регулативе Европске уније је у току и део преведеног материјала је већ доступан [56]. Са своје стране је српски тим допринео овом пројекту развојем једног другог вишејезичног ресурса који се заснива на роману Жила Верна „*Пут око света за осамдесет дана*" (у том тренутку било је укључено 16 језика). Други пројекат био је WISE – An Electronic Marketplace to Support Pairs of Less Widely Studied European Languages (Електронско тржиште за подршку паровима мање изучаваних европских језика), чији је циљ била производња не само вишејезичних лексичких ресурса обогаћених лингвистичким метаподацима већ и изградња и промоција електронског тржишта за слабије изучаване балканске језике, укључујући и српски (BSEC 009 / 05.2007, 2007 – 2008). Даље активности подразумевају, пре свега, развој поступака за синтаксичку анализу српског, која је, с обзиром на слободан ред речи и морфолошко богатство српског језика, изузетно сложен поступак. Ово подразумева развој нових ресурса, пре свега нових типова речника и корпуса, као и пратећих алата.

4.6 ДОСТУПНОСТ АЛАТА И РЕСУРСА

Табела 12 даје приказ текућег стања језичких технологија за српски језик. Рангирање постојећих алата и ресурса се заснива на процени више водећих експерата који су дали оцене на скали 0 (врло ниско) до 6 (врло високо) на основу седам критеријума.

За српски језик, стање ресурса и технологија може се описати на следећи начин:

- Што се тиче морфолошких и с њима повезаних питања, може се слободно рећи да је ниво развоја технологија и ресурса задовољавајући, углавном захваљујући постојању великог електронског речника и локалних граматика. Непосредна последица тога је да су потребни алати за проналажење информација и екстракцију информација на располагању. Неки од речника су спремни за широку употребу, док неке још треба доградити, на пример СрпНет.
- Референтни корпус савременог српског језика екавског изговора је на располагању, као и неколико поравнатих корпуса, и сви они су на располагању истраживачима српског језика. Текућа истраживања су усредсређена на доградњу референтног корпуса и његово проширивање ијекавским изговором.
- Говорне технологије су добро развијене и нашле су широке пословне примене, али се истраживања морају ширити да би се проширила и поља примене.
- Софтвер намењен повећавању продуктивности лексикографа је развијен, али недовољна спремност за нове технологије у традиционално оријентисаном лексикографском окружењу је препрека бржем развоју лексикографије.
- У неким подручјима су обављени успешни експерименти у строго истраживачком окружењу, као што је плитко парсирање, резимирање, машинско превођење, онтолошки ресурси. Међутим, добијени резултати су још увек далеко од нивоа развоја који је постигнут за развијене европске језике. Пажњу истраживача привлаче и мултимедијални и мултимодални документи, посебно у контексту дигитализације културног наслеђа.

Имајући у виду сложеност српске синтаксе, подручја заснована на дубоком парсирању једноставно не постоје: семантика реченица, семантика текста, генерисање језика. Због тога не постоји ни формализована синтакса српског, што ограничава развој синтаксички и семантички анотираних корпуса. Форма-

	Квантитет	Доступност	Квалитет	Покривеност	Зрелост	Одрживост	Прилагодљивост
Језичке технологије (алати, технологије, апликације)							
Препознавање говора	2	2	1	1	1	1	0
Синтеза говора	2	2	4	4	5	5	1
Граматичка анализа	1	1	2,5	2	2	1,5	1,5
Семантичка анализа	1	1	1	1,5	1	1	1,5
Генерисање текста	0	0	0	0	0	0	0
Машинско превођење	1	1	0	1	0	1	1
Језички ресурси (ресурси, подаци, базе знања)							
Текстуални корпуси	0,5	1	0,5	1	1	1	0,5
Говорни корпуси	1	2	4	4	3	3	3
Паралелни корпуси	3	3	3	2	2	2	3
Лексички ресурси	1	2	2	2	2	2	2,5
Граматике	1	1	0	1	0	1	1

12: Стање језичких технологија за српски језик

лизација синтаксе српског је, према томе, најхитнији задатак за даљи развој језичких технологија.

4.7 ПОРЕЂЕЊЕ ЈЕЗИКА

Текуће стање подршке језичких технологија значајно се разликује од једне језичке заједнице до друге. Да би се упоредиле ситуације у којима се налазе различити језици, овај одељак ће представити оцену засновану на два примера области примене (машинско превођење и обрада говора) и на једној технологији (анализа текста), као и на основним ресурсима неопходним за изградњу апликација језичких технологија. Језици су сврстани у групе на основу следеће скале од пет вредности:

- Одлична подршка језичким технологијама
- Добра подршка
- Умерена подршка
- Фрагментарна подршка
- Слаба подршка или без подршке

Мера подршке језичким технологијама установљена је на основу следећих критеријума:

- **Обрада говора**: Квалитет постојећих технологија за препознавање говора, квалитет постојећих технологија за синтезу говора, покривеност домена, број и обим постојећих говорних корпуса, бројност и разноврсност расположивих апликација заснованих на говору

- **Машинско превођење**: Квалитет постојећеих технологија машинског превођења, број покривених језичких парова, покривеност језичких феномена и домена, квалитет и обим постојећих паралелних корпуса, бројност и разноврсност расположивих апликација машинског превођења
- **Граматичка анализа**: Квалитет постојећих технологија за анализу текста и области које покривају (морфологија, синтакса, семантика), покривеност језичких феномена и домена, бројност и разноврсност расположивих апликација, квалитет и обим постојећих (анотираних) текстуалних корпуса, квалитет постојећих лексичких ресурса и граматика и области које покривају (нпр. WordNet)
- **Ресурси**: Квалитет и обим постојећих текстуалних, говорних и паралелних корпуса, квалитет постојећих лексичких ресурса и граматика и области које покривају

Горње табеле показују да су алати и ресурси за српски језик углавном у најнижој групи. Српски добро стоји у поређењу са језицима са малим бројем говорника, као што су хрватски, словеначки и словачки, али сви ти језици су далеко иза заступљенијих европских језика као што су немачки или француски. Па ипак, чак ни за ове последње језике, алати и ресурси језичких технологија нису достигли квалитет и покривеност одговарајућих алата и ресурса за енглески језик, који је у врху у свим областима језичке технологије. А и у енглеским језичким ресурсима постоји још увек доста празнина с тачке гледишта апликација високог квалитета.

4.8 ЗАКЉУЧЦИ

У овој серији белих књига учинили смо значајан почетни напор да оценимо подршку језичких технологија за 30 европских језика и да обезбедимо квалитетно поређење тих језика. Пошто су идентификоване потребе и недостаци, заједница европских језичких технологија, као и све заинтересоване стране сада су у прилици да осмисле програме истраживања и развоја широких размера чији је циљ изградња истински вишејезичне, технолошки оспособљене Европе.

Видели смо да постоје огромне разлике између европских језика. Док за неке језике у одређеним областима примене постоје квалитетни ресурси и одговарајући софтвер, за друге језике ту постоје значајне празнине. Многим језицима недостају основне технологије за анализу текста, као и суштински ресурси за развој тих технологија. Други имају основне ресурсе или алате, али још увек нису у прилици да инвестирају у семантичку обраду. Зато нам тек предстоји да учинимо главни напор за постизање амбициозног циља обезбеђивања високо квалитетног машинског превођења између свих европских језика. Обим ресурса и опсег алата који постоје за српски језик још увек су врло ограничени, нарочито када се упореде са алатима и ресурсима за језике као што су француски, немачки и посебно енглески, и нису довољни ни по квалитету ни по квантитету за развој оне врсте технологије која је неопходна за подршку истински вишејезичном друштву знања.

Технологије које су већ развијене и оптимизоване за енглески не могу једноставно да се пренесу на српски језик. Систем за синтаксичку анализу структуре реченице заснован на енглеском по правилу је неприкладан за примену на српском тексту. Рад на обради српског језика до сада је био концентрисан на развој ресурса и алата који су у складу са специфичним својствима српског (пре свега опис његове богате морфологије). Овај правац развоја мора обавезно да се задржи и у будућности. За скромну језичку заједницу и истраживачку средину као што је српска, сарадња у развоју ресурса, како на домаћем тако и на међународном нивоу, од пресудног је значаја. Ово гене-

рално важи за већину словенских језика, а за сарадњу су неопходне даље стимулативне мере. Посебно велике могућности за сарадњу постоје између пројеката везаних за стандардне језике штокавског порекла, као и за словенске језике уопште, имајући у виду заједничка својства тих језика.

Учешће Србије у CESAR-у и META-NET-у требало би да допринесе развоју, стандардизацији и доступности неколико важних ресурса језичких технологија и стога развоју језичких технологија за српски језик. Дугорочни циљ META-NET-а јесте да уведе технологију високог квалитета за све језике како би се постигло политичко и економско јединство кроз културну разноврсност. Технологија ће помоћи да се уклоне постојеће баријере и да се изграде мостови међу европским језицима. Ово захтева од свих заинтересованих страна – у политици, истраживању, привреди и друштву – да уједине своје напоре за будућност.

Индустрија српских језичких технологија је веома скромна. Укључено је тек неколико средњих и малих предузећа и њихов приступ је у суштини заснован на примени „грубе силе", што значи да се у основи занемарују специфичности српског језика. Наши налази показују да је једина алтернатива улагање значајних напора у стварање ресурса за језичке технологије за српски и њихово коришћење за унапређење истраживања, иновација и развоја. С обзиром на потребу за великим количинама података и екстремну сложеност система језичких технологија, од виталног је значаја развој нове инфраструктуре и кохерентније организације истраживања, која би подстакла већу сарадњу. Други кључни допринос био би успостављање мултидисциплинарног студијског програма обраде језика на мастер и докторском нивоу, што данас не постоји.

Према томе, можемо да закључимо да постоји неодложна потреба за широком, координираном иницијативом усмереном на превазилажење разлика у спремности језичких технологија за европске језике као целину.

одлична подршка	добра подршка	умерена подршка	фрагментарна подршка	слаба подршка или без ње
	енглески	немачки италијански фински француски холандски португалски шпански чешки	баскијски бугарски дански естонски галицијски грчки ирски каталонски норвешки пољски шведски **српски** словачки словеначки мађарски	исландски хрватски летонски литвански малтешки румунски

13: Обрада говора: стање подршке језичких технологија за 30 европских језика

одлична подршка	добра подршка	умерена подршка	фрагментарна подршка	слаба подршка или без ње
	енглески	француски шпански	немачки италијански каталонски холандски пољски румунски мађарски	баскијски бугарски дански естонски фински галицијски грчки ирски исландски хрватски летонски литвански малтешки норвешки португалски шведски **српски** словачки словеначки чешки

14: Машинско превођење: стање подршке језичких технологија за 30 европских језика

одлична подршка	добра подршка	умерена подршка	фрагментарна подршка	слаба подршка или без ње
	енглески	немачки француски италијански холандски шпански	баскијски бугарски дански фински галицијски грчки каталонски норвешки пољски португалски румунски шведски словачки словеначки чешки мађарски	естонски ирски исландски хрватски летонски литвански малтешки **српски**

15: Граматичка анализа: стање подршке језичких технологија за 30 европских језика

одлична подршка	добра подршка	умерена подршка	фрагментарна подршка	слаба подршка или без ње
	енглески	немачки француски холандски шведски чешки мађарски пољски италијански шпански	баскијски бугарски дански естонски фински галицијски грчки каталонски хрватски норвешки португалски румунски **српски** словачки словеначки	ирски исландски летонски литвански малтешки

16: Језички ресурси: стање подршке језичких технологија за 30 европских језика

О META-NET-У

META-NET је мрежа изврсности коју финансира Европска унија [57]. Њу тренутно чине 54 члана, који представљају 33 европске земље. META-NET подстиче технолошки савез вишејезичне Европе (Multilingual Europe Technology Alliance – META), заједницу професионалаца и организација са подручја језичких технологија из Европе. META-NET је посвећен остваривању технолошких основа за успостављање и одржавање истинског вишејезичног европског информационог друштва које:

- омогућава вишејезичну комуникацију;
- обезбеђује једнак приступ информацијама и знању на свим језицима;
- нуди напредне могућности умрежене информационе технологије.

Мрежа подржава Европу која се удружује у јединствено дигитално тржиште и информациони простор. Она стимулише и промовише вишејезичне технологије за све европске језике. Ове технологије омогућавају аутоматско превођење, генерисање садржаја, обраду информација, управљање знањем за широк распон апликација и предметних области, као и сумеђе засноване на језику за технолошке производе од кућних апарата, преко машина и возила, до рачунара и робота. META-NET је покренут 1. фебруара 2010. и већ је предузео више активности које доприносе остварењу његових циљева. META-ВИЗИЈА, META-РАЗМЕНА и META-ИСТРАЖИВАЊЕ су три правца активности ове мреже.

META-ВИЗИЈА (META-VISION) подстиче заједницу динамичних и утицајних заинтересованих страна да се удруже око заједничке визије и заједничког стратешког истраживачког плана (Strategic Research Agenda – SRA). Њен главни задатак је да изгради кохерентну и повезану заједницу за језичке технологије у Европи повезујући представнике неповезаних и разноврсних заинтересованих група. Ова бела књига припремљена је заједно са 29 томова за друге језике. Заједничка визија технологије развијена је у три групе, по секторима.

