НАУЧНИ РАД УДК 004.9:811.163.41’374

АНОТАЦИЈА КОРПУСА САВРЕМЕНОГ СРПСКОГ ЈЕЗИКА[1]

Милош Утвић[2]
Филолошки факултет,
Универзитет у Београду,
Катедра за библиотекарство и информатику

Апстракт

Овај текст описује припрему и реализацију анотације Корпуса савременог српског језика величине 113 милиона речи. Анотација је спроведена на неколико нивоа. Сваком тексту корпуса је придружена одговарајућа библиографска информација. На основу електронског морфолошког речника српског језика припремљен је скуп етикета за врсте речи, као и речник за анотацију прилагођен програму за етикетирање TreeTagger. Коришћењем програма TreeTagger и ручно анотираног корпуса INTERA величине око милион речи, извршена је аутоматска морфосинтаксичка анотација Корпуса савременог српског језика, тј. корпусним речима је придружена информација о врсти речи и леми. Применом десетоструке унакрсне провере (енг. 10-fold cross-validation) обављена је евалуација примењеног поступка.

Кључне речи: анотација, корпус, tagger, TreeTagger

---

1 Овај рад приказује резултате постигнуте током 2011. године у оквиру пројекта Српски језик и његови ресурси (178006) који финансира Министарство просвете Републике Србије и пројекта CESAR као дела шире мреже пројеката META-NET коју финансира Европска унија.

2 misko@matf.bg.ac.rs

1. Анотација корпуса - за и против

Корпус у најширем смислу се обично дефинише као колекција текстова. Иако се у корпусној лингвистици прави разлика између преелектронских и електронских корпуса у зависности од тога да ли се колекција састоји из машински читљивих текстова или не, савремени корпуси су готово искључиво електронски. Стога ће се у остатку овог чланка под корпусом углавном подразумевати електронски корпус, осим ако се не нагласи другачије. Корпусни лингвисти различито дефинишу (електронски) корпус, али се углавном слажу да је у питању „колекција машински читљивих, аутентичних текстова који представљају репрезентативни узорак појединачног језика или језичког варијетета" (Xiao, 2010).

Корпусна реч је низ карактера (корпусног) текста између два узастопна сепаратора, при чему се скуп сепаратора дефинише на различите начине, најчешће као скуп неалфанумеричких карактера. Корпусне речи и појединачни елементи скупа сепаратора се заједнички називају токенима. На основу наведене дефиниције, у реченици Ако је a=0, онда крај доказа. корпусне речи су Ако, је, a, 0, онда, крај, доказа, сепаратори су карактери размак, запета, тачка и =, док сви заједно представљају токене дате реченице.

Анотација корпуса је поступак којим се деловима корпуса (текстови, логичке целине у оквиру текста, токени) придружују додатне информације.

Придруживање информација се може остварити на неколико нивоа:

(1) тексту корпуса се могу придружити одговарајућа библиографска референца (информације о извору), податак о дужини текста (број токена и типова), подаци о креирању и ажурирању електронске верзије текста (датум настанка, да ли је у оригиналном облику текст био машински читљив или је текст из неког другог облика трансформисан у машински читљив облик и на који је начин обављена та трансформација, одговорне особе за креирање електронске верзије, исправљање грешака и сл.);
(2) у оквиру текста је могуће обележити његову логичку структуру (поглавља, наслове, пасусе, реченице);
(3) свакој корпусној речи се може придружити информација о
   (I) врсти речи (именица, придев, глагол итд.),
   (II) леми (номинатив једнине именице, инфинитив глагола итд.),
   (III) вредностима флек-тивних категорија (род, број, падеж, глаголски облик, глаголски вид итд.), односно флективној основи и наставцима;
   (IV) (творбеној) основи, префиксима, инфи-ксима и суфиксима;
   (V) начину изговора (акценат);
   (VI) границама слогова;
(4) сваком токену се може придружити ознака одговарајућег значења;
(5) низу од једне или више корпусних речи може се придружити заједничка информација о функцији у реченици (субјекат, предикат, објекат, глаголска одредба) или се такав низ може обележити као синтагма (именичка, придевска итд.).
(6) На нивоу дискурса се могу означити релације кореференције (анафо-рички и катафорички односи) између корпусних речи;
(7) говорни чин се може обележити
   (I) прагматичким и
   (II) стилистичким информацијама.

