УНИВЕРЗИТЕТ У БЕОГРАДУ

МАТЕМАТИЧКИ ФАКУЛТЕТ

Сташа И. Вујичић Станковић

ЕКСТРАКЦИЈА ИНФОРМАЦИЈА ВОЂЕНА ОНТОЛОГИЈАМА (МОДЕЛ ЗА СРПСКИ ЈЕЗИК)

докторска дисертација

Београд, 2016.

UNIVERSITY OF BELGRADE

FACULTY OF MATHEMATICS

Staša I. Vujičić Stanković

ONTOLOGY BASED INFORMATION EXTRACTION (MODEL FOR THE SERBIAN LANGUAGE)

Doctoral Dissertation

Belgrade, 2016.

Ментор:
др Душко Витас, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Гордана Павловић-Лажетић, редовни професор
Универзитет у Београду, Математички факултет

др Весна Пајић, доцент
Универзитет у Београду, Пољопривредни факултет

др Вељко Милутиновић, редовни професор
Универзитет у Београду, Електротехнички факултет

Датум одбране: _____

Посвећено мојим родитељима, Тањи и Илији,
супругу Зорану,
породици Пајић,
и свима онима који су увек веровали у мене

Желим да се захвалим свима који су уложили своје знање, стрпљење и добру вољу да ме подрже и усмере током писања ове дисертације.

На првом месту ментору, професору Душку Витасу, који ме увео у дивни свет обраде природних језика и пружио ми, дељењем свог знања и сјајних идеја, снажну мотивацију да се бавим истраживањима у овој области. Веома сам му захвална што ми се током година, на путу мог научног усавршавања, увек нашао са правим речима усмерења и помоћи.

Велику захвалност дугујем професорки Гордани Павловић-Лажетић, која ми је од основних студија представљала узор какав човек и професор, пун разумевања, топле речи и стрпљења, треба да се трудим да постанем. Захвална сам јој на свему што ме научила, на бројним сугестијама које су оплемениле моја размишљања, и конструктивним питањима, захваљујући којима је ова теза попримила постојећи облик.

Посебно се захваљујем професорки Весни Пајић на драгоценом времену које ми је посветила, која ми је у кључним тренуцима пружила помоћ и снажну мотивацију. Дугујем јој захвалност за огромну подршку и бројне савете, како у изради ове дисертације, тако и у свакодневном животу.

Захваљујем се професору Вељку Милутиновићу на указаном поверењу и прилици за заједнички рад, на многим корисним саветима, безрезервној подршци и свему што ме научио током наше сарадње.

Веома сам захвална професорки Цветани Крстев и професорки Ранки Станковић на знању које су пренеле на мене, детаљном читању и предлозима за унапређење мог рада. Захваљујем се професору Милошу Пајићу на сарадњи, саветима и безграничном оптимизму који је преносио на мене током рада на тези, као и професору Саши Малкову на корисним саветима и подршци.

Нарочито се захваљујем мојој породици, мами Тањи и супругу Зорану, који су кроз своју неизмерну љубав и разумевање увек представљали мој ослонац и пружали ми охрабрење да истрајем на свом путу.

Математички факултет Сташа Вујичић Станковић

Београд, 2016

Наслов дисертације:

Екстракција информација вођена онтологијама (Модел за српски језик)

Резиме:

Основни задатак ове докторске дисертације је истраживање различитих техника и модела који се примењују у екстракцији информација и пружање информатичке подршке за обраду текстова који припадају домену кулинарства и гастрономије. Екстракција информација је подобласт рачунарске лингвистике која обухвата скуп техника обраде природних језика, како би се у текстовима пронашле релевантне информације, одредила њихова значења и успоставиле релације међу њима. Посебна пажња је посвећена екстракцији информација вођеној онтологијама. Она обухвата задатак препознавања инстанци концепата онтологије у неструктурираним или полуструктурираним природно-језичним текстовима и резоновање над препознатим инстанцама на основу правила која су дефинисана у онтологији. Такође обухвата и задатак препознавања инстанци у тексту и њихово правилно придруживање концепту онтологије.

Главни резултат дисертације огледа се у представљању новог модела за екстракцију информација вођену онтологијама. Поред решавања задатака екстракције информација, нови модел обухвата и изградњу нових и доградњу постојећих лексичких ресурса и онтологија. Његовом применом развијен је систем за екстракцију информација из кулинарског домена, а може бити употребљен и у другим доменима. Додатно, развијена је и онтологија хране, допуњен српски *WordNet* са 1.404 синсета из домена кулинарства, а дограђен је и електронски речник српског језика са 1.248 јединица. Значај примене модела је и у томе што нови језички ресурси и доградње постојећих могу бити коришћени у другим системима обраде природних језика.

У уводној глави дисертације образложена је потреба да се обезбеди информатички модел за обраду обиља језичке грађе која припада домену кулинарства и гастрономије кроз методолошки прецизно утемељен приступ који пружа интегрисане информације о домену. Представљена је формализација основног објекта истраживања, текста у електронском облику.

Описане су апроксимације природних језика које се уводе како би савремене рачунарске технологије могле да обрађују текстове писане природним језиком. Истакнута је потреба да се направи карактеризација језика текста одговарајућим корпусом и подјезиком.

У наставку прве главе дефинисан је задатак екстракције информација и модели који омогућују да се изврши информатичка обрада полуструктурираних или неструктурираних текстова којом рачунар тумачи значење које је аутор (не обавезно човек) имао намеру да искаже приликом његовог записивања. Описане су методе које се користе у области екстракције информација – методе засноване на правилима и методе засноване на машинском учењу. Образложене су њихове предности и мане, као и разлози због којих се у овој докторској дисертацији користе технике засноване на правилима које се ослањају на лингвистичко знање. У завршном делу увода посебна пажња посвећена је онтологијама и WordNet-у, и истакнут значај његове употребе као онтологије.

У другој глави су представљени језички ресурси и алати који су коришћени у дисертацији. Описани су морфолошки речници и локалне граматике који се користе за решавање проблема екстракције информација из текстова на српском језику. Дат је преглед система за екстракцију информација и описан ток обраде текстова на српском језику при решавању задатка екстракције информација у програмским системима *Unitex* и *GATE*.

У трећој глави је представљен главни резултат дисертације, модел за решавање проблема екстракције информација интегрисањем језичких ресурса и алата, који обухвата формирање корпуса текстова, дефинисање задатака екстракције информација, изградњу коначних модела за екстракцију информација, примену развијених коначних модела, итеративну доградњу морфолошких електронских речника, проширење WordNet-а и изградњу нових онтологија. Детаљно је описан сваки од ових корака. Иако је модел првенствено разматран из угла решавања проблема који се јављају при обради српског језика, може бити примењен и за обраду текстова на другим језицима уз развој адекватних језичких ресурса.

Имплементација описаних корака приказана је у четвртом поглављу кроз систем за екстракцију информација из текстова кулинарског домена на српском језику. Описана је спрега при развоју и међусобној допуни доменских лексичких ресурса кроз кораке формирања доменског корпуса, препознавања кулинарске лексике, проширења и доградње WordNet-а и морфолошких електронских речника. Такође је описан развој доменске онтологије хране, доменске онтологије састојака који могу да се употребе као међусобне замене у кулинарском домену и доменске онтологије приближних мера у кулинарском домену. Развијени систем за екстракцију информација је послужио за реализацију система за напредно претраживање рецепата, који на основу корпуса кулинарских текстова генерише могуће одговоре на постављене упите корисника. У оквиру њега имплементиран је метод за успостављање веза између рецепата у случајевима када корисник прегледа текст рецепта и уочи да се у опису његове припреме појављује део за чију припрему постоји рецепт са додатним или другачијим објашњењем. Још један од доприноса дисертације јесте примена развијених онтологија у задацима конвертовања приближних кулинарских мера у стандардне мере и утврђивања сличности између рецепата. Сличност рецепата је за ову прилику дефинисана као сличност текстова који описују поступак припреме јела према одређеном рецепту.

У последњој глави су приказани закључци и правци даљег рада.

Кључне речи: екстракција информација, обрада природних језика, онтологије, кулинарски домен, WordNet
Научна област: Рачунарство
Ужа научна област: Рачунарска обрада текста
УДК број: [004.02+004.9]:808.61(043.3)

Title of the dissertation:

Ontology based Information Extraction (Model for the Serbian Language)

Abstract:

The basic goal of this doctoral thesis is a research into different techniques and models which are applied in information extraction, and providing an informatic support in processing of natural language texts from culinary and gastronomy domain. Information extraction is a subfield of computational linguistics which includes techniques for natural languages processing, in order to find relevant information, define their meaning and establish relations between them. A very special attention is given to ontology based information extraction. It consists of the following: recognition of instances of ontology concepts in non-structured or semi-structured texts written in natural language, reasoning over the identified instances based on the rules defined in the ontology, as well as recognition of instances and their use for instantiating the proper ontology concepts.

The main result of thesis reflects in the presentation of a new model for ontology based information extraction. Besides solving tasks of information extraction, the new model includes not only upgrade of existing lexical resources and ontologies, but also creation of the new ones. Its application resulted in development of a system for extraction of information related to the culinary domain, but this new model can be used in other fields as well. Beside this, the food ontology has been developed, *Serbian WordNet* is extended for another 1.404 synsets from the culinary domain, while electronic dictionary of Serbian is enlarged with 1.248 entries. The significance of the model application comes from the fact that the new and enriched linguistic resources can be used in other systems for natural language processing.

The opening chapter of the thesis elaborates the need of providing an informatic model for processing a huge linguistic corpus related to culinary and gastronomy domain, through methodologically precise and solid approach integrating pieces of information on the domain. Also, the formalization of the basic research subject, text in electronic form, has been presented. Further on, the chapter contains a description of the natural languages approximations introduced in order to enable modern information technologies to process texts written in natural

languages, and it emphasizes the need to make the characterisation of the text language with corresponding corpus and sublanguage.

Further on in the first chapter, the task of information extraction, and the models for informatic processing of non-structured or semi-structured texts, used by the computer to interpret the meaning that the author (not necessarily a human) has intended to give while writing the text, are defined. Additionally, this chapter contains the description of the methods used in information extraction field – methods based on rules and methods based on machine learning. Their advantages and shortcomings are listed, so as the reasons why in this thesis are used techniques based on linguistic knowledge. As a conclusion to the introduction chapter, a special attention is given to ontologies, WordNet, and the significance of its usage as ontology.

The second chapter contains the presentation of the linguistic resources and tools exploited in this thesis. It describes morphological dictionaries and local grammars used for solving the problem of information extraction from texts written in Serbian. A review of information extraction systems is given subsequently. At the end of the second chapter, the stages in processing of Serbian written texts during the information extraction in the software systems *Unitex* and *GATE* are described.

The main result of the thesis is presented in the third chapter. It is the model for solving the problem of information extraction by integrating linguistic resources and tools, which includes creation of a text corpus, definition of tasks for information extraction, establishment of finite state models for information extraction, and their application accordingly, iterative enlarging of electronic morphological dictionaries, enrichment and enhancement of WordNet, and creation of new ontologies. Each of these steps is described thoroughly. Even though the model was at first considered as a solution for problems in processing Serbian, it can be equally applied for processing texts written in other languages, with the development of suitable language resources accordingly.

The implementation of the above explained steps is described in the fourth chapter, through a system for information extraction from the culinary texts written in Serbian. Then follows the description of a bond in the development and mutual complement of lexical resources through steps in creating domain corpus,

identifying culinary lexica, expanding and upgrading of WordNet and electronic morphological dictionaries, and developing of domain ontologies – the food ontology, the approximate measure ontology, and the ontology of ingredients that can be used as mutual replacements in the culinary domain. This system, developed for information extraction, has served for creating an advanced search system which, based on a corpus of culinary texts, generates all possible answers to inquiries made by users. In the frame of this system is implemented a specific method which serves for creation of links between different recipes. This is used in case when the user reviews a text of a recipe and notices that in preparing description features some part which already had appeared in other recipe, but with additional or different explanation. Another contribution of this thesis is application of developed ontologies in tasks that convert approximate measures into standard measures, and establishment of similarities among the recipes. The similarity of the recipes is defined as similarity of texts which describe process of course preparation in accordance with a specific recipe.

The last chapter contains final conclusions and directions for future research.

Key words: Information Extracion, Natural Language Processing, Ontologies, Culinary Domain, WordNet
Scientific field: Computer Science
Scientific subfield: Text Processing
UDK number: [004.02+004.9]:808.61(043.3)

САДРЖАЈ

1 Уводна разматрања и основни појмови 1

1.1 Увод 1

1.2 Текст у електронском облику као основни предмет истраживања 16

1.2.1 Обрада текста у електронском облику 20

1.2.2 Корпуси као референтне репрезентације језика 23

1.2.3 Језик и подјезик 24

1.3 Екстракција информација 27

1.3.1 Евалуација система за екстракцију информација 33

1.4 Онтологије 34

1.4.1 Дефиниција и основни појмови 34

1.4.2 Различити типови онтологија 36

1.4.3 Онтолошки језици 38

1.4.3.1 OWL онтологије 40

1.4.3.2 SPARQL упити 42

1.4.4 WordNet као онтологија 43

2 Лексички ресурси и алати за обраду текста 50

2.1 Проблем 50

2.2 Електронски речници 54

2.3 Формати електронских речника 55

2.4 Коначни аутомати у обради текста 63

2.5 Системи за екстракцију информација 67

2.5.1 Увод 67

2.5.2 Unitex 70

2.5.2.1 Ток обраде у систему Unitex 72
2.5.2.2 Графови у систему Unitex 75
2.5.2.3 Екстракција информација у систему Unitex 78
2.5.3 GATE 86
2.5.3.1 Ток обраде у систему GATE 86
2.5.3.2 Графови у систему GATE 90
2.5.3.3 Екстракција информација у систему GATE 91
3 Модел екстракције информација вођене онтологијама 94
3.1 Екстракција информација вођена онтологијама 94
3.2 Предлог модела екстракције информација вођене онтологијама 96
3.3 Модул за развијање онтологије 100
3.3.1 Превођење онтологије са једног језика на други 100
3.3.2 Конвертовање WordNet-а у формалну онтологију 105
3.3.3 Издвајање онтологија нижег нивоа 108
4 Имплементација модела 111
4.1 Увод 111
4.2 Гастрономија и њен језик 111
4.3 Изградња корпуса кулинарског домена 117
4.4 Креирање доменске онтологије 122
4.4.1 Проширење WordNet-а и доградња доменски специфичне лексике у морфолошким речницима српског језика 122
4.4.2 Издвајање онтологије хране 129
4.4.3 Креирање доменске онтологије приближних мера у кулинарском домену 131

4.4.4 Креирање доменске онтологије замена састојака у кулинарском домену 133

4.5 Примене модела на систем за претрагу рецепата 135

4.5.1 Проширивање упита употребом онтологије 135

4.5.2 Примена онтологије приближних мера за конвертовање 141

4.5.3 Успостављање веза између рецепата 144

4.5.4 Сличност рецепата 148

5 Закључак и даљи рад 159

5.1 Закључак 159

5.2 Правци даљег рада 161

ЛИТЕРАТУРА 163

ПРИЛОГ А 186

ПРИЛОГ Б 188

ПРИЛОГ В 191

ПРИЛОГ Г 195

ПРИЛОГ Д 201

ПРИЛОГ Ђ 210

БИОГРАФИЈА АУТОРА 213

1

Уводна разматрања и основни појмови

1.1 Увод

Предмет овога рада је прибављање релевантних информација из текстова који припадају једном посебном домену, домену гастрономије и кулинарства. Ма како ово подручје изгледало општепознато, обично и свакодневно, посматрано из информатичког угла, оно је у највећој мери необрађено и неуређено. За разлику од других свакодневних активности које су добиле информатичке моделе, кулинарство се углавном своди на небројене стране на вебу са куварским рецептима. За неке домене, постоје специјализоване стране са прецизним информацијама о одређеном састојку, било да је реч о храни или напитку, али референтног, методолошки јасно утемељеног приступа, који би пружао интегрисане информације о овоме домену, још увек нема нигде. Овакво стање није необично ако се узме у обзир изузетна сложеност процеса који стоје иза феномена кувања. Наиме, у овај процес су непосредно или посредно уграђени резултати најразличитијих дисциплина почев од резултата природних наука, преко различитих примењених наука, укључујући и различите друштвене и хуманистичке науке, као и различите уметничке, али и занатске, вештине.

Ова сложеност феномена који стоје иза свакодневних активности око припреме хране јесте предмет филозофских разматрања као у „Гурманском уму“ Мишела Онфреа (Onfre, 2002) или у студији Брија-Саварена „Физиологија укуса“ (Brillat-Savarin, 1848). Онфре описује како је вештина кувања уведена

не само у науку, већ и у ред лепих уметности. Брија-Саварен, утемељивач гастрономије у савременом смислу, умеће конзумирања хране и посебну уметност „стола“ посматра као нужност која је далеко од природне потребе појединаца за исхраном, већ јој придодаје димензију промишљеног чиниоца различитих друштвених односа. По први пут се кување разматра из ширег угла, разматрају се права и обавезе домаћина, као и правила која важе за оне којима исказује гостопримство при обеду – „*подразумева се да гастроном за столом једе са извесном удубљеношћу, пробирљивошћу и чулношћу.* “ Кување је предмет и знамените антрополошке студије Клод Леви-Строса „Митологике“ (Levi-Stros, 2008) у којој се анализира, између осталог, развој концептуалног система друштава у функцији усвојених технологија припрема хране.

Поред оваквих најопштијих разматрања која осветљавају широку перспективу у развоју начина исхране, кулинарство је предмет историјских студија које осветљавају еволуцију друштава у зависности од расположивих извора хране, али и начина њеног припремања. Тако је, на пример, могуће изучавати недокументоване историјске хипотезе на основу археолошких трагова о употреби, узгоју или транспорту намирница. Бобер у књизи „Уметност, култура и кухиња“ (Bober, 1999) проучава феномен кувања ослањајући се на археологију и историју уметности које говоре о различитим кулинарским културама, традицијама, њиховим различитостима и сличностима. Теми кувања је посвећен и известан број специјализованих научних часописа (International Journal of Gastronomy and Food Science, 2015; Gastronomica, 2015; Anthropology of Food, 2015). Тек недавно је и лингвистика открила специфичности језика којим се говори о храни, где је у првом плану преглед његове етимологије, морфологије, синтаксе, граматике и семантике на енглеском и француском језику, почев од анализе ресторанских менија и кувара до телевизијских кулинарских емисија и материјала о припреми хране на интернету (Gerhardt, Frobenius i Ley, 2013). Иако су у овом зборнику, поред постављања темеља кулинарске лингвистике, описани и културолошки и историјски аспекти овог језика, само један рад се бави применом рачунара у овој области и то специфично из угла лексичке анализе, синтаксичке анализе и анализе дискурса кулинарских садржаја блогова о храни и њиховог

поређења са традиционално написаним рецептима у кулинарској литератури (Diemer i Frobenius, 2013).

Мимо научних истраживања феномена кувања из угла различитих научних дисциплина, овој теми су посвећене и студије као што је Мијоова „Речник заљубљеника у гастрономију“ (Mijo, 2012) која показује непрегледност нијанси у припремању одређеног јела, али и сложену културолошку, историјску и географску перспективу гастрономије. На ову разноврсну литературу треба додати и стручне књиге какве су код нас „Патин кувар“ (Marković, 1939) или Пелапраов „Велики кувар“ (Pelaprat, 1969), као и обиље популарних ревија или подлистака са рецептима и саветима за припрему хране.

Овом сажетом уводу уз различите аспекте припреме хране, треба додати и фрагменте књижевних и сликарских дела. Тако је, на пример, у Хашековом роману „Добри војник Швејк“ (Hašek, 2003), поред војничких невоља, незанемарљив део посвећен ишчекивању казана са *кромпир-гулашем* у трећој књизи овог романа. Код овог појма се јавља проблем баријере у разумевању, односно проблем вишејезичног разумевања кулинарског појма, јер читалац изван подручја централне и источне Европе не зна шта би могао бити *кромпир-гулаш*.

Преведена литература указује и на лексичке или концептуалне празнине у одређеним језицима: у роману „Старац и море“ Ернеста Хемингвеја, у првом српском преводу из 1952. (Hemingvej, 1952), наилази се на сегмент „*Ту су се окупљале огромне количине сардела и ситне рибе, неки пут на најдубљим местима, и јато мастиљавих риба и ове су се ноћу пеле до саме површине где су се све рибе-путнице храниле њима*.“ Мастиљаве рибе су овде превод за *squid* iz originala, дакле *лигње*, али у доба када је овај роман преведен на српски, лигња није широко код нас позната[1] као намирница. Имајући у виду присутност

[1] У „Речнику српскохрватскога књижевног језика“ (RMSMH, 1967) дефиниција лигње је преузета из „Зоолошког глосара“ из 1932. (ZTIN, 1932), без примера употреба ове речи. У „Великом народном кувару“ из 1956. (Marković, 1956) лигње се помињу у одељку „Италијанска кухиња“ у рецепту „рижото од лигања“, али је у индексу дат неубичајен номинатив *лигањ* уместо *лигња*.

лигања у данашњем српском гастрономском репертоару, овај пример указује на временску димензију у развоју кулинарских навика у једној средини. Шта више, и цео роман може бити изграђен око припрема за једну вечеру као што то приказује Карен Бликсен у „Бабетиној гозби“ (Bliksen, 2013) или из низа гурманских претеривања као у „Гаргантуи и Пантагруелу“ вејача овејане суштине Франсоа Раблеа[2] (Rable, 1989). Рабле сликовито и весело описује гозбе тако да дотакне читаочева чула и да пренесе атмосферу: „*Ту боце иду, ту шунка пљушти, ту купе круже, ту жаморе жбунови. Натегни! Додај! Редом! Мешај! Мени без воде...*“

С обзиром на свеприсутност теме исхране у најразличитијим видовима, могло би се очекивати да је ова тема привукла одавно пажњу информатичара. Ипак, информатичке студије језика гастрономије и кулинарства су у повоју. Известан број експеримената током последњих година ограничио се углавном на аутоматско одређивање супституита, осим у примеру IBM-овог система *Watson* који је посвећен састављању нових коктела[3] и јела[4] (IBM, 2014). Да би саставио нова јела која одговарају корисничким захтевима, овај систем користи велике количине енциклопедијских и других података из различитих научних области, а подржава и динамичко учење тако да на основу ранијих одговора које је давао постаје „паметнији“ кроз сваку интеракцију са новим корисницима. На основу тог система, у јуну 2014. године направљена је апликација *Chef Watson with Bon Appétit*[5]. *Watson* је прво обрадио податке из базе од 9.000 рецепата из кулинарског часописа *Bon Appétit* (Bon Appétit, 2015) како би добио увид у то који састојци се у рецептима најчешће јављају заједно, у каквој врсти јела се јављају и на који начин се та јела припремају. Потом је спојио нова знања са знањима о хемији хране којима је претходно обучен и о томе шта људи сматрају пријатним при конзумирању хране, а шта не. Као

[2] У преводу на српски Станислава Винавера.

[3] Основни механизам састављања коктела се састоји обично у одређивању односа два алкохолна пића и воћног додатка, што чини да је основна шема за састављање коктела једноставна.

[4] IBM *Watson*: http://www.ibm.com/smarterplanet/us/en/ibmwatson.

[5] *Chef Watson with Bon Appétit* апликација: https://watson.ihost.com/watson/chefwatson/page/survey.html.

резултат, у оквиру апликације корисници могу да наведу које састојке желе да користе, а које желе да избегну у јелу за које траже рецепт, као и да изаберу ког типа (на пример, италијанска кухиња, кинеска кухиња, јело које се лако припрема итд.) и које врсте (на пример, паста, пица, салата итд.) то јело треба да буде. На основу тога систем издваја све комбинације које одговарају задатим критеријумима и као резултат враћа 100 нових рецепата рангираних почев од оних који су уобичајени до оних који садрже неуобичајене комбинације састојака. Након припреме и испробавања јела према предложеним рецептима, кулинарски стручњаци су дали позитивне оцене система[6].

Како лексичку, тако и концептуалну сложеност језика хране илустровао је један од водећих светских истраживача на подручју обраде природних језика, Дан Џурафски, у својој студији „Језик хране“ (Jurafsky, 2014), приказујући еволуцију обичаја, назива, рецепата који прате говор о храни и социокултурну сложеност овог језика. Џурафски кроз призму историје, социологије и етимологије проучава повезаност кулинарских термина у различитим језицима, као и оне који се користе у једном језику са сасвим другачијим значењем у односу на језик из кога су потекли. Тако на пример изводи историјску нит која повезује називе дестилованих алкохолних пића кроз историју у разним народима почев од арапске речи *araq* која значи *слатко* и употребљавана је као метафора за кондензовани алкохол који капље из славине казана, до њених варијација које се касније јављају широм света. У Индонезији и Шри Ланки се јавља као *arrack*, црвени ликер од пиринча, односно кокоса, у Либану, Израелу, Сирији и Јордану као *Levantine arak* пиће са укусом аниса, у Турској као *raki*, у Персији као *aragh*, у Монголији као *arkhi*, а у југоисточној Европи као *ракија*. Такође, анализом ресторанских менија Џурафски даје лингвистичку потпору чињеници да пажљиво осмишљена употреба језика у њима утиче на цене јела и на то да ће нека јела бити чешће

[6] Коментари о систему *Chef Watson with Bon Appétit* дати су на адреси: http://www.bonappetit.com/tag/chef-watson.

наручивана него друга захваљујући суптилној језичкој манипулацији која укључује употребу специфичних придева и егзотичних речи.

Студија „Језик хране“ открива сву комплексност ове врсте текстова, али не даје ни најмања упутства како информатички анализирати гастрономске текстове. У овој врсти текстова преплићу се подједнако природне са друштвеним наукама, различите традиције се прилагођавају одређеном језику и поднебљу, а кулинарска вештина осцилује између занатске рутине и уметности. Ова студија открива језик кулинарства и као предмет дигиталне хуманистике, у смислу да је неопходно успоставити апарат за изучавање кулинарских садржаја методама ове дисциплине. Ово би значило да је потребан опис и изградња ресурса који су неопходни да би се кулинарски садржаји аутоматски анализирали, претраживали и повезивали, и то имајући у виду како синхрону, тако и дијахрону и вишејезичну раван. Овакав подухват је, и поред великих улагања у дигитализацију класичних текстова, сасвим маргинализован.

Наиме, и поред тога што су направљени различити каталози класичних и других текстова (Biodiversity Heritage Library, 2015; Europeana, 2015; The New York Public Library, 2015) и већи део њих је јавно доступан за прегледање, њихово претраживање је веома ограничено. Оно се у већини случајева своди на претраживање по мета-подацима којима се описују други подаци или сами материјали, њихова структура и намена (на пример, наслов, година издања, аутор итд.). На тај начин се проналазе одређени документи, али они морају да се додатно прегледају у потрази за траженом информацијом. Самим тим не постоји ни могућност да се документи који садрже тражену информацију међусобно повежу па да се пружи потпунији преглед жељених информација. Претраживање по самим информацијама које су садржане у таквим дигиталним збиркама подразумевало би да су материјали у формату који препознаје текст и омогућава његово претраживање. То би значило да се поред сканирања класичних текстова у сликовни формат, они затим обраде неким од софтвера за оптичко препознавање карактера и тиме конвертују у неки од текстуалних формата погодних за претраживање. Проблем који произилази из овог поступка, а манифестује се грешкама које се јављају у

резултујућим текстовима и које је неопходно накнадно кориговати, налази корене првенствено у квалитету штампе полазних материјала, њиховој старости и очуваности, али и у квалитету самог софтвера за оптичко препознавање карактера. Овакав процес је обиман и скуп, па се углавном своди на појединачне и ограничене истраживачке напоре у изолованим експериментима што не доприноси глобалном и универзалном решавању проблема. Када би такво решење било реализовано при проучавању говора о винима какав је представљен у књизи „Вино и конверзација“ (Lehrer, 2009), могла би се поред лексике, семантике, комуникативне и социјалне димензије која се јавља у новијем говору о винима, направити и историјска паралела како у вишејезичној француско-енглеској, тако и историјској перспективи. Примера ради, била би на располагању веза између категоризације *бордоа* из 1850. писане на француском језику (Cocks, 1850)[7] и збирке ресторанских менија писаних на енглеском језику из периода од 1800. до 2008. године[8], чије је прикупљање започела Франк Бутолп. Тренутно је претрага ових извора ограничена. Претрага првог извора није могућа, а ако се у другоме траже менији у којима се јавља податак да се служило вино *бордо* (изворно *Bordeaux*) добијају се само 3 менија из 2 кафеа и то због тога што се *Bordeaux* јавља у називима тих кафеа, а не због тога што садрже *бордо* међу винима која послужују.

Обрада оваквог обиља језичке грађе представља и информатички изазов и, по свој прилици, тек предстоји развој информатичке дисциплине која би повезала ове различите аспекте гастрономије и кувања у јединствен информатички простор. На први поглед, процес кувања, у свом упрошћеном облику, наликује на алгоритамски описив процес, па би се могло очекивати да је куварску реализацију једног рецепта могуће лако приказати моделом који би омогућио једноставну манипулацију рецептима, али, као што ће бити показано, то није случај.

[7] Категоризација *бордоа* доступна на адреси: http://www.biodiversitylibrary.org/item/71732 (Biodiversity Heritage Library, 2015).

[8] Збирка ресторанских менија доступна на адреси: http://digitalcollections.nypl.org/collections/buttolph-collection-of-menus#.

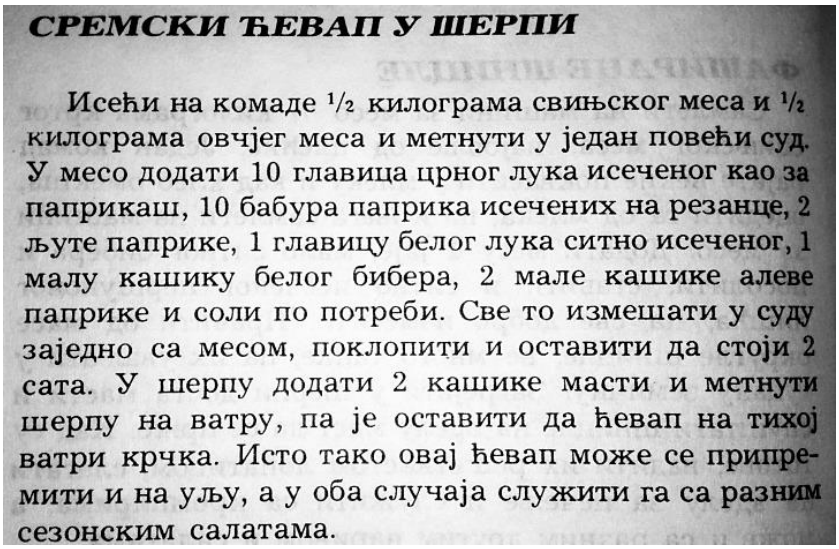
У овом уводном делу биће састављен преглед питања која би требало решити како би истовремено кување, као практична делатност, имало своју потпунију информатичку подршку, али и како би се могли пратити различити културолошки аспекти сачињавања једног јела.

Као узорак текстова на коме се могу разматрати ова питања, биће формирана колекција докумената која је састављена из хетерогених извора. Поред рецепата прикупљених са различитих веб-страница, оваквом узорку треба додати и рецепте из традиционалних куварских приручника и уџбеника кулинарства и гастрономије, али и општије, гастрономску и другу литературу, као и референце о храни из литерарних текстова. Колекција би требала да буде и вишејезична како би се могле посматрати методе лексичког и структурног трансфера из једног језика у други. Овакви трансфери нису увек једнозначни и зависе од контекста и природе самих језика, па је на пример Станислав Винавер у преводу Раблеовог дела „Гаргантуа и Пантагруел“ (Rable, 1989) у „*Net, net, à ce pyot!*“ уобичајени назив за обично вино *pyot*, уместо директног или описног превода, навео као *сласт* „*Дај ми оне сласти тамо!*“ (Đurin, 2012), што би довело до погрешног поистовећивања појма *обично вино* или *домаће вино* са појмом *сласт* при аутоматском успостављању међујезичких веза. Вишејезични аспект ће бити ограничен расположивим средствима за овакво истраживање.

За издвајање информација о храни из овако замишљене колекције докумената неопходно је, пре свега, довести рецепте на исти формат идентификацијом њихове структуре. У материјалу који се може прикупити са веб-страница, структура рецепата је углавном обележена имплицитно, коришћењем HTML-етикета:

```
<h1>Sremski ćevap u šerpi</h1><br>
<h2>Sastojci</h2><br>...
<h2>Način pripreme</h2><br>
<p>Iseći na komade meso, i svinjsko i ovčije, i staviti u veći sud.
U meso dodati 10 glavica crnog luka isečenog kao za paprikaš, 10
babura, 1 sitno isečenu glavicu belog luka, 1 malu kašiku belog
bibera, 2 male kašike aleve paprike i so po potrebi. Sve to izmešati
u sudu zajedno s mesom, poklopiti i ostaviti da stoji 2 sata. U
šerpu dodati 2 kašike svinjske masti i staviti šerpu na vatru, pa
ostaviti da ćevap na tihoj vatri krčka. Ovaj ćevap se može
pripremiti i na ulju, a uvek se služi s raznim sezonskim
salatama.</p>
```

За разлику од оваквих извора, референтни национални кувар (Marković, 1956) не садржи издвојене састојке као посебан део рецепта (слика 1).