META-РАЗМЕНА (META-SHARE) ствара отворене, широко распрострањене погодности за заједничко коришћење и размену ресурса. Мрежа репозиторијума *једнак с једнаким* (peer-to-peer) садржаће језичке податке, алате и веб услуге документоване метаподацима високог квалитета и организоване у стандардизоване категорије. Ресурсима се може у сваком тренутку приступити, а претражују се на униформан начин. Расположиви ресурси укључују материјале отвореног кода, слободне за коришћење, али и комерцијално доступне компоненте.

META-ИСТРАЖИВАЊЕ (META-RESEARCH) успоставља мостове ка релевантним сродним технолошким областима. Ова активност настоји да искористи напредак у другим областима и да употреби иновативна истраживања која могу да допринесу језичким технологијама. Посебно, она се фокусира на спровођење најсавременијих истраживања у аутоматском превођењу, прикупљању података и организовању језичких ресурса за потребе евалуације, састављање инвентара алата и метода и организовање радионица и обука за чланове заједнице.

office@meta-net.eu – http://www.meta-net.eu

EXECUTIVE SUMMARY

During the last 60 years, Europe has become a distinct political and economic structure, yet culturally and linguistically it is still very diverse. From Portuguese to Polish and Italian to Icelandic, everyday communication between Europe's citizens as well as communication in the spheres of business and politics is inevitably confronted by language barriers. The EU's institutions spend about a billion euros a year on maintaining their policy of multilingualism, i. e., translating texts and interpreting spoken communication. Yet does this have to be such a burden? Modern language technology and linguistic research can make a significant contribution to pulling down these linguistic borders. When combined with intelligent devices and applications, language technology will in the future be able to help Europeans talk easily to each other and do business with each other even if they do not speak a common language.

Language technology builds bridges for Europe's future.

Major trade partners of Serbia come from the EU, with a share of over 50% in its total trade, while exports to the EU market are free-of-customs according to the Stabilisation and Association Agreement. But language barriers can bring business to a halt, especially for SMEs who do not have the financial means to reverse the situation. The only (unthinkable) alternative to this kind of multilingual Europe would be to allow a single language to take a dominant position and end up replacing all other languages.

One classic way of overcoming the language barrier is to learn foreign languages. Yet without technological support, mastering the 23 official languages of the member states of the European Union and some 60 other European languages is an insurmountable obstacle for the citizens of Europe and its economy, political debate, and scientific progress.

The solution is to build key enabling technologies. These will offer European actors tremendous advantages, not only within the common European market but also in trade relations with third countries, especially emerging economies. To achieve this goal and preserve Europe's cultural and linguistic diversity, it is necessary to first carry out a systematic analysis of the linguistic particularities of all European languages, and the current state of language technology support for them. Language technology solutions will eventually serve as a unique bridge between Europe's languages.

Language technology as a key for the future.

The automated translation and speech processing tools currently available on the market still fall short of this ambitious goal. The dominant actors in the field are primarily privately-owned for-profit enterprises based in Northern America. Already in the late 1970s, the EU realised the profound relevance of language technology as a driver of European unity, and began funding its first research projects, such as EUROTRA. At the same time, national projects were set up that generated valuable results but never led to concerted Euro-

pean action. In contrast to this highly selective funding effort, other multilingual societies such as India (22 official languages) and South Africa (11 official languages) have recently set up long-term national programmes for language research and technology development.

The predominant actors in LT today rely on imprecise statistical approaches that do not make use of deeper linguistic methods and knowledge. For example, sentences are automatically translated by comparing a new sentence against thousands of sentences previously translated by humans. The quality of the output largely depends on the amount and quality of the available sample corpus. While the automatic translation of simple sentences in languages with sufficient amounts of available text material can achieve useful results, such shallow statistical methods are doomed to fail in the case of languages with a much smaller body of sample material or in the case of sentences with complex structures.

The European Union has therefore decided to fund projects such as EuroMatrix and EuroMatrixPlus (since 2006) and iTranslate4 (since 2010), which carry out basic and applied research and generate resources for establishing high quality language technology solutions for all European languages. Analysing the deeper structural properties of languages is the only way forward if we want to build applications that perform well across the entire range of Europe's languages.

European research in this area has already achieved a number of successes. For example, the translation services of the European Union now use MOSES open-source machine translation software that has been mainly developed through European research projects. A substantial breakthrough in the area of speech synthesis and recognition in Serbian was made by a group from the Faculty of Technical Sciences at the University of Novi Sad. Various applications in the fields of TTS and ASR have been developed based on the speech and lexical databases with accentuated word forms. Serbian speech recognition and generation has been commercialised by the AlfaNum company, a spin-off of the University of Novi Sad. The AlfaNum company has a considerable number of users among Serbian companies. The first corpus of contemporary Serbian, an electronic morphological dictionary of Serbian, aligned French-Serbian and English-Serbian corpora of literary texts, as well as different software tools were developed in the scope of joint projects of the Faculty of Mathematics and the Department of Serbian at the Faculty of Philology in Belgrade.

Language Technology helps unify Europe.

Drawing on the insights gained so far, it appears that today's 'hybrid' language technology mixing deep processing with statistical methods will be able to bridge the gap between all European languages and beyond. As this series of white papers shows, there is a dramatic difference between Europe's member states in terms of both the maturity of the research and in the state of readiness with respect to language solutions. Serbian is one of the 'smaller' European languages, and it needs further research before truly effective language technology solutions are ready for everyday use.

META-NET's long-term goal is to introduce high-quality language technology for all languages in order to achieve political and economic unity through cultural diversity. The technology will help tear down existing barriers and build bridges between Europe's languages. This requires all stakeholders – in politics, research, business, and society – to unite their efforts for the future.

This white paper series complements other strategic actions taken by META-NET (see the appendix for an overview). Up-to-date information such as the current version of the META-NET vision paper [2] or the Strategic Research Agenda (SRA) can be found on the META-NET Website: http://www.meta-net.eu.

2

LANGUAGES AT RISK: A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

The digital revolution is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology has helped to automate and facilitate many processes:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint has replaced overhead projector transparencies;
- E-mail allows documents to be sent and received more quickly than using a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- Web search engines provide keyword-based access;
- online services like Google Translate produce quick, approximate translations;
- social media platforms such as Facebook, Twitter and Google+ facilitate communication, collaboration, and information sharing.

Although these tools and applications are helpful, they are not yet capable of supporting a fully-sustainable, multilingual European society in which information and goods can flow freely.

2.1 LANGUAGE BORDERS HOLD BACK THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. However, there is a strong likelihood that the revolution in communication technology is bringing together people who speak different languages in new ways. This is putting pressure both on individuals to learn new languages and especially on developers to create new technologies to ensure mutual understanding and access to shareable knowledge. In the global economic and information space, there is increasing interaction between different languages, speakers and content thanks to new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, Google+) is only the tip of the iceberg.

> The global economy and information space confronts us with different languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language that we do not understand. According to a report from the European Commission, 57% of Internet users in Europe purchase goods and services in non-native languages; English is the most common foreign language followed by French, German and Spanish. 55% of users read content in a foreign language while 35% use another language to write e-mails or post comments on the Web [3]. A few years ago, English might have been the lingua franca of the Web – the vast majority of content on the Web was in English – but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded.

Surprisingly, this ubiquitous digital linguistic divide has not gained much public attention. Yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

2.2 OUR LANGUAGES AT RISK

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our modern languages?

> The variety of languages in Europe is one of its richest and most important cultural assets.

Europe's approximately 80 languages are one of our richest and most important cultural assets, and a vital part of this unique social model [4]. While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many languages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the goal of ensuring equal participation for every citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [5].

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investments in language preservation focused primarily on language education and transla-

tion. According to one estimate, the European market for translation, interpretation, software localisation and Website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum [6]. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport and energy needs among others.

Language technology targeting all forms of written text and spoken discourse can help people to collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us already today to:

- find information with a search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- follow the spoken directions of a navigation system;
- translate Web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core enabling technologies are for each European language.

Europe needs robust and affordable language technology for all European languages.

To maintain our position in the frontline of global innovation, Europe will need language technology, tailored to all European languages, that is robust and affordable and can be tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image of a text using a suitably powered printing press. Human beings had to do the hard work of looking up, assessing, translating, and summarising knowledge. We had to wait until Edison to record spoken language – and again his technology simply made analogue copies.

Language technology can now simplify and automate the processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management, as well as content production, in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for specialised domains, and often exhibit limited performance. However, there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, edutainment packages, libraries, simulation environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas in which language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion

trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology helps overcome the "disability" of linguistic diversity.

Language technology represents a tremendous opportunity for the European Union. It can help to address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. However, citizens need to communicate across the language borders of the European Common Market, and language technology can help overcome this final barrier, while supporting the free and open use of individual languages. Looking even further ahead, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to support their own multilingual communities. Language technology can be seen as a form of "assistive" technology that helps overcome the "disability" of linguistic diversity and makes language communities more accessible to each other. Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: Future intelligent robots with cross-lingual language capabilities have the potential to save lives.

2.5 CHALLENGES FACING LANGUAGE TECHNOLOGY

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required. Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technological challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

Technological progress needs to be accelerated.

2.6 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to process different tongues, let's look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between their parents, siblings and other family members. From the age of about two, children produce their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear.

Learning a second language at an older age requires more cognitive effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples.

Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules.

Moving now to language technology, the two main types of systems acquire language capabilities in a similar manner. Statistical (or data-driven) approaches obtain linguistic knowledge from vast collections of concrete example texts. While it is sufficient to use text in a single language for training, e. g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then learns patterns of how words, short phrases and complete sentences are translated.

This statistical approach usually requires millions of sentences to boost performance quality. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate, all rely on statistical approaches. The great advantage of statistics is that the machine learns quickly in a continuous series of training cycles, even though quality can vary randomly.

The second approach to language technology, and to machine translation in particular, is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than 20 years. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. However, due to the high cost of this work, rule-based language technology has so far only been developed for a few major languages.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focuses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology, particularly in Europe's economic and information space. Although this technology has made considerable progress in the last few years, there is still huge potential to improve the quality of language technology systems. In the next section, we describe the role of Serbian in European information society and assess the current state of language technology for the Serbian language.

3

THE SERBIAN LANGUAGE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

Standard Serbian is the standard national language of Serbs and the official language in the Republic of Serbia. It was formed on the basis of Ekavian and Ijekavian Neo-Štokavian South Slavic dialects and its form was determined by the reformer of the written language of the Serbs Vuk Karadžić (1787–1864), who at the same time reformed both the Cyrillic alphabet and orthography. In the 20th century, in the federal state of Yugoslavia, this language was officially encompassed by *Serbo-Croatian*, a name that implied a linguistic unity with Croats (and later with other nations whose languages were based on Neo-Štokavian dialects). In the last decade of the 20th century in Serbia the name Serbo-Croatian was replaced in general usage by the name Serbian. The Constitution of the Republic of Serbia from 2006 stipulates: "The Serbian language and the Cyrillic alphabet shall be in official use in the Republic of Serbia" [7].

According to the 2002 census the population of Serbia is 7,498,001, [8] and Serbian is the mother tongue of 88.3% of the population [9]. To this number one should add the ethnic Serb population in other parts of the former Yugoslavia (a number not easy to determine). The Serbian diaspora, mainly consisting of people who left the country in search of work abroad and economic migrants, lives primarily in a number of countries of Central and Western Europe, in the USA, Canada and Australia (their knowledge of Serbian is mainly determined by the generation of immigrants they belong to). According to the 2002 census the majority of Serbs abroad live in Germany (102,799), followed by Austria (87,844) and Switzerland (65,751).

> Standard Serbian is the standard national language of Serbs and the official language in the Republic of Serbia.

Serbia is a multilingual community. The ethnic minorities, [10] according to the 2002 census, are Hungarians (3.91%), Bosniaks (2.1%), Roma (1.44%), Croats (0.94%), Montenegrins (0.92%), Albanians (0.82%), Slovaks (0.79%), Yugoslavs (1.08%) and other ethnic minorities (Ashkali/Balkan Egyptians, Bulgarians, 'Bunjevci', Aromanians, Czechs, 'Gorani', Jews, Macedonians, Germans, Muslims, Romanians, Ruthenians, Slovenians, Turks, Ukrainians and Wallachians, 2.45%). The structure of the minority nationals according to language is the following: Hungarian 3.8%, Bosnian 1.8%, Roma 1.1%, Albanian 0.8%, Slovak 0.8%, Wallach 0.7%, Romanian 0.5%, Croatian 0.4%, Bulgarian 0.2% and Macedonian 0.2%. The remaining languages are spoken by 0.5% of the population, whereas for 0.8% of the population these data are unknown. In Serbia, primary and secondary school education exists in some of the minority languages, namely in Albanian (55 primary/4 secondary schools), Hungarian (108/38), Bulgarian (26/-), Romanian (27/2), Ruthenian (3/2), Slovak (15/2) and Croatian (7/1) [11]. In addition to in-

struction, textbooks and readers are published in these languages (for example, in 2005 a total of 526 textbooks for primary and 283 for secondary school were published) [9].