Иако већина аутора за сва наведена придруживања додатних информација корпусу

користи појам анотација корпуса (енг. corpus annotation), поједини аутори (Xiao, 2010) под тим појмом подразумевају само придруживање лингвистичких информација (лема, врста речи итд.), док се за остале (нелингвистичке) информације (библиографски подаци о тексту, оригиналном форматирању текста, итд.) користи термин означавање корпуса (енг. corpus markup). Такође, наведена придруживања (1)-(7) имају и посебне називе, па се тако (2) још назива структурна анотација/означавање (енг. structural annotation/markup), (3I) - етикетирање врстом речи (енг. Part of Speech tagging, скр. PoS tagging), (3II) - лематизација (енг. lemmatization), (3III) - граматичка анотација (енг. grammatical annotation), (4) – семантичка анотација (енг. semantic annotation), (5) – синтаксичка анотација (енг. parsing), (6) – анотација кореференције (енг. coreference annotation), (7I) – прагматичка анотација (енг. pragmatic annotation), (7II) – стилистичка анотација (енг. stylistic annotation). Од наведених типова анотације корпуса, најраспрострањенији су (1), (2) и (3I)-(3III); (5) је све више заступљен, док су остали још увек ретки.

Анотација корпуса пре свега има за циљ да омогући ефикаснију претрагу. У случају када постоји информација о леми, корисник може, на пример, пронаћи све облике неке именице, задајући само њен канонски облик – номинатив једнине. Морфосинтаксички опис токена омогућава прецизну спецификацију синтаксичких конструкција које корисник жели да пронађе у корпусу (нпр. све придевске синтагме облика „прилог за којим следи придев", попут веома брз). У случају када лексикографа занима неко специфично значење лексеме која се или одликује полисемијом (час као сат и час као школски час) или пак има хомониме/хомографе (коси - глагол косити, или коси - именица коса), а корпус је анотиран значењима, лексикограф ће лако филтрирати резултат претраге корпуса, тј. елиминисати „шум" који стварају хомографија и полисемија.

Пошто су резултати претраге корпуса само узорци језика извучени из ширег контекста, подаци о изворном тексту су неопходни како би се надокнадиле информације које су изгубљене са ширим контекстом.

Анотација корпуса се обавља у фази претходне обраде текстова (енг. text preprocessing). Том приликом уклањање нетекстуелних елемената (слика, табела и сл.) захтева да се уместо њих наведе текст који би указао на врсту и место изостављених делова. Слично, за транскрипцију говорних текстова неопходно је анотирати парајезичка својства (смех, пауза, одређена гласност, модулација гласа, итд.). Такође, коментари уређивача који припремају текст за корпус се могу унети у текст једино као део анотације (Xiao, 2010).

Анотација корпуса олакшава и статистичку анализу корпуса, тј. аутоматско одређивање расподеле лингвистичких својстава. Размотримо као један пример креирање листе учестаности. Листа учестаности или фреквенцијска листа представља списак корпусних речи и њиховог броја појављивања. У случају корпуса код којих није присутна информација о леми, приликом пребројавања не постоји могућност да се аутоматски утврди учестаност самих лема (нпр. учестаност именице кућа), већ се посебно рачуна број појављивања за сваки појединачни облик леме (кућа, куће, кући итд.). Ако не постоји информација о врсти речи, без ручног пребројавања немогуће је одредити учестаност именица и осталих врста речи. Такође, без информације о значењу корпусне речи не може се израчунати колика је учестаност сваког од могућих значења појединачне корпусне речи. Слично важи и за остале неанотиране лингвистичке

информације.

Сем наведених предности, анотацију карактеришу и недостаци, због којих поједини корпусни лингвисти изражавају и отпор према самој идеји. То је једна од кључних разлика између лингвиста који заступају „приступ заснован на корпусу" (енг. corpus-based linguists) и лингвиста који заступају „приступ вођен корпусом" (енг. corpus-driven linguists). Заступници првог приступа сматрају да је улога корпуса тестирање постојећих теорија, њихово кориговање и допуњавање, као и проналажење примера који их потврђују, те се стога залажу за што детаљнију анотацију корпуса. Присталице „приступа вођеног корпусом" заступају став да корпусу треба приступати без унапред изграђених теорија како би се лингвистичке категорије постулирале искључиво на основу самих података. Самим тим, потоњи приступ третира анотацију корпуса као непотребну, јер је анотација заправо једна конкретна анализа корпуса (коју су спровели анотатори), те се анализирањем анотираног корпуса добијају само поновљени резултати нечије претходне анализе корпуса (Lindquist, 2009, стр. 45).

Оно што је много важније у односу на поменуту теоријску расправу је сама реализација анотације корпуса. Наиме, анотација често захтева мукотрпан рад и значајне људске ресурсе, не само у смислу бројности, већ и кад је у питању познавање лингвистичких дисциплина (морфологије, синтаксе, семантике итд.). Иако су развијени разни софтверски алати који су стању да са тачношћу од 95-97% (Brants, 2005) аутоматски обаве поједине задатке анотације, за многе од њих је неопходан претходно припремљен ручно анотирани корпус на основу кога ће алат „увежбати" анотацију произвољног текста.