СРЕМСКИ ЋЕВАП У ШЕРПИ

Исећи на комаде ½ килограма свињског меса и ½ килограма овчјег меса и метнути у један повећи суд. У месо додати 10 главица црног лука исеченог као за паприкаш, 10 бабура паприка исечених на резанце, 2 љуте паприке, 1 главицу белог лука ситно исеченог, 1 малу кашику белог бибера, 2 мале кашике алеве паприке и соли по потреби. Све то измешати у суду заједно са месом, поклопити и оставити да стоји 2 сата. У шерпу додати 2 кашике масти и метнути шерпу на ватру, па је оставити да ћевап на тихој ватри крчка. Исто тако овај ћевап може се припремити и на уљу, а у оба случаја служити га са разним сезонским салатама.

Слика 1. Пример текста рецепта из референтног националног кувара (Marković, 1956).

Такође, у текстовима који описују припрему, рецепт или референца на припремљени оброк могу бити угњеждени било где у тексту. Тако се, на пример, у Мијоовом „Речнику заљубљеника у гастрономију“ (Mijo, 2012), под одредницом „сликарство“, наилази на фрагмент

Има ту младог лука, тај мирис чак и преовлађује. Има и зачинских трава у лонцу, а њихов мирис ми голица ноздрве. Нарочито мирис шафрана, кога не може да ту не буде. После неколико тренутака, у својој глави већ састављам рецепт: бело вино у рибљој супи с поврћем у којој ће се ража истиха кувати, свежа павлака коју ћу додати кад се укрчка, а што се тиче острига, њих пржим како се то некада радило, и стављам их преко куване рибе.

као асоцијацију на Шарденову слику *Ража* (слика 2).

Иако горњи фрагмент има форму рецепта (који потиче од Ива Пинара (Pinard, 2010)), у опис је унет значајан део неизречене куварске вештине: нема количина, ни температура, ни трајања припреме. Питање које се може поставити је онда да ли је могуће за горњи опис пронаћи у колекцији рецепата,

коректно описане рецепте који су му слични. Таквих рецепата има на десетине, али су на различитим језицима.

Слика 2. Шарденова слика *Ража*
(Jean-Baptiste-Siméon Chardin, *The Ray*, 1728, Musée du Louvre, Paris).

Полазећи од овог примера, јавља се и општи проблем класификације рецепата. Наиме, у традиционалним куварима, рецепти се класификују према типу оброка (на пример, предјело, међујело, главно јело, итд.) или према састојцима. Али сличност рецепата у смислу сличних или истоветних техника припреме, сличних састојака или могућности додатака, излази из оквира како уобичајеног приказивања рецепата, тако и из кулинарских приручника.

Ово поставља следећа општа питања. Да ли је могуће развити класификацију јела према концептуалним критеријумима који ће изразити сличност поступка припреме, не водећи рачуна о састојцима или називу самог јела? Таква класификација би морала да укључи прецизне кулинарске таксономије, као и регистре супститута за поједине намирнице. Такође, уколико таксономија треба да буде вишејезична, потребно је обезбедити лексичке еквиваленте са географском релевантношћу супститута.

Поред ових општих проблема класификовања, у рецептима се јављају различите врсте именованих ентитета. У односу на усвојене дефиниције именованих ентитета (Chinchor, 1997; Chinchor i Marsh, 1998; Sekine i Nobata, 2004), рецепти садрже специфичне класе ентитета које нису обухваћене литературом са подручја екстракције информација. Ово се огледа на више нивоа. С једне стране, у описима састојака у рецептима, користе се непрецизне мере као што су *чаша*, *кашичица*, *струк*, *веза* итд. којима се само приближно описује потребна количина. Слично је с потребним временом за припремање, температурама, итд. С друге стране, читав низ специфичних састојака, назива јела и кухињског прибора у тексту рецепата се понаша као именовани ентитет. Било да је реч о врстама алкохола (на пример, специфичне врсте рума), о биљкама (на пример, врсте кромпира), врстама меса (код којих се назначује порекло) или о називима јела, наилазимо на објекте који, ван кулинарског домена, или имају сасвим друкчије значење (на пример, *бордо* може означавати боју) или пак немају никакво (на пример, *омлет*). Неки од кулинарских појмова и у самом кулинарском домену имају више различитих значења. Такав је *ћувеч* који се користи и као назив јела од поврћа које се пече у глиненом суду, али и као назив кухињског глиненог суда.

Тако посебан тип именованих ентита чине они у чијој структури учествују властита имена, антропоними и топоними који губе функцију засебних именованих ентитета и граде именичке синтагме које овде имају посебно значење, а нису непосредан предмет садржаја опште лексике. На пример, секвенце *говеђа шницла* и *Карађорђева шницла* имају исту структуру (присвојни придев за којим следи именица). Па ипак, у првом случају таква структура припада општој лексици, где означава да је шницла направљена од говедине. Међутим, када се општи појам *шницла* јави у истој структури у примеру *Карађорђева шницла*, та структура постаје један објекат, једна именичка синтагма, која мора да се анализира као једна целина. Ова структура не носи значење да је шницла направљена од Карађорђа или да припада Карађорђу. То је именичка синтагма која представља засебан именовани ентитет чије препознавање захтева посебан приступ (Vujičić Stanković i Pajić, 2015).

Сами састојци подлежу таксономијама које омогућавају да се прецизно идентификују намирнице као објекат трговине. Али, у кулинарском смислу, ове таксономије нису довољне, посебно ако се у обзир узме тежња човека да свакодневно конзумира здраве, разноврсне и количински уравнотежене оброке у складу са својим дневним потребама и здравственим стањем, које евентуално намеће ограничења у уносу одређених намирница. Због тога је неопходно да таксономије буду на адекватан начин трансформисане тако да намирнице буду распоређене у групе у односу на различите критеријуме какви су врста, калоријска вредност, укус и слично.

У структурирању кулинарских таксономија значајну улогу има и семантика кулинарских појмова. Из угла обједињавања кулинарских појмова истог или сличног значења у лексичке концепте важну улогу има семантички лексикон *WordNet* (Fellbaum, 1998). На пример, према њему *сируп од шећера*, *шербе* и *шербет* представљају један лексички концепт – *слатку течност од воде с медом или шећером којом се заливају колачи*.

Поред дефинисања концепата WordNet успоставља различите лексичке и семантичке релације градећи својеврсну семантичку мрежу појмова. Најчешће релације које се јављају су хиперонимија/хипонимија (општији појам/ужи појам, то јест родитељ/дете, као на пример *воће*/*јабука*), меронимија/холонимија (део од/целина од, као на пример *сируп од шећера*, *шербе*, *шербет*/*оријентални колач*) и синонимија/антонимија (различите речи означавају исте/супротне појмове, као на пример *кандирано воће*, *ушећерено воће*/*хладно*, *вруће*). Како је WordNet развијен за различите језике, он пружа могућност успостављања вишејезичних паралела међу појмовима који су у њему утемељени.

Називи рецепата се такође не могу посматрати без проучавања њихове историје и географског порекла. Тако се под истим називом у различитим крајевима могу подразумевати потпуно различита јела, али и различити називи могу да се односе на исте или сличне рецепте. На пример, *пилав* у неким крајевима означава широке домаће резанце, док га у српском значењу сматрамо јелом од пиринча, поврћа и пилетине које има бројне варијанте. Наш референтни национални кувар наводи пет варијанти овог јела. На већ

примећене промене значења појма, како је то био случај са појмом *ракија*, тако се и ово јело јавља у различитим културама где носи другачије називе, као на пример *бирјани* у Индији, *полов*, *катех*, *тајин* у Персији или *плов* у Централној Азији.

Наведена питања биће размотрена на примеру рецепта који се јавља како у нашој тако и у другим националним кухињама. Сваки од ових рецепата се састоји од млевеног меса са додацима које се обликује на различите начине и на различите начине термички обрађује. Само у референтном националном кувару (Marković, 1956) појављује се 19 различитих верзија овог јела – *ћевапа* или *ћевапчића*.

Оно што њих у начелу разликује јесте да ли су направљени од јунећег, овчјег, свињског или неког другог црвеног меса, сомовине или јегуље. Може им бити додато различито поврће (од чега је додавање црног лука присутно у свим рецептима) и зачини. Потом се разликују по начину формирања смесе и, на крају, у начину термичке обраде која може бити печење на ражњу, у пергамент-папиру, на отвореној ватри, пирјање или пржење у шерпи. И поред нијанси које разликују ове рецепте, концепт је исти.

Сличан концепт јавља се између осталих и у босанској и хрватској кухињи, под истим називом, али и у бугарској где се назива *кебабче*, македонској под називом *ќебапи* или румунској под називом *mititei*, где променом облика из ваљкастог у лоптасти почињу да се претварају у ћуфте и еволуирају у такозване *pârjoale*. Различит визуелни идентитет ћевапа кроз Балканске земље приказао је Слав Петров у својеврсном путопису (Petrov, 2009). И поред тога што је ово јело општепознато, јавља се лексичка празнина и проблем његовог вишејезичног разумевања у алатима за аутоматско превођење.

Преглед вишејезичног превода његовог румунског назива *mititei* у алатима *Google translate*[9] и *Babylon free*[10] (за језике који су подржани) представљен је у табели 1.

[9] Google translate: https://translate.google.com.

[10] Babylon free: http://translation.babylon.com.

Упоредни преглед показује да једино у случају енглеског језика у оба алата назив није само презаписан, али и у том случају није дат његов локални назив или заменски назив јела које се припрема на сличан начин, већ кратак опис.

Табела 1. Упоредни вишејезични превод румунског назива јела *mititei* у алатима *Google Translate* и *Babylon free*.

Језик	**Превод** **Google Translate**	**Превод** **Babylon free**
српски	*мититеи*	-
хрватски	*mititei*	-
бугарски	*mititei*	-
француски	*mititei*	*Boulettes de viande* (срп*. куглице меса*)
енглески	*grilled minced meat rolls* (срп*. ролнице грилованог млевеног меса*)	*Meat balls* (срп*. месне лоптице*)

Овај пример води до закључка да чак и када је могуће да кулинарски појам буде исправно преведен у одговарајући концепт који постоји у другом језику то није увек случај. Поред тога, ако у другом језику концепт нема одговарајућу реализацију онда се прибегава његовом опису који не мора нужно да одговара појму који се преводи. За било какво вишејезично претраживање било би неопходно обезбедити лексичке ресурсе који на адекватан начин описују знање о кулинарском домену.

Следећи текст комплетног рецепта на српском језику у себи садржи елементе о којима је раније дискутовано – посебне именоване ентитете кулинарског домена, стандардне и нестандардне (непрецизне) мерне јединице, као и непрецизности у опису рецепта попут упутства „правити ћевапе“, где се подразумева да читалац зна каквог облика и које величине треба да буду ћевапи, „на грилу“, без наведене температуре и прецизног времена припреме, већ ослањајући се на опис да се пеку „док не постану румени и хрскави“.

„Ћевапи

600 г млевеног јунећег меса
300 г млевеног свињског меса
40 мл минералне воде
1 кашичица соде бикарбоне
со
бибер
1 беланце
2 чена белог лука

У чинију ставити месо, минералну воду, соду бикарбону, зачине, беланце и исецкан бели лук. Промешати руком смесу и правити ћевапе. Руке премазати уљем док се праве ћевапи, јер ће се лакше формирати и смеса неће лепити за длановe. Испећи ћевапе на грилу док не постану румени и хрскави."

Када се овај рецепт преведе алатом Google Translate, добија се:

„Kebabs

600g minced beef
300 g minced pork
40 ml of mineral water
1 teaspoon baking soda
salt
pepper
1 egg white
2 cloves garlic

In a bowl place the meat, mineral water, baking soda, spices, egg whites and chopped garlic. Stir the mixture by hand and make kebabs. Hands coated with oil until meatballs are made, it will be easier to form and the mixture will not stick to the hands. Bake the kebabs on the grill until they become yellow and crunchy."

У овом случају је назив преведен са *kebabs*, што није био случај у преводу са румунског на енглески језик. То указује на непотпуност коришћених ресурса овог преводиоца. Поред тога, јављају се непрецизности какав је превод чена белог лука у *cloves garlic* и дословних превода реченица са српског на енглески језик које нису исправне конструкције енглеског језика. Посебно се издваја последња реченица у којој би се читалац рецепта на енглеском сасвим довео у забуну чекајући да, према основном преводу, добије ћевапе жуте боје (енгл. *yellow*), док би према алтернативним преводима чекао да добије ружичасте (енгл. *rosy*), зајапурене (енгл. *flushed*) или пак на последњем месту, исправно, румене (енгл. *ruddy*) ћевапе. Овакви примери показују да је

неопходно да се уложе додатни напори како би, када је реч о превођењу текстова из кулинарског домена, и вишејезични аспект био вишег квалитета.

За разлику од других области које су повезане са свакодневним животом, а које су добиле своју информатичку подршку, подручје гасторномије и кулинарства је, посматрајући из информатичког угла, скоро недодирнуто и изненађујуће неуређено. Чињеница да постоје десетине сајтова на различитим језицима и у различитим друштвима, претраживих, понекада по врло сложеним критеријумима, ствара тек привидан утисак о лакој доступности информација.

У светлу уводних напомена развијена је тема овог рада. На почетку ће бити дат приказ основних појмова, области рачунарске лингвистике, рачунарских ресурса и алата. Потом ће бити описан њихов развој, интеграција и примена као информатичка подршка у подручју гастрономије.

1.2 Текст у електронском облику као основни предмет истраживања

Текст у електронском облику је један од централних објеката који су омогућили настанак савременог информационог друштва. Било да се промене које су настале у друштву приписују развоју информационих и комуникационих технологија, било променама у организацији друштва, текст суштински учествује у начину на који се одвијају ове промене. Овој улози текста је значајно допринела еволуција самог овог појма коју су одредиле како нове технолошке могућности, тако и промишљања његовог опсега у оквиру постструктуралистичких теорија.

Премда је улога текста изузетна, његове прецизне дефиниције нема. Наиме, исти објекат користе и проучавају различите дициплине које га одређују према властитим критеријумима који нису међусобно сагласни. Из угла овог рада од значаја је лингвистичка и информатичка интерпретација овог појма.

У лингвистици се текст посматра као објекат који се састоји од реченица и који представља смислену целину. На овај начин књижевно дело, уџбеник

математике, енциклопедија или преписка представљају примере текста у овом смислу. Овакво поимање текста се потпуно ослања на људску способност разумевања језика и искључује разматрање његових формалних аспеката. Тако су, на пример, у „Речнику српскохрватскога књижевног језика“ (RMSMH, 1976) наведена значења одреднице *текст* где је основно значење (а) „написане или наштампане речи, садржина (неке књиге, рукописа закона, исправе и сл.)“, (б) „речи, садржина из чијег причања, приповедања (које обично неко запише)“, (в) „написане или наштампане речи, чланак уз слике, цртеже, примедбе и сл.“ поред значења „речи уз музичко дело, уз ноте“ и „врста штампарских слова...“. Овај регистар значења, који потиче из прединформатичког доба, већ тада укључује и писани и говорени језички материјал уобличен у различите форме и праћен изванјезичком грађом (на пример, цртежима, сликама, нотама, итд.). Физички канал преноса текста ограничен је на традиционалне медијуме комуникације – на папир или звук.

У опису информационог друштва, улога текста је обично подређена појмовима информације или знања премда су они садржани у најопштије дефинисаном (или формализованом) појму текста. Наиме, у формалној равни, текст се може одредити као секвенција симбола[11] над одређеном азбуком. С друге стране, таква секвенција добија своју информатичку репрезентацију у облику секвенције карактера: на овај начин свака секвенција карактера, била она трајно сачувана на неком медијуму или не, се може схватити као примерак текста. Дакле, из информатичког угла, текст у електронском облику је формални објекат који може бити било која секвенција карактера. Из угла програмирања, оваква интерпретација појма текста налази се већ у дефиницији концепта датотеке у систему *Unix* и језику *C*, где је датотека једнодимензиони низ бајтова. У том смислу је појам датотеке информатички врло сличан појму текста, јер се датотека која садржи текст у лингвистичком смислу и било која датотека не разликују, осим по екстерним критеријумима.

[11] Симбол као означитељ.

Проблем је како да се секвенција карактера за коју се зна да је текст у лингвистичком смислу опреми информацијама које ће је чинити језичким објектом. Примера ради, ако се насумично откуца текст са тастатуре и сними у рачунару, у информатичком смислу то ће бити текст, али у лингвистичком смислу неће, јер није организован ни једним природним језиком. Оно што је битно за текст, да би био језички објекат, јесте да буде организован одређеним природним језиком. Та компонента се из информатичког угла занемарује.

Корак ка структурирању секвенције карактера која представља текст је у информатичком смислу дат рано: у језику *Pascal* појам текста је ограничен на датотеку која се састоји од редова раздвојених карактером за нови ред (Pascal ISO/IEC 7185:1990, 1990). Сиромаштво ове структуре је већ почетком 60-тих година разматрано кроз Голдфарбову дефиницију GML-а (*Generalized Markup Language*) (Goldfarb, 1980), да би еволуирало 1987. у стандард SGML (*Standardized Generalized Markup Language*) (ISO 8879:1986, 1986). Тада се по први пут прецизно дефинише језик за обележавање и уводе нова правила за структурирање секвенција карактера тиме што поједини карактери или секвенције карактера чине посебне елементе, етикете, које добијају специјалну улогу као што је дефинисање логичке или графичке структуре текста. Опис структуре је уопштен кроз дефиницију типа документа DTD (*Document Type Definition*), којом се исцрпно описују правила која структура текста одређеног типа мора да задовољи, то јест дефинише се мета-језик за означавање одређеног типа текста. Појава језика за обележавање HTML (*Hyper Text Markup Language*) једна је примена овог стандарда, то јест једна дефиниција типа документа.

Дакле, у делу информатичке литературе, текст се разликује од других секвенци карактера по томе што је подељен у редове (карактер *new line*), што садржи графичке карактере или етикете. Ипак, ако се посматра документ настао у неком програму за обраду текста, његова репрезентација у нпр. PDF-формату и његова сканирана верзија, из угла међуљудске комуникације, обављају исту функцију иако сканирана слика документа или PDF, у својој интерној репрезентацији, нису подељени (експлицитно) на редове. На сличан

начин се може гледати и на однос изворног кода једног програма и његову бинарну репрезентацију: бинарна репрезентација је добијена из текста програма формалним путем и изражава исто значење као и кôд програма, али другим средствима.

У овом светлу је занимљива једна рана дефиниција текста коју је дала Међународна организација за стандардизацију (ISO/IEC JTC1/SC 18, 1990): „текст је информација намењена људском споразумевању која може бити приказана у дводимензионалном облику, на пример на папиру или екрану, а који се састоји од графичких елемената као што су карактери, геометријски или фотографски елементи или њихове комбинације, а који чине садржај текста“. Овде се улога текста ограничава на људску писану комуникацију: текст је само посредник између особа које располажу могућношћу разумевања поруке у тексту. У смислу ове дефиниције, између сканиране слике једног документа, његовог изворног изгледа у електронском облику или верзије препознате неком од метода за оптичко препознавање карактера нема разлике. Све три верзије обављају исту улогу у људској комуникацији. Интерна структура записа документа није овде од значаја, а претпоставља се да дводимензиони објекат садржи значење које није у његовом запису, већ код оних који су га кодирали и оних који могу да га декодирају у комуникацији. Отуда ограничавање значења појма текста на посредничку улогу у људској комуникацији занемарује аутономно постојање текста. Овакво виђење текста само проширује и делимично прецизира традиционални концепт дат у чланку из „Речника српскохрватскога књижевног језика“ (RMSMH, 1976) и наговештава проширење појма текста ка концепту мултимедијалног документа.

Према раду (Hilbert i López, 2011), још 2007. године је 94% информација у свету било ускладиштено у дигиталном облику са тенденцијом даљег раста. Ово значи и да је највећи део текстова који се данас производи, у неком тренутку свог живота, у електронском облику (пре него што је отиснут на папиру или емитован на неки други начин). Другим речима, независно од традиционалне улоге текста у комуникацији, он се јавља као аутономни

објекат који потенцијално садржи значење, али за који *a priori* не располажемо начином интерпретирања његовог садржаја. Ова чињеница је у корену читавог низа информатичких дисциплина које полазећи од текста откривају (делове) значења које је њиме кодирано. Ове области, које су и по методама и начину третирања текста, веома различите повезује позитивна претпоставка да је текст састављен на одређеном језику, те да подлеже законитостима тога језика.

1.2.1 Обрада текста у електронском облику

Као основни проблем обраде текста у електронском облику поставља се питање како опремити ниску карактера која представља текст на неком природном језику информацијама које ће реконструисати објекат организован природним језиком и намеравано значење аутора.

Први корак у овом процесу је препознавање и обележавање крајева реченица текста. Овај задатак носи у себи парадокс. Наиме, реченица је једна језичка целина која је подвргнута строгим граматичким правилима. Да би се утврдило да је ниска карактера реченица, морала би да се дубље анализира и да се утврде законитости које задовољава. Са друге стране се све анализе (осим изузетно) врше на нивоу реченице. Уобичајено је и да се текст прво подели на реченице, па да се тек онда у оквиру реченица врши токенизација. Због тога се уводе формални критеријуми на основу којих се апроксимира крај реченице, без улажења у дубљу анализу.

Други корак при обради је *токенизација*, где се ниска карактера сегментира на токене у односу на сепараторски скуп карактера. Међу токенима се издвајају *формалне речи,* токени који су ограничени сепараторским знацима на узастопним позицијама и непосредно су условљени у једном природном језику правописним нормама тог језика (Vitas, 1993). Примера ради, токени су и интерпункцијски знаци, цифре, секвенце карактера из другог језика које не припадају карактерском скупу језика текста који се обрађује, али то нису формалне речи. Опционо се пре токенизације ради *нормализација*, чији је циљ да се максимално елиминише све оно што би

могло да води томе да токеном буде проглашено нешто што није формална реч. Тако би, примера ради, у фази нормализације, *да л'* или *ост'о* били преведени у *да ли* и *остао*, да се не би у фази токенизације превели у *да*, *л*, *'* и *ост*, *'*, *о*, итд.

Следећи међусобно непосредно повезани кораци су *аутоматска морфолошка анотација* и *лематизација*. Лематизација преводи формалне речи у основне облике речи, *леме.* Аутоматска морфолошка анотација придружује формалним речима морфосинтаксички статус како би могла да се врши даља анализа граматичке структуре коју они граде, с обзиром да се граматичка структура не посматра над конкретним речима већ над врстама речи које су њени конституенти. Наиме, да би одређене врсте речи могле да стоје на одређеним местима у граматичкој структури, морају да имају одређене граматичке категорије, тако да се формалним речима и њиховим лемама при анализирању граматичке структуре придружује *морфосинтаксички опис* који одговара опису њихових морфолошких, класификационих и синтаксичких карактеристика.

Уређена тројка (*облик речи*, *лема*, *морфосинтаксички опис*), где је *облик речи* неки од флективних или деривационих облика леме речи који одговара придруженом морфосинтаксичком опису, назива се *лексичка реч*. Лексичке речи које одговарају свим граматичким облицима речи граде *лексикон*. При повезивању формалних речи текста са одговарајућим лексичким речима може да се догоди да формална реч одговара реализацији облика речи у тексту или да буде непрепозната. У првом случају формална реч може да се пресликава у тачно једну лексичку реч, као на пример у случају формалне речи *брашном* којој одговара лексичка реч (*брашном*, *брашно*, *именица средњег рода у инструменталу једнине*), или у већи број лексичких речи када једном облику речи одговара већи број парова (лема, морфосинтаксички опис). Тако на пример формалној речи *уљу* одговарају две лексичке речи са једнаким лемама (*уљу*, *уље*, *нежива именица средњег рода у дативу једнине*) и (*уљу*, *уље*, *нежива именица средњег рода у локативу једнине*), а формалној речи *вода* одговара већи број лексичких речи од којих поједине имају различите леме као што су

(*вода*, *вода*, *нежива именица женског рода у номинативу једнине*), (*вода*, *вод*, *нежива именица мушког рода у генитиву једнине*) или (*вода*, *водати*, *аорист другог лица једнине глагола водати*). У другом случају, постоји више разлога због којих се јављају непрепознате речи, као што је обрада текста који садржи доменски специфичне речи или речи потекле из страних језика које нису обухваћене лексиконом.

У општем случају, обрада подразумева да се текст дели на све мање и мање делове, до дистинктивних јединица језика, како би се придруживањем знања реконструисало значење садржано у њему. *Морфолошка анализа* као претходна обрада текста обухвата четири корака – нормализацију, токенизацију, морфолошку анотацију и лематизацију, а после тога се врше даље обраде какве су *синтаксичка анализа*, *семантичка анализа* итд. Ове анализе могу бити одвојене и извршавати се једна за другом у наведеном поретку, али могу да се извршавају и истовремено или у циклусима. Имплементација процеса обраде текста, проблеми који се јављају током извршавања ових анализа и приступи њиховом решавању биће приказани у другој глави рада.

Када су у питању формални језици овај процес је недвосмислен с обзиром на ограничен број конституената који су дозвољени у формалном језику и на прецизно дефинисану структуру формалног језика коју они граде међусобним комбиновањем. Прецизније, код дефинисања формалних језика се на почетку одређује *азбука* – ограничен скуп знакова таквих да ни један од њих не може да се добије комбиновањем неких других из скупа. Након тога се формално описује које ниске над знацима азбуке припадају формалном језику који се дефинише. Формалан опис мора да буде такав да су њиме обухваћене све и само оне ниске које припадају том формалном језику. Формалне језике и начине њихове обраде изучава теорија формалних језика, где је показано да се за обраду ефикасно користе модели коначних стања какви су коначни аутомати и коначни трансдуктори (Aho i Ullman, 1972; Vitas, 2006).

С обзиром да природни језици не подразумевају строгу структуру као што је случај са формалним језицима, потребно је да се природни језик

апроксимира и сведе на своје референтне репрезентације којима ће се омогућити примена формалних модела као што су коначни аутомати и трансдуктори у његовој обради (Gross, 1989).

1.2.2 Корпуси као референтне репрезентације језика

Дефиниција корпуса у литератури није до краја прецизирана и формализована (Utvić, 2014). Према стандарду (JUS ISO 1087-2, 2005) *текстуални корпус* или *корпус* јесте систематично оформљена колекција електронских текстова или делова текстова који су припремљени, кодирани и ускладиштени према претходно утврђеним правилима. Према (Popović i Vitas, 2003) корпуси се могу делити према: *носачу* на преелектронске и електронске; *намени* на лексикографске, граматичке, намењене препознавању говора итд; *домену* на опште и специјализоване; *периоду* на синхроне и дијахроне; *извору* на корпусе писаних, разговорних и електронских текстова; *обиму* на статичке и динамичке; *начину анотирања* на корпусе анотиране с обзиром на морфолошке или корпусе анотиране с обзиром на синтаксичке анотације, али постоје и различите друге класификације.

Не постоји (електронски) корпус којим би се обухватили сви текстови икада написани на неком природном језику. Примера ради, СрпКор2013, корпус савременог српског језика је општег типа и састоји се од колекције текстова величине 122 милиона речи (Utvić, 2014). Садржи књижевне текстове српских писаца XX и XXI века, научне и научно-популарне текстове из домена природних и друштвених наука, административне и текстове преузете из дневних новина, часописа и магазина, као и текстове преузете са интернет портала на којима се објављују вести и фељтони. Међу текстовима корпуса се налазе и преводи књижевно-уметничких и општих текстова[12], али недостају текстови који припадају кулинарском домену.

Креирање свеобухватног корпуса би био тежак, ако не и немогућ задатак, а у појединим случајевима истраживање које треба спровести и не захтева да

[12] Подаци о корпусу савременог српског језика СрпКор2013 преузети су у септембру 2014. године са адресе http://korpus.matf.bg.ac.rs.

корпус садржи све могуће текстове. Из тог разлога се корпус формира само као узорак текстова, то јест као подскуп свих могућих текстова (Biber, 1993). Истраживања спроведена на корпусу имаће смисла искључиво уколико својства корпуса верно одражавају својства језика у целини или оног дела језика који је релевантан за истраживање, односно ако резултати добијени истраживањем корпуса одговарају резултатима који би се добили када би могло да се спроведе истраживање на свим текстовима од значаја. Корпус са наведеним својствима се назива *референтни корпус* (Leech, 1991; Manning i Schütze, 1999; McEnery, Xiao i Tono, 2006).

Референтни електронски корпуси се користе као основа за различита истраживања у областима рачунарске и корпусне лингвистике. Њихов значај се огледа у томе што се над њима, као референтној репрезентацији језика, развијају нове и унапређују старе методе у областима попут екстракције информација, а потом се развијају одговарајући системи у којима се примењују лингвистичка знања која су дефинисана и проверена на корпусима.

Питање референтности корпуса не може бити разматрано апсолутно, у општем случају, већ у односу на примену за коју је корпус намењен. Тако на пример, корпус кулинарских текстова не може бити референтан за обраду новинских чланака, и обрнуто.

1.2.3 Језик и подјезик

Особине језика који се користи у текстовима референтног корпуса препознају се кроз специјализовану лексику и употребу специфичних граматичких механизама. Овакве текстове карактерише употреба специфичног речника, скраћеница, симбола и реченичких конструкција (Grishman i Kittredge, 1986; Harris, 1991; Liddy, Jorgensen, Sibert i Edmund, 1993; Liddy, Symonenko i Rowe, 2006).

Део језика у целини који се лексички, синтаксички и семантички разликује од стандардног језика, назива се *подјезик* (Harris, 1960). Подјезици се уводе као један од начина да се семантички ограничи састав корпуса. Због

тога поједностављују задатак обраде текстова корпуса јер се на тај начин отклањају различите вишезначности. На пример, у подјезику кулинарског домена *шећер* у метафоричном значењу *драга особа* не може да се појави.

Херис је поставио темеље развоја теорије подјезика посматрајући језик и подјезик из угла формалне теорије језика. Према (Harris, 1968), подјезик представља подструктуре језика које су затворене за одређен скуп граматичких трансформација над језичким исказима. Под трансформацијама над језичким исказима Херис је подразумевао синтаксичке трансформације (на пример, изградње упитних облика или негација). Тако на пример за подјезик кулинарског домена у коме се описују упутства за припрему јела није карактеристична употреба упитних реченица, па трансформацију изградње упитних облика није потребно укључивати у скуп трансформација кулинарског подјезика. Посматрајући особености реченичких конструкција у одређеном домену, могу се уочити законитости које у том домену постоје, али нису карактеристичне за језик у целини, и обрнуто. Примера ради, за подјезик кулинарског домена су карактеристичне реченице са предикатом у инфинитиву које нису карактеристичне за стандардни језик у целини.

У формалном опису подјезика, Херис је навео да се они одликују скупом класа речи и ограниченим бројем начина на које се те класе могу употребљавати у изградњи реченичких конструкција. Класе речи су носиоци значења и информација одређеног домена. За реченичке конструкције се подразумева да ће бити описане граматиком подјезика, као скуп прихватљивих граматичких конструкција над класама речи подјезика. Тако би у текстовима кулинарског домена за класу речи *Нам* којима се именују намирнице, класу речи *КухПри* којима се именује кухињски прибор и класу глагола *Гла*, у општем случају појава реченичке конструкције „*Гла Нам у КухПри*.“, била у реду. Проучавањем и пописивањем класа речи које припадају семантичким категоријама и синтаксички исправних реченичких конструкција које припадају подјезику праве се формалне структуре информација које су садржане у текстовима домена.

Анализа особености одређеног подјезика омогућава да се оне формалније опишу и да се апроксимирају одређеним формалним језиком. Што је домен подјезика лексички, синтаксички и семантички ужи то је апроксимација формалним језицима боља. Описивање подјезика одређеног домена представља приближавање задатку екстракције информација, јер класе речи подјезика одговарају ентитетима, а реченичке конструкције структурама у које је потребно трансформисати информације садржане у текстовима.

Развој и специфичности подјезика анализирани су за текстове различитих домена какви су медицина (Sager, Friedman i Lyman, 1987), биомедицина (Friedman, Kra i Rzhetsky, 2002; Liddy, Jorgensen, Sibert i Edmund, 1993), метеоролошке прогнозе (Vujičić Stanković i Pajić, 2012) и други. Подјезици се примењују у различитим областима рачунарства. Један од раних система заснован на примени подјезика је TAUM-METEO систем развијен за аутоматско превођење временских прогноза са француског на енглески језик и обрнуто (Chevalier, Dansereau i Poulin, 1978).