Official use of minority languages is regulated by the Law on the Official use of Language and the Alphabet [12], which provides that laws and legal acts are issued in languages of ethnic minorities. This includes the right to address government authorities in one's own language, as well as the right to be answered in that language (depending on the size of the minority community).

Translations to and from Serbian represent an important activity. During 2010 a total of 2,549 books were translated (1,438 from English, 215 from French, 170 from German, 191 from Italian, 74 from Spanish, 149 from Hungarian). Part of the translations are from Slavonic languages (225 from Russian, 4 from Czech, 13 from Polish, 21 from Slovak, 19 from Slovenian, 18 from Macedonian, 12 from Bulgarian). As for translations from Serbian into other languages, 591 works were published in 2010.

3.2 PARTICULARITIES OF THE SERBIAN LANGUAGE

Serbian has its specific features which make its computational processing a complex task.

3.2.1 Phonetics, phonology, morphophonology

The vowel system is simple (five vowels), but the consonant system is rather complex (twenty five consonants). The vibrant *r* in some positions is pronounced as a vowel and functions as a syllable nucleus, e. g., *prst* ("finger") or *vrsta* ("species"). There is a large number of morphophonemic alternations in inflection and word formation, which are in some grammatical cases combined in such a way that two forms of a word can be very distant, e. g., the nominative singular of the noun "misao" is *misao* ('thought') whereas its instrumental singular is *mišlju* (alternations *a/ø, o/l, l+j/lj/ s/š*).

The accent system, comprising four accents, is based on two cross-related parameters: length opposition (long : short) and tone opposition (rising : falling). The distribution of rising and falling accents follow special rules. Accentual alternations are common in inflection and word formation. As accent marks are not used, written texts contain homographs. For example *luk* with a short falling accent means "onion", whereas with a long falling accent it means "arc" or "bow".

For many words and grammatical forms, the codified norm prescribes the pronunciation of post-accentual lengths, but they are increasingly disregarded in current usage. Almost all words have an accent, but clitics also exist: proclitics (the majority of conjunctions and prepositions and the negative particle *ne* before verbs) and enclitics (non-accentuated forms of pronouns and verbs and the interrogative particle *li*).

As for borrowed words (borrowings), their pronunciation is phonetically adapted to Serbian. However, combinations of phonemes (primarily consonants) in borrowings often deviate from those typical of original Serbian words, e. g., *softver 'software'*, *hardver 'hardware'*, *interfejs 'interface'*. In every day use, deviations from the normative distribution of accents in Serbian can also be found.

For a certain number of lexemes and word forms there are two different pronunciations, Ekavian and Ijekavian, etymologically related to the old Slavic vowel called *jat*, as shown in Figure 1.

3.2.2 Morphology

There are ten parts of speech (word classes), with a large number of subclasses. The systems of pronouns and numerals are especially complex. The article does not exist.

		Ekavian	Ijekavian
"flower"	**singular**	*cvet* (long **e**)	*cvijet*
	plural	*cvetovi* (short **e**)	*cvjetovi*

1: Ekavian and Ijekavian variant of pronunciation

Nouns are classified according to grammatical gender (masculine, feminine or neuter). However, classification according to semantic gender (male, female) is also relevant, e. g., the noun *gazda* (*'boss'*) declines like a feminine gender noun but designates a male person.

Verbs are classified according to verbal aspect (perfective or imperfective). A certain number of verbs have both aspects. There are several types of so called reflexive verbs. There are three types of inflection: (a) declension (nouns are inflected for number and case (as shown in Figure 2), while adjectives are inflected for gender, number, case and adjectival aspect); (b) conjugation (which is highly complex); and (c) comparison (gradable adjectives and adverbs). Within all three types of inflection there are different paradigms, with a number of exceptions. Inflection is accompanied by numerous morphophonemic and accentual alternations. The large number of identical forms, namely formal syncretism (morphological homonymy), should be pointed out. In all types of inflection, formal syncretism of certain grammatically different word forms is not uncommon. As a consequence of inflection, for a dictionary of 120,000 lemmas, at least 4.5 million inflected grammatical forms exist (however, there are fewer surface forms, as some forms in certain paradigms are identical).

Personal pronouns (including the reflexive pronoun) and the auxiliary, copulative and existential verb "jesam", as well as the auxiliary verbs "biti" and "hteti" have enclitic forms, which are used much more frequently than the corresponding stressed forms. For example, the dative singular of the masculine and neuter third person pronoun reads as follows: *njemu* (accentuated form) and *mu* (enclitic form).

Where nouns, verbs and adjectives are concerned, there is a highly developed suffixial word formation. With verbs, prefixation is also well developed (mainly related to aspectual meanings). Composition, on the whole, is less developed.

Calques and coinages, as well as so-called exocentric noun compounds, are frowned upon by language purists, as something that is not characteristic of authentic Štokavian word formation. This attitude complicates lexical and terminological elaboration through word formation, and is one of the reasons for the very large number of borrowings.

Borrowings fit into existing morphological and formational types, but there are also some exceptions, e. g., some foreign words do not inflect, such as the nouns *Meri* (Mary) or *skvo* (squaw), or the adjectives *fer* (fair) or *braon* (brown).

Well developed word formation (suffixation, prefixation, and, to a lesser extent, composition and various combined word formation processes) results in the fact that the majority of lexemes can be grouped into word families, and nested entries in dictionaries. It is very important that part of the formational relations lead to systematic (categorial) modification of the initial word, which greatly facilitates the lexicographic processing of such cases. For example, for the word "glumac" ('actor'), the diminutive is "glumčić" and the augmentative "glumčina", the female form is "glumica", and the adjectives are "glumčev", "glumičin", "glumački", etc.

	singular	paucal (2-4)	plural
"window" (masc.)	*prozor*	*prozora*	*prozori*
"egg" (neut.)	*jaje*	*jajeta*	*jaja*
"woman" (fem.)	*žena*	*žene*	
"news" (fem.)	*vest*	*vesti*	

2: Four types of nominal inflection

Borrowings are, in general, phonologically and morphologically adapted, that is, adjusted to the pronunciation and morphology of Serbian. They also form word families according to Serbian word formation rules.

3.2.3 Lexis, phraseology, terminology, onomastics

The composition of the vocabulary reflects, on the one hand, the fact that it is based on the Štokavian dialect, not only with regard to the original inventory but also with regard to new words formed according to Štokavian word formation processes. On the other hand, the vocabulary reflects the cultural and linguistic history of the Serbian people, including borrowings from Church Slavonic, Turkish ("megdan" 'battle'), Russian ("zapeta" 'comma'), German ("štrudla" 'strudel'), French ("ruž" 'lipstick'), and, especially today, English ("parking" 'parking'). In addition, there are many internationalisms based on classical languages (Greek and Latin), especially in specific fields.

In phraseology special attention should be given to idiomatic expressions and comparisons, proverbs and the like, which reflect autochthonous imagination and linguistic creativity. On the other hand, a large number of lexicalised expressions were created and are still being created by the calquing of foreign expressions, today primarily English ones.

In the field of terminology and nomenclature, Serbian has always greatly relied on foreign languages; foreign terms have either been translated, with occasional deviations from word formation norms, or borrowed, especially in the case of terminological internationalisms. Endeavours aimed at finding original Serbian solutions or adapting existing terms to Serbian have yielded some results, but cannot keep pace with the growing needs in the fields of terminology and nomenclature.

Onomastics represents an important segment of the vocabulary of Serbian, the more so as word families are also generated from these words.

3.2.4 Syntax, text linguistics

In terms of distribution of sentence constituents (subject, predicate, object, etc.), Serbian belongs to SVO languages with free word order (more precisely, with free distribution of mobile sentence constituents). This means that, in general, all permutations of mobile sentence constituents are permitted, but that the preferred order is: subject – predicate – object. However, free does not mean anarchic; on the contrary, the selection of a particular order is based on a very complex functional system, i. e., regulated by combinations of various syntactic, semantic, pragmatic and stylistic factors. Consider, for example, the sentence:

Marija dade Jovanu jabuku. [Mary gave John an apple.]

In Serbian, this idea can be expressed in 24 = 4! = 1*2*3*4 (number of permutations of four words) different ways:

- *Marija dade Jovanu jabuku.*
- *Marija dade jabuku Jovanu.*
- *Marija Jovanu dade jabuku.*
- *Marija jabuku dade Jovanu.*
- *Jovanu dade Marija jabuku.*
- *Jovanu Marija dade jubuku.*
- *Jabuku Marija dade Jovanu.*
- *Jabuku Jovanu dade Marija.*
- *Dade Marija jabuku Jovanu.*
- *Dade Jovanu jabuku Marija*, etc.

Certain constituents are also expressed by enclitics, which are distributed according to very specific rules.

Subject pronouns need not be expressed; instead, they can be implied (the so-called zero subject). For example; ***Ja** se zovem Marko* vs. *Zovem se Marko* (*'My name is Marko'*). A considerable number of sentence patterns are formed with various types of semantic subjects.

Besides the active and passive voice, there is another special way of formulating sentences with a non-specified human agent by using a reflexive form of the verb.

Negation is applied both to the verb and to the pronominal constituent (so-called double negation), e. g., *Ovde **ne** poznajem **nikog*** (*'I don't know anybody here'*).

There are seven cases: nominative, genitive, dative, accusative, vocative, instrumental and locative (see Figure 3). There are five oblique cases in Serbian, which can all be combined with prepositions (the locative always is). All these cases and prepositional phrases are polysemous. Conversely, the same meaning can occasionally be expressed by different cases or prepositional phrases (case synonymy). There are also a number of expressions functioning as prepositions, e. g., *prilikom* (+ genitive) *'on the occasion of'*.

In Serbian, there is a well-developed system of personal verb forms for expressing temporal and modal meanings (the aspect is the classification category); all these forms are polysemous. One of the features of the verb system is that the construction *da* + present tense increasingly tends to supplant the infinitive.

Agreement in gender, number, case and person is one of the characteristic aspects of Serbian syntax, and it is also important for establishing textual cohesion. Categorisation of agreement controllers (especially certain types of nouns, constructions with numerals and coordinated noun phrases), as well as the ways this control is expressed in different agreement positions, represents an extremely complex area.

The majority of subordinate clauses (especially relative, temporal, conditional and causal) have several formal and semantic subtypes. In the case of coordinated clauses, the inventory of conjunctions for copulative and for adversative relations is especially rich.

Relations between expressions in a text are established by various kinds of textual coordinators and textual

	singular	**paucal**	**plural**
Nominative	*prozor*	*prozora*	*prozori*
Genitive	*prozora*		*prozora*
Dative	*prozoru*		*prozorima*
Accusative	*prozor*	*prozora*	*prozore*
Vocative	*prozore*	*prozora*	*prozori*
Instrumental	*prozorom*		*prozorima*
Locative	*prozoru*		*prozorima*

3: An example of noun declension

connectors. The choice of the order of sentence constituents is important for topic-comment distribution and focus prominence. The so-called zero subject and enclitic pronoun forms are important tools for sentence contextualisation.

3.2.5 Orthography

The traditional Serbian alphabet is Cyrillic, which consists of thirty graphemes. Today the Latin alphabet is also increasingly used. It also consists of thirty graphemes (three of them digraphs) which stand in a bijective (one-to-one) relation to Cyrillic graphemes. However, the official alphabet is only Cyrillic (see Figure 4). As to the relation between the graphemic and the phonemic systems, graphemes and phonemes stand in a bijective relation to each other.

At the level of coding schemes, the Latin alphabet digraphs *lj, nj, dž* can be coded either as ligatures or as digraphs. In the first case, Unicode [13] provides special codes, for example, for the ligatures *LJ, Lj* and *lj*, whereas in the second case, as digraphs, they represent a combination of two ASCII codes, for example for *L* and *J*. This can lead to problems in transliteration, which, in general, can nevertheless be performed automatically in the majority of cases. For example, in the Serbian Wikipedia each article can be displayed both in the Cyrillic and the Latin alphabet.

The Latin alphabet does not envisage the use of the Latin characters *q, x, y, w*, nor the use of Latin characters for writing Roman numerals, which can lead to a distortion of the message when a text is transliterated from Latin to Cyrillic. Thus, for example *www* can become *ььь*, and Latin *Petar II* may become *Петар ИИ* instead of *Петар II*. Both alphabets are used in contemporary publishing. According to the data of the National Library of Serbia, a total of 12,574 monographs were published in 2010. Out of this number, 6,459 were in Cyrillic, 6,050 in Latin and 65 in other alphabets. As for daily newspapers with a wider circulation, *Politika* and *Večernje novosti* are published in Cyrillic, whereas the majority of other daily newspapers (*Blic*, *Kurir*, *Danas*, etc.) are published in the Latin alphabet.