Две најчешће тешкоће које прате поступак аутоматске анотације су разрешавање вишезначности и анотација „непознатих" речи. Прва тешкоћа се јавља у случају када се токену могу придружити две или више информација које се међусобно искључују (нпр. токену море се, као информације за лему и врсту речи, могу придружити и именица море и глагол морити). Човек разрешава вишезначност користећи више механизама који се служе разним изворима информација (контекст на нивоу реченице или шире логичке јединице текста, домен текста, здраворазумско знање, итд.). Програми за аутоматску анотацију располажу ограниченом количином информација у односу на човека, и још увек нису у стању да те информације повезују и из њих изводе закључке ни са приближном ефикасношћу коју постиже човек.

Другу тешкоћу представљају „непознате речи", токени са којима се програм за анотацију није претходно „сусрео" током своје обуке, тј. о којима нема никакву расположиву информацију. Типичан пример „непознатих" речи су хапакси (грч. απαξ λεγομενον), речи које је неко сковао (користећи неке познате деривационе механизме) и једном употребио.

С обзиром да резултати анотације најчешће представљају улазне податке за даље анализе (нпр. синтаксички анализатор користи резултате морфолошке анотације), „непознате" речи се не могу игнорисати, већ програм за анотацију мора располагати неком хеуристиком која ће и њима доделити неопходне информације. Такође, у случају када је анотација само један од првих корака обраде, разрешавање вишезначности се може препустити наредним корацима, поготово ако ће њима бити доступна нека додатна сазнања која ће олакшати одлуку о избору „праве" анотације. Тада се токену, уместо једнозначне анотације, придружује скуп свих могућих анотација или пар

највероватнијих, а разрешавање више-значности се одлаже (Guengoer, 2010).

У овом раду акценат ће бити, пре свега, на етикетирању врстом речи и лематизацији, и то тако да се свакој речи у тексту придружује тачно једна лема и једна врста речи.

2. Стандарди за анотацију корпуса

Један општеприхваћени и примењени стандард за анотацију корпуса још увек не постоји. Размотрићемо неке тренутно најзаступљеније „незваничне стандарде" које је прихватила стручна јавност.

2.1. TEI

Потреба за стандардизацијом кодирања и анотације машински читљивог текста постоји од самог почетка обраде текста на рачунару. Први покушај успостављања таквог стандарда кога су корисници масовно прихватили, односно применили у пракси, представљају Смернице Иницијативе за обележавање текста (енг. Text Encoding Initiative Guidelines, скр. TEI Guidelines). Први предлог (TEI P1) појавио се 1990. године, а актуелни пети предлог (TEI P5) 2007. године (TEI, 2009). TEI је потекао у академској средини и у почетку су бригу о њему водиле асоцијације ACH (Association of Computer in the Humanities), ALLC (Association of Literary and Linguistic Computing) и ACL (Association for Computational Linguistics), а током 1999/2000. године је формиран посебан непрофитни конзорцијум (TEI Consortium) са циљем да развија, одржава и промовише TEI.

Смернице TEI спецификују скуп етикета које могу да се уметну у електронску репрезентацију текста како би обележиле структуру текста и друга својства од значаја (библиографске информације о тексту, информације о лингвистичким елементима текста, итд.). Као језик за означавање, прве верзије TEI су користиле SGML, док су верзије настале после 2002. године (TEI P4 и касније) дефинисане помоћу XML.

Смернице TEI настоје да обухвате анотацију што више врста електронског текста. Стога су Смернице прилично опширне (на око хиљаду и по страна предлога TEI P5 је објашњена употреба скоро пет стотина етикета), али и сувише опште за поједине примене (нпр. корпусну лингвистику). Због своје општости је анотација коју предлаже TEI организована модуларно са хијерархијом наслеђивања елемената и атрибута, што омогућава да корисници прилагоде анотацију својим потребама модификацијом - додавањем, брисањем, преименовањем итд. - назива елемената и атрибута, ажурирањем модела садржаја елемената или променом вредности атрибута.

Неки од учесника у стварању анотације TEI су учествовали и у анотацији Британског националног корпуса (енг. British National Corpus, скр. BNC), и то је свакако најзначајнији пример примене те анотације у корпусној лингвистици.

Међутим, општост и опширност анотације TEI су утицали на многе корпусне истраживаче да се не определе за анотацију помоћу TEI. Додатни проблем је што усаглашеност са Смерницама TEI не значи и да је анотација спроведена конзистентно, поготово што Смернице нуде различите начине анотације истих феномена.

Као покушај да се почетницима приближе Смернице TEI, а тиме и да се прошири

њихова употреба, издвојен је стандардизовани подскуп најчешћих или најважнијих елемената TEI под називом TEI-Lite („лака верзија TEI").