Поред апроксимирања природног језика којим је текст писан његовим подјезиком, сваки од процеса обраде текста на природном језику има своје специфичности, које су последица различитих намена и корисничких захтева. Сложеност задатка обраде једног текста или колекције текстова огледа се већ у проблему аутоматске идентификације језика на коме је текст написан (енгл. *Language Identification*) (Padró i Padró, 2004; Baldwin i Lui, 2010; Milne, O'Keefe i Trotman, 2012). Ово питање данас нема позитивно решење за већину писаних језика у свету, а посебан проблем представљају блиски језици (Zečević i Vujičić Stanković, 2013; Vitas, 2014). С друге стране овог задатка јесте проблем аутоматског превођења с једног језика на други (енгл. *Machine Translation*), који је још увек далеко од довољно добре апроксимације решења (Slocum, 1985; Li, 2013). У сваком од ових процеса обраде, без обзира којој области припада, појављује се потреба за откривањем и лоцирањем мањих делова текста који су носиоци неких конкретних информација, њиховим издвајањем, класификовањем и успостављањем односа са другим информацијама. Овакве

и сличне задатке истражује област звана екстракција информација (енгл. *Information Extraction*).

1.3 Екстракција информација

Екстракција информација је подобласт рачунарске лингвистике која обухвата скуп техника обраде природних језика (енгл. *Natural Language Processing*), за проналажење информација од интереса, одређивање њиховог значења и успостављање релација међу њима употребом информатичких ресурса који описују језик. У том смислу је екстракција информација дефинисана као процес препознавања специфичних информација у неструктурираном извору података, као што је текст на природном језику, и њихове поступне или истовремене класификације у семантичке класе, чиме се постиже представљање информација у структурираном облику и омогућава да те информације буду погодне за рачунарску обраду (Moens, 2006). Другим речима, информације се издвајају, означавају и организују у строге структуре какве су базе података или онтологије, тако да им је јасно одређено значење и везе које постоје међу њима.

Треба истаћи разлику између задатака екстракције информација и претраживања информација (енгл. *Information Retrieval*). Док задатак екстракције информација обухвата анализирање информација садржаних у текстовима, задатак претраживања информација подразумева да се из скупа текстова, издвоје они који су релевантни, то јест, који одговарају информацијском упиту корисника. Иако различите, ове две области су уско повезане, јер у процесу издвајања релевантних текстова најчешће учествује и процес екстракције информација (Manning, Raghavan i Schütze, 2008).

Основни појмови, методи и технике области екстракције информација су детаљно приказани у (Grishman i Sundheim, 1996; Cowie i Lehnert, 1996; Chinchor i Marsh, 1998; Riloff i Lorenzen, 1999; Moens, 2006; Pajić, 2012). Овде ће бити представљани само они који су од интереса за овај рад, какви су ентитети и њихове релације и атрибути.

Ентитет је објекат чија је реализација на семантичком нивоу одређена значењем које је од интереса за корисника, а реализација на синтаксичком нивоу одређена нискама текста које именују тај објекат. Реализација ентитета на семантичком нивоу назива се *тип ентитета*, док се његова реализација у тексту назива *вредност ентитета* (пример 1).

ПРИМЕР 1. У реченици

Марко и Дарко путују у Београд.

су реализовани уређени парови (*тип*, *вредност*) ентитета:

$ЕНТ_1$(*лично име*, *Марко*)

$ЕНТ_2$(*лично име*, *Дарко*)

$ЕНТ_3$(*назив града*, *Београд*).

Према дефиницији представљеној током конференција *Message Understanding Conferences* MUC-6 (1995.) и MUC-7 (1998.) (Grishman i Sundheim, 1996; Chinchor, 1997; Chinchor i Marsh, 1998) основни типови ентитета су они којима се именују особе, организације, географске локације, временски изрази (изрази којима се описују датум и време) и мере (изрази за представљање новчаних вредности и процената). Ова дефиниција је током каснијих MUC, ACE[13], TAC[14], CoNNL[15] и LREC[16] конференција проширена тако да обухвата преко 100 различитих типова ентитета, као што су: наслов књиге (Brin, 1999), назив производа (Sekine i Isahara, 2000), адреса електронске поште и број телефона (Maynard, Tablan, Ursu, Cunningham i Wilks, 2001), филм и научник (Etzioni i sar., 2005) или хијерархија именованих ентитета која обухвата именоване ентитете који се најчешће јављају у текстовима новинских чланака чији је извод приказан у прилогу А (Sekine, Sudo i Nobata, 2002).

[13] ACE (*Automated Content Extraction*): http://www.itl.nist.gov/iad/mig/tests/ace.

[14] TAC (*Text Analysis Conference*): http://www.nist.gov/tac.

[15] CoNNL (*Conference on Computational Natural Language Learning*): http://www.conll.org.

[16] LREC (*Language Resources and Evaluation Conference*): http://www.lrec-conf.org.

Поред дефинисања ентитета који су значајни за решавање одређеног задатка екстракције информација, уколико је потребно, унапред се дефинише и задатак одређивања *релација* које постоје између екстрахованих ентитета. Релације могу да буду једноструке или вишеструке. Једноструке релације повезују тачно два ентитета (пример 2), док се код вишеструких прави веза између већег броја ентитета. Типичан пример вишеструких релација је екстракција информација о догађајима, где се повезују различити ентитети који означавају место, време, врсту догађаја, учеснике и слично. Додатно, задаци екстракције могу да обухвате и екстракцију *атрибута* ентитета, односно описа ентитета у тексту и њихово нормирање.

ПРИМЕР 2. Нека је дефинисан задатак екстракције информација типова ентитета *назив_десерта*, *назив_намирнице* и *састојак*[17], као и релација *је_потребно*, између њих. Из текста рецепта

> *За шненокле је потребно додати 5 јаја, 1 л слатког млека, 1 кесица ванилин шећера, 7 супених кашика шећера, 2 супене кашике шећера у праху, 2 супене кашике брашна.*

се онда процесом екстракције информација екстрахују ентитети:

ЕНТ$_1$(*назив_десерта*, *шненокле*),
ЕНТ$_2$(*назив_ намирнице*, *јаја*) који учествује у изградњи
ЕНТ$_3$(*састојак*, *5 јаја*),
ЕНТ$_4$(*назив_намирнице*, *слатко млеко*) који учествује у изградњи
ЕНТ$_5$(*састојак*, *1 л слатког млека*),
ЕНТ$_6$(*назив_намирнице*, *ванилин шећер*) који учествује у
ЕНТ$_7$(*састојак*, *1 кесица ванилин шећера*),

[17] У изградњи ентитета типа *састојак* поред ентитета типа *назив_намирнице*, учествују и ентитети којима се описује количина намирнице (на пример, *5*, *1 л*, *1 кесица* итд.), о којима ће бити више речи у поглављу 2.5.2.3.

ЕНТ$_8$(*назив_намирнице*, *шећер*) који учествује у изградњи
ЕНТ$_9$(*састојак*, *7 супених кашика шећера*),
ЕНТ$_{10}$(*назив_намирнице*, *шећер у праху*) који учествује у изградњи
ЕНТ$_{11}$(*састојак*, *2 супене кашике шећера у праху*) и
ЕНТ$_{12}$(*назив_намирнице*, *брашно*) који учествује у изградњи
ЕНТ$_{13}$(*састојак*, *2 супене кашике брашна*),

као и релације:

ЕНТ$_1$ *је_потребно*
ЕНТ$_3$ (*шненокле је_потребно 5 јаја*)
ЕНТ$_1$ *је_потребно*
ЕНТ$_5$ (*шненокле је_потребно 1 л слатког млека*)
ЕНТ$_1$ *је_потребно*
ЕНТ$_7$ (*шненокле је_потребно 1 кесица ванилин шећера*)
ЕНТ$_1$ *је_потребно*
ЕНТ$_9$ (*шненокле је_потребно 7 супених кашика шећера*)
ЕНТ$_1$ *је_потребно*
ЕНТ$_{11}$ (*шненокле је_потребно 2 супене кашике шећера у праху*)
ЕНТ$_1$ *је_потребно*
ЕНТ$_{13}$ (*шненокле је_потребно 2 супене кашике брашна*).

Задатак екстракције информација подразумева да се анализом текста екстрахују унапред прецизно дефинисане семантичке класе информација, односно типова ентитета. Начелно, за конкретне проблеме у екстракцији информација у различитим доменима могу се дефинисати нови типови ентитета, а поред тога, треба посебно разматрати случајеве у којима ентитети једног типа учествују у изградњи ентитета другог типа где задржавају или губе своје основно семантичко значење, као што је на пример случај ентитета *Беч* којим се именује топоним а који губи семантичко значење у оквиру ентитета *бечка шницла* којим се именује јело.

У подјезику кулинарског домена, опште именице добијају статус ентитета. Тако су *шненокле* из претходног примера ентитет ЕНТ$_1$ једног типа, *јаја* ентитет ЕНТ$_2$ другог типа који учествује у изградњи конструкције којом се описује састојак рецепта *5 јаја*, што представља ентитет трећег типа ЕНТ$_3$ итд. Када се у опису рецепта реализације ентитета замене, добија се структура:

„*За* ЕНТ$_1$ *је потребно додати* ЕНТ$_3$, ЕНТ$_5$, ЕНТ$_7$, ЕНТ$_9$, ЕНТ$_{11}$, ЕНТ$_{13}$.“

или општије:

„*За* ТО *је потребно додати* ТОГА, ТОГА, ТОГА, ТОГА, ТОГА, ТОГА.“

Оваква структура је предмет изучавања дependенцијалне граматике, где се глаголске допуне замењују заменицама, а онда се анализира како се реализују заменице. У кулинарском домену такве структуре су на пример „ТО *се сипа* ТУ“ или „ТО *се додаје* ТУ“ или „ТО *се измеша са* ТИМ“. Код припреме шненокли ТО = *шненокле*, ТОГА = *5 јаја*, итд. Отуда се проблем екстракције информација своди на списак глагола који се користе у рецептима, опис њихових допуна и структура које граде.

Поред прецизног дефинисања које информације у тексту систем треба да екстрахује и да ли треба да се утврде релације између екстрахованих информација, у овим системима се одређује и формат излазних података, *образац* (енгл. *template*). Образац се састоји од поља (енгл. *slots*) која представљају ентитете и њихове атрибуте, релације између ентитета или догађаје у којима они учествују. Структура обрасца треба јасно да представља и дефинише значење екстрахованих информација, да поља у излазној структури могу да се генеришу на основу улазних података и да буде погодна за даљу обраду и анализу (Pajić, 2012).

Образац који одговара задатку екстракције информација приказаном у примеру 2 о релацији ЈЕ_ПОТРЕБНО чине поља НАЗИВ_ДЕСЕРТА и САСТОЈАК, а одговарајући попуњени образац је:

ЈЕ_ПОТРЕБНО
НАЗИВ_ДЕСЕРТА: шненокле

САСТОЈАК: 5 јаја
1 л слатког млека
1 кесица ванилин шећера
7 супених кашика шећера
2 супене кашике шећера у праху
2 супене кашике брашна

Методе које се користе за опремање текста информацијама о језику и језичким информацијама се веома разликују. Ипак, различити приступи се могу груписати у две основне групе. Једну групу метода чине оне којима се настоји да се успоставе модели језика на основу његових експлицитних система граматичких правила: ове методе почивају на обимним лексичким ресурсима као и на формализованим моделима граматичких структура различитог нивоа. Другу групу метода представљају методе које се изводе из квантитативних показатеља о језику: ову групу чине пре свега методе изведене из метода машинског учења.

Системи за екстракцију информација који користе технике засноване на правилима ослањају се на лингвистичко и експертско знање. Поред унапред дефинисаних образаца који треба да буду препознати, ови системи захтевају да се унапред дефинишу и правила по којима ће да буде извршена претрага и њихово препознавање (Cowie i Lehnert, 1996; Poibeau, 2000; Chang, Kayed, Girgis i Shaalan, 2006; Sarawagi, 2008). Правила се обично дефинишу регуларним изразима, коначним трансдукторима, каскадама трансдуктора и рекурзивним мрежама прелаза (енгл. *Recursive Transition Networks* – RTN) (Friburger i Maurel, 2004; Sarawagi, 2008). Изградња ових система подразумева стручно знање из домена и лингвистичко познавање природног језика за који се метода екстракције развија.

Системи који користе технике машинског учења засновани су на теорији вероватноће (Nadeau, Turney i Matwin, 2006; Tatar i Cicekli, 2011; Liu, Zhang, Wei i Zhou, 2011). Они из скупова за учење уче да препознају жељене обрасце и уче правила за њихово препознавање која ће примењивати над новим текстовима. Могу да се примене на различите домене и природне језике без посебног прилагођавања и не захтевају стручно знање за развој правила за екстракцију информација. Њихов недостатак је у томе што подразумевају

постојање корпуса анотираних текстова (скупова за учење) који нису развијени за сваки језик и за сваки домен и примену. С обзиром да је такав случај и са српским језиком, у овом раду ће разматрање бити ограничено на системе за екстракцију информација који користе технике засноване на правилима.

1.3.1 Евалуација система за екстракцију информација

Изграђени системи за екстракцију информација имају различиту успешност у примени. Ефикасност система за екстракцију информација се мери са две основне мере. То су *прецизност* (енгл. *precision*), која мери тачност система и *одзив* (енгл. *recall*), који мери потпуност система (Lehnert i sar., 1994).

Прецизност је мера релевантних информација[18] које су екстраховане у односу на све информације које су екстраховане. Одзив је мера односа релевантних информација које су екстраховане према укупном броју релевантних информација у тексту, то јест, колико је оних информација из текста које је требало да буде екстраховано, заиста и екстраховано.

Податак о томе колики је број релевантних информација у тексту најчешће није познат, тако да се у таквим случајевима прецизност и одзив не могу тачно израчунати. Тада се врши њихова процена на основу вредности добијених из скупа текстова изабраних на случајан начин.

Ефикасност система за екстракцију информација се повећава повећањем мера прецизности и одзива, али с обзиром да оне теже да се мењају обрнуто пропорционално њихова истовремена оптимизација често није могућа. Примера ради, да би се повећао одзив треба да се повећа број релевантних информација које су екстраховане. Повећање броја релевантних информација које су екстраховане постиже се повећањем броја свих информација које су екстраховане. Тиме расте вероватноћа да ће бити екстраховано све више оних информација које нису релевантне, чиме се смањује прецизност. Због тога се

[18] *Релевантне информације* су информације које одговарају неком унапред дефинисаном критеријуму.

уводи *Ф-мера* која представља хармонијску средину између прецизности и одзива (Jurafsky i Martin, 2000).

Систем за екстракцију информација мора прво да издваја ентитете и релације које су експлицитно садржане у тексту који се обрађује, али допуна и корекција резултата може да се постигне ограничавањем примене на поједине подјезике и увођењем додатних ресурса који би носили информацију о значењу речи и њиховим односима, и омогућили откривање релација које нису изражене у самом тексту. Један од таквих ресурса су онтологије.

1.4 Онтологије

1.4.1 Дефиниција и основни појмови

Појам *онтологија* преузет је из филозофије. У свом изворном, филозофском, значењу представља науку о бићу[19], о томе који све *концепти* (типови ствари) постоје и какви су њихови међусобни односи. У информатици се појам онтологије користи за структуре које описују концепте,[20] заједно са релацијама и ограничењима која међу њима постоје.

У литератури се као референтне дефиниције онтологија у информатици наводе Груберова дефиниција (Gruber, 1993): „*онтологије су експлицитне спецификације концептуализације*“ и Борстова прецизнија варијанта (Borst, 1997): „*онтологије су формалне експлицитне спецификације дељене концептуализације*“.

Концептуализација подразумева апстрактан поглед на свет који је потребно представити из одређених, унапред дефинисаних разлога, употребом концепата и њихових међусобних веза, а *дељена концептуализација* означава да унутар заједнице која уводи или користи ту онтологију постоји договор о томе која је намена онтологије и које заједничко знање се њоме представља (Studer, Benjamins i Fensel, 1998).

[19] Етимолошки, појам *онтологија* изведен је из грчких речи *о́нтос* – биће, стварност и *логία* – наука.

[20] У информатици се концепти називају и *класе*.

Основна сврха онтологија је да се омогући аутоматизовано дељење и поновна употреба знања, како између човека и рачунара, тако и између више рачунара, кроз формализацију семантике реалног света. Приликом дељења, односно размене знања, потребно је да обе стране које учествују у том процесу, поседују одређени ниво „разумевања“ информација које размењују. У том смислу, предуслов да је онтологија *формална* означава да је потребно да онтологија буде машински читљива и „разумљива“ рачунарима.

Експлицитна спецификација онтологије подразумева представљање концептуализације у одређеном облику употребом неког од језика за имплементацију онтологија и основних елемената онтологије – класа, инстанци, релација, функција и аксиома (Gruber, 1993).

Када се посматра знање које треба представити онтологијом, на почетку се уочавају сви они објекти, све *инстанце* које чине основ тог знања. Потом се инстанце организују у *таксономије*, хијерархије *класа* у којима се знање наслеђује, тако да све инстанце које чине одређену класу имају заједничка *својства*. Ако све инстанце једне класе имају све особине друге класе и још су додатно одређене неким специфичним особинама, прва класа је *подкласа* друге класе. Појединачне инстанце се разликују различитим *вредностима својстава* која их карактеришу. Међу класама, као и међу инстанцама класа, успостављају се различите *релације*, дефинишу се *ограничења* релација и својства која морају да важе (попут транзитивности или кардиналности), као и *аксиоме* која обично имају облик „Ако је тачно *тврђење1* онда мора да буде тачно и *тврђење2*“ (Nirenburg i Raskin, 2004).

Формално, онтологија *О* се дефинише као уређена петорка *О* = (*К*, *Р*, *И*, *Ф*, *А*), где су *К*, *Р*, *И*, *Ф* и *А* редом коначни скупови класа, релација дефинисаних над скупом класа, инстанци, функција дефинисаних над скупом класа и аксиома које се користе за проверу конзистентности онтологије и знања које је у њој садржано. Аксиоме су описане правилима извођења којима се на основу знања које је експлицитно наведено у онтологији закључује ново знање (пример 3) (Bergamaschi, Guerra i Vincini, 2005).

ПРИМЕР 3. Нека постоје класе *Локација*, *Држава*, *Град*, *Река* и конкретне инстанце *Србија*, *Београд*, *Сава*, класа *Држава*, *Град*, *Река* редом. Међу њима могу да се дефинишу релације *Држава је Локација*, *Град је Локација*, *Град је_део Држава*, *Река протиче_кроз Град*, *Река протиче_кроз Држава,* функција *Град је_главни_град Држава* и аксиома „ако за инстанце *р*, *г* и *д* класа *Река*, *Град* и *Држава* важи *р протиче_кроз г* и *г је_део д*, онда важи *р протиче_кроз д*“.

Ако се дефинише да је *Београд главни_град Србија*, онда се из релације *Сава протиче_кроз Београд* и *Београд је_део Србија* може закључити да *Сава протиче_кроз Србија*.

1.4.2 Различити типови онтологија

Како би нека онтологија била применљива за одређену намену, неопходно је да се направи договор између корисника, дизајнера и доменских експерата о нивоу знања које ће бити обухваћено онтологијом (Lassila i McGuinness, 2001; Gómez Pérez, Fernández López i Corcho Garcia, 2004b; Borgo, 2007). У зависности од знања које се представља у онтологијама оне се деле на *онтологије највишег нивоа* (енгл. *top-level ontologies*), које описују опште концепте, независне од било којег домена или задатка, *доменске онтологије* (енгл. *domen ontologies*), које описују концепте везане за одређени домен и *апликацијске онтологије* (енгл. *application ontologies*), које описују концепте везане за одређене задатке (Guarino, 1998).

Онтологије се разликују по томе колико детаљно описују знање из одређене области – колико детаљно описују класе, њихова ограничења и релације. Онтологијама се сматрају и једноставни лексикони са малим бројем релација, али и богате онтологије које садрже знања о великом броју појмова и њихових релација везаних за неку област. Поред онтологија које су везане за одређени домен или прилагођене одређеним врстама апликација, постоје и веома „широке“ онтологије којима је обухваћен већи број појмова из различитих области. Додатно, оне се разликују и према начину представљања

онтолошког знања чиме се уједно представља и степен формализације (изражајности) тог знања (Gómez Pérez, Fernández López i Corcho Garcia, 2004a; McCuinness, 2005; Cimiano, 2006).

У класификацији онтологија према начину представљања онтолошког знања (слика 3), као онтологије најмање изражајности наводе се *листе појмова* са одговарајућим описима на природном језику и *контролисани речници* као листе појмова које формирају стручњаци из области за коју је контролисани речник намењен. У контролисаним речницима дефиниције којима се појмови описују нису увек једнозначне. Када се листама појмова и дефиницијама контролисаних речника дода семантика, тако што се између појмова успоставе релације синонимије, добијају се *тезауруси*. *Неформална таксономија* представља надградњу тезауруса увођењем експлицитне хијерархије појмова у којој важе генерализација и спецификација, али без строго дефинисаног наслеђивања (инстанца подкласе није нужно инстанца надкласе). Строгим дефинисањем наслеђивања формирају се *формалне таксономије*, где важи да су све инстанце подкласе уједно и инстанце надкласе. *Оквири* поред формалне таксономије укључују и дефинисање инстанци и релација међу класама. Додатно се над њима могу дефинисати различита ограничења вредности употребом математичких, логичких формула или примера ради логике првог реда, чиме се постиже највећа изражајност онтологија.

Слика 3. Начини представљања онтолошког знања (McCuinness, 2005).

Поред нивоа изражајности онтологија, на слици 3 представљена је разлика између „лаких“ и „тешких“ онтологија. Према (Corcho, Fernández-López i Gómez-Pérez, 2003), „лаке“ онтологије укључују концепте, особине које описују концепте, релације међу концептима и таксономије концепата, док „тешке“ онтологије такође обухватају и аксиоме и ограничења.

1.4.3 Онтолошки језици

За моделирање онтологија је развијен велики број језика (Gómez Pérez i Corcho, 2002). У почетку су језици развијани првенствено за примену у области вештачке интелигенције и нису били засновани на XML-у. Такви су: KIF[21] (*Knowledge Interchange Format*) заснован на предикатском рачуну првог реда, *Loom*[22] заснован на дескриптивној логици и *Shoe*[23] који представља доградњу HTML језика.

Новији онтолошки језици и модели су: XML (*eXtensible Markup Language*), RDF (*Resource Descriptiom Framework*), RDFS (*RDF Schema*), OIL (*Ontology Inference Layer*), DAML+OIL и OWL (*Web Ontology Language*).

XML[24] језик је направљен за описивање, структурирање, складиштење и слање информација. Пружа могућност прецизног дефинисања конкретног језика за обележавање. XML дозвољава корисницима да додају произвољну структуру документима, али не каже ништа о томе шта структуре значе.

RDF[25] је модел којим се првенствено представљају подаци на вебу. Развијен је како би се њиме изразило значење структура докумената. RDF описује семантичке везе између електронских извора, то јест електронских описа о стварима, конкретним или апстрактним, и релација које постоје међу њима.

RDFS је надградња над RDF-ом и омогућава дефинисање типова података за RDF.

---

[21] KIF: http://www.ksl.stanford.edu/knowledge-sharing/kif.
[22] Loom: http://www.isi.edu/isd/LOOM.
[23] Shoe: https://www.cs.umd.edu/projects/plus/SHOE.
[24] XML: http://www.w3.org/XML/ i http://www.w3.org/TR/2006/REC-xml-20060816.
[25] RDF: http://www.w3.org/2001/sw/RDFCore.

OIL језик заснован је на описној логици[26] и системима заснованим на оквирима. OIL из система заснованих на оквирима наслеђује оквире са одређеним својствима која се називају атрибути, док из описне логике наслеђује формалну семантику и ефикасну подршку за закључивање (Fensel, van Harmelen i Horrocks, 2003).

DAML+OIL[27] језик настао је као последња верзија *DAML* (*DARPA Agent Markup Language*) језика заснованих на RDFS-у, комбиновањем са OIL језиком.

OWL[28] језик изведен је из DAML+OIL стандарда у сврху креирања и размене онтологија. Намењен је програмским апликацијама, које уместо једноставне презентације информација за људско коришћење, треба да обрађују значење тих информација. OWL има неколико подјезика различитог нивоа изражајности: *OWL Lite*, *OWL DL* и *OWL Full*. У великој мери се заснива на RDF-у и RDFS-у. Синтакса *OWL Lite* и *OWL DL* подјезика OWL-а се заснива на RDF-у.

OWL Lite пружа подршку за једноставну хијерархијску класификацију са једноставним ограничењима.

OWL DL има сложенији формализам од OWL Lite-а. Формализам се заснива на описној логици, па OWL DL пружа максималну изражајност, обезбеђујући комплетност израчунавања (сви закључци су сигурно израчунљиви) и одлучивост (сва израчунавања ће се извршити у коначном времену).

OWL Full има максималну изражајност, синтаксно је независан од RDF-а, али не даје никакве гаранције по питању одлучивости. За разлику од OWL Lite и OWL DL, главна карактеристика OWL Full језика је да једна класа, која је по дефиницији колекција инстанци, може и сама да буде инстанца. Управо ова

[26] *Описне логике* (енгл. *description logics*) представљају класу језика за представљање знања уз могућност његовог претраживања и поновне употребе. Теоријска основа описних логика је предикатски рачун. Оне су одлучиви подскуп логике првог реда. Логика првог реда није одлучива, што значи да није могуће направити алгоритам који би за сваку формулу логике првог реда вратио њену вредност. Када језик није одлучив, није могуће унапред утврдити да ли формула може да буде доказана или не. Израчунавање може да буде бесконачно, без одговора да ли је формула тачна или није.

[27] DAML+OIL: http://www.w3.org/TR/daml+oil-reference.

[28] OWL Web Ontology Language Reference: http://www.w3.org/TR/owl-features.

карактеристика чини израчунавања, односно било које обраде модела заснованих на OWL Full језику, неодлучивим (потенцијално бесконачним).

1.4.3.1 OWL онтологије

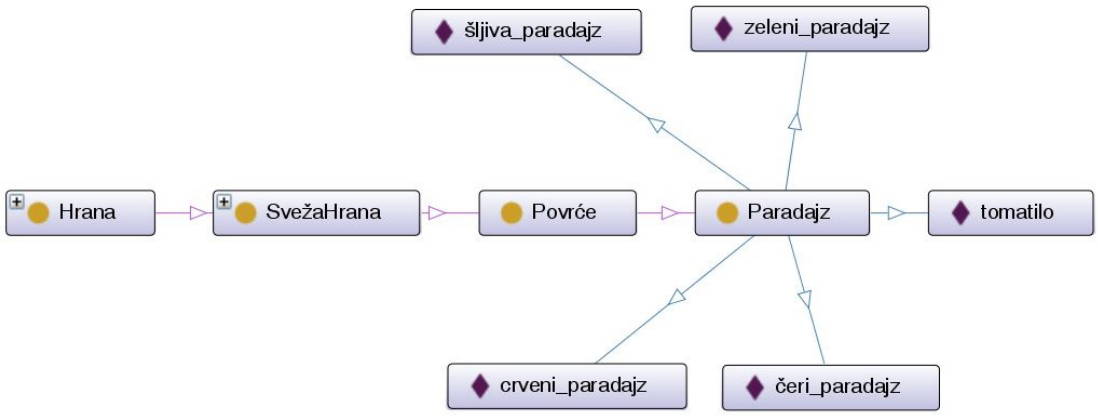
Према (Horridge, Knublauch, Rector, Stevens i Wroe, 2004) *главни елементи OWL онтологија* су класе, инстанце и својства. Класе су описане формалним описима који садрже прецизне услове којима се одређује које инстанце припадају класи. Обезбеђено је да класе и својства могу да се организују у таксономије.

Својства су бинарне релације које могу да буду инверзне, рефлексивне, ирефлексивне, симетричне, антисиметричне, транзитивне, функционалне или инверзно функционалне. Функционалне релације једној инстанци придружују тачно једну вредност, док за релацију важи да је инверзно функционална ако је њена инверзна релација функционална.

Разликују се три врсте OWL својстава: *својства објеката*, *својства типова података* и *својства анотација*. *Својства објеката* су релације између инстанци, где се за дефинисани домен релације одређује одговарајући опсег који ће чинити кодомен. *Својства типова података* су релације између инстанци и вредности података у смислу одређивања типова података или ограничења вредности података. Ограничења могу да буду квантификаторска (типа „за сваки“ или „постоји“), ограничења кардиналности (на пример, „минимална вредност“ или „максимална вредност“) или ограничења типа „има вредност“. *Својства анотација* су релације између елемената онтологије (класа, инстанци и особина) и њихових описа.

На слици 4 је приказан део таксономије онтологије хране – класа *Храна*, њена потклаcа *СвежаХрана*, са својом потклaсом *Поврће* и следећом потклaсом *Парадајз*, као и инстанце класе *Парадајз*: *шљива_парадајз*, *зелени_парадајз*, *црвени_парадајз* и *чери_парадајз*. Број инстанци класе ограничен је степеном развијености онтологије, па тако у овом случају нису наведене све могуће инстанце класе какве су примера ради *парадајз јабучар* или *воловско срце*. Пример 4 приказује одговарајући део онтологије у OWL језику који садржи и информације о њеним својствима као што су својства типова података –

дефиниција и литерал инстанци који ће бити детаљније описани у поглављу 1.4.4.



Слика 4. Део таксономије онтологије и пример инстанци.

ПРИМЕР 4.

```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF [
<!ENTITY owl "http://www.w3.org/2002/07/owl#" >
<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
<!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
<!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
<!ENTITY oh "http://www.semanticweb.org/stasa/ontologies/2014/7/oh#" >]>

<rdf:RDF xmlns="http://www.semanticweb.org/stasa/ontologies/2014/7/oh#"
xml:base="http://www.semanticweb.org/stasa/ontologies/2014/7/oh"
          xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
          xmlns:owl="http://www.w3.org/2002/07/owl#"

xmlns:oh="http://www.semanticweb.org/stasa/ontologies/2014/7/oh#"
          xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
          xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

<owl:Ontology
rdf:about="http://www.semanticweb.org/stasa/ontologies/2014/7/oh">
<rdfs:label>Ontologija hrane</rdfs:label>
</owl:Ontology>
<!-- Svojstva tipova podataka -->
...
<owl:DatatypeProperty rdf:about="&oh;definicija">
       <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:about="&oh;literal">
       <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- Klase -->
...
<owl:Class rdf:about="&oh;Hrana"/>
```

```
<owl:Class rdf:about="&oh;SvežaHrana">
       <rdfs:subClassOf rdf:resource="&oh;Hrana"/>
</owl:Class>
<owl:Class rdf:about="&oh;Povrće">
       <rdfs:subClassOf rdf:resource="&oh;SvežaHrana"/>
</owl:Class>
<owl:Class rdf:about="&oh;Paradajz">
       <rdfs:subClassOf rdf:resource="&oh;Povrće"/>
</owl:Class>

<!-- Instance -->
...
<owl:NamedIndividual rdf:about="&oh;crveni_paradajz">
       <rdf:type rdf:resource="&oh;Paradajz"/>
       <oh:definicija>blago kiselo crveno voće koje se jede kao
povrće</oh:definicija>
       <oh:literal>crveni paradajz</oh:literal>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="&oh;tomatilo">
       <rdf:type rdf:resource="&oh;Paradajz"/>
       <oh:definicija>mali, jestiv paradajz sa ljuskom; žute je do
ljubičaste boje sličan voću</oh:definicija>
       <oh:literal>tomatilo</oh:literal>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="&oh;zeleni_paradajz">
       <rdf:type rdf:resource="&oh;Paradajz"/>
       <oh:definicija>kiselo zeleno voće koje se jede kao povrće i
koristi za pripremu turšije</oh:definicija>
       <oh:literal>zeleni paradajz</oh:literal>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="&oh;čeri_paradajz">
       <rdf:type rdf:resource="&oh;Paradajz"/>
       <oh:definicija>mali, jestiv paradajz sa ljuskom crvene
boje</oh:definicija>
       <oh:literal>čeri paradajz</oh:literal>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="&oh;šljiva_paradajz">
       <rdf:type rdf:resource="&oh;Paradajz"/>
       <oh:literal>šljiva paradajz</oh:literal>
       <oh:definicija>duguljast čeri paradajz</oh:definicija>
</owl:NamedIndividual>
</rdf:RDF>
```

1.4.3.2 SPARQL упити

Над овако структурираним онтологијама могуће је постављати упите у SPARQL формату (Segaran, Evans i Taylor, 2009; Yu, 2011).

SPARQL (*SPARQL Protocol and RDF Query Language*) је упитни језик за податке ускладиштене у RDF формату који је 2008. године стандардизовала *W3C SPARQL* радна група[29]. SPARQL упити пружају могућност издвајања, уметања и брисања података.

[29] W3C SPARQL: http://www.w3.org/2009/sparql/wiki/Main_Page.

Постоје четири врсте SPARQL упита: SELECT, ASK, CONSTRUCT и DESCRIBE. SELECT упит је намењен претрази података. ASK је намењен провери да ли наведени упит има резултат, али не враћа информацију о резултујућим подацима. CONSTRUCT је намењен креирању нових података у одговарајућем формату на основу постојећих. DESCRIBE служи за издвајање информација које описују податке (DuCharme, 2013).

Примери 25 и 26 приказују неколико SPARQL упита који су коришћени у оквиру рада на овој дисертацији.

1.4.4 WordNet као онтологија

WordNet је семантичка мрежа која се састоји од чворова који описују концепте и лукова који приказују семантичке везе између тих концепата.