The orthography is of a quasiphonemic type: with a few exceptions, a word is written as it is pronounced (according to the rule "Write as you speak!"), more precisely, according to its phonemic composition. The punctuation is of a logical, rather than grammatical type (akin to French and English). According to the orthographic norm, foreign words are written both in the Cyrillic and Latin alphabets the way they are pronounced, i. e., they are transcribed. Foreign names are also transcribed (e. g., instead of "Shakespeare", the proper way to write, and pronounce the name, is *Шекспир* and *Šekspir*).

Cyrillic	А	Б	В	Г	Д	**Ђ**	Е	**Ж**	З	И	Ј	К	Л	**Љ**	М
	а	б	в	г	д	**ђ**	е	**ж**	з	и	ј	к	л	**љ**	м
Latin	A	B	V	G	D	**Đ**	E	**Ž**	Z	I	J	K	L	**Lj**	M
	a	b	v	g	d	**đ**	e	**ž**	z	i	j	k	l	**lj**	m
Cyrillic	Н	**Њ**	О	П	Р	С	Т	**Ћ**	У	Ф	Х	Ц	**Ч**	**Џ**	**Ш**
	н	**њ**	о	п	р	с	т	**ћ**	у	ф	х	ц	**ч**	**џ**	**ш**
Latin	N	**Nj**	O	P	R	S	T	**Ć**	U	F	H	C	**Č**	**Dž**	**Š**
	n	**nj**	o	p	r	s	t	**ć**	u	f	h	c	**č**	**dž**	**š**

4: Serbian letters

3.2.6 Serbian and other languages of Štokavian origin

The common Štokavian basis, mutual influences and coexistence within a common state and – conceptually – within the common Serbo-Croatian language resulted in the fact that computational processing of other languages of Štokavian origin (Croatian, Bosnian, Montenegrin) has to solve similar problems. This opens great possibilities for synergy, or at least productive cooperation, as well as for a rational and economical approach to solving common problems. It is also supported by the existence of considerable resources for the former common Serbo-Croatian language (grammars and dictionaries), where, truth be told, due attention had not been paid to differences within the Štokavian standard language field. In general, the issue here is not translation from one foreign language to another, but rather *adaptation* of texts composed in standard languages with the same dialectical basis and strongly interconnected in their development. The main problems pertain, in fact, to the phenomena related to the elaboration of the Štokavian core, and especially, to the terminology.

The standard languages of Štokavian origin have to solve similar problems. This opens great possibilities for productive cooperation.

3.3 RECENT DEVELOPMENTS

The developments at the end of the 20th and the beginning of the 21st century include the following:

- Instead of common standard Serbo-Croatian there are now four national standard languages. More specifically, the official language in Serbia is now Serbian, and no longer Serbo-Croatian. Due to recent migrations resulting form wartime circumstances, the dialect picture in Croatia and Bosnia and Herzegovina (in the parts affected by war) has changed.
- Increasing changes in lexis and phraseology as well as in terminology can be observed, related to political, social and economic changes in Serbia, its opening towards the world, but also due to the harmonisation of legal acts, standards and terminology with those existent in the European Union. The influence of English can especially be observed, not only due to cultural and economic factors, which is true for other countries as well, but also due to the fact that in harmonisation with the European Union the source texts used are texts in English.
- The use of the Latin alphabet is increasing (except in official texts).
- Texts in Serbian are increasingly realised in digital form (use of computers, electronic publishing, the Internet, text messages).

3.4 OFFICIAL LANGUAGE PROTECTION IN SERBIA

3.4.1 Work on standardisation and protection of the language

We will mention here the following activities:

- In 1997 an inter-academy and inter-university body was formed as the Board for Standardisation of Serbian, [14] composed of representatives from relevant institutions from Serbia, Montenegro and the Republic of Srpska (in Bosnia and Herzegovina).
- Instead of the former Serbo-Croatian standard, the standard of Serbian is now being specified.
- There is no purism towards Croatisms (words borrowed from Croatian).
- A new Serbian orthography has been produced.

- The use of the Cyrillic alphabet is supported, as it is viewed as endangered by the Latin alphabet, especially with younger generations.
- Curricula and textbooks in primary and secondary schools are harmonised with the new language situation.

The standardisation of Serbian is institutionalised through the Board for the Standardisation, an inter-academy and inter-university body.

3.4.2 Modernisation of language norms

The Board for the Standardisation of Serbian has organised the production of a series of descriptive-normative monographs with the aim of presenting the actual state of the language and offering standardised solutions (to date the following topics have been processed: word formation, syntax and phonology). A number of standardizing recommendations have been issued. The official orthography has twice been modernised.

3.4.3 Protection of language usage

The Board for the Standardisation of Serbian (by way of its recommendations), the Society for Serbian Language and Literature (by way of its publications and by organising Serbian language competitions for students of primary and secondary schools), Matica Srpska (by organising work on the production of orthography, through its publications and by organising round tables and conferences on the Serbian language), the Foundation of Vuk Karadžić (by way of its publications and by organising round tables and conferences on the Serbian language) and various other institutions, some publishing houses, editorial boards of daily newspapers and editorial boards of radio and TV stations, as well as language experts and mother tongue enthusiasts are endeavouring to contribute to the preservation of the regularity and purity of Serbian in its written and oral usage.

3.4.4 Response to the rising influence of English

A need for substitution of English words and expressions by Serbian ones is emphasised, as well as of calqued translations from English by (authentic) Serbian words and expressions. (In a wider context, the resistance towards the increasing use of the Latin alphabet is also part of this resistance.)

3.4.5 Improvement of the situation in the field of lexicography

More and more attention is being given to lexicography, both monolingual and bilingual. A much-needed large one-volume dictionary of modern Serbian has been published. The work on the compilation of the large Serbian Academy of Sciences and Arts dictionary of Serbian is being modernised. European Union laws and regulations are being translated [15], as well as international standards, [16] including terminological standards.

3.5 LANGUAGE IN EDUCATION

The subject *Serbian Language and Literature* is one of the most important subjects in primary and secondary school. However, instruction is focused on proper writing and speech, knowledge about the language (grammar and lexis), knowledge about the history of the literary (written) languages of the Serbs and about the origin of standard Serbian. Mother tongue competitions (starting from the upper primary school grades) are based on this type of instruction. So, insufficient attention is given to the practical use of language and functional literacy. The wish to bring the goals and standards

of instruction closer to the instruction in the European Union, as well as the unsatisfactory results of students on PISA testing, serve as impulses for the modernisation of language instruction and for putting a greater emphasis on functional literacy and communication skills. This is being reflected both in the current educational reform (goals of language instruction, standards to be reached, syllabi), and in the improvement of the quality of textbooks. At the university level, there is a general shortage of courses in Serbian that would systematically prepare future experts for successful professional communication and develop appropriate functional literacy. The application of language technology methods could certainly contribute to the modernisation of instruction, for example, by way of computer-assisted language learning (CALL) systems.

3.6 INTERNATIONAL ASPECTS

The official use of the Serbian language and its instruction in neighbouring countries with Serbian ethnic minorities are regulated by the laws of these countries. The disappearance of the common Serbo-Croatian language and the official existence of distinct languages of Štokavian origin is reflected in the organisation of instruction of the former Serbo-Croatian language at universities abroad, as well as in the names of university departments where this instruction was formerly held: for these languages, hence for the Serbian language (and literature) as well, distinct curricula and diplomas now exist, with various combinations of subjects, whereas departments now have collective names. The practice of organising summer schools for foreigners continues in Serbia, but now for Serbian instead of Serbo-Croatian. Teachers from Serbia are also being sent to work as language instructors at departments abroad. Supplementary mother tongue instruction is organised in some countries for children of Serbian origin. The need for harmonisation of legal systems and terminology with those in the European Union, the influence of Anglo-American culture in the field of entertainment and the media, as well as the effects of globalisation, are contributing to increasingly closer relations between Serbian and other languages, especially English, thus giving an even greater impetus and importance to the field of translation.

3.7 SERBIAN ON THE INTERNET

A survey [17] from 2010 showed that 50.8% of the population uses the computer and the Internet on a regular basis, whereas 43.7% of the population has never used a computer. According to another source, [18] as much as 55.9% of the population uses the Internet with an increase rate of 926.8% in the period 2000–2010. According to the same source, there were 2,237,680 Facebook users in Serbia on August 31, 2010 which represents 30.5% of the total population. E-government public services are used by only 13.2% of the population, whereas 38.5% claimed they would never use such services. Trading via the Internet has been used by only 13% of the population. According to the Statistical Office of the Republic of Serbia [19], the usage of ICT equipment shows the growth.

According to the same source, the number of companies using the Internet was 96.8% in 2010 (compared to 90.2% in 2006); the number of companies having their own Website was 67.5% in 2010 (compared to 52.9% in 2006). In 2010, 70.6% of them used e-government services.

The data of the Statistical Office of the Republic of Serbia from a 2010 survey on a sample of 2,400 households and the same number of individuals aged from 16 to 74, show that 39% of respondents have an Internet connection, the highest percentage of 51% being in Belgrade [20]. Access to the Internet is income dependent, as 83% of households with a monthly income over 600 euro have Internet, while for households with a monthly

income less than 300 euro the percentage decreases to only 29%. The majority of the population accesses the global Web from desktop computers, one fifth from cell phones, and a little less from laptops.

As for connection type, almost one half of the households in Serbia that use the Internet have an ADSL connection, one quarter have cable Internet, whereas 29% of the respondents use mobile devices for connection. In the majority of cases access is from home (84%), then from work, from another person's home, from school or university, and as little as 3.8% from Internet cafés. Students are the most largely represented category on the Web, with as much as 95%. Other than for business purposes, the Internet is most commonly used for e-mail (78%), then for entertainment (games, movies, music – 55%), for reading the electronic press (41%) and for learning (23%).

The most popular Serbian Websites are Serbian news portals (Blic, [21] B92, [22] Naslovi, [23] RTS [24]). The most visited domestic portal is *Krstarica* [25], which includes a search engine, up-to-date daily news from Serbia, a directory of local sites grouped by topics and a variety of other content. An experiment initiated in 2005 with the introduction of a local search engine *Pogodak*, where the search was adjusted to the morphology of Serbian, was terminated in 2010 as unprofitable. The Serbian Wikipedia represents a source of various language data. It contains a little over 142,000 articles, and it holds the 28th position [26] in the world regarding the number of articles. The alternative Wikipedia in Serbo-Croatian [27] is smaller and contains about 40,000 articles. Free content language data projects can also be found within the portals *Rastko*, [28] *Antologija srpske književnosti* [29] (Anthology of Serbian Literature) and *Transpoetika* [30] where primarily literary texts are stored.

The visibility of a number of pages with content in Serbian has dramatically fallen during 2010, due to the change of the domain from .yu to .rs.

The most commonly used Web application is Web search, which involves automatic processing of language on multiple levels, as will be described in more detail in the second part of this paper. It involves sophisticated language technology, differing for each language. For Serbian, as we have already mentioned, the problem arises from the relation between the Cyrillic and Latin alphabets, Ekavian and Ijekavian variations, graphemic variations in the form of the lemma, as well as morphological richness.

Internet users and providers of Web content can also profit from language technology in less obvious ways, e. g., if it is used to automatically translate Web content from one language into another. In spite of the high costs associated with manually translating this content, comparatively little usable language technology is developed and applied, compared to the anticipated need. This may be due to the complexity of Serbian and the number of technologies involved in typical language technology applications. In the next chapter, we will present an overview of language technology and its core application areas as well as an evaluation of the current situation of language technology support for Serbian.

4

LANGUAGE TECHNOLOGY SUPPORT FOR SERBIAN

Language technology is used to develop software systems designed to handle human language and are therefore often called "human language technology". Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 5 illustrates the LT landscape.

When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies.

In this section, we will discuss the main application areas of language technology, i. e., language checking, Web search, speech interaction, and machine translation. These applications and basic technologies include

- spelling correction
- authoring support
- computer-assisted language learning
- information retrieval
- information extraction
- text summarisation
- question answering
- speech recognition
- speech synthesis

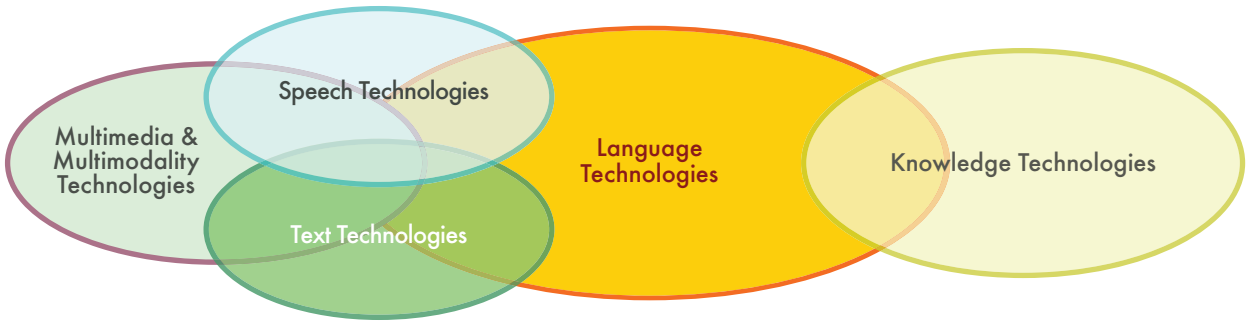
Language technology is an established area of research with an extensive set of introductory literature. The interested reader is referred to the following references: [31, 32, 33, 34].