2.2. CES/XCES

Један од деривата језика за означавање TEI, прилагођен потребама корпусне лингвистике, је Стандард за кодирање корпуса (енг. Corpus Encoding Standard, скр. CES). Овај стандард, усаглашен са Смерницама TEI, је 1996. године објавило саветодавно тело EAGLES (енг. Expert Advisory Groups on Language Engineering Standards), као део својих смерница. EAGLES је формирала Европска заједница (ЕЗ) 1993. године са циљем да, на основу постојеће праксе кодирања и анотације званичних језика ЕЗ, развије стандарде који би се користили у будућим европским пројектима.

Као и TEI, и CES је најпре дефинисан помоћу SGML-а (Ide, 1998), а развојем XML-технологија настаје XCES (Ide et al, 2000), као XML-верзија (дефиниција) стандарда CES.

CES и XCES користе подскуп TEI елемената чије је значење прецизније описано, а модел садржаја редукован. Посебна пажња је посвећена морфосинтаксичкој анотацији.

Важна одлика стандарда XCES је могућност да се текст и анотација чувају у одвојеним датотекама (енг. stand-off annotation) и повезују показивачима (XPointer). Одвојеност омогућава да се тексту придружи више различитих анотација у посебним датотекама које не морају истовремено да се користе. На тај начин се током аутоматске обраде у природно-језичким апликацијама смањује временска сложеност искључивањем слојева анотације који у датом тренутку нису од значаја. Показивачи су посебно значајни за креирање паралелизованих корпуса, јер се помоћу њих повезују одговарајуће јединице превођења изворних и циљних текстова. Слојевитост анотације решава и проблем преклапања различитих анотација карактеристичан за чување текста и анотација у истој датотеци.

2.3. MULTEXT-East

Једна од првих примена стандарда CES је анотација вишејезичног корпуса JOC (званични часопис Европске заједнице, енг. Official Journal of European Community), креираног у оквиру низа пројеката Вишејезични алати и корпуси (енг. Multilingual Tools and Corpora), познатијег под скраћеним именом MULTEXT (Ide and Veronis, 1994). Пројекти MULTEXT су имали за циљ развој стандарда и спецификација за кодирање и обраду лингвистичких корпуса, као и алата и ресурса који би применили те стандарде. Ресурси пројеката MULTEXT се састоје од текстова на пет западноевропских језика (енглески, француски, шпански, италијански и немачки). Паралелизација је обављена на нивоу реченице, а корпусне речи су етикетиране врстом речи.

Проширивање ресурса, алата, методологија и искустава пројеката MULTEXT на средњоевропске и источноевропске језике постао је циљ пројекта MULTEXT-East (Erjavec, 2010). Иако је пројекат званично трајао од 1995. до 1997. године, његови резултати у виду ресурса, спецификација и алата су до сада објављивани 1998, 2002, 2004. и 2010. године, сваки пут кориговани и проширени ресурсима додатих језика, укључујући и српски (Krstev et al, 2004).

Морфосинтаксичка анотација примењена у оквиру пројекта MULTEXT-East (Табела 1) је позициона анотација, тј. свака позиција у опису представља један атрибут. За ознаке вредности атрибута MULTEXT-East користи слова и цифре, као и специјалан знак (-) који указује на одсуство атрибута за дати токен. Иста ознака се може употребити на више позиција, при чему позиција одређује значење ознаке. Нпр, ознаке у опису Afpmsnn редом означавају да је у питању придев (A), и то описни (f), у позитиву (p), мушког рода (m), у једнини (s), у номинативу (n), неодређеног вида (n).

токен	морфосинт. опис
Bio	Vmps-sman-n---p
je	Va-p3s-an-y---p
vedar	Afpmsnn
i	C-s
hladan	Afpmsnn
aprilski	Aopmpn
dan	Ncmsn—n

Табела 1 MULTEXT-East (морфосинтаксичка анотација почетка прве реченице српске верзије Орвеловог романа 1984)

Спецификација анотације коју препоручује MULTEXT-East је један од кандидата за анотацију корпуса текстова на српском језику. Нажалост, и овај предложени стандард има своје недостатке. Принципи на основу којих су поједини атрибути и вредности укључени у морфосинтаксички опис који предлаже MULTEXT-East нису увек јасни и конзистентни, а важне информације, попут оних које одређују сложене услове конгруенције у српском језику, не могу се изразити тим морфосинтаксичким описом (Vitas et al, 2007).