Први WordNet, пристонски WordNet, развијен је за енглески језик[30]. Његов развој започет је 1985. године у истраживачком центру *Cognitive Science Laboratory* на Принстонском универзитету. Пројектом је руководио психолог Џорџ Милер, са намером да се направи „ментални лексикон“, својеврсну лингвистичку базу података чије организација и структура треба да буду такве да подражавају начин на који људски ум складишти и користи језичке појмове. Милер и његови сарадници су имали намеру да овакву лингвистичку базу података користе првенствено у психолингвистичким пројектима, али је њена употреба проширена у различитим рачунарским областима (Fellbaum, 1998a). Верзија 3.1 принстонског WordNet-а за енглески језик садржи 155.287 лема (именица, глагола, придева и прилога) организованих у 117.659 синсетова[31].

Чворови WordNet-а се називају *синсетови*. То је реч потекла од енглеског појма *synset*, односно *synonymous set*, којим се означавају скупови *блискозначних речи* (енгл. *near synonym*). Појам синонимије који је коришћен при изградњи синсетова WordNet-а (Miller, 1995), подразумева да су два израза блискозначна у одређеном језичком контексту ако замена једног

[30] Принстонски WordNet (PWN): http://wordnet.princeton.edu.

[31] Статистика преузета са http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html о PWN у мају 2015. године.

израза другим не мења истинитосну вредност језичког контекста (Zgusta, 1971). То значи да скупове блискозначних речи не чине само они чланови чије значење је потпуно исто у одређеном контексту, већ и они чије се значење разликује у нијансама. У примеру 5 је приказан синсет коме припадају речи *потрошач* и *муштерија,* са значењем *особа која купује или користи добра или услуге*. Док ова дефиниција добро одговара опису *потрошача*, *муштерија* описује незнатно шири појам с обзиром да се особа сматра муштеријом и у случају када само показује интересовање да купи или користи добра или услуге.

ПРИМЕР 5.

```
LNOTE: N28+Hum
LNOTE: N601+Hum+MG+FG
POS: n      ID: ENG30-09612848-n     BCS: 2
Synonyms: potrošač:1, mušterija:1a

Definition: osoba koja kupuje ili koristi dobra ili usluge
Last Edit: Cvetana 2004/07/27

-->> [hypernym] +[n] korisnik:1
```

Сваки синсет садржи јединствену идентификациону ознаку (`ID`). Тако у примеру 5, синсет има јединствену идентификациону ознаку „ENG30-09612848-n“. Идентификационе ознаке концепата принстонског WordNet-а почињу ниском `ENG`. Поред тога, сваки синсет садржи речи које се називају *литерали* и чије значење одговара значењу синсета који се дефинише.

Једна реч може да припада већем броју синсетова, с обзиром да једној речи може да одговара више значења. У том случају се иза литерала наводи ознака (*sense*) којом се он на јединствен начин повезује са синсетом који одговара његовом значењу у одређеном контексту. Исти литерал коме је придружена друга ознака биће повезан са синсетом који дефинише његово друго значење. Ове ознаке се записују цифрама и словима. Тако је у примеру 6 приказано да литерал *кора* са придруженом ознаком 7 има значење описано кратком дефиницијом (`Definition`) „Развијено танко тесто за пите и гибанице“, док исти литерал са придруженим ознакама 4 и 2 има редом

значења „Природни спољни омотач плода (обично се уклони пре јела)“ и „Тврд спољашњи омотач око нечега.“.

ПРИМЕР 6. Синсет за литерал *кора* и ознаку 7:

```
POS: n      ID: SRP-1498988473
Synonyms: kora:7, jufka:1

Definition: Razvijeno tanko testo za pite i gibanice.
Usage: Jedan Jugosloven iz Švedske se preko Unije raspitivao
 kako  bi mogao u Skandinaviju da uvozi kore za gibanicu.

Note: Thinly spread dough used for "pita" and "gibanica".
Last Edit: Cvetana 2004/07/18
-->> [hypernym] $[n] hrana:1a
-->> [holo_part]
-->> [holo_part] [n] burek:1
-->> [holo_part] [n] pita od visxanxa:1
-->> [holo_part] [n] gibanica:1
```

за литерал *кора* и ознаку 4:

```
POS: n      ID: ENG30-07670731-n
Synonyms: kora:4, ljuska:1x

Definition: prirodni spoljni omotoač ploda
(obično se ukloni pre jela)
Last Edit: Cvetana 2013/06/20

-->> [hypernym] $[n] građa:1x, materija:2x
<<-- [hyponym] [n] kožica:1
```

за литерал *кора* и ознаку 2:

```
POS: n      ID: ENG30-09260218-n   BCS: 2
Synonyms: kora:2

Definition: Tvrd spoljašnji omotač oko nečega.
Last Edit: Cvetana 2004/06/21

-->> [hypernym] $[n] prirodni pokrivač:X
```

Сваки синсет садржи и податак о томе да ли припада именичкој, глаголској, придевској или прилошкој групи синсета (`POS`). Примера ради, припадност именичкој групи је означена са „POS: n“. Синсет опционо садржи примере употребе литерала из синсета за означавање тог концепта (`Usage`), напомене (`Note`) и податке о томе ко је и када направио последње измене

синсета (`Last Edit`). У случају првог наведеног синсета примера 6 је „Usage: Jedan Jugosloven iz Švedske se preko Unije raspitivao kako bi mogao u Skandinaviju da uvozi kore za gibanicu.“, „Note: Thinly spread dough used for "pita" and "gibanica".“, и „Last Edit: Cvetana 2004/07/18“.

Подаци о релацијама синсета са другим синсетовима наведени су између заграда „`[`“ и „`]`“. Преглед основних релација између именичких синсетова приказан је у табели 2.

Табела 2. Основне именичке релације у WordNet-у.

Релација	**Опис**	**Пример**
хиперонимија	ако је А општији појам[32] од Б, онда је А хипероним од Б	А – *храна* Б – *кора*, *јуфка*
хипонимија	ако је А ужи појам од Б, онда је А хипоним од Б	А – *доручак* Б – *оброк*
синонимија	ако је А исти појам као и Б, онда је А синоним од Б	А – *кора* Б – *јуфка*
антонимија	ако је А супротан појам појму Б, онда је А антоним од Б	А – *истина* Б – *лаж*
меронимија[33]	ако је А део од Б, онда је А мероним од Б	А – *жуманце* Б – *јаје*
холонимија	ако је А целина од Б, онда је А холоним од Б	А – *јаје* Б – *жуманце*

У примеру 7 приказан је синсет којим је дефинисано значење „Stanje čiste osobe; bez prljavštine ili drugih nečistoća.“ за литерал „čistoća:1“. Он је означен као именички синсет („n“) са јединственом идентификационом ознаком „ENG30-14496193-n“. Описано је да је његов хипероним појам „higijenski uslovi:1, sanitarni uslovi:1“, а скоро антоним „prljavština:1, nečistoća:1“. Поред наведених релација између именичких синсетова постоје и оне које повезују синсетове који припадају различитим деловима WordNet-a (Miller, Beckwith, Fellbaum, Gross i Miller, 1990; Fellbaum, 1998b). Тако су у примеру 7 означене

[32] Појам А је *општији појам* појма Б, ако Б има сва својства која има и А, али има и нека специфична својства.

[33] Код релација меронимије/холонимије се разликује неколико врста односа *део*/*целина*: однос *порција*,*парче*/*целина* (teleći odrezak:1/teletina:1, teleće meso:1), однос *члан*/*група* (jelovna kašika:1/escajg:1) и однос *састојак*/*целина* (protein:1, belančevina:1/jaje:1).

релације типа „be_in_state“ (бити у стању, стање нечега) између именичког синсета са дефинисаним значењем „Stanje čiste osobe; bez prljavštine ili drugih nečistoća.“ са придевским („a“) синсетовима „čist:1a“ (дефиниција: „Bez prljavštine ili nečistoće; onaj koji ima naviku da bude čist.“) и „prljav:1“ (дефиниција: „Koji na sebi ima prljavštinu ili nečistoću.“).

ПРИМЕР 7. Синсет са значењем „stanje čiste osobe; bez prljavštine ili drugih nečistoća“:

```
LNOTE:N600
POS: n ID: ENG30-14496193-n
Synonyms: čistoća:1
Definition: Stanje čiste osobe; bez prljavštine ili drugih
 nečistoća.
Last Edit: User 2003/08/01

–» [hypernym] *[n] higijenski uslovi:1, sanitarni uslovi:1
–» [near_antonym] +[n] prljavština:1, nečistoća:1
–» [be_in_state] +[a] čist:1a
–» [be_in_state] +[a] prljav:1
«– [state_of] +[a] čist:1a
«– [state_of] +[a] prljav:1
«– [near_antonym] +[n] prljavština:1, nečistoća:1

DOMAIN: medicine
```

WordNet такође садржи податке о семантичким доменима којима припадају синсетови. На тај начин су успостављене додатне семантичке релације између значења речи које припадају синсетовима. Такве релације могу да се допуњују и користе у различитим доменима обраде природних језика. Примери семантичких домена су *гастрономија*, *медицина*, *лингвистика* или *метеорологија*, итд. (Magnini i Cavaglia, 2000; Bentivogli, Forner, Magnini i Pianta, 2004).

На основу описане структуре изграђене су семантичке мреже WordNet и за друге језике. Први међународни пројекат за развој семантичких мрежа WordNet је био пројекат *EuroWordNet* у оквиру кога су развијене одговарајуће семантичке мреже WordNet за седам европских језика: холандски, италијански, шпански, француски, немачки, чешки и естонски (Vossen, 1998). У оквиру овог пројекта остварена је вишејезичност тако што су синсетови из

различитих језика, којима је описан исти концепт, повезани преко *међујезичког индекса* (енгл. *Interlingual Index*, скраћено *ILI*) који је представљен описаном етикетом `ID`.

Изградња српског WordNet-а[34] започета је у оквиру пројекта *BalkaNet*[35] (2001–2004) и одговара структури која је успостављена у пројекту *EuroWordNet*, док је повезана са принстонским WordNet-ом преко међујезичког индекса. У оквиру пројекта BalkaNet уведена је и етикета `BSC` (*Balkan Specific Concept*), која означава концепте који су специфични у неком од балканских језика или су потекли из неког од балканских језика а прихваћени у другим језицима (на пример, *ћевапчић* или *баклава*) (Krstev, Obradović i Vitas, 2008). Поред тога, они концепти који су специфични само за балканске језике а нису постојали у принстонском WordNet-у, добили су `ID` етикету која почиње ниском `BILI`. За оне концепте који су специфични искључиво за српски језик (какав је на пример *кајмак*) уведено је правило да `ID` етикета почиње ниском `SRP`.

У оквиру пројекта BalkaNet такође је започет и развој WordNet семантичких мрежа за бугарски, грчки, румунски и турски језик. На крају BalkaNet пројекта је повезивање било извршено са верзијом 2.1 принстонског WordNet-а.

Све семантичке мреже WordNet су у пројекту BalkaNet развијене у складу са *моделом проширења* предложеним у раду (Fellbaum, 2010). Наиме, синсетови из принстонског WordNet-а су преведени и очуване су релације међу њима. По завршетку BalkaNet пројекта, 2004. године, српски WordNet имао је 7.000 синсетова (Tufiş, Cristea i Stamou, 2004). Поред раније описаних података у српском WordNet-у свакој речи из синсета придружене су и информације о коду флективне класе (етикета `LNOTE`).

После завршетка пројекта BalkaNet развој српског WordNet-а је настављен, али знатно споријим темпом, јер се углавном ослањао на волонтерски рад главног уредника, професорке Цветане Крстев и студената

[34] Српски WordNet: http://sm.jerteh.rs/.
[35] BalkaNet пројекат: http://www.dblab.upatras.gr/balkanet/index.htm.

магистарских и докторских студија. Због таквих околности, избор синсетова за даље проширење био је усмерен на додавање оних који припадају одређеним доменима, какви су домен лингвистике, домен биомедицине, домен религија, домен литературе, домен права, домен библиотекарства и издаваштва или домен кулинарства (Krstev i sar., 2008; Vujičić Stanković, Krstev i Vitas, 2014). Као резултат тога, српски WordNet је повезан са принстонским WordNet-ом 3.0, проширен, и садржи приближно 21.200 синсетова[36].

Српски WordNet је постао драгоцени ресурс за развој апликација заснованих на обради природних језика за српски, као што су класификација текстова (Pavlović-Lažetić i Graovac, 2010), постављање вишејезичних упита над дигиталним библиотекама (Stanković, Obradović i Trtovac, 2012), прикупљање полилексичких јединица (Krstev, Stanković, Obradović, Vitas i Utvić, 2010), развој доменски специфичних онтологија и система (Mladenović i Mitrović, 2013).

Иако по својој стуктури WordNet није стриктно онтологија, јер нема систем за закључивање, велики је број радова који описују како је могуће користити WordNet као онтологију или креирати онтологију на основу WordNet-a (Brickley, 1999; van Assem, Menken, Schreiber, Wielemaker i Wielinga, 2004; Graves i Gutierrez, 2006; Hu, Du, Liu i Ouyang, 2006; van Assem, Gangemi i Schreiber, 2006).

С обзиром да је српски WordNet један од најразвијених електронских ресурса који описују српски језик, а да је његово трансформисање у формалну онтологију могуће, у оквиру овог рада WordNet је допуњен синсетовима из домена кулинарства, уместо креирања посебне онтологије. На тај начин, повезивањем појмова из WordNet-a са појмовим који нису из домена кулинарства, читав домен је стављен у шири контекст него што би то било да је креиран као одвојен ресурс.

[36] Број синсетова у септембру 2014. године.

2

Лексички ресурси и алати за обраду текста

2.1 Проблем

Текст у електронском облику, као што је раније речено, јавља се као ниска карактера у којој његова језичка организација није експлицитно назначена. Проблем обраде текста се изражава као проблем опремања ове ниске информацијама које ће га описати као језички објекат. Проблем додавања језичких информација се разлаже на више независних подпроблема. Неки од њих се темеље на постојећим теоријским и конкретним описима појединих природних језика, а неки су језички независни. Општа архитектура система за обраду текста подразумева да ће се, пре свега другог, из текста издвојити оне ниске које су у одређеном језику носиоци значења. Овакве ниске условно одговарају појму речи у уобичајеном, свакодневном значењу. Ипак, ово поклапање је само делимично па ће се оне називати *формалне речи*. Један од почетних задатака је да се одреди које су то ниске и које им се информације могу доделити.

Појам формалне речи зависи, пре свега, од система писања у одређеном језику, а затим и од самих својстава језика. Из тог разлога, опис формалне речи је условљен конвенцијама које важе за одређени језик.

Из угла лингвистике, уобичајни појам речи се сматра недовољно прецизним, па се уводе други термини који би описали овај језички објекат. Његова анализа тада подразумева анализу елемената од којих се речи састоје, као што су графеме и морфеме као и правила која важе за комбиновање

морфема. Анализа речи, која има своје лингвистичко оправдање, али је неприкладна за информатичке потребе, полази од морфема. Морфеме се реализују у речима једног језика као аломорфи, па структура речи изражена преко њених аломорфа није одређена на такав начин да се може поуздано имплементирати.

Из тог разлога, уместо анализе која би приказала морфемску структуру формалне речи (ако постоји), у аутоматској анализи текста се обично бира опортунистички приступ који се састоји у идентификацији формалне речи без улажења у питање њене морфемске структуре. Овакав поступак подразумева да ће бити дефинисана процедура која текст, као непрекинуту ниску карактера, раставља на мање ниске које би могле да представљу речи анализираног језика имајући у виду његов графемски састав. Резултат процедуре је тада низ јединица, које се називају *токени*, од којих само неке представљају речи анализираног језика у уобичајеном смислу. Друге, као на пример, интерпункцијски знаци, низови цифара или сегменти из других језика, су такође су токени. Међу овако одређеним токенима је у следећем кораку потребно препознати формалне речи, оне ниске које припадају вокабулару анализираног језика.

Један начин се састоји у томе да се вокабулар језика (као списак токена који јесу у анализираном језику) састави пре саме анализе одређеног текста. Тада се критеријум којим се одлучује да ли неки токен јесте реч у анализираном језику своди на питање да ли се токен јавља у вокабулару или не. Ако су елементима листе која чини вокабулар додељене и друге језичке информације, онда се оне могу приписати и токену у случају када је препознат. Питање доделе других информација токену се може решити и прилазима који су засновани на методама машинског учења. Тада се обично на токене пројектују информације које су, према одређеним принципима, прикупљене из скупова за обуку. Недостатак овог приступа је у томе што са повећањем скупа информација које је потребно придружити токенима прецизност опада, а такође је потребно доследно ручно обележити токене текстова који припадају скуповима за обуку. Наиме, квалитет који се постиже применом метода машинског учења умногоме зависи од тога колико је информација

потребно придружити токенима, колико су прецизно примењени дефинисани принципи придруживања и колико је репрезентативан тако добијен узорак који чини скуп за обуку. На пример, уколико се ради морфолошко обележавање текстова једне врсте, а обележени узорак садржи текстове друге врсте, вероватноћа придруживања одговарајућих информација токенима опада јер се рачуна на основу раније добијених резултата статистичких анализа заснованих на нерепрезентативном узорку.

Какве ће информације бити додељене токену зависи од природе анализе коју треба обавити, али и од различитих других фактора. У најједноставнијем случају, додела информације о врсти речи омогућује да се врши део даљих анализа над текстом. У сложеним случајевима, токену који је реч језика, треба доделити што више информација које дефинишу његову улогу као језичке јединице (лема, врста речи, морфограматичке категорије, значења, итд).

Поред доделе граматичког статуса, треба решити читав низ питања која се у класичној анализи текста не постављају. Извесни токени, иако припадају језику, резултат су различитих конвенција и немају атрибуте који се додељују речима. Међу овакве токене се сврставају бројевни придеви, комбинације бројева и речи, различити симболи, скраћенице, акроними и слично. Тако би, примера ради, за случај токена *17-ти* требало установити да то нису две речи, број *17* и заменица *ти*, већ да њихова комбинација чини недељиву целину, реч која је редни број. За овакве ниске неопходно је обезбедити додатне механизме који ће им обезбедити статус речи у језику. Један од приступа је пописивање оваквих токена у речнику на сличан начин како је то урађено за остале речи језика. Други приступ је описивање формалних речи које припадају посебним класама какве су, примера ради, датуми или лична имена, *локалним граматикама* (Gross, 1993), којима се одређују локални услови и ограничења које треба да задовоље суседне ниске у прихватљивим језичким исказима.

У овом раду се проблем опремања токена информацијама које га чине језичким објектом решава методом *лексичког препознавања* (Silberztein, 1989). Ова метода се темељи на специфичној структури речника која је позната као LADL-формат (Courtois, 1989). За разлику од морфолошких

анализатора, у овој методи морфемска структура речи остаје скривена. Ту се посматрају облици речи који се могу реализовати у тексту и њихове релације према одредницама речника. Ове релације су спецификоване граматичким правилима посматраног језика. Ово значи да ће се сви могући облици једне леме наћи у речнику без обзира на то да ли су се икада реализовали у тексту. Неки облици се не употребљавају, иако су морфолошки могући, и њихова фреквенција је 0 и на великим корпусима. Ипак, и такви облици су део речника. Овакав приступ, који није присутан у другим методама, подразумева да ће бити темељно описан флективни систем језика, а не репертоар облика речи из одређеног корпуса, што ову методу суштински разликује од метода заснованих на машинском учењу. Поступак се разликује и од метода аутоматске морфолошке анализе тако што се статус формалне речи из текста не утврђује на основу њене морфемске структуре као што је то случај, на пример, у Кимоовом дворазинском моделу (Koskenniemi, 1983; Karttunen, 1983).

Успешност метода лексичког препознавања зависи од свеобухватности речника, као и од информација које су придружене леми. Израда оваквог речника је спор и скуп ручни посао, али се тиме значајно повећавају могућности које он доноси у обради текста у односу на друге моделе опремања токена језичким информацијама. У даљем излагању ће бити размотрена основна својства структура података које су придружене концепту електронског речника, као и њихова имплементација у оквиру система *Unitex* (Paumier, 2014). Затим ће бити представљена архитектура над којом је развијен систем за обраду текстова *GATE* (Cunningham i sar., 2011), који је првенствено намењен екстракцији информација. Биће размотрени основни принципи над којима се у оквиру овог система токени опремају језичким информацијама, као и проблеми који се јављају када се они примене при обради текстова на српском језику. На крају ће бити приказан један од начина да се описани проблеми превазиђу кроз адекватну примену језичких ресурса и функционалности развијених у систему Unitex и тиме омогући обрада текстова на српском језику.

2.2 Електронски речници

Под *електронским речником* се подразумева речник намењен аутоматској обради текста. Овакви речници се разликују од машински читљивих речника, речника на дигиталном носиоцу, који су намењени људској употреби, а који су обично настали из папирног издања речника.

Конверзија машински читљивих у електронске речнике није изводљива јер се разликују по информацијама које садрже и по прецизности граматичке обраде. На пример, у машински читљивим речницима српског језика граматичка информација је понекада непотпуна и непрецизна. Тако, на пример, речници српског језика не обележавају да ли су именице I врсте обележене као живо или неживо, па информација о облику акузатива ових именица није забележена. Овај податак је од значаја не само за одређивање облика акузатива, већ и у случају када аниматност одређује значење одреднице (Vitas, 1993). Назнака о разлици између природног и граматичког рода и броја у српском је такође несистематично назначена, па се род, а тиме и флективно понашање леме, не може одредити (Popović Lj. , 2000). Изостају и поједина значења од значаја за обраду одређених подјезика. У раду (Krstev i Lazić, 2015) су наведени бројни глаголи и облици трпног придева којих нема у „Речнику српскохрватскога књижевног језика“ (RMSMH, 1976), а од значаја су за обраду кулинарских садржаја.

Наведене непрецизности, уколико се репродукују у електронском речнику, онемогућавају да се токену који је реч у анализираном језику доделе ваљане граматичке информације. Такав је случај поједностављених електронских речника, *листа назива* (енгл. *gazetteer*), који садрже листе лема назива који се одликују заједничким семантичким карактеристикама, али без придружених граматичких информација. Такве су на пример листе назива градова, организација или река. При обради текстова се у овом случају одсуство граматичких информација надомешта применом метода заснованих на машинском учењу.

Отуда је израда електронског речника пројекат у коме се, на основу традиционалних извора, врши одабир и утврђивање лема, а затим се исцрпно

спецификују граматичке информације које ће (а) омогућити генерисање флективних облика (речи које се сравњују са токенима) и (б) прецизно описати граматичка својства леме (Vitas, 1997). Део ових задатака се може извршити аутоматски, али други део захтева ручну интервенцију, па је стварање електронског речника дуготрајан процес.

Електронски речник би, поред доделе информација токену, морао да омогући и различите екстраполације којима се апроксимира статус оних токена који се не налазе у електронском речнику. Ово је од посебног значаја за језик какав је српски, чија је деривациона морфологија врло богата. Наиме, токен који није описан у речнику има статус *непознате речи* (Vitas, 2007) и таквом токену се не може доделити граматичка информација. Адекватним формалним описом творбених правила, могуће је развити механизме њиховог коректног препознавања. Применом одговарајућих симулација творбених модела за сложене придеве на корпус правних текстова у раду (Vitas, Vasiljević i Krstev, 2014), број непознатих речи смањен је за 15%, препознавањем сложених придева какви су на пример *јавно-бележнички* или *пореско-прекршајни*. На овај начин је омогућено да се већи број токена који су речи у анализираном језику снабде одговарајућим граматичким информацијама неопходним за даљу обраду, али и да се обогати процес изградње електронских речника дописивањем нових одредница добијених оваквим формалним описима.

2.3 Формати електронских речника

Електронски речници имају различите формате у зависности од информације које су у њима представљене и начина на који ће бити употребљене (Vitas, Krstev i Sabo, 2003). Разликују се две основне класе представљања лексичких информација.

Једну класу чине *позициони системи кодирања* код којих позиција одређене информације у запису одреднице речника одређује њено значење. Такви су формати предложени у оквиру пројеката MULTEXT и MULTEXT-East

(Erjavec i sar., 2003), као и у ISO стандарду LMF (*Language Resource Management – Lexical Markup Framework*) (ISO 24610-2:200, 2007).

У речницима заснованим на MULTEXT-опису, сваком облику је додељен кôд који се састоји од одређеног броја позиција, а вредност на одређеној позицији описује одређену морфосинтаксичку категорију. Тако, на пример, кôд

```
Ncmpi--y
```
(1)

описује у српском језику заједничку (`c`) именицу (`N`) мушког (`m`) рода у инструменталу (`i`) множине (`p`) која је обележена као живо (`y`). Непопуњене позиције користе други језици. Приликом обележавања текста, ови кодови се користе у складу са TEI-препорукама[37] како би се у тексту описала граматичка вредност појединачног токена.

У оквиру стандарда LMF дефинисана је XML репрезентација лексичких информација која укључује морфолошке, синтаксичке и семантичке категорије. LMF оквир за опис лексичког ресурса (коме одговара XML етикета *LexicalResource*) у основи пружа могућност да у њему буде описан један или више речника (којима одговара XML етикета *Lexicon*). Лексичке одреднице речника (са XML етикетом *LexicalEntry*) у себи садрже детаље описа леме, њених облика и опционо наведеног значења (са XML етикетама *Lemma*, *WordForm* и *Sense*) употребом угњеждених XML етикета, атрибута и вредности. Мана основног LMF оквира је што значење дефинисано за једну одредницу не може да буде придружено другој одредници, па су за успостављање семантичких веза уведена проширења. Пример речника у LMF формату дат је у прилогу Б.

Непозициони систем кодирања подразумева да ће сваки елемент кôда имати унапред дефинисано значење без обзира на позицију у кôду. Овакав начин кодирања се користи у LADL-формату где су морфосинтаксичке

[37] TEI-препоруке: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html#DITPFO.

вредности кодиране свака по једним словом. Запис (1) се у овом систему кодира на један од следећих начина

`N:mp6q` или `N:m6pq` или `N:6mpq`, итд. (2)

где је `N` ознака врсте речи, `m` ознака за мушки род, `p` за множину, `6` за инструментал и `q` за неживо независно од позиције у кôду. Ознака да је именица заједничка се записује у пољу иза ознаке врсте речи, нпр. `N+C`.

За разлику од позиционих система, где се исто слово може употребити за обележавање различитих вредности у зависности од позиције, код овог формата то није случај. Ово својство кодова у LADL-формату је тесно везано за представљање садржаја речника посредством коначних аутомата које ће бити описано у наредном поглављу.

Општа структура података у LADL-речнику је

`oblik,lema.K+SinSem:msk*` (3)

где је `oblik` – флективни облик леме, `lema` – канонски облик (на пример, номинатив једнине за именице), `K` – кôд врсте речи, `SinSem` – списак синтаксичких и семантичких својстава леме, а `msk` кôд који описује однос облика према леми преко морфосинтаксичких категорија (пример 8).

ПРИМЕР 8. На пример, одреднице речника

```
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:mp2q
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:ms2q
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:mw2q
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:mw4q
```

описују да је облик `kolača` леме `kolač` конкретна (`Conc`) нежива (`q`) именица (`N`) мушког рода (`m`), у генитиву множине (`p`), једнине (`s`) и паукала[38] (`w`), као и у акузативу (`4`) паукала.

[38] *Паукал* је облик некадашње двојине и јавља се само уз бројеве два, три и четири (на пример, две *крушке*, три *крушке*, четири *крушке*).

Овај формат је јединствен за све врсте речника у LADL-формату који се примењују у анализи текста.

Основну разлику и предност LADL-формата у односу на позиционе системе кодирања електронских речника представља поље синтаксичких и семантичких својстава које омогућава да се опишу атрибути леме који нису предвиђени у позиционим системима. Захваљујући овој могућности, у морфолошком електронском речнику српског својства доменске лексике кулинарства су кодирана тако што су у оквиру дела `+SinSem` уведени нови семантички маркери представљени у табели 3.

Табела 3. Преглед семантичких маркера предложених за кулинарски домен.

Семантички маркери	**Опис**
+CULINARY	кулинарски домен
+FOOD	храна (нпр. *сенф +Conc+Food+Prod+DOM=Culinary*)
+ALIM	намирница (нпр. *млеко +Alim+Conc+Drink+Food+DOM=Culinary*)
+PROD	производ (нпр. *сирће +Conc+Food+Prod+DOM=Culinary*)
+MEAL	оброк (нпр. *доручак +Conc+Food+Meal+DOM=Culinary*)
+COURSE	јело (нпр. *пудинг +Conc+Course+Food+Prod+DOM=Culinary*)
+DRINK	пиће (нпр. *пиво +Conc+Drink+Food+Prod+DOM=Culinary*)
+UTEN	кухињски прибор (нпр. *виљушка +Conc+Uten+DOM=Culinary*)
+ERG	заштићени назив производа (нпр. *рокфор +Conc+Erg+Food+Prod+DOM=Culinary*)
+MESAPP	приближна мера (нпр. *кашичица +MesApp+DOM=Culinary*)
+CONT	контејнер (нпр. *супена кашика*)
+POR	порција (нпр. *кришка*)
+PART	део од (нпр. *главица*)
+WH	целина (нпр. *штапић*)
+SET	скупина (нпр. *веза*)

+TASTE	укус (нпр. *слаткокисео*)
+WOP	начин припреме (нпр. *динстати*; *динстање*)
+COND	стање (нпр. *бајат*)

Семантички маркер +CULINARY је придружен свим лемама из домена кулинарства. Сви остали маркери се користе у комбинацији са маркером +CONC, осим маркера +MESAPP за означавање приближних мера које се користе у кувању. Слично, маркер +FOOD се јавља уз све остале маркере, осим уз +UTEN, који се користи за означавање прибора за припрему и сервирање хране. Маркер +ERG се додељује речима које имају статус заштићеног назива. То може да се односи и на храну (на пример, *табаско*) и на прибор (на пример, *тефлон*). Овај маркер се користи и изван кулинарског домена (на пример, *Ролс-Ројс*). Ови семантички маркери могу да се користе појединачно или у комбинацији (пример 9).

ПРИМЕР 9. У примеру 8 је семантичким својствима описано да се именицом `kolač` именује храна (`Food`) која припада домену кулинарства (`DOM=Culinary`) и представља јело (`Course`).

Мимо разлике у могућности описа синтаксичких и семантичких информација, позициони и непозициони формати су еквивалентни у смислу да је речник из једног формата могуће конвертовати у други формат (Krstev, Stanković i Vitas, 2010).

За разлику од позиционог система који је подесан за ручну анотацију текста који се припрема као скуп за обуку обележивача, LADL-формат располаже могућношћу генерисања речника.

Уколико за одређени језик постоји прецизна спецификација односа облика и леме, описана посредством регуларних израза, тада се електронски речник чији су редови облика (3) може генерисати тако што се леми придружи одговарајући кôд за врсту речи и број флективне класе добијен нумерацијом ових регуларних израза (пример 10).

ПРИМЕР 10. Кôд `N27` морфолошке класе описује опште именице мушког рода које имају множину истог рода, обележене су као неживе и седмог су деклинационог типа, што значи да у номинативу сингулара немају наставак, да у генитиву сингулара имају наставак *а*, у вокативу сингулара наставак *у*, у инструменталу сингулара наставак *ем* а у генитиву плурала наставак *а*.

Један од представника ове класе је *колач*, чији су облици:

	једнина (*s*)	*множина* (*p*)	*паукал* (*w*)
номинатив (*n*)	колач***ø***	колач***и***	-
генитив (*g*)	колач***а***	колач***а***	колач***а***
датив (*d*)	колач***у***	колач***има***	-
акузатив (*a*)	колач***ø***	колач***е***	колач***а***
вокатив (*v*)	колач***и***	колач***и***	-
инструментал (*i*)	колач***ем***	колач***има***	-
локатив (*l*)	колач***и***	колач***има***	-

што се може представити регуларним изразом:

`колач(ø/`$_{ns,as}$ `+ а/`$_{gs,gp,gw,aw}$ `+ у/`$_{ds,vs,ls}$ `+ ем/`$_{is}$ `+ и/`$_{np,vp}$ `+ има/`$_{dp,ip,lp}$ `+ е/`$_{ap}$`).`

Тако проширен кôд врсте речи описује тада трансформацију коју треба применити на лему да би се добили њени појединачни облици. Формат таквог речника је онда

`lema,Kn+SinSem` (4)

где је `Kn` опис регуларног израза који описује флективне облике леме.

Из формата (4) је даље могуће аутоматски генерисати следећи облик електронског речника

`oblik,lema.K+SinSem` (5)

Уколико се регуларни израз прошири тако да, поред облика, описује и одговарајуће морфосинтаксичке категорије, онда се, као трансдукторски излази генеришу редови облика (3) (пример 11).

ПРИМЕР 11. Линијом

```
kolač,N27+Conc+Course+Food+DOM=Culinary
```

речника означено је да именица *колач* припада флективној класи `N27`.