Before discussing the above application areas, we will briefly describe the architecture of a typical LT system.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. Figure 6 shows a highly simplified architecture that can be found in a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, and detects the input language. In Serbian it can also help in resolving the Latin and Cyrillic alphabets duality, as well as the Ekavian – Ijekavian duality.



5: Language technologies

2. Grammatical analysis: finds the verb, its objects, modifiers and other sentence elements; detects the sentence structure.

3. Semantic analysis: performs disambiguation (i. e., computes the appropriate meaning of words in a given context); resolves anaphora (i. e., which pronouns refer to which nouns in the sentence) and substitute expressions; represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups. This is a simplified and idealised description of the application architecture, and illustrates the complexity of LT applications.

In the remainder of this section, we firstly introduce the core application areas for language technology, and follow this with a brief overview of the state of LT research and education today, and a description of past and present research programmes. Finally, we present an expert estimate of core LT tools and resources for Serbian in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Serbian language is summarised in Figure 11 (p. 71) at the end of this chapter. This table lists all tools and resources that are boldfaced in the text.

4.2 CORE APPLICATION AREAS

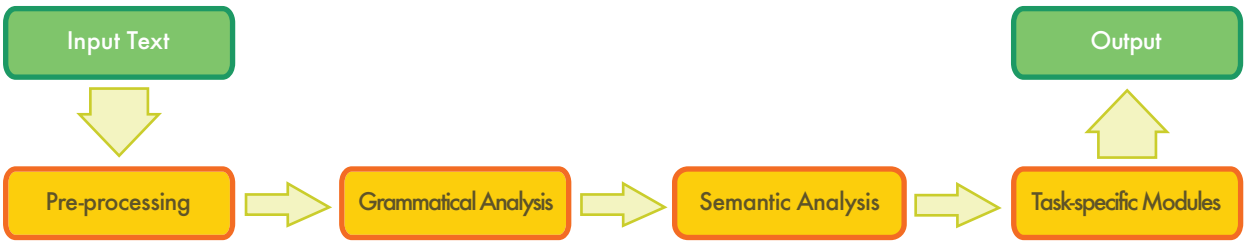
In this section, we focus on the most important LT tools and resources, and give an overview of LT activities in Serbia.

4.2.1 Language checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spell checker that highlights spelling mistakes and proposes corrections. The first spelling correction programs compared a list of extracted words against a dictionary of correctly spelled words. Today these programs are far more sophisticated. Using language-dependent algorithms for **grammatical analysis**, they detect errors related to morphology (e. g., plural formation) as well as syntax–related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spell checkers will not find any errors in the following text [35]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Handling these kinds of errors usually requires an analysis of the context. For example, whether a word needs to be capitalised in Serbian or not:



6: A typical text processing architecture

- Divio se *Ruži.* [He admired *Rose*.]
- Divio se *ruži.* [He admired *the rose*.]

This type of analysis either needs to draw on language-specific **grammars** laboriously coded into the software by experts, or on a statistical language model. In this case, a model calculates the probability of a particular word occuring in a specific position (e. g., between the words that precede and follow it). For example, *plava laguna* (blue lagoon) is a much more probable word sequence than *plava Laguna* (where Laguna is the name of a publishing house). A statistical language model can be automatically created by using a large amount of (correct) language data, a **text corpus**. These two approaches have been mostly developed around English language data. Neither approach can be transferred easily to Serbian, because the language has a flexible word order and rich inflection.

Language checking is not limited to word processors but also applies to authoring systems.

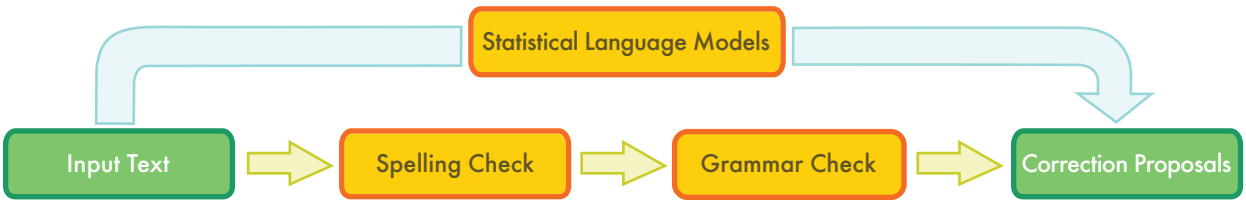
The first attempts to develop spelling checking software for Serbian dates back to the end of the 1970s [36], motivated by problems confronted by large publishing houses. To date, free spelling checking modules for Serbian are available for OpenOffice [37] on different operating systems, and there exists also a custom-made product, the RAS package, [38] developed by the Srbosof company (individualised installation).

Language checking is not limited to word processors; it is also used in "authoring support systems", i. e., software environments in which manuals and other documentation are written to special standards for complex IT, healthcare, engineering and other products. To offset customer complaints about incorrect use and damage claims resulting from poorly understood instructions, companies are increasingly focusing on the quality of technical documentation while targeting the international market (via translation or localisation) at the same time. Advances in natural language processing have led to the development of authoring support software, which helps the writer of technical documentation to use vocabulary and sentence structures that are consistent with industry rules and (corporate) terminology restrictions.

Besides spelling checkers and authoring support, language checking is also important in the field of computer-assisted language learning. And language checking applications also automatically correct search engine queries, as found in Google's *Did you mean...* suggestions.

4.2.2 Web Search

Searching the Web, intranets or digital libraries is probably the most widely used yet largely underdeveloped language technology application today. The Google search



7: Language checking (top: statistical; bottom: rule-based)

engine, which started in 1998, now handles about 80% of all search queries [39]. The verbs *guglati/izguglati* are in common use in Serbian. The Google search interface and results page display has not significantly changed since the first version. However, in the current version, Google offers spelling correction for misspelled words and incorporates basic semantic search capabilities that can improve search accuracy by analysing the meaning of terms in a search query context [40]. The Google success story shows that a large volume of data and efficient indexing techniques can deliver satisfactory results using a statistical approach to language processing.

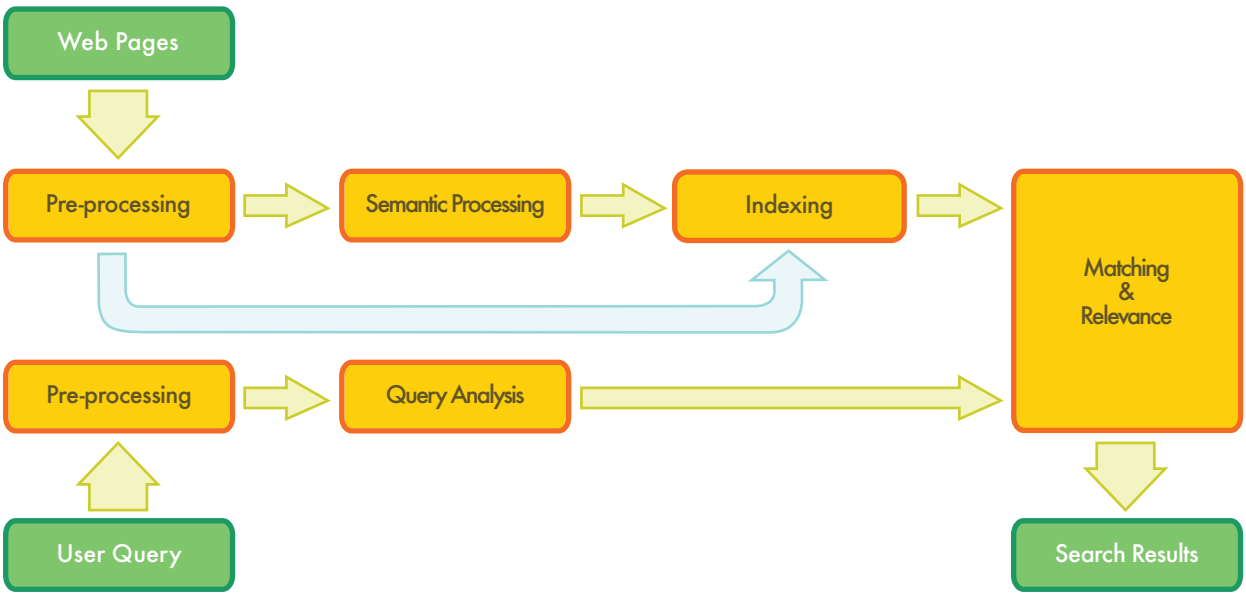
For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge to facilitate semantical analysis. Experiments using **lexical resources** such as machine-readable thesauri or ontological language resources (e. g., WordNet for English or SrpNet for Serbian) have demonstrated improvements in finding pages using synonyms of the original search terms, such as *atomska energija* (atomic energy) and *nuklearna energija* (nuclear energy), or even more loosely related terms, such as *beli luk* and *češnjak* (synonyms for garlic).

The next generation of search engines will have to include much more sophisticated language technology.

The next generation of search engines will have to include much more sophisticated language technology, especially to deal with search queries consisting of a question or other sentence type rather than a list of keywords. For the query, *Give me a list of all companies that were taken over by other companies in the last five years*, a syntactic as well as **semantic analysis** is required. The system also needs to provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, rather than companies that have acquired other companies. For the expression *last five years*, the system needs to determine the relevant range of years, taking into account the present year. The query then needs to be matched against a huge amount of unstructured data to find the pieces of information that are relevant to the user's request. This process is called information retrieval, and involves searching and ranking relevant documents. To generate a list of companies, the system also needs to recognise a particular string of words in a document represents a company name, using a process called named entity recognition.

A more demanding challenge is matching a query in one language with documents in another language. Cross-lingual information retrieval involves automatically translating the query into all possible source languages and then translating the results back into the target language.

Now that data is increasingly found in non-textual formats, there is a need for services that deliver multime-



8: Web search

dia information retrieval by searching images, audio files and video data. In the case of audio and video files, a speech recognition module must convert the speech content into text (or into a phonetic representation) that can then be matched against a user query.

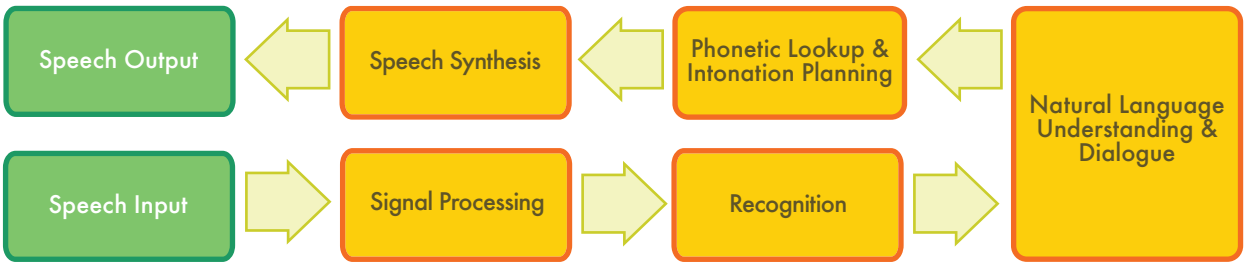
Popular sites in Serbia offering search capabilities, such as B92 and Krstarica, mostly rely on Google services [41]. An attempt to introduce a search engine which would perform exclusively a top-down search of the .rs domain, and which would partly be adjusted to the specific features of Serbian, was abandoned in 2010 as unprofitable. A certain number of SMEs is working on the enhancement of search services, albeit mainly for foreign partners and for English.

For research purposes, experiments have been performed with query expansion, by sending queries expanded on the basis of morphological dictionaries and multilingual semantic networks to search engines. The experiments yielded interesting and useful results in various domains.

4.2.3 Speech Interaction

Speech interaction is one of many application areas that depend on speech technology, i. e., technologies for processing spoken language. Speech interaction technology is used to create interfaces that enable users to interact in spoken language instead of a graphical display, keyboard and mouse. Today, these voice user interfaces (VUI) are used for partially or fully automated telephone services provided by companies to customers, employees or partners. Business domains that rely heavily on VUIs include banking, supply chain, public transportation, and telecommunications. Other uses of speech interaction technology include interfaces to car navigation systems and the use of spoken language as an alternative to the graphical or touch-screen interfaces in smartphones. Speech interaction technology comprises four technologies:

1. Automatic **speech recognition** (ASR) determines which words are actually spoken in a given sequence of sounds uttered by a user.



9: Speech-based dialogue system

2. Natural language understanding analyses the syntactic structure of a user's utterance and interprets it according to the system in question.
3. Dialogue management determines which action to take given the user input and system functionality.
4. **Speech synthesis** (text-to-speech or TTS) transforms the system's reply into sounds for the user.

One of the major challenges of ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from **speech corpora**, i. e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly – prompted by a *How may I help you?* greeting – tend to be automated and are better accepted by users.

Speech interaction is the basis for interfaces that allow a user to interact with spoken language.