3. Корпус савременог српског језика

Прва верзија Корпуса савременог српског језика (скр. СрпКор), расположива online од 2002. године, најпре на старој адреси http://www.korpus.matf.bg.ac.yu, а сада http://www.korpus.matf.bg.ac.rs, пред-стављала је колекцију неанотираних текстова, величине око 22 милиона речи, без информација о изворима (Krstev and Vitas, 2005). Током прве деценије свог постојања СрпКор је постепено мењао свој изглед. Најпре је допуњен информацијама о изворима, а корисницима је омогућено да током претраге имају увид у библиографске референце текстова из којих су издвојене конкорданце.

Неопходност проширивања корпуса диктира његов даљи развој у два смера. Једно усмерење је ка постепеном проширивању корпуса тако да се сачува баланс између појединих функционалних стилова и регистара заступљених у текстовима корпуса. Друго усмерење је ка креирању опортунистичког корпуса српског језика величине бар 100 милиона речи. Осим обезбеђивања библиографске информације о новим текстовима корпуса, оба развојна усмерења разматрају етикетирање врстом речи и лематизацију, као додатни облик анотације.

Јула 2011. настаје нова верзија СрпКор чија је величина 113 милиона речи. За све текстове је обезбеђена информација о изворима. Сем тога обављена је диференцијација текстова по функционалним стиловима (књижевно-уметнички, научни, публицистички, административни и остало). Више о овом корпусу и условима приступа може се наћи на адреси

http://www.meta-net.eu/meta-share.

У даљем тексту овог чланка биће представљени досадашњи резултати анотације Корпуса савременог српског језика.

4. Етикетирање врстом речи и лематизација

Као што је већ поменуто у одељку 1, етикетирање врстом речи (енг. PoS tagging) је процес којим се токенима корпуса придружује информација о врсти речи или ознака неке друге лексичке класе. Ова врста анотације се примењује и на интерпункцију која се обележава или заједничком ознаком или се, у зависности од намене корпуса, уводе посебне ознаке за знаке интерпункције који су од значаја.

Да би се уопште приступило етикетирању врстом речи најпре је неопходно:

- прецизно дефинисати скуп етикета које се придружују појединачним токенима (енг. tagset).
- одабрати програм за етикетирање врстом речи (енг. PoS-tagger) који ће се користити за аутоматску анотацију.
- припремити помоћне ресурсе који су потребни програму за етикетирање врстом речи, најчешће ручно анотиран корпус на основу ког ће се програм обучити за етикетирање, и евентуално речник у виду скупа свих могућих етикетирања лексичких речи.

4.1. Скуп етикета

Уместо креирања новог скупа етикета за анотацију СрпКор, прилагођени су постојећи морфосинтаксички описи, примењени у електронском морфолошком речнику српског језика (Krstev and Vitas, 2005). Формат морфосинтаксичких описа који се користи у том речнику је познат као LADL/DELA и најпре је примењен у морфолошком електронском речнику француског језика (Courtois and Silberztein, 1990). Формат LADL/DELA (Табела 2) омогућава, поред навођења самог облика лексичке речи, запис информација о леми, врсти речи, флективним категоријама (род, падеж, број итд.), а садржи и синтаксичко-семантичке маркере на основу којих се може екстраховати информација о именованим ентитетима, изговору (дијалект), типу деривације (творба деминутива, присвојни/релациони придев, моција рода), семантичким улогама (агент, инструмент, итд.).

Приликом избора скупа етикета мора се направити равнотежа између величине скупа, то јест детаљности информација које пружају етикете, с једне стране, и ефекта вишезначности и његовог утицаја на прецизност анотације, са друге стране. Богатији скуп етикета омогућава више информација, али зато отежава задатак прецизног придруживања етикета и обрнуто.

Пример једног записа у морфолошком електронском речнику српског језика		
korisnikovog,korisnikov.A+Hum+Pos+Der:adms4v		
Објашњење морфосинтаксичког записа		
korisnik*ovog*	лексичка реч (елемент речника)	
korisnik*ov*	лема (канонски облик лексичке речи)	
A	врста речи (придев)	
+Hum+Pos+Der	синтаксичко-семантички маркери	
	+Hum	особа
	+Pos	присвојни придев
	+Der	изведена реч
:adms4v	флективне категорије	
	a	позитив (степен поређења)
	d	одређени (вид)
	m	мушки (род)
	s	једнина (број)
	4	акузатив (падеж)
	v	анимат ност инанимат ност

Табела 2: формат LADL/DELA

С обзиром да је један од главних циљева етикетирања постићи што већу прецизност, у првим експериментима је тестиран базични скуп од свега 16 етикета које обухватају врсте речи у српском језику, као и неке специфичне токене који захтевају посебан третман (римски бројеви, скраћенице, префикси, суфикси):

1. N (именица)
2. A (придев)
3. V (глагол)
4. PRO (заменица)
5. NUM (број)
6. PREP (предлог)
7. CONJ (везник)
8. INT (узвик)
9. PAR (речца или партикула)
10. ADV (прилог)
11. PREF (префикс)
12. ABB (скраћеница)
13. RN (римски број)
14. PUNCT (знак интерпункције)
15. SENT (ознака краја реченице)
16.? (ознака за остало: стране речи у тексту, суфиксе попут их у 1990-их итд.).