Из ње се генеришу редови:

```
kolač,.N+Conc+Course+Food+DOM=Culinary:ms1q
kolač,.N+Conc+Course+Food+DOM=Culinary:ms4q
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:mw2q
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:mw4q
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:mp2q
kolača,kolač.N+Conc+Course+Food+DOM=Culinary:ms2q
kolačem,kolač.N+Conc+Course+Food+DOM=Culinary:ms6q
kolači,kolač.N+Conc+Course+Food+DOM=Culinary:mp1q
kolači,kolač.N+Conc+Course+Food+DOM=Culinary:mp5q
kolačima,kolač.N+Conc+Course+Food+DOM=Culinary:mp3q
kolačima,kolač.N+Conc+Course+Food+DOM=Culinary:mp6q
kolačima,kolač.N+Conc+Course+Food+DOM=Culinary:mp7q
kolaču,kolač.N+Conc+Course+Food+DOM=Culinary:ms3q
kolaču,kolač.N+Conc+Course+Food+DOM=Culinary:ms5q
kolaču,kolač.N+Conc+Course+Food+DOM=Culinary:ms7q
```

Овај механизам се примењује како на *просте леме* (низ алфабетских карактера између два сепаратора), тако и на *полилексичке речи* (као контингентне ниске простих речи) (пример 12).

ПРИМЕР 12. Линијом

```
čajni(čajni.A2:adms1g) kolač(kolač.N27:ms1q),
NC_AXN+Comp+Conc+Food+Prod+DOM=Culinary
```

речника означено је да је именица *чајни колач* полилексичка реч (`Comp`) која је конкретан прехрамбени производ (`+Conc+Food+Prod`) из кулинарског домена (`DOM=Culinary`). Она припада флективној класи `NC_AXN` и из ње се генеришу редови:

```
čajne kolače,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:mp4q
čajni kolač,
      .N+Food+Conc+Prod+DOM=Culinary+Comp:ms1q
čajni kolač,
      .N+Food+Conc+Prod+DOM=Culinary+Comp:ms4q
čajni kolači,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:mp1q
čajni kolači,čajni kolač
```

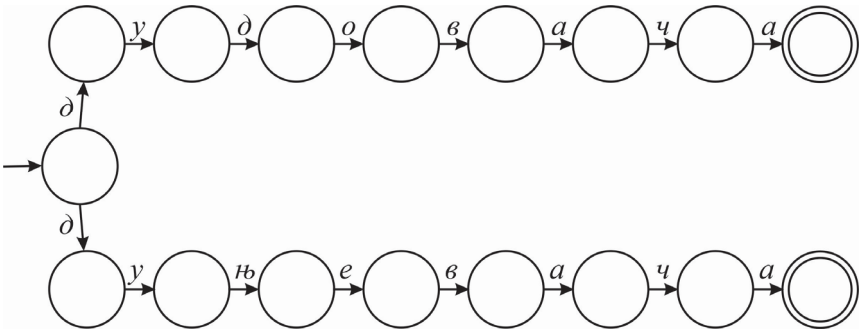
```
      .N+Food+Conc+Prod+DOM=Culinary+Comp:mp5q
čajni kolaču,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:ms5q
čajnih kolača,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:mp2q
čajnim kolačem,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:ms6q
čajnim kolačima,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:mp3q
čajnim kolačima,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:mp6q
čajnim kolačima,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:mp7q
čajnog kolača,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:ms2q
čajnom kolaču,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:ms3q
čajnom kolaču,čajni kolač
      .N+Food+Conc+Prod+DOM=Culinary+Comp:ms7q
```

Уколико за један језик није развијен опис флективне морфологије преко регуларних израза, електронски речник може бити развијен на основу парцијалних описа односа облика и леме (на пример, у случају Персеусовог речника латинског (Crane, 2015) за који је састављен речник облика (3) али са непотпуним парадигмама), или коришћењем независног генератора за формирање речника у формату (3) (Vetulani, 2000).

Српски морфолошки речник лема простих речи, DELAS (*Dictionnaires électroniques des mots simples*), садржи 130.000 лема, од чега се већина односи на општу лексику. Приближно 28.5% лема представљају различите врсте личних имена – имена људи, геополитичких имена, назива организација итд. (Gucul-Milojević, 2010; Krstev, Vitas, Obradović i Utvić, 2011). DELAF (*Dictionnaires électroniques des formes fléchies*) речници облика лема из речника DELAS садрже приближно 1.450.000 различитих облика речи. Величине DELAC (*Dictionnaire électronique des mots composés*) и DELACF (*Dictionnaire électronique des mots composés fléchis*) речника лема полилексичких речи и њихових облика су редом приближно 10.500 и 54.000 лема, односно облика речи (Krstev, 2008).

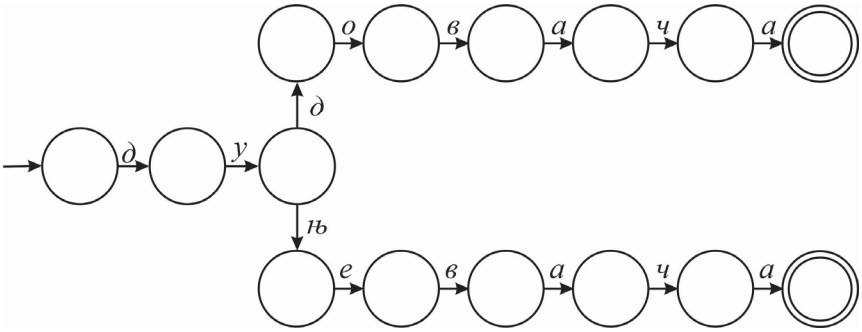
2.4 Коначни аутомати у обради текста

Коначни аутомати се јављају као природан алат у обради различитих природнојезичних феномена, а посебно на морфолошком нивоу (Gross, 1988). Према раду (Langendoen, 1981), може се сматрати да се сваки скуп (можда бесконачан) речи природног језика може описати овим формализмом. На пример, ако се претпостави да језик има само две речи *дуњевача* и *дудовача*, онда је коначни аутомат који описује овај скуп речи приказан на слици 5.

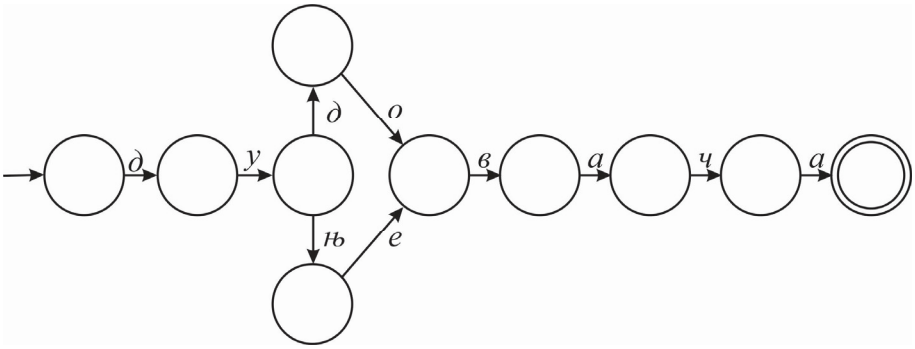


Слика 5. Недетерминистички коначни аутомат за језик који садржи само речи *дуњевача* и *дудовача*.

С обзиром да се из почетног стања може прећи по улазном знаку *д* у два различита стања која припадају различитим путањама, овај коначни аутомат је недетерминистички. Односно, приликом препознавања речи, читањем знака по знака са улаза, могућ је улазак у проверу по погрешној путањи, а самим тим неефикасно враћање и поновна провера по осталим путањама. Да би се ово избегло, за сваки недетерминистички коначни аутомат се уводи њему еквивалентан детерминистички аутомат у коме се из једног стања по једном знаку са улаза може прећи у највише једно стање. Детерминистички коначни аутомат који одговара коначном аутомату са слике 5 приказан је на слици 6. На крају се за речи језика минимизацијом овог коначног аутомата добија коначни аутомат приказан на слици 7.



Слика 6. Детерминистички коначни аутомат за језик који садржи само речи *дуњевача* и *дудовача*.



Слика 7. Минимални детерминистички коначни аутомат за језик који садржи само речи *дуњевача* и *дудовача*.

Различите флективне парадигме и творбени процеси такође се могу описати на овај начин. Потврда овог тврђења долази и из искуства са различитим имплементацијама морфолошких анализатора заснованих на овом моделу. Како се речи, као јединице говора, изговарају (и пишу) у непрекинутим секвенцама гласова (и слова), њихова линеарна природа намеће овакав формални модел описивања.

Еквивалентност између регуларних језика и коначних аутомата (Kleene, 1956) успоставља додатно директну везу између уобичајених начина описивања морфолошких феномена и аутомата. С једне стране, регуларни изрази пружају могућност да се њима опишу морфолошки феномени, док се с друге стране коначним аутоматима утврђује припадност токена језику аутомата. Тако су, на пример, табеле којима се обично описују флективни

облици једне класе речи заправо другачија репрезентација за регуларне изразе (пример 10).

Полазећи од рада (Kaplan i Kay, 1981), ова веза је проширена и на коначне трансдукторе како би се успоставила веза између формалне речи и њене канонске репрезентације у речнику. Овај полазни модел је кроз Кимоов модел касније имплементиран за већину светских језика (укључујући и семитске језике).

Успостављање веза између формалне речи и леме посредством коначних аутомата и трансдуктора може се обавити на више начина у зависности од тога како ће бити распоређене информације између граматике и речника. Наиме, у моделу какав је Кимоов модел, већина информација је енкодирана у правилима којима се успоставља веза између текстуалне и лексичке речи, док је улога речника ограничена на „изузетке“, оне облике речи који одступају од правила. Дефинисање и одржавање оваквих правила која описују како морфолошке процесе којима се текстуална реч (потенцијално неједнозначно) своди на лексичку, тако и оне којима се из лексичке речи генерише текстуална, сложен је и захтеван поступак, а њихов број се повећава са количином информација које је потребно енкодирати при опису језика. Поред тога, повећање броја правила утиче на експоненцијално повећање временске сложености повезивања формалне речи и леме (Ritchie, 1992). Због тога се овакав модел не сматра одговарајућим за изградњу система за морфолошку обраду текста на, морфолошки богатом, српском језику (Vitas, 1993).

У моделу електронских речника у LADL-формату, репрезентација морфолошких феномена посредством коначних аутомата и трансдуктора има централну улогу у описивању како флективних, тако и творбених феномена, али и у проширивању скупа токена на полилексичке јединице и сложенице.

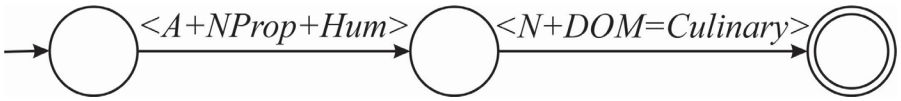
Поред тога, преко појма локалне граматике, која комбинује лексичке и граматичке информације, они се употребљавају како би се препознале комплексније структуре у тексту као што су, на пример, именовани ентитети.

Тако су за решавање задатка аутоматског препознавања оних именованих ентитета типа назива јела у чијој структури учествују властита имена у текстовима кулинарског домена у раду (Vujičić Stanković i Pajić, 2015),

конструисане локалне граматике засноване на морфолошким електронским речницима српског језика који садрже 31.748 лема простих речи које одговарају властитим и посебно означеним личним именима, и 1.653 одреднице које су полилексичке јединице које одговарају властитим именима[39].

У електронским речницима су облици именице која припада кулинарском домену означени са `N+DOM=Culinary`, топоними су означени са `N+NProp+Top`, облици властитих имена су означени са `N+NProp+Hum`, а из њих изведени присвојни придеви са `A+NProp+Hum`. Анализом текстова кулинарског домена издвојене су четири доминантне структуре именованих ентитета у чијој изградњи учествују властита имена. Прве две граде облици именице која припада кулинарском домену и облици топонима у различитом редоследу (које се означавају са `<N+DOM=Culinary><N+NProp+Top>` и `<N+NProp+Top><N+DOM=Culinary>`), трећу граде облици властитог имена и именице из кулинарског домена која га следи (означена са `<N+NProp+Hum><N+DOM=Culinary>`), док четврту гради присвојни придев за којим следи именица из кулинарског домена (означена са `<A+NProp+Hum><N+DOM=Culinary>`).

За сваку од ових структура су креиране локалне граматике у форми коначних аутомата. Пример коначног аутомата који одговара структури присвојног придева за којим следи именица из кулинарског домена приказана је на слици 8.



Слика 8. Коначни аутомат за препознавање кандидата за именоване ентитете са структуром: присвојни придев за којим следи именица која припада кулинарском домену.

Описани коначни аутомати препознају именоване ентитете као што су „Америчке крофнице“, „Дижон сенф“, „Бечка шницла“, „Сарајевски бурек“,

[39] Величина речника 10. XII 2014. године.

„Турска кавурма“, „Лесковачки ћевап“, „Велингтон филе“, „Велингтон стек“ или „говедина Велингтон“. Овакви именовани ентитети који се појављују у текстовима кулинарског домена, а у чијој структури учествују властита имена, постављају различите проблеме какви су проблем синонимије или проблем промене по падежима. Примера ради, у именованом ентитету „филе Велингтон“ приликом промене кроз падеже, само ће се општи појам „филе“ мењати по падежима, а не обе компоненте полилексичке јединице како важи у општем случају. Зато је потребно да се овакви именовани ентитети посебно препознају и опишу у лексичким ресурсима српског језика.

Локалне граматике омогућују да се опише синтаксичка структура и додели граматичко значење низовима токена који не могу експлицитно да се наведу у електронским речницима.

У овом приступу, анализа се врши првенствено на основу садржаја речника, док се аутоматима описује како сама конструкција речника, тако и његове примене у анализи текста. Шта више, и анализирани текст и речници су представљени у облику ацикличних аутомата што омогућава да се у конципирању анализе текста користе бројна корисна својства ове класе препознаваоца, а посебно чињеница да је пресек два регуларна језика регуларан језик.

2.5 Системи за екстракцију информација

2.5.1 Увод

Како је описано у поглављу 1.3 системи за екстракцију информација се према приступу екстракцији деле на две основне групе, где се у првој користи приступ заснован на правилима, а у другој приступ заснован на машинском учењу.

Код приступа заснованог на правилима, да би се употребом система за екстракцију информација екстраховале информације из доступних текстова, правила за екстракцију морају бити написана ручно. Особа која прави такву врсту система или је одговорна за писање правила мора бити стручњак у области знања изабраног за екстракцију или барем мора бити блиско

упозната са њим. Поред тога, мора да зна формализме за писање правила које одређени систем користи. Обично, при том, постоји велики број текстова који се односе на изабрани домен. Анализом тих текстова проналазе се заједнички обрасци у њима и на основу њих пишу правила која представљају важан чинилац у прављењу система са високим нивоом перформанси (Appelt i Israel, 1999). Правила се затим интерпретирају одговарајућим компонентама система и користе у екстракцији информација. Прављење система за екстракцију информација употребом овог приступа дуготрајан је итеративан процес јер се једном креирана правила, након примене над расположивим текстовима и провере добијених информација, према потреби мењају и поново тестирају док се не постигну жељени резултати. Представници ових система су *Finite State Automaton Text Understanding System* – FASTUS (Appelt, Hobbs, Bear, Israel i Tyson, 1993; Hobbs i sar., 1997), GE NLTOOLSET (Jacobs i Rau, 1990; Krupka, Jacobs, Rau, Childs i Sider, 1992), *Probabilistic Language Understanding Model* – PLUM (Ayuso i sar., 1992; Weischedel i sar., 1996), PROTEUS (Yangarber i Grishman, 1998), CICERO (Harabagiu i Maiorano, 2000) итд. С друге стране, код приступа заснованог на машинском учењу не постоји потреба да се ручно креирају правила, већ се тај процес обавља аутоматски употребом алгоритама машинског учења. Ови алгоритми морају да имају велики број унапред ручно обележених текстова са примерима из којих могу да науче правила. Међу овим системима су WHISK (Soderland, 1999), RAPIER (Califf i Mooney, 1999; Califf i Mooney, 2003), SRV (Freitag, 1998), IEPAD (Chang i Lui, 2001), OLERA (Chang i Kuo, 2004), RoadRunner (Crescenzi, Mecca i Merialdo, 2001), DEPTA (Zhai i Liu, 2005) итд.

Различити прегледи развијених система за екстракцију информација користе широк спектар критеријума за поређење, па се поред приступа који се користе при екстракцији, као критеријуми за поређење наводе и тип улазних текстова које системи користе или тип излазних резултата које системи производе.

Према типу улазних ресурса Муслеа (Muslea, 1999) дели системе на оне који екстрахују информације из неструктурираног (слободног) текста и из полуструктурираних или структурираних докумената на мрежи. Системи који

користе слободан текст као улазни, попут система AutoSlog (Riloff, 1993), LIEP (Huffman, 1995), PALKA (Kim i Moldovan, 1995) или CRYSTAL (Soderland, Fisher, Aseltine i Lehnert, 1995), могу да користе само технике за обраду природних језика за креирање правила за екстракцију. Други системи, као што су WHISK, RAPIER, SRV или STALKER (Muslea, Minton i Knoblock, 1998), нису ограничени техникама за обраду природних језика, већ поред лингвистичких могу да користе и структурне карактеристике докумената за имплицитну класификацију екстрахованих информација (Laender, Ribeiro-Neto, da Silva i Teixeira, 2002; Chang, Kayed, Girgis i Shaalan, 2006).

Када је у питању подела према формату излазних података, један од доминантних стандарда за њихово представљање је XML који користе системи као што су DEByE (Laender, Ribeiro-Neto i da Silva, 2002), LAPIS (Kuhlins i Tredwell, 2002), NoDoSE (Adelberg, 1998), RoadRunner или XWRAP (Liu, Pu i Han, 2000). Поред њега су заступљени и други излазни формати као што су слободни текст, који на пример може бити излазни формат система RoadRunner, или формати погодни за употребу у базама података који се користе на пример у систему DEByE.

У најранијим фазама развоја области екстракције информација за сваки нови проблем екстракције овакви системи су креирани из почетка. Није разматрана могућност да се делови једном направљеног система поново употребе у изградњи новог система па су развијани независно, на различитим програмским платформама. Касније се јавила идеја да се, при изградњи нових, искористе делови развијених система или да се интегришу делови различитих система, па су тако развијене нове програмске платформе које омогућавају да се ови задаци обаве. Између осталих су развијени програмски системи Unitex, *General Architecture for Text Engineering* (GATE), *Unstructured Information Management Architecture* (UIMA)[40], *Natural Language Toolkit* (NLTK)[41], OpenNLP[42].

[40] UIMA: https://uima.apache.org/.
[41] NLTK: http://www.nltk.org/.
[42] OpenNLP: https://opennlp.apache.org/.

GATE и UIMA имају сличну архитектуру у оквиру које се документ представља као текст са додатим обележјима која описују његове различите карактеристике које се употребљавају при обради. Док се UIMA заснива на Java и C++ оквирима, GATE се заснива на Java оквирима. За разлику од GATE-а који је од почетка развијан као пројекат отвореног кода, развој платформе UIMA започет је у IBM-у, а потом је постала *Apache* пројекат отвореног кода. Постоји интероперабилност између ове две платформе, па је тако на пример, омогућено да се у GATE апликацијама користе UIMA компоненте. Они подржавају приступ заснован на правилима и приступ заснован на машинском учењу, док Unitex у основи подржава само први приступ. Unitex програми су писани у програмским језицима C/C++ док је једино графички кориснички интерфејс писан у програмском језику Java. Unitex је настао као слободан софтвер на темељима затворене *Intex* платформе са циљем да се превазиђу недостаци који ограничавају научну заједницу да слободно користи његове ресурсе и алате при експериментима у области обраде природних језика.

Примери платформи које су засноване на машинском учењу су OpenNLP која под својим окриљем окупља хетерогену колекцију пројеката отвореног кода и NLTK која се развија на Универзитету у Мелбурну као пројекат отвореног кода у Python-у и састоји се од модула за хеуристички и статистички засноване природнојезичне обраде.

Како је у овом раду разматрање ограничено на системе за екстракцију информација који користе технике засноване на правилима, а да су за Unitex и GATE развијени језички ресурси и алати потребни за решавање задатака екстракције информација из текстова на српском, у наставку ће детаљније бити приказана ова два система.

2.5.2 Unitex

За комплексну обраду текстова неопходно је на формалан начин исказати знања о природном језику на коме су написани. При томе је потребно

кроз формализацију експлицитно дефинисати све језичке податке и правила која важе у граматици одређеног језика (Harris, 1968; Harris, 1982).

Unitex је скуп програма који допушта обраду текстова на природном језику користећи језичке ресурсе. Ови ресурси се састоје од електронских речника, граматика и табела лексикон-граматике (ако постоје). Систем је резултат радова које је започео Морис Грос у оквиру Лабораторије за аутоматизовану документалистику и лингвистику Универзитета Париз VII средином осамдесетих година 20. века. Његова основна идеја се заснива на захтевима да се детаљно, формално и комплетно лингвистички опишу речници, граматике и лексикон-граматике језика, како би могли да се употребљавају у разноврсним рачунарским апликацијама без обзира на област примене. Таква систематска формализација на првом месту пружа могућност синтаксне и семантичке анализе, као и генерисања текста. Иако оваква исцрпна формализација не може у потпуности да буде постигнута, јер се непрекидно јављају нове речи, неопходно је да се обезбеди за непроменљиви и најважнији део језика. Првобитни такав рад на француском и енглеском језику проширен је на бројне друге језике кроз мрежу лабораторија RELEX[43].

Речници које користи систем Unitex су у LADL-формату. Њихов обим и начин конструкције се разликују од једног до другог језика. Речници српског језика су описани у поглављу 2.2. Граматике у овом систему су формалне репрезентације језичких феномена посредством коначних трансдуктора и рекурзивних мрежа прелаза, формализма који је близак формализму коначних аутомата. Ове граматике, које су се показале као адекватан начин да се опишу морфолошки и синтаксички феномени природних језика, представљене су у облику графова које корисник може лако да направи и одржава.

Таблице лексикон-граматике су формализам који описује структуру елементране реченице. Овај формализам, који се показао изузетно подесним

[43] RELEX: http://infolingu.univ-mlv.fr/Relex/Relex.html.

за опис романских језика, тешко је применљив на словенске језике због слободног реда речи.

Unitex омогућава експлоатцију језичких ресурса расположивих за одређени језик. У техничком погледу, овај систем карактеришу преносивост, модуларност, могућност обраде језика са посебним системима писања (као што су арапски или корејски) и отвореност у смислу да се дистрибуира као отворени код. Језичка својства овог система су она која су мотивисала развој ресурса: прецизност, исцрпност и могућност обраде феномена фиксних језичких структура, а посебно полилексичких јединица.

2.5.2.1 Ток обраде у систему Unitex

Током почетне обраде текста у систему Unitex, врши се нормализација предефинисаних сепаратора[44], који се, према контексту, замењују или једним размаком или знаком за крај реда, и растављање улазног текста на токене. Токени су формирани од преосталих карактера који се појављају у тексту. У предефинисаној датотеци *alphabet.txt* (табела 4) дефинисани су карактери који учествују у формирању формалних речи српског језика писаних латиничним писмом: свака непрекидна ниска карактера из датотеке *alphabet.txt* образује једну формалну реч. Остали карактери (карактери који нису описани у *alphabet.txt*) чине токене сваки за себе. На тај начин, Unitex разлаже улазне карактере у ниске које потенцијално представљају речи у обрађиваном језику и на остале токене – неалфабетске карактере који могу бити цифре, интерпункцијски знаци итд. Трансформисани текст се чува у датотеци са екстензијом *snt* над којом се врше даље обраде.

Табела 4. Садржај датотеке *alphabet.txt* у модулу за српски језик система Unitex.

Aa	Dd	Hh	Mm	Rr	Vv	Žž
Bb	Đđ	Ii	Nn	Ss	Ww	
Cc	Ee	Jj	Oo	Šš	Xx	
Čč	Ff	Kk	Qq	Tt	Yy	
Ćć	Gg	Ll	Pp	Uu	Zz	

[44] Предефинисани сепаратори су размак, табулатор и знак за крај реда.

За текст

G-đa Jovanović mi je dala 2015. godine recept za brze piroške. Rekla je da g-đici jovanović najbolje uspeju kada zamesi 400 gr brašna, 250 ml jogurta, 2 jaja, 1 kesicu praška za pecivo i malo soli, tako da se ne lepi za ruke. Dalje razvije testo na debljinu od 1-2cm, iseče ga na rombove i peče u zagrejanom ulju. Posle doda sir, kajmak itd. Dal' ti se čini dobar?

Unitex формира листу од 68 токена где је предефинисани сепаратор (размак) најчешћи. Како је предефинисани алфабет латинични, Unitex издваја токене попут *g*, *đa*, *đici*,... који су формалне речи и остале токене (', -, ., *0*, *?*...). Токени могу бити модификовани током ове фазе тако што ће се извршити превођење неких ниски карактера на нормализовани облик. На пример, ниска *dal'* (која садржи неалфабетски карактер ') се трансдукторским излазом преводи на облик *da li*, а ниска *itd.* у развијену скраћеницу *i tako dalje*. Такође, током ове фазе се, на основу појављивања карактера који потенцијално обележавају крај реченице (као што су ., *!*, *?* итд.) може уметнути сепаратор реченица *{S}* тако што ће се анализирати одговарајућим трансдуктором контекст у коме се појавио индикатор краја реченице.

Горњи пример добија облик:

gospođa Jovanović mi je dala 2015. godine recept za brze piroške.{S} Rekla je da gospođici jovanović najbolje uspeju kada zamesi 400 gr brašna, 250 ml jogurta, 2 jaja, 1 kesicu praška za pecivo i malo soli, tako da se ne lepi za ruke.{S} Dalje razvije testo na debljinu od 1-2cm, iseče ga na rombove i peče u zagrejanom ulju.{S} Posle doda sir, kajmak i tako dalje{S} da li ti se čini dobar?

где су скраћенице развијене, иза 2015. није уметнут сепаратор краја реченице, као ни иза знака питања јер наредна ниска почиње малим словом. Иза развијене скраћенице *i tako dalje* тачка је уклоњена, али је претходно уметнут реченични сепаратор.

Уколико се током почетне обраде укључи и сравњивање са одређеним речником (или речницима), онда Unitex сравњује алфабетске токене са садржајем речника. Тада су могућа два случаја:

(а) Формална реч одговара првој компоненти речника у изразу (3). Тада се у речник текста додаје ред облика (3).
Ова реч може бити проста (када је уазбучени списак у датотеци *dlf*) или сложена (када је резултат у датотеци *dlc*).

(б) Формална реч се не налази ни у једном од речника.
Таква реч се назива *непозната реч* и смешта се у засебну датотеку чији је генерички назив *err*.

Укључивање свих речника ће генерисати и следеће редове у овим датотекама:

```
dlf: dalje,.ADV+Adj+Comp
     dalje,dalek.A:befw2g
     lepi,lepiti.V+Imperf+Tr+Iref+Ref+Ek:Pzs
dlc: 2015,.NUM+C+v5+NVAL=2015
```

(где је низ неалфабетских токена `2`, `0`, `1`, `5` окупљен у лексички токен и додељен му граматички статус (број `NUM`) који је сложен (`C`). Параметар `v5` указује да у конгруенцији захтева облике множине, док `NVAL` дефинише атрибут који садржи вредност броја у декадном сисему.)

```
err: jovanović
```

(налази се у речнику као властито име, па мора бити записано великим почетним словом)

У изводу из речника текста *dlf* облик *dalje* је вишезначан (може бити придев или прилог). Уклањање оваквих вишезначности се у систему Unitex врши дефинисањем локалних синтаксичких услова, када је то могуће, посредством ELAG-граматика. Овај проблем, који се не јавља у методама машинског учења, могуће је решавати и ручном елиминацијом неодговарајућих редова, али је и у том случају у питању спор посао. Могуће је и ограничити се на одређене речнике или слојеве језика чиме се део

вишезначности може отклонити. С друге стране, формирањем локалних граматика које моделирају шире делове реченице, контекст такве граматике елиминише нежељена граматичка значења.

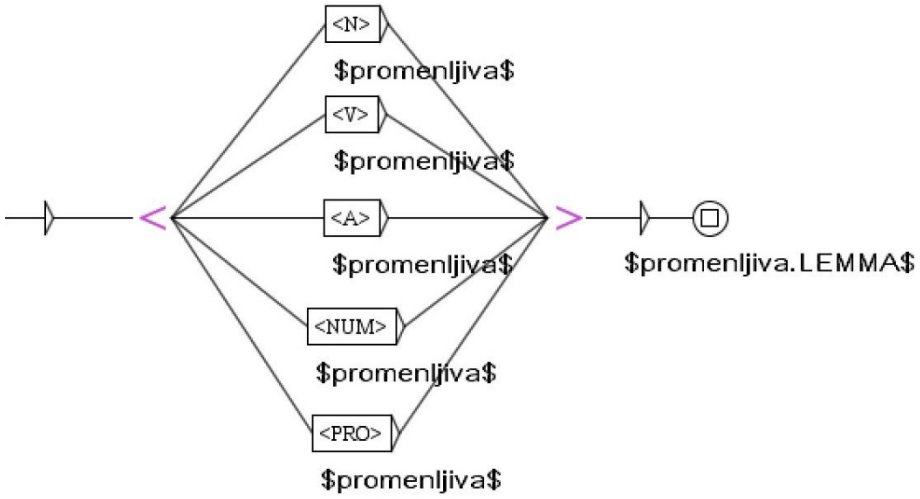
2.5.2.2 Графови у систему Unitex

Репрезентација коначних аутомата и подаутомата је у систему Unitex реализована кроз графове и подграфове. За разлику од уобичајене нотације за коначне аутомате, у Unitex-у се стања представљају луковима графа, а прелази чворовима графа који се називају *кућице*.

У систему Unitex су имплементирани аутомати коначних стања који одговарају графовима различитих намена као што су графови за претходну обраду текста, графови за аутоматско генерисање речника облика променљивих врста речи на основу речника њихових лема (*флективни трансдуктори*), графови за претрагу текста (*синтаксички графови*), графови за нормализацију аутомата текста, параметризовани графови и графови за разрешавање вишезначности лексичких симбола у аутоматима текста (*ELAG граматике*) (Paumier, 2014).

Пример једног синтаксичког графа који је коришћен у овом раду као коначни трансдуктор за лематизацију текстова, приказан је на слици 9. У његовим кућицама су наведене лексичке маске којима се препознају различите променљиве врсте речи текста као што су именице (означене са `<N>`), глаголи (означени са `<V>`), заменице (означене са `<PRO>`), придеви (означени са `<A>`) и бројеви (означени са `<NUM>`) чије се вредности које одговарају прелазу дефинисаном у кућици смештају у променљиву `$promenljiva$`. Вредност променљиве може да се користи у осталим путањама графа, где се могу користити њен флективни облик (`$promenljiva.INFLECTED$`), лема (`$promenljiva.LEMMA$`) и кодови (`$promenljiva.CODE$`). У овом случају се као излаз из графа користи лема препознате речи (`<E>/$promenljiva.LEMMA$`) која се употребљава у посебном „Replace mode“ режиму обраде текста како би се свака препозната

реч заменила својом лемом у излазном тексту, док би непрепознате речи биле дословно преписане.

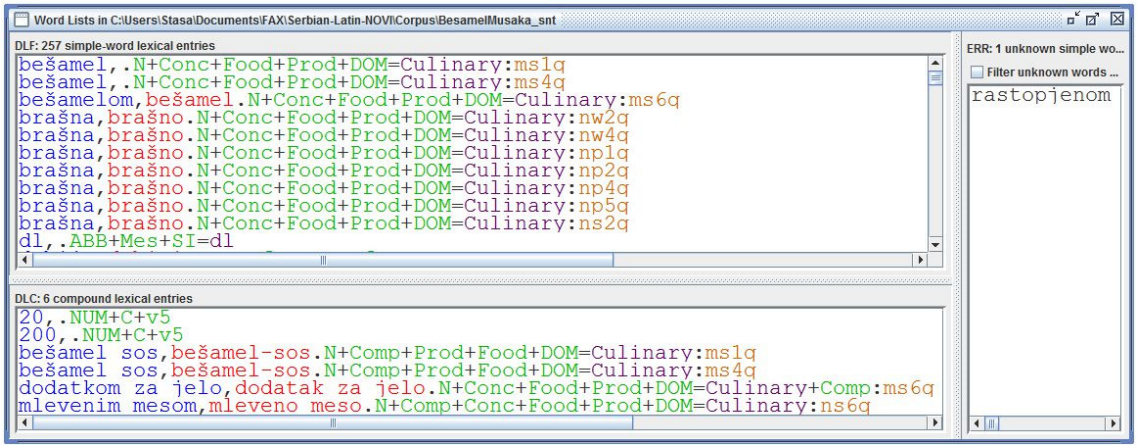


Слика 9. Коначни трансдуктор за лематизацију текста.

Применом коначног трансдуктора за лематизацију текста са слике 9 над текстом:

Makarone obariti. Luk izdinstati sa mlevenim mesom i začiniti dodatkom za jelo. Bešamel sos napraviti na sledeći način: upržiti 2 kašike brašna na rastopjenom puteru, dodati 2 dl mleka i kad provri i dobije određenu gustinu, dodati parče sira za topljenje, žumance i malo senfa. Kačkavalj izrendati. U pouljen pleh staviti bešamel, zatim polovinu pripremljenih makarona, meso, rendani kačkavalj, preostale makarone, preliti bešamelom i preostalim rendanim kačkavaljem. Musaku staviti u rernu i peći na 200 stepeni oko 20 minuta. Kad se prohladi musaku iseći na kocke i poslužiti toplu Jelo se može jesti i hladno.