Companies tend to use utterances pre-recorded by professional speakers for generating the output of the voice user interface. For static utterances where the wording does not depend on particular contexts of use or personal user data, this can deliver a rich user experience. But more dynamic content in an utterance may suffer from unnatural intonation because different parts of audio files have simply been strung together. Today's TTS systems are getting better at producing natural-sounding dynamic utterances.

Interfaces in speech interaction have been considerably standardised during the last decade in terms of their various technological components. There has also been strong market consolidation in speech recognition and speech synthesis. The national markets in the G20 countries (economically resilient countries with high populations) have been dominated by just five global players, with Nuance (USA) and Loquendo (Italy) being the most prominent players in Europe. In 2011, Nuance announced the acquisition of Loquendo, which represents a further step in market consolidation.

The speech synthesis and recognition methods in Serbia (and in the countries of the former Yugoslavia) were developed mainly in electrical engineering environments in cooperation with phonetics experts. These early endeavours were focused on recognition of isolated phonemes. A substantial breakthrough in this area was made by a group from the Faculty of Technical Sciences

at the University of Novi Sad, when they developed, in addition to speech databases, a lexical database with more than 4,000,000 accentuated word forms for Serbian and more than 3,000,000 word forms for Croatian. Various applications in the fields of TTS and ASR have been developed based on these resources. Serbian speech recognition and generation has been commercialised by the AlfaNum company, a spin-off of the University of Novi Sad. This company is successfully conducting business activities in other countries of the former Yugoslavia as well (Croatia, Macedonia, Bosnia and Montenegro). The AlfaNum company has a considerable number of users among Serbian companies.

When translating to Serbian, Google translator also offers an elementary TTS for translation results (albeit without built-in accents).

Looking ahead, there will be significant changes, due to the spread of smartphones as a new platform for managing customer relationships, in addition to fixed telephones, the Internet and e-mail. This will also affect how speech interaction technology is used. In the long term, there will be fewer telephone-based VUIs, and spoken language apps will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepwise improvements in the accuracy of speaker-independent speech recognition via the speech dictation services already offered as centralised services to smartphone users.

4.2.4 Machine translation

The idea of using digital computers to translate natural languages can be traced back to 1946 and was followed by substantial funding for research during the 1950s and again in the 1980s. Yet **machine translation** (MT) still cannot deliver on its initial promise of providing across-the-board automated translation.

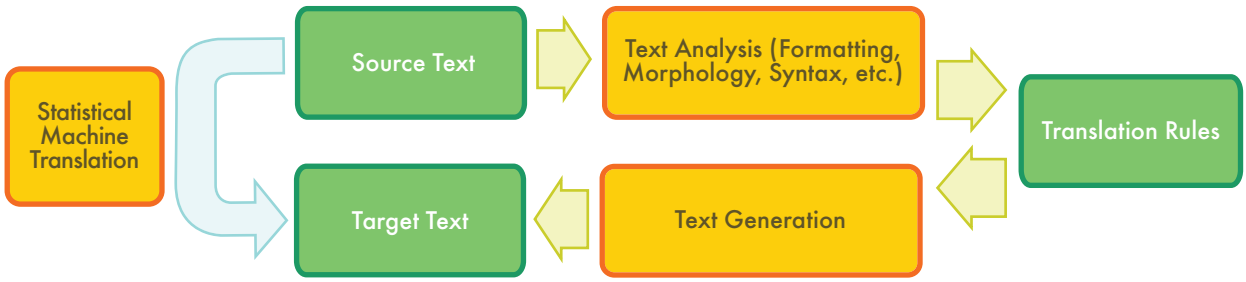
The most basic approach to machine translation is the automatic replacement of the words in a text written in one natural language with the equivalent words of another language. This can be useful in subject domains that have a very restricted, formulaic language such as weather reports. However, in order to produce a good translation of less restricted texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language.

> At its basic level, Machine Translation simply substitutes words in one natural language with words in another language.

The major difficulty is that human language is ambiguous. Ambiguity creates challenges on multiple levels, such as word sense disambiguation at the lexical level (a *jaguar* is a brand of car or an animal) or the assignment of case on the syntactic level, for example:

- *Policajac je uspeo da primeti čoveka bez dvogleda.* (The policeman caught sight of the man without binoculars.)
- *Policajac je uspeo da primeti čoveka bez revolvera.* (The policeman caught sight of the man without the revolver.)

One way to build an MT system is to use linguistic rules. For translations between closely related languages, a translation using direct substitution may be feasible in cases such as the above example. However, rule-based (or linguistic knowledge-driven) systems often analyse the input text and create an intermediary symbolic representation from which the target language text can be generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by skilled linguists. This is a very long and therefore costly process.



10: Machine translation (left: statistical; right: rule-based)

In the late 1980s when computational power increased and became cheaper, interest in statistical models for machine translation began to grow. Statistical models are derived from analysing bilingual text corpora, **parallel corpora**, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages.

Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text by processing parallel versions and finding plausible patterns of words. Unlike knowledge-driven systems, however, statistical (or data-driven) MT systems often generate ungrammatical output. Data-driven MT is advantageous because less human effort is required, and it can also cover special particularities of the language (e. g., idiomatic expressions) that are often ignored in knowledge-driven systems.

The strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid approaches that combine both methodologies. One such approach uses both knowledge-driven and data-driven systems, together with a selection module that decides on the best output for each sentence. However, results for sentences longer than, say, 12 words, will often be far from perfect.

A more effective solution is to combine the best parts of each sentence from multiple outputs; this can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

Machine translation is particularly challenging for the Serbian language.

When it comes to the relation between Serbian and other foreign languages, the problems depend on the nature of the specific language (whether its morphology is developed or not, whether it has a free or fixed distribution of sentence constituents, whether it possesses an article or not, whether it is written in the Cyrillic or Latin alphabet, whether it uses logical or grammatical punctuation, etc.) However, there is not only an issue of problems here, but also of possibilities for cooperation in solving similar problems. In that sense, cooperation with projects related to computational processing of other Slavonic languages is especially useful. However, lexical-terminological relations are also important, namely, the extent to which a foreign language has influenced the elaboration of Serbian. In this field, cooperation should be sought with projects aimed at computational processing of languages which have served and are still serving as the backbone for the elaboration of Serbian, notably, English, French, German and Russian.
It should also be added that contrastive research on the relation between Serbian and some foreign languages is also taking place. However, there is unfortunately insuf-

ficient cooperation between linguists dealing with Serbian as mother tongue and those who engage in contrastive research as experts for foreign languages. Another problem is the insufficient number of large bilingual dictionaries.

The greatest need for LT in Serbia is in the area of translation. There are some specialised associations (e. g., the Association of Literary Translators of Serbia, the Association of Technical and Scientific of Serbia), some local SMEs (e. g., Elitence and Proverbum) and some foreign companies (e. g., WorldLingo) that offer professional translation services or free, phrase-based machine translation (e. g., Google Translate, WorldLingo). Some of them use proprietary electronic dictionaries in their work, while WorldLingo also offers enhanced machine translation services (Websites, texts, documents, emails, APIs, etc.).

Apart from the well-known freely available Google statistical translation systems which also include Serbian, no other MT systems have been produced for Serbian, with the exception of some preliminary work (e. g., done as part of the SEE-ERA project) and toy experimental systems.

However, generic statistical MT systems such as Google Translate support Serbian to a considerable degree, especially in translation from and into English. Nevertheless, for other language pairs, the performance is low and the results far from comprehensible, sometimes even ridiculous. This is due to the scarcity of parallel corpora that are used to train statistical MT.

The quality of MT systems is still considered to have huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories.

Evaluation campaigns help compare the quality of MT systems, the different approaches and the status of the systems for different language pairs. Figure 11 (p. 30), which was prepared in the course of the EC Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 official EU languages (Irish was not compared). The results are ranked according to a BLEU score, which indicates higher scores for better translations [43]. A human translator would achieve a score of around 80 points.

The best results (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programs and from the existence of many parallel corpora (e. g., English, French, Dutch, Spanish and German). Languages with poorer results are shown in red. These languages either lack such development efforts or are structurally very different from other languages (e. g., Hungarian, Maltese and Finnish).

4.3 OTHER APPLICATION AREAS

Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but they provide significant service functionalities "behind the scenes" of the system in question. They all form important research issues that have now evolved into individual sub-disciplines of computational linguistics.

Question answering, for example, is an active area of research for which annotated corpora have been built and scientific competitions have been initiated. The concept of question answering goes beyond keyword-based searches (in which the search engine responds by delivering a collection of potentially relevant documents) and enables users to ask a concrete question to which the system provides a single answer. For example:

> *Question: How old was Neil Armstrong when he stepped on the moon?*
>
> *Answer: 38.*

While question answering is obviously related to the core area of Web search, it is nowadays an umbrella term

for such research issues as which different types of questions exist, and how they should be handled; how a set of documents that potentially contain the answer can be analysed and compared (do they provide conflicting answers?); and how specific information (the answer) can be reliably extracted from a document without ignoring the context.

Question answering is in turn related to information extraction (IE), an area that was extremely popular and influential when computational linguistics took a statistical turn in the early 1990s. IE aims to identify specific pieces of information in specific classes of documents, such as the key players in company takeovers as reported in newspaper stories. Another common scenario that has been studied is reports on terrorist incidents. The task here consists of mapping appropriate parts of the text to a template that specifies the perpetrator, target, time, location and results of the incident. Domain-specific template-filling is the central characteristic of IE, which makes it another example of a "behind the scenes" technology that forms a well-demarcated research area, which in practice needs to be embedded into a suitable application environment.

Language technology applications often provide significant service functionalities behind the scenes of larger software systems.

Text summarisation and **text generation** are two borderline areas that can act either as standalone applications or play a supporting role. Summarisation attempts to give the essentials of a long text in a short form, and is one of the features available in Microsoft Word. It mostly uses a statistical approach to identify the "important" words in a text (i. e., words that occur very frequently in the text in question but less frequently in general language use) and determine which sentences contain the most of these "important" words. These sentences are then extracted and put together to create the summary. In this very common commercial scenario, summarisation is simply a form of sentence extraction, and the text is reduced to a subset of its sentences. An alternative approach, for which some research has been carried out, is to generate brand new sentences that do not exist in the source text. This requires a deeper understanding of the text, which means that so far this approach is far less robust. On the whole, a text generator is rarely used as a stand-alone application but is embedded into a larger software environment, such as a clinical information system that collects, stores and processes patient data. Creating reports is just one of many applications for text summarisation.

Within the aforementioned areas, highly successful experiments for Serbian are underway related to named entity extraction as a part of the information extraction problem. A speedy development of IE and QA is expected, given the extent of developed morphological dictionaries and local grammars.

There are other fields in which linguistic technology is being applied. One of them is plagiarism detection, which uses language-independent technologies, but may be enhanced with search for simple paraphrases of the text. A research along these lines for scientific articles in Serbian has been realised by CEON [44].

4.4 EDUCATIONAL PROGRAMMES

Language Technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists, among others. As a result, it has not yet acquired a fixed place in the Serbian higher education system and is largely limited to isolated courses within more general post-graduate study programmes. Paradoxically, despite this state of affairs, short research seminars on topics related to computational linguistics

for high school students are organised within the Petnica science centre [45] each year.

At the level of university studies, topics from the field of computational linguistics are present within computer science, electronics, library science, linguistics and psychology studies at the Universities of Belgrade and Novi Sad. Courses offered to students cover the basic concepts of natural language processing, but they aim to educate students for other professions. As part of undergraduate studies at the Faculty of Mathematics in Belgrade, courses in lexical analysis and text mining are offered, in addition to courses providing basic mathematical knowledge necessary in the field of natural language processing (especially statistics, algebra, and logic), whereas a greater choice of courses in the HLT field exist at the level of doctoral studies. The most comprehensive education in the HLT field is offered to students at the Department of Library Science at the Faculty of Philology, whereas at other departments students take at most one introductory course. Within Serbian language studies, education in the field of NLP is not envisaged. The Faculties of Philosophy in Belgrade and Novi Sad offer courses in psycholinguistics, where students can get acquainted with methods of statistical text processing. Methods of interest for speech processing are studied at technical faculties. None of the faculties offer a curriculum giving expertise in the field of computational linguistics or language technologies.

4.5 NATIONAL PROJECTS AND INITIATIVES

Due to various reasons the LT industry in Serbia is relatively undeveloped compared to the leading EU economies. The main driving force behind the development of LT in Serbia are mainly domestic SMEs but also some foreign companies, which sometimes provide support for the Serbian language in various LT-related applications. Since a national programme to support the development of language technologies does not exist, their development and application are often realised in an uncoordinated manner. The introduction of language technologies in Serbia follows at least three different directions: (a) through state supported scientific and technology development projects (b) through (mainly) foreign companies which, in addition to computer equipment, also offer some sort of language support, and (c) through in-house development within domestic organisations such as publishing houses and translation agencies. Except in rare cases, these three lines of activities are realised independently from each other.