4.2. Алати за етикетирање врстом речи

Од избора конкретног програма за етикетирање (енг. Part-of-Speech tagger, скр. PoS tagger) зависи како ће се адаптирати морфосинтаксички описи у електронском речнику српског језика. С обзиром на резултате приказане у (Поповић, 2010) и функционалности описане у (Paumier, 2008), у ужем избору су се нашла три алата: Unitex (Paumier, 2008), TnT (Brants, 2000) и TreeTagger (Schmid, 1994).

Unitex је систем за обраду корпуса заснован на технологији коначних аутомата и рекурзивних мрежа преласка. Систем користи лексичке ресурсе (електронске речнике и граматике) за обраду текстова и креирање корпуса. Корпус креиран системом Unitex је могуће претраживати не само коришћењем регуларних израза (у којима је дозвољено и коришћење свих морфосинтаксичких категорија присутних у електронским речницима), већ и применом сложених графова помоћу којих се могу описати и морфолошки и синтаксички феномени.

С обзиром да речници система Unitex користе LADL/DELA формат, избор тог система као програма за етикетирање је представљао логичан први корак. Приликом обраде корпуса системом Unitex, један од најважнијих резултата је аутомат текста (енг. text automaton). Аутомат текста се састоји од сегмената (најчешће реченица) и представља све могуће лексичке интерпретације корпусних речи у оквиру сваког сегмента. Пошто за сваки аутомат постоји еквивалентан граф, сваки пут од почетног до завршног чвора у графу којим је представљена једна реченица (или, у општем случају, сегмент) описује једно могуће етикетирање реченице. Суштински проблем који се при том јавља јесте разрешавање вишезначности (енг. handling ambiguity). Изузимајући ручно отклањање „лажних“ понуђених могућности, једно од расположивих решења које користи Unitex јесте формализам ELAG (Laporte, 1998). У основи, ELAG користи ручно креирана правила за етикетирање (такође имплементирана у облику аутомата) која разрешавају вишезначност тако што спецификују дозвољени или забрањени контекст корпусне речи етикетиране на одређени начин.
Важну предност система Unitex представља чињеница да се токену током етикетирања придружују искључиво оне лексичке интерпретације које постоје у речнику. Међутим, неопходно је уложити значајан напор у креирање граматика формализма ELAG које би постигле бар исту прецизност као и статистички програми за етикетирање (нпр. TnT

и TreeTagger). Осим тога, развој таквих граматика за српски језик је још увек у почетној фази, те је тестирање система Unitex као програма за етикетирање остављено за неки будући експеримент.

Евалуација описана у (Поповић, 2008), заснована на десетострукој унакрсној провери (енг. 10-fold cross-validation), показује да TnT и TreeTagger постижу сличне резултате над истим корпусима за обуку (TnT: 93,86%, TreeTagger: 91,78%), при чему је TnT бољи у етикетирању „непознатих" (боље речено, непрепознатих) речи (TnT: 58,36%, TreeTagger: 36,71%). При том је коришћен скуп од 908 етикета - морфосинтаксичких описа за српски језик дефинисаних у оквиру пројекта MULTEXT-East (http://nl.ijs.si/me/V3/msd/html/). Величина тестираног корпуса је око 105 хиљада корпусних речи (18 хиљада корпусних типова, тј. различитих корпусних речи и око 7,6 хиљада лема).

Наведени резултати у етикетирању „непознатих" речи, као и потреба да етикетирани корпус садржи информацију о леми, свели су даљи избор на TreeTagger.

4.3. Помоћни ресурси

Оба статистичка програма за етикетирање, TnT и TreeTagger, захтевају ручно анотирани корпус за обуку (енг. training set), док је за TreeTagger такође неопходан и тзв. потпуни речник (енг. full lexicon).

Као ручно анотирани корпус за обуку искоришћен је корпус INTERA. Овај корпус је добио име по истоименом пројекту (Gavrilidou, 2006) током кога је настао СЕЛФЕХ (енг. Serbian-English Law Finance Education and Health, скр. SELFEH), паралелизовани српско-енглески корпус текстова из области економије, здравства, права и образовања (http://www.korpus.matf.bg.ac.rs/prezentacija/selfeh.html). СЕЛФЕХ је креирала Група за природно језичке технологије на Математичком факултету Универзитета у Београду и он се састоји од 150 докумената у формату TMX. Српска верзија овог корпуса (INTERA) је прилагођена формату који захтева TreeTagger, садржи информације о врсти речи и леми за сваки од 1.100.281 токена. Скуп етикета се састоји од 16 етикета наведених у одељку 4.1. INTERA се састоји од 907.633 корпусних речи, односно 55.488 корпусних типова.