Unitex креира датотеке *dlf*, *dlc* и *err* чији су делови приказани на слици 10



Слика 10. Приказ дела *dlf*, *dlc* и *err* датотека у систему Unitex.

и формира облик текста:

makaroni obariti.{S} luk izdinstati sa mleveno meso i začiniti dodatak za jelo.{S} bešamel-sos napraviti na sledeći način: upržiti 2 kašika brašno na rastopjenom puter, dodat 2 dl mleko i kad provreti i dobiti određen gustina, dodat parče sir za topljenje, žumance i malo senf.{S} kačkavalj izrendati.{S} u pouljen plesti staviti bešamel, zatim polovina pripremljen makaroni, meso, rendan kačkavalj, preostali makaroni, preliti bešamel i preostali rendan kačkavalj.{S} musaka staviti u rerna i peći na 200 stepen oko 20 minut.{S} kad se prohladiti musaka iseći na kocka i poslužiti topao jesti se moći jesti i hladno.

где су препознате речи текста замењене лемама из датотека морфолошких речника *dlf* и *dlc*, док је непрепозната реч *растопјеном* датотеке *err* преписана у изворном облику.

У систему Unitex је омогућено да се коначни трансдуктори примењују секвенцијално како би се излазне ниске једног трансдуктора користиле као улазне ниске коначних трансдуктора који се примењују након њега. Оваква организација, где сваки граф прави измене на тексту које се користе у даљој обради наредним графовима, назива се *каскада коначних трансдуктора* (Friburger i Maurel, 2004). На тај начин се моделирају сложеније везе између целина које одговарају појединачним графовима. Поред тога се за креирање библиотека сложених графова користе рекурзивне мреже прелаза, колекције

графова организованих у мрежу, где сваки граф може да позива друге графове, па и самог себе (Woods, 1970; Gross, 1975). У том поступку се прво креирају основни графови за препознавање делова текста који чине сложеније структуре које треба екстраховати, а онда се на основу њих креирају графови који одговарају сложенијим структурама. У сваком кораку изградње у графовима који се тренутно креирају могу да се позову раније креирани графови. Рекурзивне мреже прелаза које су допуњене променљивима које могу да чувају вредности препознатих делова текста називају се проширене рекурзивне мреже прелаза. Такве променљиве се потом користе за формирање излаза проширених рекурзивних мрежа тако што се над њиховим вредностима врше различите операције брисања, замена или допуна. Примене описаних графова у екстракцији информација биће представљене у наредном поглављу.

2.5.2.3 Екстракција информација у систему Unitex

Даље трансформације текста у процесу екстракције информација остварују се употребом описаних ресурса, како ће бити представљено на примеру екстракције приближних мера карактеристичних за кулинарски домен (Krstev, Vujičić Stanković i Vitas, 2014).

Приближне мере које су карактеристичне за кулинарске текстове представљају једну класу именованих ентитета. Поред њих, у кулинарским текстовима се појављују и мере представљене стандардним јединицама. Текстуелна секвенца која означава меру може у себи да садржи и друге именоване ентитете којима се описују износи и временске одреднице (попут *четири шоље* или *тридесет минута*).

За српски језик је развијен систем за препознавање именованих ентитета (Krstev, Obradović, Utvić i Vitas, 2013). То је ручно израђен систем заснован на правилима која се ослањају на језичке ресурсе за српски имплементиране у систему Unitex. Он препознаје главне типове именованих ентитета: имена особа, називе локација и организација, временске изразе и нумеричке изразе, укључујући и мере, новац, износе и проценте. За препознавање појединих врста именованих ентитета, на пример, личних

имена и назива локација, значајни су електронски речници и информације у њима. За друге, као што су временски изрази, значајне су локалне граматике у форми коначних трансдуктора да би се препознали сви синтаксички облици у којима може да се појави именовани ентитет. За све њих су развијене локалне граматике које користе шири контекст да би се разрешиле вишезначне појаве.

Систем за препознавање именованих ентитета у српском језику је организован као каскада коначних трансдуктора. Сваки од њих препознаје неки од подтипова именованих ентитета и додаје у текст одговарајућу етикету коју следећи коначни трансдуктори могу да користе. Употреба каскада омогућава, између осталог, разликовање израза којима се описују износи и других израза у којима се користе бројеви, као што су изрази којима се исказују мере.

Да би се препознале јединице за изражавање мера које се користе у српском језику у кулинарству анализиран је доменски корпус рецепата[45] употребом електронских речника и каскада коначних трансдуктора. Америчке стандардне јединице мера попут *инч*, *унца*, *фунта*, *степен Фаренхајт* итд. се не користе у рецептима на српском језику. Што се тиче јединица дужине користи се само *центиметар* (*cm*) у деловима рецепата који описују поступак припреме или деловима у којима се набрајају потребни састојци (пример 13).

ПРИМЕР 13. Примене јединице за дужину *центиметар* (*cm*) у деловима текстова рецепата који описују поступак припреме

Тесто развити на 1cm дебљине

Плех величине 20cm x 28cm подмазати уљем

или у деловима у којима се наводе потребни састојци:

7 котлета дебљине око 2cm

један комад ребара широк 10 до 20cm

[45] Својства кулинарског корпуса описана су у поглављу 4.3.

За навођење температуре употребљавају се само степени Целзијуса (ни једном се нису јавили степени Фаренхајта) за описивање припремне фазе спремања хране или начина очувања припремљене хране. Пун облик *Целзијус* употребљен је само шест пута у кулинарском корпусу, док се чешће користи његова скраћеница *C* (пример 14).

ПРИМЕР 14. Примене јединице за температуру *Целзијус* и скраћено *C* у деловима текстова рецепата који описују припремну фазу спремања хране:

Угрејати пећницу на 200 степени Целзијуса

у рерни загрејаној на 200°C све док се сир не истопи

или начина очувања припремљене хране

Идеална је температура чувања око 10 степени C

За описивање времена припреме или чувања хране користе се јединице као што су: *минут*, *сат*, *час*, *дан* (пример 15).

ПРИМЕР 15. Примене јединица за означавање времена:

Чувајте га у фрижидеру 2-3 дана

Кору сушити 100 минута

оставити да одстоје око сат времена

Оставити на топлом месту пола часа

Јединице пребројавања се користе као јединице мере за означавање тачне количине или као мере за означавање приближне количине (пример 16).

ПРИМЕР 16. Примене јединица пребројавања за тачне количине:

2 велика кромпира

три цела јајета

или за приближне количине:

неколико црних маслинки

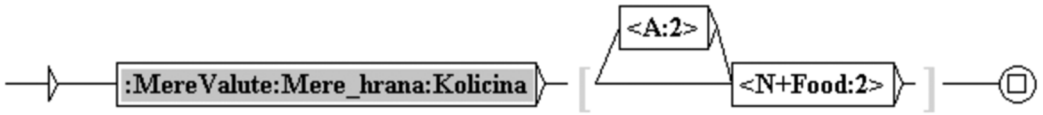
У овој тези пажња је посвећена приближним мерама којима се означавају количине састојака, а које се изражавају неформално у рецептима и нису наведене у стручним приручницима. Да би се оне препознале у кулинарском корпусу направљена је разлика између пребројивих и непребројивих јединица мера.

У случају непребројивих јединица мера поштован је следећи приступ: уколико именици не претходи број нити јединица мере, а иза ње следи именица у генитиву која се односи на храну (којој опционо претходи придев у одговарајућем роду, броју, падежу и категорији живо/неживо), закључује се да она представља непребројиву приближну меру. У овом случају су издвојене само две непребројиве јединице мере – *прстохват* и *на врх ножа*, па ће поступак издвајања бити приказан на случају пребројивих приближних мера.

У случају пребројивих приближних мера примењен је следећи приступ: ако након израза којим се описује износ следи именица у генитиву која се односи на неку врсту хране (и опционо јој претходи придев у одговарајућем падежу, роду, броју и категорији живо/неживо), онда се може претпоставити да се именица која се јавља у изразу којим се описује износ односи на пребројиву приближну меру (на пример, *веза*, *струк*).

Коначни трансдуктор којим је имплементиран овај приступ приказан је на слици 11. Приказани граф поред почетног и завршног стања садржи: позив подграфа `Kolicina` (који се налази у репозиторијуму графова, у подкаталогу `Mere_hrana` подкаталога `MereValute`), кућицу `<A:2>` која препознаје придев у генитиву (тај придев није обавезан) и `<N+Food:2>` која препознаје именицу, означену као храна, такође у генитиву. Образац `<N+Food:2>` одговара именицама у генитиву које се односе на храну и у случају простих речи (на пример, *винобран*) и у случају полилексичких јединица (на пример, *млади лук*). Кућице `<A:2>` и `<N+Food:2>` налазе се унутар великих заграда,

`[` и `]`, и представљају позитиван десни контекст. То значи да оне нису део препознавања, већ се оно што препознаје кућица са позивом подграфа `Kolicina` препознаје само ако следи одговарајући контекст – фраза „придев за којим следи именица у генитиву“. Другим речима, фраза коју чине опциони придев и именица у генитиву на које се јединица мере односи, користи се за препознавање, али не чини део препознате секвенце. Резултујућа секвенца је оно што препознаје позив подграфа `Kolicina`. Подграф `Kolicina` део је система за препознавање именованих ентитета који је описан у (Krstev, Obradović, Utvić i Vitas, 2013) и препознаје бројчане износе којима се изражава количина нечега. Применом овог трансдуктора произведено је 15.521 линија конкорданци (пример 17).



Слика 11. Коначни трансдуктор за препознавање приближних мера[46].

ПРИМЕР 17. Део конкорданци насталих применом трансдуктора са слике 11:

3 **везице** црног младог лука
једну **везицу** исецканог першуновог листа
1 **везу** сецканог першуновог листа
1 **врећица** Винобрана
½ **врећице** прашка за пециво
1 **вршна кашичица** прашка за пециво
5-6 **зрна** бибера у зрну
два **зрна** сувог грожђа
8-10 **зрнаца** црног бибера

[46] Услови слагања придева и именице по роду, броју, падежу и категорији живо/неживо су изостављени због једноставности.

Због употребе контекста у трансдукторима једино кандидати за избор јединица мера могу да чине део резултујућих секвенци, приказаних подебљано у линијама конкорданци, што олакшава проверу великог броја кандидата. Исто важи и за бројеве који су ограничени применом контекста у подграфу `Kolicina`.

Анализом произведених конкорданци изабране су јединице приближних мера и означене оне које су синоними или се користе само са неком одређеном врстом хране. Као резултат овог процеса, добијено је 105 јединица приближних мера од којих су 95 просте речи, а 10 полилексичке јединице.

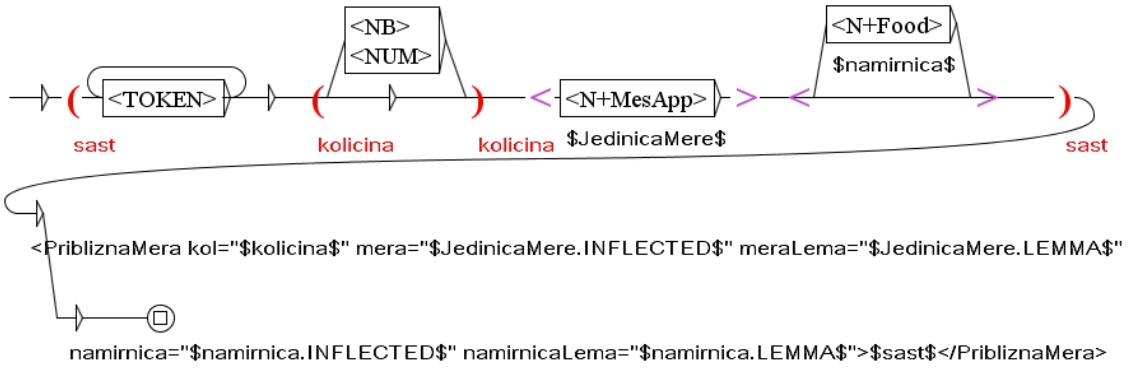
У погледу евалуације у кулинарском корпусу употребом одговарајућих графова пребројани су изрази у којима се јављају јединице приближних или стандардних мера уз именице које се односе на храну (табела 5). Из табеле се види да готово 45% од ових израза користе јединице приближних мера.

Добијена прецизност у случају употребе приближних мера је висока (приближно 1), што је очекивано с обзиром да су за њихово препознавање креиране локалне граматике које препознају само оне информације које одговарају локалном контексту који је тим граматикама описан. Како је кулинарски корпус обиман и како није позната тачна количина релевантних информација у њему није могуће тачно израчунати одзив. Зато је извршена процена квалитета одзива на основу вредности добијених из случајног узорка од 100 текстова кулинарског корпуса. На тај начин је добијена вредност одзива 0,96.

Табела 5. Статистика употребе мерних јединица у кулинарском корпусу.

Јединице	**Са бројевима**	**Без бројева**	**Укупно**
Стандардне јединице	12.966	16	12.982
Јединице приближних мера	7.431	2.933	10.364

Како би се омогућило постављање упита над онтологијама где су инстанце приближних мера и намирница једнаке лематизованим формама даље је креиран коначни трансдуктор приказан на слици 12. Обележавање израза у кулинарском корпусу у којима се употребљавају раније екстраховане приближне мере врши се етикетом `<PriblizнаMera>` која има атрибуте `kol`, `mera`, `meraLema`, `namirnica` и `namirnicaLema` у које трансдуктор уписује издвојене информације о количини (којој одговарају лексичке маске `<NB>` и `<NUM>` којима се препознају редом непрекидне ниске цифара и алфанумерички бројеви), врсти мере (којој одговара лексичка маска `<N+MesApp>`), леми мере, намирници (којој одговара лексичка маска `<N+Food>`) и леми намирнице, редом. Описани коначни трансдуктор за екстракцију информација уједно врши и лематизацију назива приближне мере и намирнице.



Слика 12. Коначни трансдуктор за обележавање израза у којима се употребљавају препознате приближне мере.

Програмски систем Unitex се састоји од више модула. Графичка корисничка сумеђа је написана у програмском језику Java и у оквиру ње корисник бирањем различитих опција предузима одређене акције и подешава параметре рада система. Међутим, главне функције које омогућавају саму обраду текста и примену локалних граматика и других ресурса налазе се у посебним модулима који су написани на програмском језику C++. Ови модули се називају *спољашњи програми* система Unitex. Спољашње програме позива модул корисничке сумеђе, редоследом који је неопходан да би се извршио

одређени задатак који је корисник задао. На тај начин се трансформације текста у процесу екстракције информација остварују употребом радног окружења система, кроз његову корисничку сумеђу. Екстракција информација је у овом случају ограничена на унапред дефинисан редослед примене ресурса и програма и подразумева да корисник мора да предузме низ акција како би спровео све потребне кораке процеса екстракције (нормализацију, токенизацију, примену речника, примену локалних граматика, снимање резултата и др.). Као резултат екстракције информација добијају се текстови у које су уметнуте различите етикете, као што су у случају екстракције приближних мера етикете семантичких класа. У зависности од примене система, често је потребно додатно обрадити овакве текстове (на пример, да би се текстуални подаци трансформисали у нумеричке податке над којима могу бити вршене различите анализе), за шта је неопходно користити друге програме осим система Unitex.

Модуларна архитектура и раздвајање спољашњих програма од корисничке сумеђе омогућавају да спољашњи програми система Unitex буду позивани и из других програма, чиме се омогућава аутоматизација процеса екстракције информација. Такви, посебно развијени системи за екстракцију информација у одређеним доменима, употребљавају ресурсе развијене за систем Unitex и резултате спољашњих програма у комбинацији са другим ресурсима (нпр. онтологијама) и алатима (Stanković, 2009; Pajić, 2012; Vasiljević, 2014).

Спољашњи програми система Unitex који су коришћени за екстракцију информација у овом раду описани су детаљније у прилогу B. Позиви описаних програма су интегрисани са онтологијама како би се омогућило да се резултати употребе у процесима одлучивања и процесима извођења нових закључака.

2.5.3 GATE

GATE[47] је развојно окружење за ресурсе и апликације које решавају различите задатке обраде природних језика, какав је задатак екстракције информација. GATE развија група за обраду природних језика Универзитета Шефилд од 1995. године (Cunningham, Maynard i Bontcheva, 2002; Cunningham i sar., 2011).

Архитектура GATE система се састоји од компоненти које граде три основна типа ресурса – језички ресурси, ресурси за обраду и ресурси за визуелизацију. Језички ресурси представљају језичке податке који обухватају документе, корпусе и онтологије. Ресурси за обраду су алгоритамски ресурси, компоненте за обраду погодне за примену и прилагођавање за нове језике или домене. Ресурси за визуелизацију представљају графичку корисничку сумеђу која обезбеђује приказивање и мењање језичких ресурса и ресурса за обраду. GATE компоненте носе различите функционалности и могу да буду састављене и модификоване тако да одговарају потребама решавања одређеног задатка. Међу њима има језички независних и језички зависних компоненти.

Поред енглеског језика, GATE укључује подршку за бројне језике. Језици за које постоји подршка користе ресурсе енглеског језика. Српски језик, као језик са високо развијеном морфологијом, не може да користи апликације развијене за језике других група што представља мотивацију за развој подршке за обраду текстова писаних на српском језику у систему GATE.

2.5.3.1 Ток обраде у систему GATE

Главни начин складиштења података у систему GATE су корпуси докумената са одговарајућим анотацијама и карактеристикама. Формат улазних и излазних докумената у GATE-у може да буде чист текстуални TXT формат, XML, HTML, SGML, PDF или MS Word. GATE обезбеђује складиштење

[47] GATE: http://www.gate.ac.uk.

језичких ресурса за каснију употребу у интерној GATE репрезентацији документа у XML формату, што је корисно при обради великих корпуса.

Претходна обрада текстова у систему GATE аналогна је описаној обради у систему Unitex и обухвата поделу текста на токене и реченице уз свођење неких ниски карактера на нормализовани облик.

Централна структура у систему GATE су анотације повезане са документима. Све лингвистичке информације у документима су уобличене у форми анотација које су груписане у скупове анотација. Свака анотација се састоји од: типа анотације, њене почетне позиције у тексту који се обрађује, завршне позиције у тексту који се обрађује и скупа карактеристика које описују одговарајући тип анотације (енгл. *features*).

Након претходне обраде тексту се придружује скуп анотација `AnnotationSet`, којима одговарају етикете `Annotation` типа `Token`, `SpaceToken` и `Sentence`. Карактеристика анотације `Token` је врста токена, којом се прецизира да ли је обележени токен реч, број, симбол или знак интерпункције, док се непрекидне ниске сепаратора препознате као `SpaceToken` анотације према карактеристикама деле на размаке и контролне сепараторе.

За текст

Razmutite kvasac u pola šolje mlakog mleka, sa kašičicom šećera i brašna i ostavite da nadođe. Od brašna, nadošlog kvasca i ostalih sastojaka zamesite srednje meko testo, i ostavite da odmara 10 - 15 minuta. Testo podelite na 10 loptica, svaku oblikujte u "glistu", i podelite na 3 jednaka dela. Ispletite pletenice i ostavite da još malo stoje dok se ne zagreje rerna. Pecite na 220° C oko 20 minuta, a pred kraj pečenja premažite ušpinovanim šećerom i pospite krupnim kristal šećerom. Za premazivanje kratko prokuvajte vodu i šećer, 2-3 minuta, da se dobije sirup.

GATE проналази 208 токена где је предефинисани сепаратор са 94 појављивања најчешћи. GATE након претходне обраде интерно представља текст у XML формату. Део ове репрезентације са примерима анотација `Token`, `SpaceToken` и `Sentence` је

```
<TextWithNodes>
<Node id="0"/>Razmutite<Node id="9"/> <Node id="10"/>kvasac<Node
id="16"/> ...
</TextWithNodes>

<AnnotationSet>
<Annotation Id="33728" Type="Token" StartNode="0" EndNode="9">
<Feature>
  <Name className="java.lang.String">length</Name>
  <Value className="java.lang.String">9</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">orth</Name>
  <Value className="java.lang.String">upperInitial</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">string</Name>
  <Value className="java.lang.String">Razmutite</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">kind</Name>
  <Value className="java.lang.String">word</Value>
</Feature>
</Annotation>

<Annotation Id="33729" Type="SpaceToken" StartNode="9"
EndNode="10">
<Feature>
  <Name className="java.lang.String">string</Name>
  <Value className="java.lang.String"> </Value>
</Feature>
<Feature>
  <Name className="java.lang.String">length</Name>
  <Value className="java.lang.String">1</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">kind</Name>
  <Value className="java.lang.String">space</Value>
</Feature>
</Annotation>

<Annotation Id="33730" Type="Token" StartNode="10" EndNode="16">
<Feature>
  <Name className="java.lang.String">length</Name>
  <Value className="java.lang.String">6</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">orth</Name>
  <Value className="java.lang.String">lowercase</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">string</Name>
  <Value className="java.lang.String">kvasac</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">kind</Name>
  <Value className="java.lang.String">word</Value>
</Feature>
</Annotation>
```

```
...
<Annotation Id="33942" Type="Sentence" StartNode="0" EndNode="94">
</Annotation>
</AnnotationSet>
```

Између етикета <TextWithNodes> и </TextWithNodes> наведене су етикете <Node id="BR"/> које садрже атрибуте са вредношћу BR о почетним и завршним позицијама ниски у тексту. Тако је ниска *Razmutite* приказана између етикета <Node id="0"/> и <Node id="9"/>, чиме је означено да је њена почетна позиција у тексту 0, а завршна 8. Размак се налази између етикета <Node id="9"/> и <Node id="10"/> што означава да је његова позиција 9. Слично је означено да је почетна позиција ниске *kvasac* у тексту 10, а завршна 15.

Ове позиције се заједно са јединственим бројем додељеним свакој ниски користе при дефинисању других етикета какви су различити подаци о анотацијама наведени између етикета <AnnotationSet> и </AnnotationSet>. Тако је означено да је ниска *Razmutite* на почетној позицији (StartNode) и завршној позицији (EndNode) са јединственим идентификационим бројем (Id) 33.728 типа (Type) токен (Token), да јој је дужина 9, да је у питању реч чији запис почиње великим словом. Размаку на позицији 9 придружен је јединствени идентификациони број 33.729 и означен је као размак дужине 1 типа SpaceToken. На сличан начин је реченици која почиње на позицији 0 и завршава се на позицији 93, придружен јединствени идентификациони број 33.942 и означена је типом анотације Sentence.

Даља обрада текстова у систему GATE подразумева познавање лема и морфосинтаксичких описа алфабетских токена који су формалне речи. Алати за обављање ових задатака који се оригинално налазе у GATE систему, заснивају се на Бриловом тагеру и одговарају обради текстова писаних на енглеском језику (Brill, 1992; Hepple, 2000). Ови ресурси за обраду зависе од језика на коме се примењују. Према радовима (Popović Z. , 2008; Popović Z. , 2010), Брилов тагер не даје добре резултате за српски језик, па је потребно да се ова компонента прилагоди српском језику како би могла да се ради даља обрада заснована на граматичким правилима.

2.5.3.2 Графови у систему GATE

Даља обрада текста у систему GATE заснива се на листама назива, које обезбеђују препознавање именованих ентитета, и скуповима граматичких правила којима се дефинишу локалне граматике за екстракцију осталих информација потребних за конкретан задатак екстракције.

Листе назива садрже засебне скупове назива као што су имена људи, имена градова, називи организација, дана у недељи, као и листе ознака компанија или звања. За енглески језик је дато око осамдесет листа са по шездесетак елемената у листи. Ове листе се користе да се анотирају појаве ставки листа у текстовима. Овај ресурс је језички зависан с обзиром да за сваки језик листе које су у општем случају на енглеском језику треба заменити одговарајућим листама назива на језику који се обрађује. Препрека која се јавља при обради морфолошки богатих језика какав је српски је што се у текстовима поред леме појављује велики број њених флективних облика који на овај начин не би били анотирани, осим у случају да се формирају листе свих могућих морфолошких облика лема, што одговара формализму електронских речника облика лема DELAF и DELACF.

Скупови граматичких правила се дефинишу *JAPE* (*Java Annotations Pattern Engine*) граматикама (Cunningham i sar., 2011). JAPE граматике су коначни трансдуктори који се користе за описивање регуларних израза над анотацијама и додатним локалним контекстуалним језичким информацијама да би се препознао именовани ентитет или његов тип.

Да би се омогућило да се текстови на српском језику обрађују у систему GATE алатима који нису језички зависни, развијена је подршка за српски језик (Vujičić Stanković, 2012; Vujičić Stanković, Kojić, Rakočević, Vitas i Milutinović, 2012; Vujičić Stanković, 2013). При развоју подршке за српски језик направљена је веза између описаних ресурса система Unitex, електронских речника и локалних граматика, са листама назива и JAPE граматикама система GATE. Један пример њене употребе за унапређење резултата рада система бежичних сензорских мрежа описан је у (Vujičić Stanković, Rakočević i Milutinović, 2011; Vujičić Stanković, Kojić, Rakočević, Vitas i Milutinović, 2012).

На првом месту се употребом спољашњих Unitex програма креирају морфолошки речници текста који се обрађује, а онда се на основу њих производе одговарајуће листе назива, које поред лема садрже и све њихове морфолошке облике. Овакве листе назива садрже само оне називе који се појављују у тексту који се обрађује, чиме се њихова дужина ограничава и решава проблем евентуалног формирања исцрпних листа свих назива и њихових морфолошких облика који се јављају у српском језику. Поред тога, с обзиром на еквивалентност између коначних аутомата, на чијем формализму су засноване локалне граматике и механизам постављања упита у систему Unitex, и регуларних израза, на којима су засноване JAPE граматике, могуће је успоставити везу између ове две компоненте различитих система.

Већ развијени ресурси и функционалности система Unitex за формирање електронских речника текста који се обрађује, употребљени су и у компонентама за означавање граматичких категорија и морфолошку анализу у систему GATE, где су на тај начин свакој текстуалној речи придружене граматичка категорија и одговарајућа лема потребне за даље обраде (прилог Г), тако што јој је придружена анотација типа `POS` са карактеристикама `category` (врста речи), `string` (текстуална реч) и `lemma` (одговарајућа лема текстуалне речи). Тиме се обезбеђује да семантички тагери система GATE, који користе податке о врстама речи и листе назива, у својим граматичким правилима могу да препознају и класификују информације од интереса за конкретан задатак екстракције информација.

2.5.3.3 Екстракција информација у систему GATE

Задатак екстракције информација у систему GATE обављају компоненте за токенизацију и поделу текста на реченице, означавање граматичких категорија, морфолошку анализу, препознавање именованих ентитета и разрешавање кореференци, интегрисане у систем *ANNIE* (*A Nearly-New Information Extraction*) (Cunningham, Maynard i Bontcheva, 2002).

Када се систем ANNIE примењује за решавање задатка екстракције информација из текстова на српском језику користе се оригиналне GATE

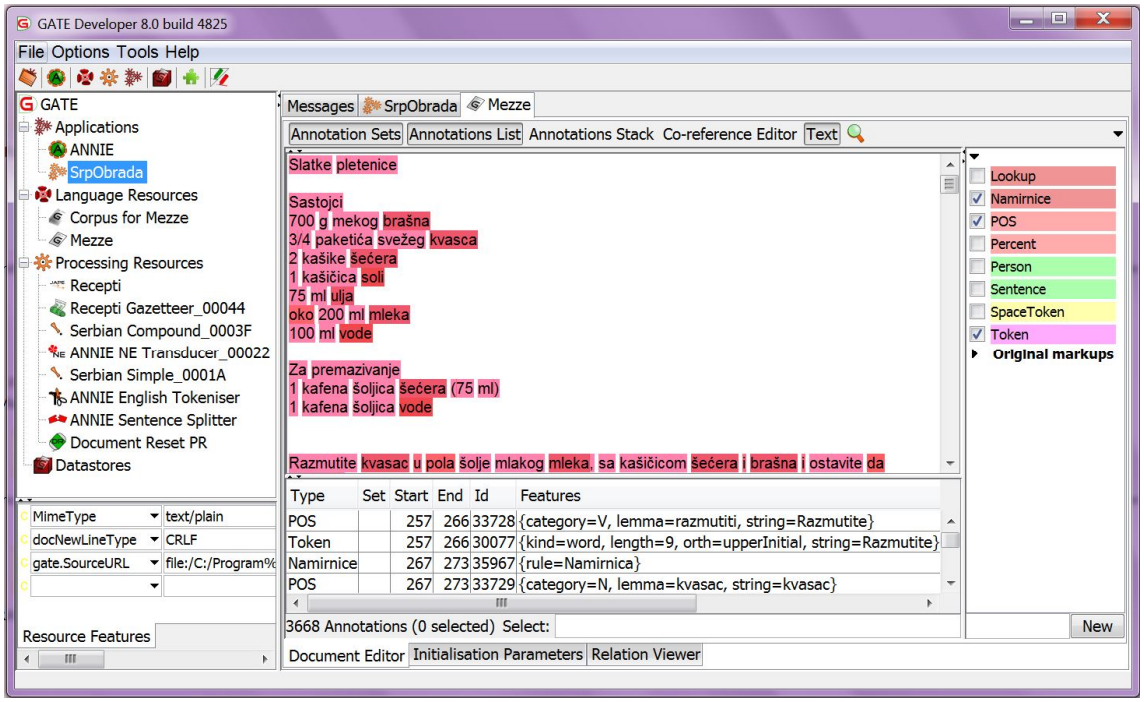
компоненте за поделу текста на реченице и токене, док се као основа за даљу обраду која обухвата дефинисање локалних граматика примењује специјално развијен додатак за српски језик.

Један од задатака екстракције информација из текстова кулинарског корпуса је екстракција информација о намирницама. На слици 13 приказан је текст рецепта на српском језику, који је учитан из улазног документа у чистом текстуалном формату и над којим је извршен задатак екстракције информација о намирницама употребом компоненти приказаних у делу *Processing Resources* с леве стране прозора. Међу примењеним компонентама које су део основног ANNIE система, какве су компоненте за токенизацију (*ANNIE English Tokeniser*) и поделу текста на реченице (*ANNIE Sentence Splitter*), употребљене су компоненте развијене за српски језик за означавање граматичких категорија (*Serbian Simple* и *Serbian Compound*) и листа назива (*Recepti Gazetteer*). Над обележеним нискама текста направљена је одговарајућа локална граматика за екстракцију информација о намирницама *JAPE Transducer Recepti*.

Након обраде текста приказаног на слици произведене су анотације `Token`, `Sentence`, `SpaceToken`, `POS` и `Namirnice` (приказане на десној страни прозора).

У доњем делу прозора су приказани детаљи неколико анотација – тип, подаци о почетној и крајњој позицији дела текста коме анотација одговара, идентификациони број и додатне карактеристике.

У интерном XML формату обрађеног текста ниска *kvasac* је, поред раније описаних етикета, обележена анотацијом типа `POS` са карактеристикама `category`, `string` и `lemma` и одговарајућим *N*, *kvasac* и *kvasac* вредностима, редом. Ове вредности описују да је ниска која се појављује у тексту у облику *kvasac* именица чија је лема *kvasac*, а анотација типа `Namirnice` која има карактеристику `rule` са вредношћу *Namirnica* описује да је у питању намирница.



Слика 13. Текст рецепта на српском језику обрађен у систему GATE.

Систем GATE је алтернативни модел система Unitex, али с обзиром да се описана интеракција између ова два система заснива на подвлачењу Unitex ресурса у GATE и да је Unitex у смислу архитектуре боље теоријски заснован и тренутно има боље развијене ресурсе за српски језик, у овом раду је изабрана употреба система Unitex.

3

Модел екстракције информација вођене онтологијама

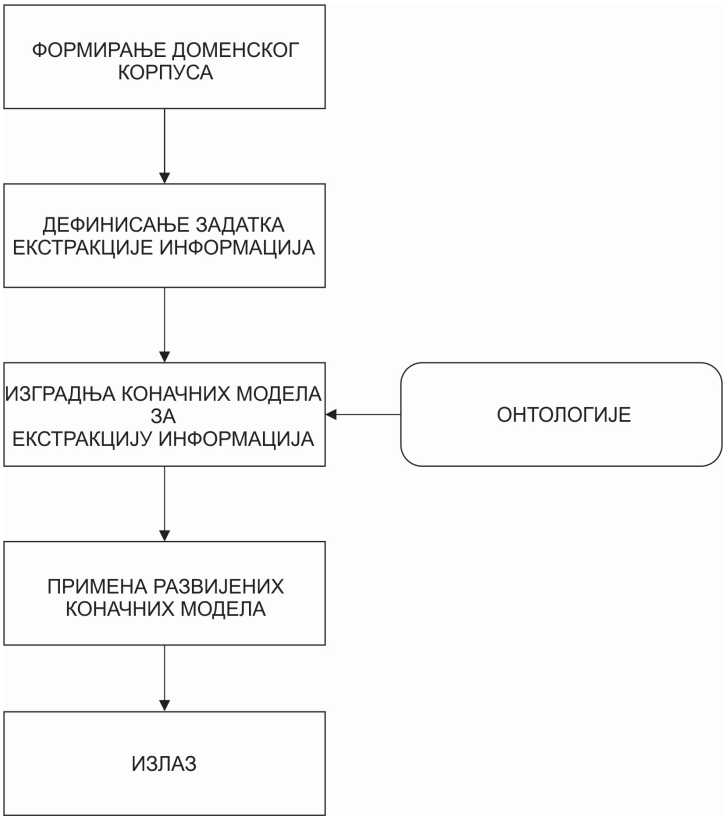
3.1 Екстракција информација вођена онтологијама

Под екстракцијом информација вођеном онтологијама подразумева се процес екстракције информација који поред правила издвајања која су описана локалним граматикама користи и знање садржано у некој онтологији како би побољшао ефикасност правила издвајања и/или класификовања издвојених информација. У већини случајева онтологијом се описују концепти одређеног домена. Како је и сам процес екстракције информација везан за препознавање информација из унапред дефинисаног домена, формално и експлицитно дефинисање концепата и њихових релација кроз онтологије је од помоћи.