On the other hand, the computer-literate population in Serbia is accustomed to using English GUIs even though some of them may not speak English. They often find the localised versions awkward and imprecise, so they are reluctant to use them. The only applications that massively use Serbian GUI are various business, financial and accountant applications including the SAP ERP system. However, there are also some examples of GUI localised by other renowned software vendors like Microsoft (e. g., MS Windows, MS Office), Google or Oracle (localisation of OpenOffice, funded in the 2008–11 period by the Ministry for Telecommunications and Information Society through a project at the Faculty of Mathematics [46]).

Interdisciplinarity has been recognised only in the latest cycle of scientific projects (for the 2011–2014 period) funded by the Ministry of Education and Science. Until 2010 scientific projects (and hence criteria for their evaluation) have been strictly divided among the fields of mathematics (including computer science as its part), language, and technological disciplines. In such a setting, it was hard to realise the natural combination of disciplines which form the basis of language technology development. In this context, it was necessary to estab-

lish connections between research in the field of Serbian language and informatics.

The first project along these lines entitled "Interactions between text and dictionaries" was conceived in 2002 as a joint project of the Departments of Serbian at the Faculty of Philology in Belgrade and the Faculty of Philosophy in Novi Sad, as well as the Faculty of Mathematics in Belgrade. In the scope of this project, the first corpus of contemporary Serbian was developed, [47] accessible via the Web, currently having more than 300 registered users from different Serbian and foreign universities and institutes. Development of an electronic morphological dictionary of Serbian following the so-called LADL format was also initiated within the scope of this project [48]. The project was later continued as a joint project of the Department of Serbian at the Faculty of Philology and the Faculty of Mathematics in the period from 2006 to 2010 under the name "A theoretical and methodological framework for the modernisation of Serbian" and from 2011 to 2014 under the name "Serbian and its resources: theory, description and applications". Within the scope of these projects, the development of the electronic dictionary of simple words was finalised, the development of a dictionary of compounds was initiated. Aligned French-Serbian and English-Serbian corpora of literary texts were developed, as well as local grammars for certain segments of Serbian (especially for named entities). Different software tools were also developed, among which special attention should be given to LeXimir, a workstation which enables integration and transformation of heterogeneous lexical resources.

Parallel with this research in the field of language, a project was funded within the social sciences field under the name "Fundamental cognitive processes and functions", realised by the Department of Psychology at the Faculty of Philosophy in Belgrade. The aim of this project, among other things, was to investigate the possibility of the automatic annotation of texts based on an annotated corpus, [49] developed during the 1950s and converted to electronic form in the 1990s.

Speech synthesis and recognition is being realised at the Faculty of Technical Sciences of the University of Novi Sad through projects of technological development from 2005, namely "Development of speech technologies in Serbian and their application in Telekom Serbia" (2005–2007), "Man-machine speech communication" (2008–2010), "Development of dialogue systems for Serbian and other South-Slavic languages" (2011–2014). They provide support for different TTS and ASR applications and services including IVR systems, private branch exchanges, call centres, audio logging, track commercials, word spotter, etc.

Other single resources of interest for HLT have been developed within other scientific areas, albeit without any direct interaction with the aforementioned projects. Let us just mention a few examples such as the Serbian-English geological thesaurus [50] and the folkloristic database DABI of the Institute of Balkan studies SASA [51].

In addition to national projects, Serbian scientific institutions have also taken part in various international projects related to the HLT field. A certain level of activities was maintained during the UN sanctions due to the participation in projects TELRI I and II [52]. Although Serbian research groups could not participate at that time in the project MULTEXT-East [53], they nevertheless produced useful resources in formats defined by that project: a morphosyntactic description of Serbian, an aligned version of the Serbian translation of Orwell's *1984*, its lemmatised morphosyntactically tagged version and a comprehensive dictionary covering *1984*'s lexicon.

After the sanctions were lifted, of particular importance was the BalkaNet [54] project which enabled the development of a WordNet type semantic network for

Serbian. The Serbian part of the multilingual lexical database of proper names Prolex [55] was developed within the scope of bilateral cooperation with France, whereas a one-million aligned English-Serbian project, lemmatised and morphologically annotated, was developed within the scope of the Intera project. This corpus was used for tagger training, as well as for experiments in alignment at the word level and in automatic translation.

The situation in various domains related to processing of Serbian differs, but there is definitely a considerable improvement in corpus development, morphological analysis, electronic dictionaries as well as NE extraction.

Serbian participants were also involved in two regional projects. One of them was SEE-ERA.NET – Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages (ICT 10503 RP, 2007–2008). Its main contribution was the development of unidirectional translation models that rely on large-scale multilingual resources, namely *The Acquis Communautaire*. However, since documents that are the base of this resource had not yet been translated into Serbian at that time no translation model was produced for Serbian. Translation of EU legislation is underway, and part of the translated material is already available [56]. For its part, the Serbian team contributed by developing another multilingual aligned resource based on Verne's novel *Around the World in 80 Days* (in 16 languages at that time). The other project was WISE – An Electronic Marketplace to Support Pairs of Less Widely Studied European Languages (BSEC 009 / 05.2007, 2007–2008) with the aim not only to produce cross-lingual lexical resources enriched with linguistic meta-data but also to develop and promote an electronic marketplace for the less widely studied Balkan languages, including Serbian.

Further activities encompass, in the first place, the development of procedures for the syntactic analysis of Serbian, which, due to the free order of words and morphological richness, represents an extremely complex task. This means that new resources need to be developed, above all, new types of dictionaries and corpora, as well as accompanying tools.

4.6 AVAILABILITY OF TOOLS AND RESOURCES

Figure 11 summarises the current state of language technology support for the Serbian language. The rating for existing tools and resources was generated by leading experts in the field who provided estimates based on a scale from 0 (very low) to 6 (very high) according to seven criteria. For Serbian, the state of resources and technologies could be described as follows:

- Where morphological issues and issues related to them are concerned, it is safe to say that the level of development of technologies and resources is satisfactory, mainly due to the existence of large electronic dictionaries and local grammars. An immediate consequence of this fact is that necessary tools for information retrieval and information extraction are available. Some of the dictionaries are ready for wider use, whereas some need to be upgraded, as for example SrpNet.
- A reference corpus of contemporary Serbian in Ekavian dialect is available, as well as several parallel aligned corpora, all of which are available to researchers of Serbian. Current research is focused on upgrading the reference corpus and expanding it with the Ijekavian variant.
- Speech technologies are well developed, and they have found wide use in business, but research needs to be further expanded, in order to expand the area of their usability.

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies and Applications)							
Speech Recognition	2	2	1	1	1	1	0
Speech Synthesis	2	2	4	4	5	5	1
Grammatical analysis	1	1	2,5	2	2	1,5	1,5
Semantic analysis	1	1	1	1,5	1	1	1,5
Language generation	0	0	0	0	0	0	0
Machine translation	1	1	0	1	0	1	1
Language Resources (Resources, Data and Knowledge Bases)							
Text corpora	0,5	1	0,5	1	1	1	0,5
Speech corpora	1	2	4	4	3	3	3
Parallel corpora	3	3	3	2	2	2	3
Lexical resources	1	2	2	2	2	2	2,5
Grammars	1	1	0	1	0	1	1

11: State of language technology support for Serbian

- Software aimed at enhancing the productivity of lexicographical work has been developed, but the issue of accepting new technologies in traditionally oriented lexicographic environments is an impediment to the speedier development of lexicography.
- Successful experiments have been performed in some areas, such as shallow parsing, summarisation, machine translation, ontological resources, in a strictly research environment. However, the results obtained are still far from the level of development reached for developed European languages. The attention of researchers is also drawn to multimedia and multimodal documents, especially in the context of the digitisation of cultural heritage.

Given the complexity of Serbian syntax, areas based on deep parsing simply do not exist: sentence semantics, text semantics, and language generation. This results in the absence of a formalised syntax of Serbian and restricts the development of syntactically and semantically annotated corpora. The formalisation of Serbian syntax is thus the most urgent task for the further expansion of HLT.

4.7 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The

languages were clustered using the following five-point scale:

- Excellent LT support
- Good support
- Moderate support
- Fragmentary support
- Weak or no support

LT support was measured according to the following criteria:

- Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications
- Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications
- Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars
- Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars

The relevant tables show that the tools and resources available for Serbian are mostly in the bottom cluster. Serbian compares well with other languages with a small number of speakers, such as Croatian, Slovenian and Slovak but these languages lag far behind more widely spoken European languages such as German and French. However, even for the latter languages LT resources and tools clearly do not yet reach the quality and coverage of comparable resources and tools for English, which is in the lead in all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications.

4.8 CONCLUSIONS

In this series of white papers, we have provided the first high-level comparison of language technology support across 30 European languages. By identifying the gaps, needs and deficits, the European language technology community and its related stakeholders are now in a position to design a large scale research and development programme aimed at building truly multilingual, technology-enabled communication across Europe.

The results of this white paper series show that there is a dramatic difference in language technology support between European languages. While there are good quality software and resources available for some languages and application areas, other (usually smaller) languages have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources.

Others have basic tools and resources, but there is little chance of implementing semantic methods in the near future. This means that a large-scale effort is needed to reach the ambitious goal of providing support for all European languages, for example through high quality machine translation.

The scope of the resources and the range of tools available for Serbian are still very limited, especially when compared to the resources and tools for languages like French, German, and especially English, and they are not sufficient in quality and quantity to develop the kind of technologies required to support a truly multilingual knowledge-based society.

Technologies already developed and optimised for English cannot be simply transferred to handle Serbian.

English-based systems for syntactic analysis of sentence structure are in general unsuitable for Serbian texts. The work on language processing for Serbian has been concentrated so far on the development of resources and tools that comply with the specific features of Serbian (in the first place a description of its rich morphology). This line of development should by all means be followed in the future.

For a rather modest language community and research environment such as the Serbian one, cooperation both on the national and international level in developing language resources is of vital importance. This is true in general for the majority of Slavic languages, and this cooperation asks for further stimulative measures. There are especially great possibilities for cooperation among projects related to standard languages of Štokavian origin, as well as Slavic languages in general, given the common specific features shared among them.

Serbia's participation in CESAR and META-NET is expected to contribute to the development, standardisation and availability of several important LT resources and thus to the development of language technology for Serbian. META-NET's long-term goal is to introduce high-quality language technology for all languages in order to achieve political and economic unity through cultural diversity. The technology will help tear down existing barriers and build bridges between Europe's languages. This requires all stakeholders – in politics, research, business, and society – to unite their efforts for the future.

The Serbian language technology industry is extremely modest. There are just a few SMEs involved and their approach is basically founded on the application of "brute force", which means that they are basically ignoring the specific features of Serbian. Our findings show that the only alternative is to make a substantial effort to create LT resources for Serbian, and use them to drive forward research, innovation and development. The need for large amounts of data and the extreme complexity of language technology systems makes it vital to develop a new infrastructure and a more coherent research organisation to stimulate greater sharing and cooperation. Another key contribution would be the establishment of multidisciplinary studies related to language processing at the master and doctoral levels, which are currently not available.

We can therefore conclude that there is a desperate need for a large, coordinated initiative focused on overcoming the differences in language technology readiness for European languages as a whole.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	German Italian Finnish French Dutch Portuguese Spanish Czech	Basque Bulgarian Danish Estonian Galician Greek Irish Catalan Norwegian Polish Swedish **Serbian** Slovak Slovenian Hungarian	Icelandic Croatian Latvian Lithuanian Maltese Romanian

12: Speech processing: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	German Italian Catalan Dutch Polish Romanian Hungarian	Basque Bulgarian Danish Estonian Finnish Galician Greek Irish Icelandic Croatian Latvian Lithuanian Maltese Norwegian Portuguese Swedish **Serbian** Slovak Slovenian Czech

13: Machine translation: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	German French Italian Dutch Spanish	Basque Bulgarian Danish Finnish Galician Greek Catalan Norwegian Polish Portuguese Romanian Swedish Slovak Slovenian Czech Hungarian	Estonian Irish Icelandic Croatian Latvian Lithuanian Maltese **Serbian**

14: Grammatical analysis: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	German French Dutch Swedish Czech Hungarian Polish Italian Spanish	Basque Bulgarian Danish Estonian Finnish Galician Greek Catalan Croatian Norwegian Portuguese Romanian **Serbian** Slovak Slovenian	Irish Icelandic Latvian Lithuanian Maltese

15: Speech and text resources: State of support for 30 European languages

ABOUT META-NET

META-NET is a Network of Excellence partially funded by the European Commission. The network currently consists of 54 research centres in 33 European countries [57]. META-NET forges META, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots. Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and Web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

ЛИТЕРАТУРА REFERENCES

[1] Aljoscha Burchard, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] Aljoscha Burchardt, Georg Rehm, and Felix Sasaki. The Future European Multilingual Information Society: Vision Paper for a Strategic Research Agenda, 2011. http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf.

[3] Directorate-General Information Society & Media of the European Commission. User Language Preferences Online, 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[4] European Commission (Europäische Kommission). Multilingualism: an Asset for Europe and a Shared Commitment, 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[5] Directorate-General of the UNESCO. Intersectoral Mid-term Strategy on Languages and Multilingualism, 2007. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

[6] Directorate-General for Translation of the European Commission. Size of the Language Industry in the EU, 2009. http://ec.europa.eu/dgs/translation/publications/studies.