Број корпусних речи у корпусу INTERA обележених етикетом ? (остало/непознато) износи 4404, што је мање од 0,5% укупног броја корпусних речи.

Потпуни речник програма TreeTagger представља текстуалну датотеку чија свака линија садржи (као колоне) једну лексичку реч (токен) и њена могућа етикетирања. Колоне су међусобно раздвојене табулаторима. Свако могуће етикетирање је уређени пар (врста речи, лема) чије су компоненте раздвојене размаком (Табела 3).

токен	етик.	етик.	етик.	етик.
bacili	N bacili	V baciti		
vrelo	Nvrelo	V vreti	Avrelo	ADV vrelo

Табела 3 - Извод из потпуног речника за два токена (корпусне речи)

Један од тежих проблема кога је требало решити пре саме обуке програма TreeTagger, свео се на припрему потпуног речника на основу постојећег морфолошког електронског

речника српског у формату LADL/DELA. При том треба истаћи да TreeTagger није „прави“ лематизатор, већ да се његов рад заправо своди само на избор највероватније етикете за врсту речи, после чега програм само „дописује” лему из потпуног речника која одговара изабраној врсти речи. Стога се у потпуном речнику не смеју наћи токени којима одговарају исте етикете за врсту речи, али различите леме (Табела 4).

токен	вр. речи/ лема	вр. речи/ лема	вр. речи/ лема
kapi	N kap	N kapa	N kapo
Brankom	N Branko	N Branka	
donesen	V doneti	V donijeti	

Табела 4 - Недозвољени уноси у потпуном речнику

Овакво ограничење је имало значајан утицај на разликовање хомографа чије (међусобно различите) леме представљају различите дијалекте или слична властита имена различитог рода. Такође, није могуће разликовати хомографе чије су и саме леме хомографи. Уместо увођења нових етикета које би биле „клонови” постојећих, али са другачијом ознаком, одлучено је да се сви дупликати у виду уређених парова (токен, врста речи) елиминишу. Критеријум који је при том коришћен се свео на задржавање екавског изговора и властитих имена мушког рода, док сви хомографи чије су леме такође хомографи, имају јединствену заједничку репрезентацију у потпуном речнику.

После избацивања дупликата (токен, врста речи), у потпуном речнику и даље постоји вишезначност (Табела 5). Најчешћи случај вишезначности је могућност да се токен интерпретира и као придев и као глагол (84,77%).

5. Евалуација

Евалуација извршене анотације обављена је десетоструким унакрсним тестом (енг. 10-fold cross validation test). Тест се састоји у томе да се неанотирана верзија корпуса за обуку подели на 10 делова, програм за анотацију користи девет делова корпуса за обучавање, а затим аутоматски анотира преосталу десетину корпуса. Избор десетине корпуса која ће бити анотирана једнозначно одређује преосталих девет десетина корпуса које ће представљати корпус за обуку. На тај начин се, за сваку поделу на 10 делова, описана обука коришћењем девет десетина корпуса и анотација преостале десетине може обавити на 10 начина и сваки пут упоредити резултати ручне и аутоматске анотације преостале десетине корпуса. Резултати упоређивања су низови од по 10 вредности (за сваку десетину корпуса) који представљају редом величину десетине (укупан број токена), као и токена који су истоветно анотиране ручно и аутоматски. Сви токени који нису етикетирани истоветно (ручно и аутоматски), а при том их нема у потпуном лексикону, третирани су као непознате речи.

релативна учестаност	секвенце етикета		
84,77%	A	V	
7,02%	N	V	
2,96%	A	N	
2,75%	A	ADV	
0,32%	?	N	
0,31%	A	N	V
0,29%	A	ADV	V
0,23%	ADV	N	

Табела 5 - Расподела вишезначности у потпуном речнику

На основу добијених података израчуната је прецизност појединачне анотације, процентуални удео непознатих речи у односу на укупан број неслагања између ручне и аутоматске анотације, а затим су за добијене резултате израчунате минимална, максимална и просечна вредност, као и варијанса, односно стандардна девијација (Табела 6).