Екстракција информација вођена онтологијама (енгл. *Ontology Based Information Extraction*) обухвата задатак препознавања инстанци концепата онтологије у неструктурираним или полуструктурираним природнојезичним текстовима, које се потом користе за резоновање на основу правила која су дефинисана у онтологији или за њихово правилно придруживање концептима онтологије, што се назива аутоматска популација онтологија (енгл. *ontology population*) (Wimalasuriya i Dou, 2009).

Системи за екстракцију информација вођени онтологијама као улазне податке узимају корпус текстова које је потребно обрадити, одговарајуће језичке ресурсе који описују језик корпуса (на пример, електронске речнике) и једну или више онтологија које садрже одређено доменско знање, док су

излазни подаци представљени структурираном листом издвојених информација са једнозначно одређеним значењем. Додатно, и резултат процеса према потреби може да се структурира као онтологија (Wimalasuriya i Dou, 2010). Систем може, а не мора, да буде организован тако да у себе интегрише и модул за израду корпуса, па је општа архитектура система дата на слици 14.



Слика 14. Модел за решавање проблема екстракције информација употребом постојећих онтологија.

Овакви системи подразумевају да већ постоји развијена доменска онтологија којом се моделира знање из домена коме припадају и информације које се издвајају. Један од таквих система је, на пример, систем који употребом онтологија екстрахује догађаје везане за регулацију и експресију гена из текстова научних радова из биологије (Kim i Rebholz-Schuhmann, 2011). Овај систем значајно унапређује процес екстракције информација о регулацији и експресији гена из научних радова, које су раније биле ручно екстраховане и похрањиване у базе података. У систему је употребљена онтологија *Gene*

Regulation Ontology (GRO) (Beisswanger i sar., 2008), како би се на основу придружених семантичких информација из основних препознатих догађаја дошло до информација о сложеним догађајима о регулацији гена.

Онтологије попут GRO и сличних, због значаја њихових примена, израђује и развија велики део научне заједнице на светском нивоу, па су и системи који користе такве онтологије ефикасни и развијени. Међутим, за велики број примена и језика (међу којима је и српски језик), онтологије нису или су недовољно развијене. Због тога је добро и значајно интегрисати процес изградње онтологије у цео систем, као и њену итеративну допуну и примену у правилима екстракције информација.

3.2 Предлог модела екстракције информација вођене онтологијама

У светлу претходних разматрања, основни проблем који је решаван у оквиру овог истраживања јесте како побољшати ефикасност система за екстракцију информација из текстова на природном језику употребом онтологија у правилима издвајања. Иако су наредни примери разматрани и дати за случај српског језика, овај приступ се може применити на било који природни језик, уз коришћење одговарајућих ресурса.

Модел који је развијен у оквиру овог истраживања у себи интегрише основне модуле система за екстракцију информација приказане у оквиру модела на слици 14, са посебним модулом за изградњу и допуну онтологије, чији се излаз користи поново у процесу екстракције информација. Архитектура система за екстракцију информација заснованог на овом моделу приказана је шематски на слици 15.

Предложени модел се састоји од модула за формирање корпуса текстова, модула за дефинисање задатака екстракције информација, модула за изградњу коначних модела за екстракцију информација, модула за примену развијених коначних модела, модула за доградњу морфолошких електронских речника, модула за проширење WordNet-а и модула за изградњу нових онтологија. Сваки од ових модула је независан од осталих, али поступак

њихове примене може да се организује итеративно у корацима тако да се резултати једног модула користе за постизање бољих резултата у другом модулу.

Слика 15. Модел за решавање проблема екстракције информација интегрисањем представљених језичких ресурса и алата.

Први корак је формирање корпуса текстова који ће да се обрађују. За решавање конкретног проблема одлучује се да ли ће да се употребљавају текстови општег корпуса или је потребно да се изгради доменски корпус. Изградња доменског корпуса обухвата прикупљање текстова у штампаној или електронској форми. У оба случаја је неопходно да се текстови додатно обраде пре употребе у наредним модулима.

Када су текстови у штампаној форми у питању, потребно је да се сканирају и да се применом неког од програма за оптичко препознавање знакова преведу у текст. Овај поступак је подложан грешкама које је потребно да се исправе полуаутоматски, применом морфолошких речника и накнадним ручним проверама.

Поред електронских текстова добијених преводом из штампаног облика, извори текстова у електронској форми су и текстови преузети са веба или електронска издања, као што су уџбеници, часописи, енциклопедије или литерарна дела. Ови текстови се налазе у неструктурираној или полуструктурираној форми, па је потребно да се структурирају у формат погодан за даљу обраду. Погодан формат за даљу обраду подразумева да се у зависности од проблема који се решава текст ускладишти у чистом текстуалном формату или структурира у формате као што су XML формат или слогови базе података. Један од начина решавања задатка преузимања текстова са веба и њиховог структурирања који се користи у овом раду јесте и употреба посебно писаних Java програма заснованих на регуларним изразима.

Следећи корак је дефинисање задатка екстракције информација, односно одређивање ентитета који треба да се екстрахују и анализа њихових доминантних структура које се јављају у формираном корпусу. Поред структура којима се представљају основни ентитети, сложене структуре обухватају и групе једног или више ентитета истог или другог типа. На основу описаних доминантних структура праве се одговарајући коначни трансдуктори. При препознавању реализација доминантних структура се такође употребљавају морфолошки електронски речници и локалне граматике које се обрађују у системима као што су Unitex или GATE.

Примена развијених коначних трансдуктора над текстовима корпуса поред екстракције информација обезбеђује развој различитих језичких ресурса који учествују у процесу екстракције. Развој подразумева допуну постојећих и изградњу нових језичких ресурса. Они се затим употребљавају за итеративну међусобну допуну, и измене и допуне коначних трансдуктора како би се добили бољи резултати екстракције информација.

Тако се прегледањем листа непознатих речи које систем Unitex произведе након претходне обраде и примене коначних трансдуктора, издвајају речи које су погрешно излистане као непознате, иако представљају информације које је било потребно екстраховати. Такве речи се допуњавају одговарајућом врстом речи, кôдом флективне класе и опционо низом граматичких, деривационих, дијалектних, доменских или семантичких маркера, а затим се уносе у морфолошке електронске речнике. Из угла екстракције информација придружени доменски и семантички маркери су значајни јер могу да се употребе у коначним трансдукторима и на тај начин да допринесу смањењу вишезначности, у случајевима када појам припада већем броју домена.

Проучавањем речи из електронских речника добијених применом коначних трансдуктора, могу да се направе хијерархије речи, установе њихове релације и да се оне организују у онтологије. Примера ради, појам може да се употреби за допуну хијерархије WordNet-а тако што се направи нови синсет. При томе се води рачуна да ли тај синсет одговара већ постојећем синсету у енглеском WordNet-у како би се у српски WordNet пресликала његова позиција у хијерархији. У супротном се његова позиција одређује на основу постојеће хијерархије српског WordNet-а, а он се опционо означава као синсет који је својствен српском језику.

Са друге стране, анализом доменски специфичних грана из енглеског и српског WordNet-а, издвајају се појмови које је потребно пресликати у српски WordNet и које је потребно прво допунити одговарајућим одредницама па додати у морфолошке електронске речнике српског језика. При допуни електронских речника, доменски и семантички маркери се додају на тај начин да одговарају хијерархији која важи у WordNet-у.

На овај начин систем се користи и за изградњу ресурса и за побољшање квалитета екстракције информација. Језички ресурси добијени кроз овај поступак се употребљавају за проширење резултата екстракције информација синонимијом и другим релацијама. Детаљан процес имплементације модела

биће представљен у поглављу 4.5, а у наредном поглављу ће бити посебно размотрен модул за развијање онтологије.

3.3 Модул за развијање онтологије

Онтологије које би се користиле у процесу екстракције информација, уколико не постоје, могу бити креиране на више начина, у зависности од потребе и расположивих ресурса. Наравно, увек је могуће започети са изградњом онтологије од почетка, описујући скуп класа, својстава и релација, а затим попуњавањем инстанци. Ипак, велики број онтологија је развијен, па је добро искористити знање већ садржано у њима, као и знање које је садржано у другим типовима електронских језичких ресурса, какви су електронски речници.

3.3.1 Превођење онтологије са једног језика на други

За неке светске језике, као што су енглески и француски, постоји велики број већ развијених онтологија, какве су *Prolex*, *AGROVOC*, *Friend Of A Friend* или *Dublin Core*.

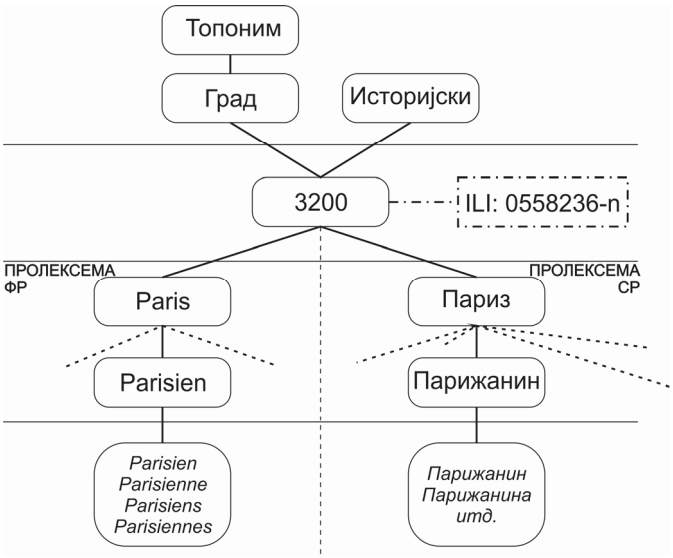
Prolex онтологија је настала у оквиру Prolex пројекта[48] који је започет 1990. године, мотивисан решавањем проблема препознавања властитих имена. У вишејезичним апликацијама је опис властитих имена електронским речницима недовољан због сложених семантичких релација које их повезују, па се у вишејезичном контексту користи репрезентација онтологијама (Gruber, 1995). Општа структура онтологије властитих имена предложена је у (Krstev, Vitas, Maurel i Tran, 2005) и организована у језички независним и језички зависним нивоима.

У језички независним нивоима концепт властитог имена је представљен идентификатором (јединственим у различитим језицима). Он је доведен у везу са надтиповима који описују класе властитих имена у складу са њиховим семантичким и синтаксичким особинама, и типовима који дају финију

[48] Prolex: http://www.cnrtl.fr/lexiques/prolex.

класификацију надтипа. На тај начин је на језички независном нивоу направљена разлика између историјских, религијских и фиктивних имена. У језички зависним нивоима су описане реализације властитих имена у конкретним језицима, њихове варијације у правопису, скраћене форме, акроними итд. и успостављене су њихове везе са одговарајућим флективним и деривационим облицима. Тиме је обезбеђена структура за прецизно превођење концепата са једног језика на други. На слици 16 приказан је пример имплементације дела онтологије за властито име *Париз* за случај српског и француског језика, преузет и прилагођен из рада (Krstev, Vitas, Maurel i Tran, 2005).

На концептуалном нивоу концепту властитог имена *Париз* додељена је `ID` вредност `3200` и придружена веза са међујезичким индексом из WordNet-а са вредношћу `0558236-n`. На мета-концептуалном нивоу додељене су му релације са надтиповима `Историјски` и `Топоним`, и типом `Град`. На лингвистичком нивоу је концепт реализован пролексемама `Paris` у француском и `Париз` у српском делу и повезан са деривационим облицима `Parisien` и `Парижанин` који означавају становника Париза редом на француском и српском језику.



Слика 16. Концепт Prolex онтологије властитог имена *Париз* на француском и српском језику.

AGROVOC[49] је онтологија коју је развила Организација за храну и пољопривреду Уједињених нација (*Food and Agriculture Organization*[50]) раних 1980-их. То је вишејезична онтологија која обухвата 21 језик[51]. Намењена је систематизацији терминологије свих области пољопривреде, шумарства, рибарства, исхране и неколико еколошких домена какви су квалитет животне средине, загађење животне средине, итд. Састоји се од преко 32.000 појмова који се користе за индексирање, претраживање, екстракцију и организовање података у пољопривредним информационим системима и веб странама. Пример је онтологије која је добијена као резултат споразума о доменској терминологији у заједници стручњака из те области. Међутим, српски језик није заступљен у AGROVOC-у иако постоје материјали који су преведени на српски језик (Zaid, Hughes, Porceddu i Nicholas, 2007).

Friend Of A Friend ontologija[52] (FOAF) описује особе, њихове активности (на пример, фотографије, календаре, веб блогове) и везе са другим особама и објектима. Сваки корисник интернета може да употреби FOAF при обележавању своје личне стране како би описао податке о себи. FOAF је једноставна технологија која олакшава дељење и употребу информација о особама и њиховим активностима, пренос података између веб страна, као и аутоматско проширење, спајање и поновно коришћење информација на вебу. Пројекат израде FOAF онтологије, са циљем стварања мреже машински читљивих веб страна које описују особе, везе између њих и ствари које стварају и раде, започет је 2000. године, а данас постоји више милиона FOAF профила особа објављених на вебу и тај број свакодневно расте.

Dublin Core ontologija[53] (DC) се користи за опис дигиталних докумената. Појмови DC онтологије могу да се користе за описивање низа веб ресурса: видео снимака, фотографија, веб страница итд., као и физичких ресурса (на

[49] AGROVOC: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus.
[50] FAO: http://www.fao.org/home/en/.
[51] Податак преузет са званичне стране у децембру 2014. године.
[52] Friend Of A Friend онтологија – спецификација: http://xmlns.com/foaf/spec.
Friend Of A Friend онтологија – страна пројекта: http://www.foaf-project.org/.
[53] Dublin Core онтологија: http://dublincore.org/.

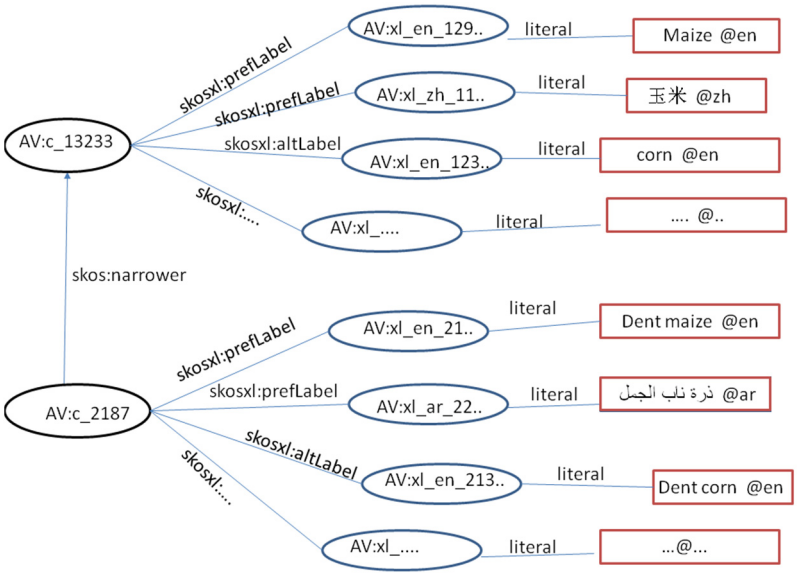
пример, књига) и објеката (на пример, уметничких дела). DC је 2006. године постао ISO 15836 стандард и користи се као основни елемент скупа података за опис ресурса за учење.

Како је наведено, неке од ових онтологија су вишејезичне, у смислу да је за сваки концепт предвиђена могућност уношења његовог превода на већи број језика. У пракси, иако та могућност постоји, често није искоришћена.

Директно превођење концепата једног језика на други може да послужи као вид изградње онтологије. Овде треба имати у виду да по природи самих језика то није увек једнозначно, нити могуће. Поједини концепти су специфични искључиво за неке језике и културе, док се у другима не јављају, тако да за њих не постоји одговарајући превод. Са друге стране, за концепт једног језика у другом језику може да постоји већи број одговарајућих концепата који дубље описују његово значење. Тако су у енглеском језику концепти који описују родбинске везе сиромашнији него у српском језику. За енглеску реч *aunt* при превођењу на српски језик без проучавања ширег контекста није сасвим јасно да ли се мисли на тетку, стрину или ујну, и слично, *uncle* може да означава течу, стрица или ујака.

Пример два концепта AGROVOC онтологије дат је на слици 17 (AIMS Agricultural Information Management Standards, 2014). Приказани су шири концепт *кукуруз* (енгл. *maize*) и једна његова подврста, ужи концепт, *кукуруз зубан* (енгл. *dent maize*).

У елипсама су класе и инстанце приказане својим јединственим идентификаторима на вебу, URI-јима (*Uniform Resource Identifier*), док су у правоугаоницима приказани литерали који одговарају реализацији ових класа у различитим језицима. Скраћенице AV, SKOS и SKOS-XL одговарају веб адресама http://aims.fao.org/agrovoc, http://www.w3c.org/2008/05/skos-xl#, и http://www.w3c.org/2004/02/skos/core#, редом.



Слика 17. Пример реализације концепата у AGROVOC онтологији (AIMS Agricultural Information Management Standards, 2014).

Реализације концепта *maize* у различитим језицима приказане су на слици 18.

Слика 18. Реализације концепта *maize* у AGROVOC онтологији.

У сваком случају, постојеће онтологије и начин организације њихових класа и инстанци могу да послуже као полазна основа за изградњу онтологија на другим језицима.

3.3.2 Конвертовање WordNet-а у формалну онтологију

Трансформација WordNet-а у формалну онтологију, осим организације основног знања у таксономију, пружа могућност да се дефинишу додатна правила одлучивања, да се унапреде постојеће релације и произведу нова знања.

Први од таквих приступа конвертовања приказан је у (Brickley, 1999), где није очувана структура оригиналног WordNet-а јер су само именички синсетови преведени у RDFS класе и задржана релација хиперонимије.

Приступ представљен у раду (van Assem, Menken, Schreiber, Wielemaker i Wielinga, 2004) конвертује WordNet у OWL репрезентацију. Он је организован у четири фазе: припремна фаза, синтаксичка конверзија, семантичка конверзија и стандардизација. Поступак је документован смерницама које су постале званична препорука W3C конзорцијума[54] за конверзију принстонског WordNet-а у RDF/OWL (van Assem, Gangemi i Schreiber, 2006).

Конверзија представља литерале речи и придружене бројеве којима се назначавају њихова различита значења као засебне класе са јединственим идентификаторима ресурса који омогућавају да се директно реферише на њих приликом дефинисања релација. Конверзија обухвата представљање свих релација WordNet-а, дефинисање инверзних релација, својстава објеката и ограничења.

Овај приступ је широко прихваћен и направљене су његове различите модификације и апликације. Неке од њих су представљене у (Graves i Gutierrez, 2006) и (Huang i Zhou, 2007).

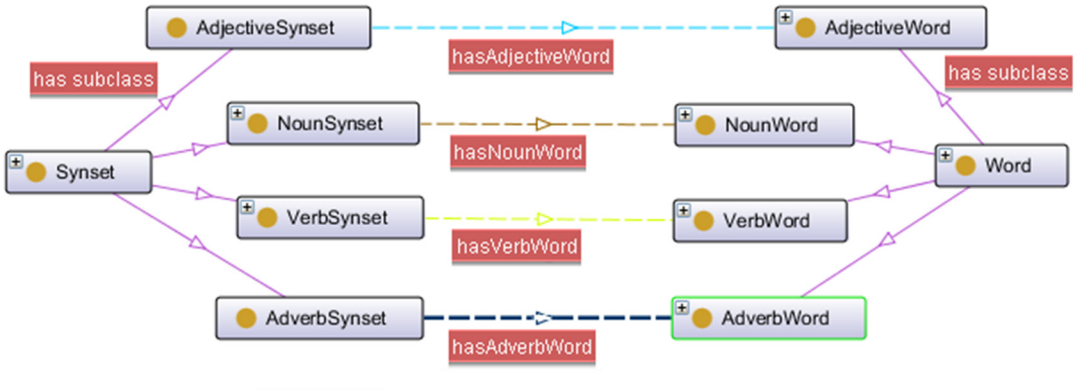
Структура српског WordNet-а одговара структури BalkaNet-а, која је изведена из EuroWordNet структуре. Поред основне структуре WordNet-а при конверзији EuroWordNet-а у OWL репрезентацију било је неопходно узети у обзир и вишејезичност, обратити пажњу на међујезички индекс ILI (описан у поглављу 1.4.4 уводне главе), који повезује синсетове који одговарају истом

[54] W3C конзорцијум: http://www.w3.org/.

значењу појма у различитим језицима, као и на додатне семантичке релације (De Luca, Eul i Nurnberger, 2007).

Међутим, BalkaNet структура има своје специфичности у односу на EuroWordNet структуру. Једна специфичност је *проблем различитих врста речи* (енгл. *cross-part of speech problem*), када литерал који у енглеском WordNet-у припада синсету који описује једну врсту речи, у језику BalkaNet-а припада синсету који одговара другој врсти речи. Ова врста проблема је разматрана у раду (Krstev, Pavlović-Lažetić, Vitas i Obradović, 2004), где је решавана увођењем нових релација као што су *eq_xpos_synonym*, *be_in_state*, *derivedVN*, *derivedPos,* итд. Друга специфичност, *непостојање лексичке подударности* (енгл. *lexical gap*) разматрана у радовима (Bentivogli i Pianta, 2000; Tufiş, Ion, Bozianu, Ceauşu i Ştefănescu, 2008), постоји када концепт не може бити изражен лексичком јединицом, већ само слободном комбинацијом речи. Трећа специфичност односи се на неједнозначно и необавезно обележавање значења литерала унутар синсета (енгл. *literal sense unjustification*) (етикета „Sense"), у језицима који сачињавају BalkaNet. Проблем је разматран у радовима (Koeva, Mihov i Tinchev, 2004; Erjavec i Fišer, 2006; Fišer, 2009; Tufiş, Verginica, Ştefănescu i Ion, 2013), где се приказују различите технике уједначавања обележавања значења литерала на нивоу језика који је предмет интересовања када обележавање из пристонског WordNet-а не може да се примени, као што је у случајевима када је потребно обележити језички специфичне литерале.

У складу са наведеним специфичностима, за потребе овог истраживања развијена је BalkaNet WordNet онтологија на основу BalkaNet семантичке мреже, чија је таксономија представљена на слици 19, са две главне класе – *Synset* и *Word*. Изведене класе су креиране у зависности од вредности врсте речи и дефинисане су као међусобно дисјунктне (пример 18).



Слика 19. Таксономија предложеног BalkaNet OWL модела.

ПРИМЕР 18. Део OWL кôда BalkaNet OWL модела дисјунктних изведених класа *VerbSynset*, *AdjectiveSynset* и *AdverbSynset*:

```
<rdfs:subClassOf rdf:resource="&swn30;Synset"/>
  <owl:disjointWith rdf:resource="&swn30;VerbSynset"/>
  <owl:disjointWith rdf:resource="&swn30;AdjectiveSynset"/>
  <owl:disjointWith rdf:resource="&swn30;AdverbSynset"/>...
</owl:Class>
```

Дефинисано је 18 релација као што су: синонимија, хиперонимија, меронимија, антонимија, итд., које су описане са пет општих параметара: *тип, домен, кодомен, ограничења домена и кодомена релације, смер у хијерархији класа које учествују у релацији,* као и помоћу четири бинарна параметра којима се описује природа релације (*транзитивна*, *симетрична*, *рефлексивна, инверзна*).

На крају су генерисане инстанце. Инстанце класе *Word* описане су својствима типа података (*Data Properties*), док су инстанце класе *Synset* описане својствима типа података (*Data Properties*) и својствима објеката (*Object Properties*). Такође су дефинисана ограничења кардиналности својстава (пример 19).

ПРИМЕР 19. Следећим делом кода је дефинисано да својство *synsetId* има тачно једну *string* вредност.

```
<owl:Restriction>
<owl:onProperty rdf:resource="&swn30;synsetId"/>
<owl:qualifiedCardinality rdf:datatype="&xsd;nonNegativeInteger">
 1</owl:qualifiedCardinality>
<owl:onDataRange rdf:resource="&xsd;string"/>
</owl:Restriction>
```

Предложена конверзија је имплементирана у оквиру система SWNE 2.0[55] (Mladenović, Mitrović i Krstev, 2014) за српски WordNet. У овом систему се у основној форми WordNet складишти као релациона база података и омогућена је његова серијализација у XML формат. Његова доградња поред тога омогућава да се српски WordNet серијализује у OWL формат у коме се користи предложена структура BalkaNet WordNet OWL онтологије за креирање формалне OWL онтологије српског WordNet-а. Ова врста интеграције у SWNE 2.0 систем доприноси да корисник може да допуњава и побољшава српски WordNet кроз графички кориснички интерфејс система, да постојеће апликације које користе XML формат српског WordNet-а (Krstev, Stanković, Vitas i Obradović, 2006) и даље могу да користе допуњене верзије, али и да нове апликације, базиране на онтологијама, могу да користе овај ресурс у формалном облику. Овакав приступ омогућује брз и поуздан развој формалне онтологије WordNet-а, али и онтологија нижег нивоа, оних које настају њеним трансформацијама.

3.3.3 Издвајање онтологија нижег нивоа

Уколико постоји већ развијена онтологија ширег домена, могуће је аутоматски издвојити само један њен део који описује подјезик текстова који се обрађујy (Ding, Lonsdale, Embley, Hepp i Xu, 2007; Grau, Horrocks, Kazakov i Sattler, 2007; Stecher, Niedere, Nejdl i Bouquet, 2008; Lonsdale, Embley, Ding, Xu i

[55] SWNE 2.0: http://sm.jerteh.rs.

Hepp, 2010). Такав приступ је употребљен и у оквиру овог истраживања, за издвајање онтологије хране нижег нивоа (описано у поглављу 4.4.2).

Приликом издвајања подонтологије најпре се за конкретан задатак одређују релевантне класе, инстанце и релације које је потребно издвојити из шире онтологије. Овај поступак зависи од намене и домена подонтологије, па у том смислу није једнозначан. Ипак, без обзира на специфичности, аутоматско издвајање подонтологије се састоји из два корака. У првом кораку се креирају класе онтологије, односно таксономија онтологије, а у другом инстанце класа (слика 20). Оба корака се заснивају на хијерархијској структури семантичких концепата онтологије ширег домена. Концепти који се налазе на вишем нивоу хијерархије имају општије семантичко значење него концепти на нижем нивоу.

Слика 20. Аутоматско издвајање подонтологија из онтологије ширег домена.

У зависности од структуре саме онтологије, процедуре за издвајање класа и инстанци су специфичне за сваку онтологију. Тако на пример, у WordNet-у истовремено постоје синсетови који да се односе на општи концепт и они које се односе на инстанце концепата (Gangemi, Guarino, Masolo, Oltramari i Schneider, 2002).

У српском WordNet-у су ови концепти и њихове релације означене увођењем етикета *Sumo* и *SumoType* (Mladenović i Mitrović, 2014), при чему је успостављена веза између WordNet-а и формалних онтологија какве су SUMO[56] и MILO[57] (Niles i Pease, 2003). *Sumo* етикетом је означена релација синсета и SUMO концепта, док је тип везе описан етикетом *SumoType*. Сваки синсет је реферисан као: еквивалентан одговарајућем SUMO концепту (када је вредност

[56] SUMO онтологија: http://www.adampease.org/OP/.
[57] MILO онтологија:
http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/KBs/Mid-level-ontology.kif.

SUMO етикете *SumoType* „=“), инстанца тог концепта (када је вредност SUMO етикете *SumoType* „@“) или концепт нижег нивоа (када је вредност SUMO етикете *SumoType* „+“).

Српски WordNet је интегрисан и са другим лексичким и семантичким ресурсима, какав је *WordNet Domains*[58] (Magnini, Strapparava, Pezzulo i Gliozzo, 2002), који се користи за именовање домена. Сваки синсет за који постоји успостављена релација пресликавања између класе или инстанце класе посматране онтологије и синсета, и званично су је објавили аутори онтологија *Domains* и *Sumo* за верзију PWN 3.0, у српском WordNet-у садржи семантичку етикету *Domain* (Krstev i sar. 2008; Graovac, 2013) и етикету *Sumo*.

Вредности ових етикета се користе, у првом кораку, за генерисање класа онтологије. За унапред дефинисани домен за који се креира онтологија, издваја се подскуп синсетова српског WordNet-а који према вредности етикете *Domains* припадају том домену, а све различите вредности *Sumo* етикета издвојеног подскупа се користе као класе онтологије. У другом кораку се креирају инстанце класа на основу синсетова који имају исту вредност етикете *Sumo* као и класа, али су различитог нивоа у хијерархији, тј. бирају се само синсетови чије су вредности етикете *SumoType* нижег нивоа у хијерархији SUMO онтологије од вредности *SumoType* етикете саме класе, чиме се избегава да синсет који представља класу постане истовремено и себи самом инстанца.

[58] WordNet Domains: http://wndomains.fbk.eu.

4

Имплементација модела

4.1 Увод

У примени формалних модела који омогућавају да се изврши информатичка обрада текста писаног на природном језику којом рачунар „тумачи“ значење које је аутор, који није нужно човек, имао намеру да искаже приликом његовог записивања, у овом раду су истраживани текстови кулинарских рецепата на српском језику. За ту намену је развијен систем за екстракцију информација заснован на онтологијама чије ће изградња и примене бити описани у наставку. Овај систем екстрахује релевантне информације из текстова кулинарских рецепата преузетих са веб страна. Екстраховане информације се обрађују како би им се придружило одговарајуће значење. Процес придруживања значења вођен је онтологијама. Над таквим информацијама снабдевеним значењем корисник може да поставља напредне упите.

4.2 Гастрономија и њен језик

Гастрономија се налази на тромеђи науке, уметности и заната. По први пут се систематско истраживање ове области јавља у Француској почетком XVII века. Термин *гастрономија* први пут је употребио поета Жозеф Бершу 1801. године, за означавање умећа припреме доброг јела. У то време су проучавани и у новијој историји по први пут систематски записивани традиционални рецепти, начини припреме хране, начини понашања домаћина и госта за трпезом, али и увођени нови укуси и манири. Једним од зачетника проучавања гастрономије и гурманске учтивости сматра се и Гримо

де Ла Рејниер, који, који поред осталог кроз серију „Гурманских алманаха“[59], објављиваних од 1803. до 1812. године, као и дело „Приручник за домаћине“[60], уводи модерну гастрономску критику, нове начине понашања, нове начине припреме, послуживања хране, као и нове називе јела. Његове идеје преузима и допуњује филозоф Жан Брија-Саварен у књизи „Физиологија укуса“ (Brillat-Savarin, 1848). Почев од тада припрема јела се посматра као својеврсна уметност, пре него пуко задовољавање потребе за храном – *„јести, то значи хранити свој дух, мисли, способност расуђивања*“ (Onfre, 2002).

У данашње време се, поред широког знања о различитим светским кухињама, под гастрономијом подразумева поседовање бројних додатних знања као што су: знања о системима набавке, системима процене квалитета, протоколу, организационим способностима, о историји, географији, језику. *Кулинарство* или *куварство* је ближе изворном схватању гастрономије и подразумева вештину припремања јела (Vukić i Portić, 2009), мада не постоје строге разлике између ова два термина.

С обзиром на пораст популарности и доступности кулинарских портала на вебу, различитих кувара, као и кулинарских телевизијских емисија, и у овом домену долази до глобализације, развоја нове гастрономске интернационалне културе и речника хране (Jurafsky, 2014). Интересовање за изучавање кулинарства је присутно у различитим областима науке, а интензивно се изучава и из угла рачунарства и лингвистике у смислу помоћи рачунара у кулинарству (Gerhardt, Frobenius i Ley, 2013; IBM, 2014; Szatrowski, 2014). Неки од примера су препоручивање и модификовање рецепата засновано на изградњи мреже састојака (Teng, Lin i Adamic, 2012), интелигентно препоручивање здравих рецепата прилагођено појединцима (Freyne i Berkovsky, 2010), обележавање рецепата подржано семантичким мрежама на основу састојака, поступака припреме и оброка (Zhang, Hu, Mac Namee i Delany, 2008), екстракција куварских поступака припреме из кулинарских рецепата употребом машинског учења (Mori, Sasada, Yamakata i

[59] Оригинални назив *L'Almanach des gourmands.*
[60] Оригинални назив *Manuel des amphitryons* (1808).

Yoshino, 2012), аутоматска екстракција заменских састојака употребом статистичких приступа (Nedović, 2013; Boscarino, Koenderink, Nedović i Top, 2014).