[7] Constitution of the Republic of Serbia. http://www.srbija.gov.rs/cinjenice_o_srbiji/ustav.php?change_lang=en.

[8] Popis stanovništva, domaćinstava i stanova u 2002.: STANOVNIŠTVO (Census of population, households and dwellings in 2002.: POPULATION). http://webrzs.stat.gov.rs/axd/Zip/VJN3.pdf.

[9] Human Development Report – SERBIA 2005: The Strength of Diversity. http://hdr.undp.org/en/reports/national/europethecis/serbia/Serbia_nhdr_2005.pdf.

[10] http://www.ombudsman.rs/pravamanjina/index.php/sr_YU/podaci.

[11] OBRAZOVANJE (EDUCATION). http://webrzs.stat.gov.rs/WebSite/repository/documents/00/00/18/48/god2010pog22.pdf.

[12] Službeni glasnik RS, br. 45/91, 53/93, 67/93, 48/94, 101/2005 – dr. zakon i 30/2010 (Official Gazette of the Republic of Serbia, no. 45/91, 53/93, 67/93, 48/94, 101/2005 – state law and 30/2010).

[13] Unicode: Latin Extended-B. http://unicode.org/charts/PDF/U0180.pdf.

[14] Board for Standardization of the Serbian Language. http://en.wikipedia.org/wiki/Board_for_Standardization_of_the_Serbian_Language.

[15] Government of the Republic of Serbia – European Integration Office. http://www.seio.gov.rs/home.50.html.

[16] Institut za standardizaciju Srbije (Serbian Institute for Standardization). http://www.iss.rs.

[17] Republički zavod za statistiku: Upotreba informaciono-komunikacionih tehnologija (Republic Institute for Statistics: Use of information-communication technologies). http://webrzs.stat.gov.rs/WebSite/Public/PageView.aspx?pKey=204.

[18] Internet World Stats – Usage and Population Statistics: Serbia. http://www.internetworldstats.com/europa2.htm#rs.

[19] Republički zavod za statistiku (Republic Institute for Statistics). http://webrzs.stat.gov.rs/WebSite/.

[20] Republički zavod za statistiku: Upotreba informaciono-komunikacionih tehnologija u Republici Srbiji, 2010. (Republic Institute for Statistics: Use of information-communication technologies in the Republic of Serbia, 2010.). http://webrzs.stat.gov.rs/WebSite/repository/documents/00/00/10/40/PressICT2010.pdf.

[21] BLIC online. http://www.blic.rs.

[22] B 92. http://www.b92.net.

[23] naslovi.net. http://www.naslovi.net.

[24] РТС: Радио-телевизија Србије (RTS: Radio-Television Serbia). http://www.rts.rs.

[25] cruiser. http://www.krstarica.com.

[26] Wikipedia metadata. http://meta.wikimedia.org/wiki/List_of_Wikipedias.

[27] Vikipedija (Wikipedia). http://sh.wikipedia.org.

[28] Пројекат Растко: библиотека српске културе (Project Rastko: library of Serbian culture). http://www.rastko.rs.

[29] Учитељски факултет Универзитета у Београду: Антологија српске књижевности (Faculty of Teacher Education in Belgrade: Anthology of Serbian literature). http://www.ask.rs.

[30] Транспоетика (Transpoetics). http://transpoetika.org.

[31] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2 edition, 2009.

[32] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[33] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1998.

[34] Language Technology World (LT World). http://www.lt-world.org.

[35] Jerrold H. Zar. Candidate for a Pullet Surprise. *Journal of Irreproducible Results*, page 13 (Fist Verse), 1994.

[36] Zoran Urošević. *Statistička metoda otkrivanja i korekcije slovnih grešaka supstitucionog tipa u tekstu na srpskohrvatskom jeziku (Statistical method for detection and correction of typos of substitutional type in a text in Serbo-Craotian)*. BIGZ, 1975.

[37] OpenOffice: Serbian (Cyrillic and Latin) Spelling and Hyphenation. http://extensions.services.openoffice.org/en/node/1572/releases.

[38] KOPEKTOP за Word (CORRECTOR for Word). http://www.rasprog.com/html/3_0_korektor.html.

[39] Spiegel Online. Google zieht weiter davon (Google is still leaving everybody behind), 2009. http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.

[40] Juan Carlos Perez. Google Rolls out Semantic Search Capabilities, 2009. http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html.

[41] Alexa – The Web Information Company. http://www.alexa.com/topsites/countries/CS.

[42] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 Machine Translation Systems for Europe. In *Proceedings of MT Summit XII*, 2009.

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA, 2002.

[44] Centre for Evaluation in Education and Science (CEON/CEES). http://ceon.rs/index.php?option=com_content&task=view&id=224&Itemid=106.

[45] Istraživačka stanica Petnica (ISP) (Petnica Science Center). http://www.petnica.rs.

[46] Open Office. http://ooo.matf.bg.ac.rs.

[47] Resursi srpskog jezika (Serbian language resources). http://www.korpus.matf.bg.ac.rs.

[48] Cvetana Krstev. *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade, 2008.

[49] Đorđe Kostić. Corpus of Serbian Language (CSL). http://www.serbian-corpus.edu.rs/ns/eindex.htm.

[50] GeolISSTerm, Geološki Informacioni Sistem Srbije – Geološka Terminologija i nomenklatura (GeolISSTerm, Geology Information System of Serbia – Geology Terminology and nomenclature). http://www.rgf.bg.ac.rs/geolissterm/Index.aspx.

[51] Балканолошки институт, Српска академија наука и уметности (Balkans Institute, Serbian Academy of Sciences and Arts). http://www.balkaninstitut.com/srp/projekti/sikimic/stratifikacija_balkana.html.

[52] TELRI, Trans-European Language Resources Infrastructure. http://telri.nytud.hu/.

[53] MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. http://nl.ijs.si/ME/.

[54] EUROPA CORDIS: A wordnet for the Balkans. http://cordis.europa.eu/ictresults/index.cfm?section=news&tpl=article&ID=73737.

[55] Centre National de Ressources Textuelles et Lexicales (CNRTL): Prolex. http://www.cnrtl.fr/lexiques/prolex/.

[56] ЕВРОТЕКА – Енглеско-српски паралелни корпус (EVROTEKA – English-Serbian aligned corpus). http://prevodjenje.seio.gov.rs/evroteka/index.php?jezik=srpc.

[57] Georg Rehm and Hans Uszkoreit. Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):51–52, April/May 2011.

B

ЧЛАНИЦЕ META-NET
META-NET-A MEMBERS

Аустрија	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Белгија	Belgium	Computational Linguistics and Psycholinguistics Research Centre, Univ. of Antwerp: Walter Daelemans
		Centre for Processing Speech and Images, Univ. of Leuven: Dirk van Compernolle
Бугарска	Bulgaria	Institute for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
УК	UK	School of Computer Science, Univ. of Manchester: Sophia Ananiadou
		Institute for Language, Cognition and Computation, Centre for Speech Technology Research, Univ. of Edinburgh: Steve Renals
		Research Institute of Informatics and Language Processing, Univ. of Wolverhampton: Ruslan Mitkov
Грчка	Greece	R.C. "Athena", Institute for Language and Speech Processing: Stelios Piperidis
Данска	Denmark	Centre for Language Technology, Univ. of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Естонија	Estonia	Institute of Computer Science, Univ. of Tartu: Tiit Roosmaa, Kadri Vider
Ирска	Ireland	School of Computing, Dublin City Univ.: Josef van Genabith
Исланд	Iceland	School of Humanities, Univ. of Iceland: Eiríkur Rögnvaldsson
Италија	Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Language Technology Research Unit, Fondazione Bruno Kessler: Bernardo Magnini
Кипар	Cyprus	Language Centre, School of Humanities: Jack Burston
Летонија	Latvia	Tilde: Andrejs Vasiļjevs
		Institute of Mathematics and Computer Science, Univ. of Latvia: Inguna Skadiņa
Литванија	Lithuania	Institute of the Lithuanian Language: Jolanta Zabarskaitė
Луксембург	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Мађарска	Hungary	Research Institute for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Dept. of Telecommunications and Media Informatics, Budapest Univ. of Technology and Economics: Géza Németh, Gábor Olaszy

Малта	Malta	Dept. Intelligent Computer Systems, Univ. of Malta: Mike Rosner
Немачка	Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
		Human Language Technology and Pattern Recognition, RWTH Aachen Univ.: Hermann Ney
		Dept. of Computational Linguistics, Saarland Univ.: Manfred Pinkal
Норвешка	Norway	Dept. of Linguistic, Literary and Aesthetic Studies, Univ. of Bergen: Koenraad De Smedt
		Dept. of Informatics, Language Technology Group, Univ. of Oslo: Stephan Oepen
Пољска	Poland	Institute of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
		Univ. of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Dept. of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz Univ.: Zygmunt Vetulani
Португалија	Portugal	Univ. of Lisbon: António Branco, Amália Mendes
		Spoken Language Systems Laboratory, Institute for Systems Engineering and Computers: Isabel Trancoso
Румунија	Romania	Research Institute for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiș
		Faculty of Computer Science, Univ. Alexandru Ioan Cuza of Iași: Dan Cristea
Словачка	Slovakia	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Словенија	Slovenia	Jožef Stefan Institute: Marko Grobelnik
Србија	Serbia	Univ. of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
		Pupin Institute: Sanja Vraneš
Финска	Finland	Computational Cognitive Systems Research Group, Aalto Univ.: Timo Honkela
		Dept. of Modern Languages, Univ. of Helsinki: Kimmo Koskenniemi, Krister Lindén
Француска	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur and Institute for Multilingual and Multimedia Information: Joseph Mariani
		Evaluations and Language Resources Distribution Agency: Khalid Choukri
Холандија	Netherlands	Utrecht Institute of Linguistics, Utrecht Univ.: Jan Odijk
		Computational Linguistics, Univ. of Groningen: Gertjan van Noord
Хрватска	Croatia	Institute of Linguistics, Faculty of Humanities and Social Science, Univ. of Zagreb: Marko Tadić

Чешка	Czech Republic	Institute of Formal and Applied Linguistics, Charles Univ. in Prague: Jan Hajič
Швајцарска	Switzerland	Idiap Research Institute: Hervé Bourlard
Шведска	Sweden	Dept. of Swedish, Univ. of Gothenburg: Lars Borin
Шпанија	Spain	Barcelona Media: Toni Badia, Maite Melero
		Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel
		Aholab Signal Processing Laboratory, Univ. of the Basque Country: Inma Hernaez Rioja
		Center for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
		Dept. of Signal Processing and Communications, Univ. of Vigo: Carmen García Mateo

Кључне резултате и поруке серије белих књига продискутовало је и усвојило око сто експерата језичких технологија – представника земаља и језика представљених у META-HET-у, на састанку META-HET-а у Берлину, Немачка, 21-22. октобра 2011. – About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

МЕТА-НЕТ СЕРИЈА БЕЛИХ КЊИГА
THE META-NET WHITE PAPER SERIES

баскијски	Basque	euskara
бугарски	Bulgarian	български
галицијски	Galician	galego
грчки	Greek	ελληνικά
дански	Danish	dansk
енглески	English	English
естонски	Estonian	eesti
ирски	Irish	Gaeilge
исландски	Icelandic	íslenska
италијански	Italian	italiano
каталонски	Catalan	català
летонски	Latvian	latviešu valoda
литвански	Lithuanian	lietuvių kalba
мађарски	Hungarian	magyar
малтешки	Maltese	Malti
немачки	German	Deutsch
норвешки бокмал	Norwegian Bokmål	bokmål
норвешки нинорск	Norwegian Nynorsk	nynorsk
пољски	Polish	polski
португалски	Portuguese	português
румунски	Romanian	română
словачки	Slovak	slovenčina
словеначки	Slovene	slovenščina
српски	Serbian	српски
фински	Finnish	suomi
француски	French	français
холандски	Dutch	Nederlands
хрватски	Croatian	hrvatski
чешки	Czech	čeština
шведски	Swedish	svenska
шпански	Spanish	español

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the Serbian language. It is part of a series that analyzes the available language resources and technologies for 31 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations, non-governmental organisations, language communities and European universities. META-NET's vision is high-quality language technology for all European languages.

Грађани Европе, као и пословни свет и политичари суочавају се у својој свакодневној комуникацији са језичким препрекама. Оно што доносе језичке технологије је превазилажење таквих препрека и обезбеђивање нове сумеђе ка технологијама и знању уопште. Ова бела књига описује актуелни ниво подршке језичких технологија у обради српског језика. Она је део серије која анализира расположиве језичке ресурсе и технологије за 31 европски језик. Анализа је спроведена у оквиру META-NET-а, мреже изврсности коју је основала Европска комисија. META-NET повезује 54 истраживачка центра из 33 земље, који сарађују са заинтересованим странама из економије, владе, истраживачких организација, невладиних организација, језичких заједница и универзитета. Визија META-NET-а је језичка технологија високог квалитета за све европске језике.