	прецизност	„непознате“ речи
мин.	95,38%	7,68%
макс.	96,98%	16,40%
прос.	96,57%	11,93%
станд. дев.	0,43%	2,20%

Табела 6 - Резултати евалуације

Прецизност појединачне анотације (rj) је рачуната као количник укупног броја ручно и аутоматски истоветно анотираних токена и укупног броја токена у десетини која је анотирана. Просечна прецизност $\bar{r}$ и варијанса var су рачунате по следећим формулама

$$\bar{r} = \frac{1}{0} \sum_{j=1}^{0} r_j$$

$$var = \frac{1}{0} \sum_{j=1}^{0} \left(\bar{r} - r_j \right)^2$$

6. Закључак

Резултати евалуације указују да је прецизност етикетирања врстом речи у границама тачности коју постижу програми за анотацију који користе скуп етикета истог реда величине. Корпус за обуку INTERA је искоришћен за анотацију нове верзије Корпуса савременог српског језика у којој ће, поред библиографске информације о текстовима, бити доступна и информација о врсти речи и леми за сваки токен. Нова верзија СрпКор која ће користити и ново корисничко сучеље (енг. user interface) биће представљена као један од резултата пројекта CESAR (http://www.meta-net.eu/projects/cesar/), једног од пратећих пројеката пројекта META-NET током 2011/12. године.

Литература

Brants, Thorsten. 2000. TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.

Brants, Thorsten. 2005. Part-of-speech tagging. In The Encyclopedia of Language and Linguistics. (Second ed.), Volume 1-14, Ed. Brown, K., 221-230. Oxford: Elsevier.

Courtois, Blandine and Max Silberztein. 1990. Dictionnaires électroniques du français. Paris: Larousse.

Erjavec, Tomaž. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proceedings of the LREC 2010, Malta, 19-21 May 2010.

Gavrilidou, Maria, Penny Labropoulou, Stelios Piperidis, Voula Giouli, Nicoletta Calzolari, Monica Monachini, Claudia Soria, and Khalid Choukri. 2006. Language Resources Production Models: the case of the INTERA multilingual corpus and terminology, In: Proceedings of the Fifth International Conference on

Language Resources and Evaluation-LREC2006, 24-26 May 2006, Genoa, Italy, 609 – 614.

Guengoer, T. 2010. Part-of-speech tagging. In, Handbook of Natural Language Processing (Second ed.), Eds. Nitin Indurkhya and Fred J. Damerau, Machine Learning and Pattern Recognition, Chapter 10, 205-235. Boca Raton, London, New York: Chapman & Hall/CRC, Taylor & Francis Group.

Ide, Nancy. 1998. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In First International Conference on Language Resources and Evaluation, LREC'98, Granada, 463-470. ELRA.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. In Second International Conference on Language Resources and Evaluation, LREC'00.

Ide, Nancy and Jean Véronis. 1994. Multext (Multilingual Tools and Corpora). In Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, 90-96. ACL.

Krstev, Cvetana and Duško Vitas. 2005. Corpus and Lexicon - Mutual Incompleteness. In Proceedings of the Corpus Linguistics Conference, Eds. Pernilla Danielsson, P., Wagenmakers, M., 14-17 July 2005, Birmingham, http://www.corpus.bham.ac.uk/PCLC/.

Krstev, Cvetana., Duško Vitas, and Tomaž Erjavec. 2004. MULTEXT-East Resources for Serbian. In Zbornik 7. mednarodne multikonference „Informacijska druzba IS 2004" Jezikovne tehnologije, Eds. Tomaž Erjavec and Jerneja Z. Gros, 9-15 Oktober 2004, Ljubljana, Slovenija. Institut „Jozef Stefan".

Laporte, Eric and Anne Monceaux, A. 1998. Elimination of lexical ambiguities by grammars: The ELAG system. Lingvisticæ Investigationes, 22:341–367. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Lindquist, Hans. 2009. Corpus Linguistics and the Description of English. Edinburgh University Press.

Paumier, Sébastien 2008. Unitex 2.1 User Manual. http://www-igm.univmlv.fr/unitex/UnitexManual2.1.pdf.

Поповић Зоран. 2010. Програми за етикетирање текста на српском језику. Инфотека 11(2):19–36.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

TEI Consortium (Ed.). 2009. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Vitas, Duško, Cvetana Krstev and Svetla Koeva. 2007. Towards a Complex Model for Morpho-Syntactic Annotation. Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages, Eds. E. Paskaleva and M. Slavcheva, 26 September 2007, Borovets, Bulgaria, 65-71.

Vitas, Duško, Cvetana Krstev, Ivan Obradović, Ljubomir Popović and Gordana Pavlović-Lažetić. 2003. Processing Serbian Written Texts: An Overview of Resources and Basic Tools. Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki, Greece, Eds. Piperidis, S., Karkaletsis, V., 97-104.

Xiao, Richard. 2010. Corpus Creation. In, Handbook of Natural Language Processing, Eds. N. Indurkhya and F. Damerau, Machine Learning & Pattern Recognition Series, Chapter 7, 147-165. CRC Press, Taylor and Francis Group.