Језичке специфичности кулинарске лексике српског језика нису темељно истражене. Њене поједине карактеристике представљене су у раду (Tomić, 2014), али подјезик кулинарског домена није темељно описан. Пример 20 приказује неколико описа припреме кулинарских рецепата где се уочавају неке од значајнијих карактеристика овог подјезика. У оваквим описима често је изостављен субјекат у реченици. Карактеристична је употреба глаголских именица (на пример, *цеђењем*) и придева насталих од трпних глаголских придева (на пример, *исечен*), а глаголи су углавном дати у инфинитиву, императиву другог лица множине или презенту првог лица множине (на пример, *ставити*, *додајте, уситнимо*) (Krstev i Lazić, 2015).

ПРИМЕР 20.

Пола литре воде и кафену шољицу уља ставите да прокува. Када је прокувало додајте 300 г брашна и пола прашка за пециво и кувајте уз мешање док се не згусне. Тачније, то се одмах згусне, али мешајте на рингли док не постане глатко.

Када се мало прохлади додајте једно по једно 5 јаја уз мешање. Добро сједините све да тесто буде скроз глатко.

Шприцем за тулумбе у врело уље истискујте тулумбе жељене величине. Пола литре воде, ванилин шећер и 750 г шећера прокувајте. Када се шећер отопи прелијте тулумбе. Лимун потопите да одстоји у води, како би бар донекле из коре изашла прљавштина и сл., а затим га исеците на мање комадиће и додајте сирупу и тулумбама.

***

Направити маринаду од соја-соса, ракије, вина, жутог шећера и сока од ђумбира, који се добија цеђењем ољуштеног наренданог ђумбира.

Маринирати бифтек, извадити га из исте и оставити да се оцеди. Затим га изгриловати.

Размутити брашно, скроб, прашак за пециво, со и воду у смесу гушћу од смесе за палачинке. У њу умакати поврће и пржити у дубокој масноћи на умереној ватри.

Преосталу маринаду прокувати са скробом размућеним у води, неколико минута. Тим сосом прелити грилован и исечен бифтек, а може се преливати и док се грилује. Бифтек исећи на трачице и аранжирати га са темпура поврћем и пиринчем. Прелити теријаки сосом.

⁂

Чајне колачиће уситнимо и одвојимо око 100 грама уситњене масе, преко остатка прелијемо белу кафу у количини да се добије маса лака за обликовање. Масу утиснемо у калупи и у масу утиснемо парчиће чајних колутића, које смо раније одвојили.

Шећер упржимо у карамел, уклонимо са ватре и додамо слатку павлаку. Мешамо на лаганој ватри, да се карамел отопи од прилике ће течност почети да ври, када се карамел отопи ако не свакако чим почне да ври, уклоните са ватре и у топлу течну крему, умешајте маргарин и лагано мешајте док се маргарин не отопи, као и остатци карамеле.
Да би се крема брже стврдла и охладила, распоредите је у тацну и ставите тацну фрижидер, на око 10 минутата да се стегне.
Кикирики печени неслани, још мало препечите у тефлонском тигању, да се добије јача арому, а након тога мало га уситните у авану.
Шлаг крем од ваниле улупајте, по упутству са кесице, додавајући нес кафу у праху.

У улупан шлаг, додајте скроб кроз ситно сито и полако мешајте миксером, да се не створе грудвице. На овај начин шлаг ће бити чвршћи.

На подлигу од чајних колачића распоредимо крему од карамела, преко креме поспемо кикирики, преко кикирикија нанесемо шлаг са кафом. А од горе торту поспемо струганом чоколадом и одложимо торту у фрижидер на сат два.

Једна од карактеристика кулинарских рецепата је употреба израза којима се описују мере. Од израза којима се описују мере зависе, примера ради, контрола трошкова приликом припреме хране, конверзија величина у рецептима, измене постојећих и креирање нових рецепата (Blocker i Hill, 2007). На основу њих се израчунавају калорије у храни која се припрема (Marcus, 2013) и омогућује да се испуне посебни критеријуми у вези количине и нутритивних вредности за исхрану постављени пред кухиње у различитим окружењима какви су ресторани, школе, болнице, итд. (Edelstein, 2008).

За њихову аутоматску обраду неопходно је да се зна шта су јединице мере и како се међусобно односе. Врсте јединица мера које се користе у кулинарству су према (Edelstein, 2008) јединице дужине, масе и запремине (метричке и америчке јединице и њихови односи), јединице температуре (Целзијуса и Фаренхајта и њихови односи) и односи између величина које се мере стандардним кашикама и шољама. У књизи „Кулинарски прорачуни: поједностављена математика за кулинарске професионалце“ (Jones, 2008) као јединица мере такође је укључено пребројавање.

Када корисници који нису професионални кувари пишу кулинарске рецепте за друге кориснике, наводе јединице мере на специфичне начине. Они користе стандардне јединице мера и пребројавање заједно са нестандардним (неформалним) јединицама (на пример, *на врх ножа*). За обраду овако наведених мера неопходно је да оне буду садржане у ресурсима како би се утврдило њихово значење и како би се мере написане у различитим форматима, конвертовале у прецизне мере.

Различити народи имају специфичне начине припреме хране. Иако границе у гастрономији нису оштре, разликује се неколико типова кухиња – западни (европски) тип, источни (јапански и кинески) тип, малоазијски

(оријентални) тип, афрички тип и остале егзотичне гастрономије (Portić, 2011b). Често се говори и о домаћој, националној и интернационалној кухињи. При томе се подразумева да домаћа кухиња обухвата јела регионалног карактера, специфична за ужи регион, а да национална кухиња обухвата јела специфична за једну земљу. Како се обично регионална јела прихватају у целој земљи као „домаћа“, то се термин домаће кухиње полако губи и потпада под термин националне кухиње. Интернационална кухиња обухвата јела различитих светских кухиња која су прихваћена и афирмисана широм света.

Трендом глобализације и широке распрострањености интернационалне кухиње, јавља се проблем код увођења нових или превођења старих назива намирница, јела, прибора и начина припреме. Јавља се питање који појам у једном језику одговара појму другог језика, да ли у оба језика постоје адекватни појмови и ако не, на који начин треба увести нови појам. У раду (Epstein, 2009) се разматра проблем који се јавља приликом превођења састојака рецепата у случајевима када неки од суштинских састојака за припрему рецепта не може да се набави у земљи за чије се говорно подручје прави превод. Један од начина да се превазиђе проблем јесте да се наведе састојак који је замена. У том случају се поставља питање квалитета превода рецепата ако се сличан проблем јавља за већину састојака, јер постоји могућност да се од једног рецепта, заменом великог броја његових састојака, добије сасвим други. Као решење проблема се сугерише да се наведе оригиналан састојак и листа његових могућих замена.

Поред сугерисања измене рецепта употребом заменског састојка, што може да буде корисно када неко ко прави рецепт нема одређени састојак или не сме да га конзумира, формирање група заменских састојака може да послужи и при утврђивању сличности рецепата. Тако се анализом рецепата може утврдити да су јела различитог назива суштински иста или слична (на пример, *рижото* и *пилав*).

Гримо је описао како су се у време зачетка гастрономије одржавали састанци на којима су јела пажљиво дегустирана, како би им се потом давала оцена или назив. Тако су нека јела понела називе заслужних личности (на пример, *филе од листа Орли*, *бујон Креси* или *пиринач Конде*), описа начина

припреме, поступака припреме или порекла (на пример, *грашак на француски* или *артичоке на италијански*) (Onfre, 2002). Слично се и у називима јела на српском језику јављају имена познатих личности или топоними (на пример, *Карађорђева шницла, Ужички медаљони* или *Краљевачки котлет*).

Поред изворних назива јела, у кулинарству се јављају и називи који су преведени или преузети из других језика, па је потребно да приликом обраде текстова кулинарског корпуса буду третирани као посебне лексичке јединице. За поједине називе интернационалних јела је јасно због чега су у српском језику на одређени начин преведени или дословно преузети из других језика (на пример, *сахер торта* или *тирамису* су изворно преузете из оригиналног назива), док за друге није јасно због чега се тако зову (на пример, *руска салата* се у Русији назива *оливиерова салата*). Такође се код неких јела користе називи потекли из других језика иако у српском језику постоје одговарајући називи попут *помфрит* за *пржени кромпир*, *бекендекс* за *јаја са сланином* или *хемендекс* за *јаја са шунком*. У српском језику овакви називи приликом аутоматске обраде текстова постављају различите проблеме, па је потребно да се посебно препознају и опишу у лексичким ресурсима (Vujičić Stanković i Pajić, 2015).

Језичке технологије, ресурси и алати за српски језик нису довољно развијени за специфични кулинарски домен. Како би се обезбедила могућност аутоматске обраде кулинарских садржаја потребно је развити одговарајуће електронске ресурсе какви су корпус кулинарског домена, електронски речници и онтологије који садрже исцрпна знања о кулинарству.

4.3 Изградња корпуса кулинарског домена

Полазна тачка за аутоматску обраду кулинарских садржаја (кулинарских рецепата, савета, дефиниција, описа итд.) јесте формирање корпуса кулинарске лексике (Vujičić Stanković i Pajić, 2013). Текстови с оваквом лексиком омогућавају доменску допуну електронских речника, проширивање WordNet-а и даљи развој кулинарских онтологија и на њима заснованих

апликација. Као извор текстова за формирање корпуса у раду на овој дисертацији изабрано је неколико веб-сајтова са кулинарским садржајима.

Пре самог преузимања кулинарских рецепата са веб-сајтова креирана је XML шема којом су рецепти обележавани. Анализом садржаја рецепата уочено је да се састоје из следећих елемената: наслов рецепта, категорија (на пример, *доручак*, *предјело*, *главно јело*, *хлеб и пецива*, *посластица*, *вечера* итд.), опис рецепта, време припреме, састојци (понекад раздвојени на посебне целине типа *фил*, *кора* итд.), начин припреме (на пример, *маринирање*, *печење*, *кување*, *без термичке обраде* итд.), текстуалан опис припреме, тежина припреме, број порција. Не садрже сви рецепти све наведене елементе, али је свакако неопходно да се обележе одговарајући делови рецепата који су препознати као неки од ових елемената.

За обележавање су коришћене следеће етикете: за наслове `<NASLOV>`, за категорију `<KATEGORIJA>`, за опис рецепта `<OPIS>`, за време припреме `<VREME>`, за списак потребних састојака `<SASTOJCI>`, а за појединачне састојке `<SASTOJAK>`, за начин припреме `<NACIN_PRIPREME>`, за опис припреме `<PRIPREMA>`, за тежину припреме `<TEZINA>` и за број порција `<BROJ_PORCIJA>`.

Рецепти су преузимани аутоматски помоћу посебно креираних Java програма (енгл. *wrapper*) са веб-сајтова као што су Рецепти[61], Кухињица[62], Велики кувар[63], Coolinarika[64], Гастрономад[65], Моје грне[66], Mezze[67], Миљина кухињица[68], Кутлача[69] и Recepti.org[70]. Један од рецепата приказан је на слици 21.

[61] Рецепти: http://www.recepti.com.
[62] Кухињица: http://www.kuhinjica.rs.
[63] Велики кувар: http://velikikuvar.com.
[64] Coolinarika: http://www.coolinarika.com.
[65] Гастрономад: http://www.gastronomad.rs.
[66] Моје грне: http://moje-grne.com.
[67] Mezze: http://mezze.rs.
[68] Миљина кухињица: http://www.minjina-kuhinjica.com.
[69] Кутлача: http://www.kutlaca.com.
[70] Recepti.org: http://www.recepti.org.

Pileći bataci na grčki način

Osobe koje redovno koriste maslinovo ulje, naročito kada ga upotrebljavaju umesto drugih vrsta masnoća, ređe boluju od srčanih bolesti, ateroskleroze, šećerne bolesti, karcinoma debelog creva, ili astme.

Vreme pripremanja: 60 min.

Sastojci

- 800 g pilećih bataka
- sok od limuna
- 2 kašike ulja
- bosiljak
- menta (nana)
- so
- biber
- 2 čena belog luka
- 150 g feta sira
- crne masline
- 400 g barenog krompira
- 600 g paradajza
- 200 g integralnog hleba

Priprema

Pileće batake staviti u vatrostalnu posudu i preliti marinadom napravljenom od limunovog soka, belog luka, maslinovog ulja, začina i začinskog bilja. Staviti u rernu da se peče na 200°C oko 30 minuta. Tada izvaditi piletinu i na nju izrendati feta sir, dodati masline i vratiti da se zapeče još 10 minuta. Gotove batake poslužiti uz bareni krompir i kriške paradajza. Servirati i krišku integralnog hleba.
Ukus marinade najviše zavisi od vrste mesa. Marinadu je najbolje napraviti sa maslinovim uljem, a umesto seckanog treba upotrebiti zdrobljeni beli luk. Umesto običnog limuna, probajte limetu – izvrsna je za marinade.

Слика 21. Пример рецепта на кулинарској веб страни.

Програми за преузимање рецепата, креирани за потребе овог истраживања, користе регуларне изразе којима се описују ниске карактера које се у тексту препознају као делови рецепта. Регуларни изрази су имплементирани уз помоћ пакета Java класа *java.util.regex*. Регуларни израз се задаје као ниска карактера (објекат Java класе *String*). Он се компилира у инстанцу класе *Pattern*, која се даље користи за креирање објекта *Matcher* за сравњивање ниски карактера текста са регуларним изразом.

Помоћу ових програма се неструктурирани или делимично структурирани подаци са веб страна рашчлањују и пресликавају у структуриране. Овакво рашчлањивање се ослања на структуру веб страна (на пример, постојеће HTML етикете којима је одређен изглед приказа страна) и регуларне изразе којима се препознају делови те структуре. Предност овако креираних програма, који су на посебан начин прилагођени изворном садржају веб стране и њеној структури, лежи у одличним резултатима и високој прецизности. Са друге стране, недостатак им је што без додатног

прилагођавања регуларних израза не могу да се употребе ни при најмањој промени структуре изворне веб стране за коју су направљени, а непримењиви су на веб стране које имају другачију структуру.

Делови рецепта на веб странама (као што су наслов, састојци, време припреме и слично), нису били обележени посебним етикетама које би означавале њихово значење, али сваки од тих делова јесте био посебно форматиран, па је начин њиховог форматирања искоришћен за препознавање одређених делова рецепта. Тако су, на пример, наслови рецепата на једној од веб страна били означени етикетом <h1>, па је њен садржај превођен регуларним изразом у садржај етикете <NASLOV> у XML фајлу (пример 21).

ПРИМЕР 21. Регуларни израз "<h1>([\\d\\D]*?)</h1>" преводи наслов рецепта који је у изворној веб страни обележен етикетама <h1> и </h1> у садржај етикета <NASLOV> и </NASLOV> у резултујућем XML фајлу.

```
// pretraga regularnim izrazom
// NASLOV
String regx_naslov = "<h1>([\\d\\D]*?)</h1>";

Pattern pat1 = Pattern.compile(regx_naslov,Pattern.MULTILINE);
Matcher mat1 = pat1.matcher(strana);
String naslov = null;
if(mat1.find())
{
  naslov = mat1.group(1).trim();
  rezultat = rezultat+"<NASLOV> "+naslov+" </NASLOV>\r\n";
}
else
{
  System.out.println("Nisam nasao naslov!");
  System.exit(1);
}
```

Анализом изворног HTML кôда стране рецепта који је приказан на слици 21 направљен је један од описаних програма за преузимање рецепата. Његовом применом је добијен XML кôд који је приказан у примеру 22.

ПРИМЕР 22.

```
<?xml version=" 1.0 " encoding="UTF-8" ?>
<RECEPT>
       <ID>1714</ID>
       <IZVOR>Kuhinjica</IZVOR>
       <LINK>http://www.kuhinjica.rs/recepti/glavno-jelo/pileci-bataci-na
       -grcki-nacin</LINK>
       <KATEGORIJA>Glavno jelo</KATEGORIJA>
       <NASLOV>Pileći bataci na grčki način</NASLOV>
       <OPIS>Osobe koje redovno koriste maslinovo ulje, naročito kada ga
       upotrebljavaju umesto drugih vrsta masnoća, ređe boluju od srčanih
       bolesti, ateroskleroze, šećerne bolesti, karcinoma debelog creva,
       ili astme</OPIS>
       <VREME jedinica="min">60</VREME>
       <SASTOJCI>
              <SASTOJAK>800 g pilećih bataka</SASTOJAK>
              <SASTOJAK>sok jednog limuna</SASTOJAK>
              <SASTOJAK>2 kašike ulja</SASTOJAK>
              <SASTOJAK>bosiljak</SASTOJAK>
              <SASTOJAK>nana</SASTOJAK>
              <SASTOJAK>so</SASTOJAK>
              <SASTOJAK>biber</SASTOJAK>
              <SASTOJAK>2 čena belog luka</SASTOJAK>
              <SASTOJAK>150g feta sira</SASTOJAK>
              <SASTOJAK>10 crnih maslina</SASTOJAK>
              <SASTOJAK>400g barenog krompira</SASTOJAK>
              <SASTOJAK>600g paradajza</SASTOJAK>
              <SASTOJAK>200g integralnog hleba</SASTOJAK>
       </SASTOJCI>
       <PRIPREMA> Pileće batake staviti u vatrostalnu posudu i preliti
       marinadom napravljenom od limunovog soka, belog luka, maslinovog
       ulja, začina i začinskog bilja. Staviti u rernu da se peče na
       200°C oko 30 minuta. Tada izvaditi piletinu i na nju izrendati
       feta sir, dodati masline i vratiti da se zapeče još 10 minuta.
       Gotove batake poslužiti uz bareni krompir i kriške paradajza.
       Servirati i krišku integralnog hleba. Ukus marinade najviše zavisi
       od vrste mesa. Marinadu je najbolje napraviti sa maslinovim uljem,
       a umesto seckanog treba upotrebiti zdrobljeni beli luk. Umesto
       običnog limuna, probajte limetu – izvrsna je za marinade.
       </PRIPREMA>
</RECEPT>
```

Приликом прикупљања текстова, аутоматски су нормализоване одређене ниске карактера. Ово се посебно односи на представљање карактера као HTML ентитета. Наиме, данас је уобичајено да се за кодирање веб страна користи UTF-8 стандард, али неке старије верзије сајтова задржавају употребу HTML ентитета, попут * * уместо размака, *°* уместо карактера ° или *&scaron* уместо *š*.

Описаним поступком је формиран анотирани корпус кулинарских рецепата. У таквом корпусу сваком рецепту је придружен број који га јединствено идентификује (обележен етикетом <ID>), наведен је наслов веб

стране са које је текст преузет (обележен етикетом <IZVOR>), као и тачна путања изворног текста рецепта (обележена етикетом <LINK>) и јасно су издвојени елементи рецепата.

Уклањањем ознака из корпуса, добија се корпус чистих текстова од 11.670 рецепата, који су коришћени за даље истраживање.

4.4 Креирање доменске онтологије

4.4.1 Проширење WordNet-а и доградња доменски специфичне лексике у морфолошким речницима српског језика

У оквиру овог истраживања WordNet је, у складу са разматрањима приказаним у поглављу 1.4.4 и поглављу 3.3.2, употребљен као онтологија. У ту сврху, српски WordNet је допуњен доменски специфичним терминима и концептима, који су касније употребљавани као онтологија нижег нивоа. Слична проширења WordNet-а за специфичне домене за друге језике представљена су у радовима (Navigli i Velardi, 2002; Vintar i Fišer, 2011), али су размотрени различити приступи у односу на онај који ће бити приказан у овом поглављу.

Српски WordNet, са приближно 21.200 синсетова, један је од најважнијих и најсвеобухватнијих електронских ресурса за српски језик. Додатно, овај ресурс користе бројни истраживачи, па је самим тим и најбоље одржаван. Па ипак, кулинарски домен није до сада био детаљније проучаван у српском језику, те је као такав, неадекватно представљен у српском WordNet-у. Од концепата који су постојали у српском WordNet-у пре овог истраживања, њих 393 је припадало кулинарском домену (од чега 91 специфичних балканских или српских, попут *бурек*, *кајмак*, *гибаница* или *зељаница*), али није обележено одговарајућим маркерима (DOMAIN=CULINARY).

Слична ситуација је и са представљањем кулинарског домена у електронским морфолошким речницима за српски језик. У њима је 218 простих речи и 217 полилексичких јединица било означено семантичким маркером +FOOD који означава категорију хране. Међутим, постојале су недоследности у означавању тих појмова другим маркерима (на пример,

изостао је семантички маркер +CONC за означавање конкретних објеката као општије категорије код 32 просте и 20 полилексичких речи). Такође, на почетку овог истраживања није био познат број оних речи које припадају кулинарском домену, а нису биле означене ни једним од поменутих маркера.

Како би се ови ресурси проширили и како би се омогућила њихова употреба над текстовима из кулинарског домена на српском језику, за означавање концепата и терминологије специфичних за кулинарски домен, уведени су нови доменски маркери, приказани у табели 3 (страна 58).

Поступак проширења семантичке мреже WordNet и доградње доменски специфичне лексике у морфолошким речницима српског језика извршен је у неколико корака:

1. *Превод синсетова из принстонског WordNet-а који припадају кулинарском домену.*

   У овом кораку је циљ био да се истраже специфичне гране принстонског WordNet-а и да се концепти који припадају тим гранама, а раније нису били укључени у српски WordNet, преведу и укључе у њега. За сваки од новоуведених синсетова додата је и одговарајућа дефиниција концепта.

   Гране које су обрађене су: грана *food, nutrient* (хранљива материја), везана за намирнице, производе, пића, јела и оброке, и гране *kitchen utensil* (кухињски прибор и посуђе) и *tableware* (стони прибор и посуђе). Приликом овог поступка, поједини концепти нису могли да буду преведени јер за њих у српском језику не постоје одговарајуће речи.

2. *Проверавање непознатих речи које потичу из примена српских електронских морфолошких речника на корпус рецепата у потрази за новим речима из кулинарског домена.*

   Након примене електронског речника на корпус кулинарских текстова, формирана је листа од 9.100 непрепознатих речи (речи које нису биле садржане у речнику). У овом кораку је циљ био да се из те листе препознају оне речи које припадају кулинарском домену у српском језику и за које се познаје значење. Таквих речи је било 390 и свима су придружени одговарајући маркери и кôдови који означавају врсту речи и класе лема које деле иста флективна својства. На овај начин су

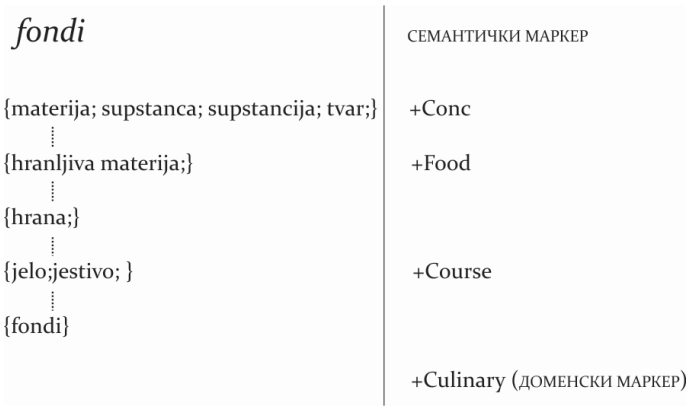
електронски морфолошки речници допуњени неким од значајних и честих појмова у кулинарству, попут *црна чоколада* или *едамер*.

3. *Полуаутоматска производња нових простих и полилексичких речи за електронске речнике са семантичким маркерима изведеним из синсетова српског WordNet-а који припадају кулинарском домену.*

Овај корак се састојао из два подзадатка. Прво су аутоматски произведени нови кандидати за електронске речнике анализом синсетова који припадају хијерархијама раније поменутих грана избором оних који нису већ припадали електронским речницима. Кандидатима су аутоматски придружени одговарајући семантички маркери у зависности од позиције синсета у хијерархији WordNet-а (пример 23).

ПРИМЕР 23. Нови кандидат *fondi* припада хијерархији (слика 22): *{jelo;jestivo;}*, *{hrana;}*, *{hranljiva materija;}*, *{materija; supstanca; supstancija; tvar;}* и због тога су му придружени доменски семантички маркер +CULINARY и семантички маркери +CONC, +FOOD и +COURSE. Додавањем одговарајућих кодова добијена је речничка одредница:

fondi,N1063+Conc+Course+Food+Culinary.



Слика 22. Пример везе између хијерархије концепата у WordNet-у и семантичких маркера у електронском речнику.

Други део се састојао из ручне провере свих нових кандидата и њихових семантичких маркера:

- Део кандидата је одбачен јер су се појавили вишеструко, па је задржан само један од њих – на пример, *бризле* се у WordNet-у јављају у оквиру два синсета дефинисана као *јестиве грудне жлезде животиња* и *јестиве жлезде животиња*, који се налазе у истој хијерархији, односно упућују на додавање истих семантичких маркера.
- Поједини кандидати су одбачени јер нису припадали кулинарском домену иако су се јавили у хијерархијама наведених грана – на пример, *Последња вечера*, *Господња вечера* из гране *хранљива материја*.
- Извршена је провера и исправљање придружених семантичких маркера – на пример, *помфрит* у хијерархији WordNet-а припада грани *површе*, па му је придружен семантички маркер +ALIM за означавање непрерађених намирница, али с обзиром да је помфрит прерађен кромпир тај семантички маркер је замењен семантичким маркером +PROD за означавање производа.

4. *Полуаутоматско додавање семантичких маркера који недостају речима у електронским речницима, на основу синсетова српског WordNet-а који припадају кулинарском домену.*

Четврти корак је извршен на сличан начин као трећи корак, с тим што су разматране само речи из електронских речника из кулинарског домена, којима су недостајали неки или сви семантички маркери изведени из WordNet хијерархије.

5. *Полуаутоматско додавање нових кулинарских концепата у српски WordNet, где су посебно означени они који су специфични за српски језик; њихова ручна провера и исправљање.*

У овом кораку су употребљене нове речи, препознате у кораку 2, како би се креирали нови синсетови за проширење српског WordNet-а и убацили на одговарајућу позицију у хијерархији српског WordNet-а. Овако додати синсетови и њихове позиције су ручно проверени и према потреби су грешке исправљене (пример 24).

ПРИМЕР 24. Међу новим појмовима из кулинарског домена су препознати и они специфични за српски језик, као на пример врста грожђа *афусали*. С обзиром да су новим појмовима у кораку 2 већ придружени одговарајући семантички маркери, они су искоришћени како би се нови синсетови исправно позиционирали у хијерархији српског WordNet-а.

Код полилексичких јединица, речи које учествују у њиховој изградњи су у појединим случајевима указивале на одговарајући хипероним у хијерархији. На пример *ватростална чинија* је врста чиније, а *бућави сир* или *плави сир* представљају врсту сира, па нове синсетове треба позиционирати као синсетове чији су хипероними синсетови *чинија*, односно *сир*, редом.

Српски WordNet је описаним поступком проширен за 1.404 синсета из домена кулинарства и садржи укупно 1.797 таквих синсетова. Додато је 450 концепата из кулинарског домена специфичних за српски језик.

Електронски речници простих речи су увећани за 636 речи, од којих је 246 добијено из српског WordNet-а, а 390 из кулинарског корпуса. Електронски речници полилексичких јединица допуњени су са 612 речи, од којих је 514 преузето из српског WordNet-а, а 98 је добијено преузимањем из кулинарског корпуса. Постојећи семантички маркери су код 735 простих речи и 125 полилексичких јединица допуњени одговарајућим семантичким маркерима уведеним за кулинарски домен. Извод из електронских речника са примерима куварских термина дат је у прилогу Д.

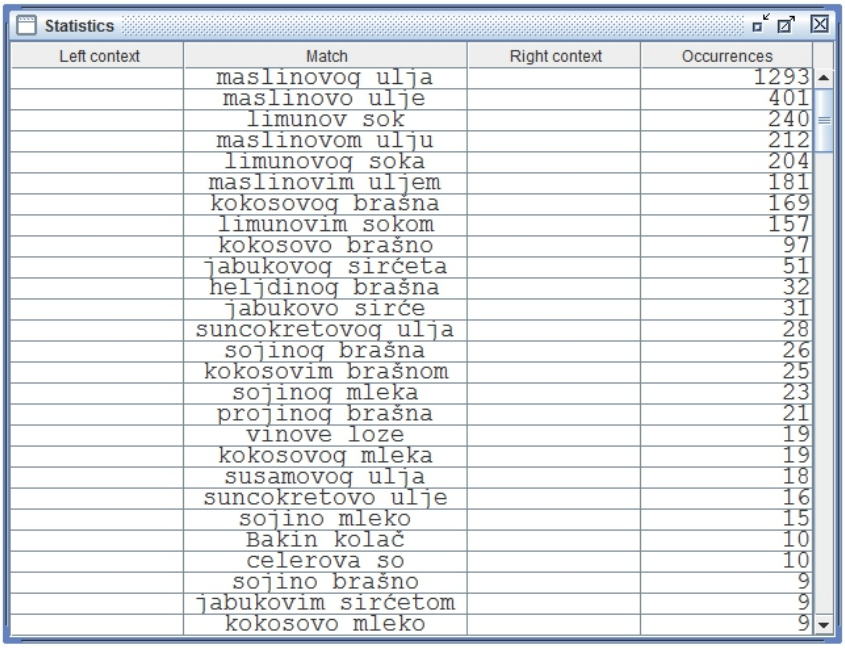
Препознавање нових полилексичких јединица и њихово полуаутоматско издвајање из корпуса кулинарских текстова вршено је анализом доминантних структура и употребом одговарајућих регуларних израза који те структуре препознају.

Примера ради, када се изврши претрага кулинарског корпуса у систему Unitex применом регуларног израза `<A+Pos><N+Food>`, који одговара

структури „присвојни придев за којим следи именица означена семантичким маркером +FOOD“, екстрахују се укупно 3.629 ниски (слика 23).

Слика 23. Део конкорданци добијених претрагом кулинарског корпуса регуларним изразом <A+Pos><N+Food>.

Део резултујућих ниски и њихових фреквенција приказан је на слици 24.

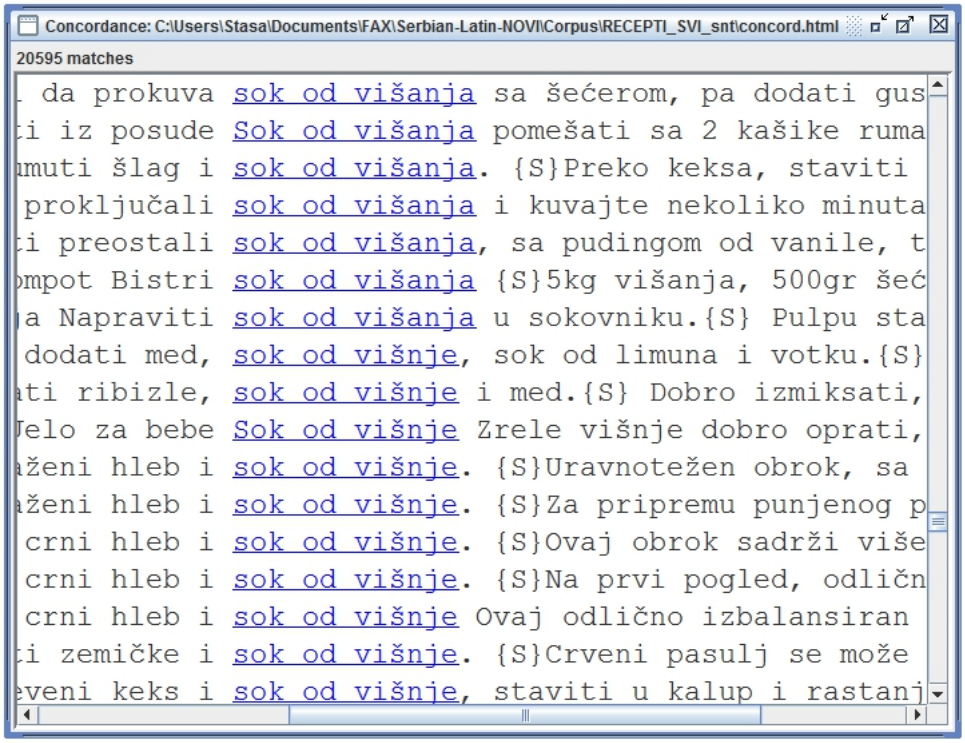


Left context	Match	Right context	Occurrences
	maslinovog ulja		1293
	maslinovo ulje		401
	limunov sok		240
	maslinovom ulju		212
	limunovog soka		204
	maslinovim uljem		181
	kokosovog brašna		169
	limunovim sokom		157
	kokosovo brašno		97
	jabukovog sirćeta		51
	heljdinog brašna		32
	jabukovo sirće		31
	suncokretovog ulja		28
	sojinog brašna		26
	kokosovim brašnom		25
	sojinog mleka		23
	projinog brašna		21
	vinove loze		19
	kokosovog mleka		19
	susamovog ulja		18
	suncokretovo ulje		16
	sojino mleko		15
	Bakin kolač		10
	celerova so		10
	sojino brašno		9
	jabukovim sirćetom		9
	kokosovo mleko		9

Слика 24. Статистички приказ фреквенција структура <A+Pos><N+Food> у кулинарском домену.

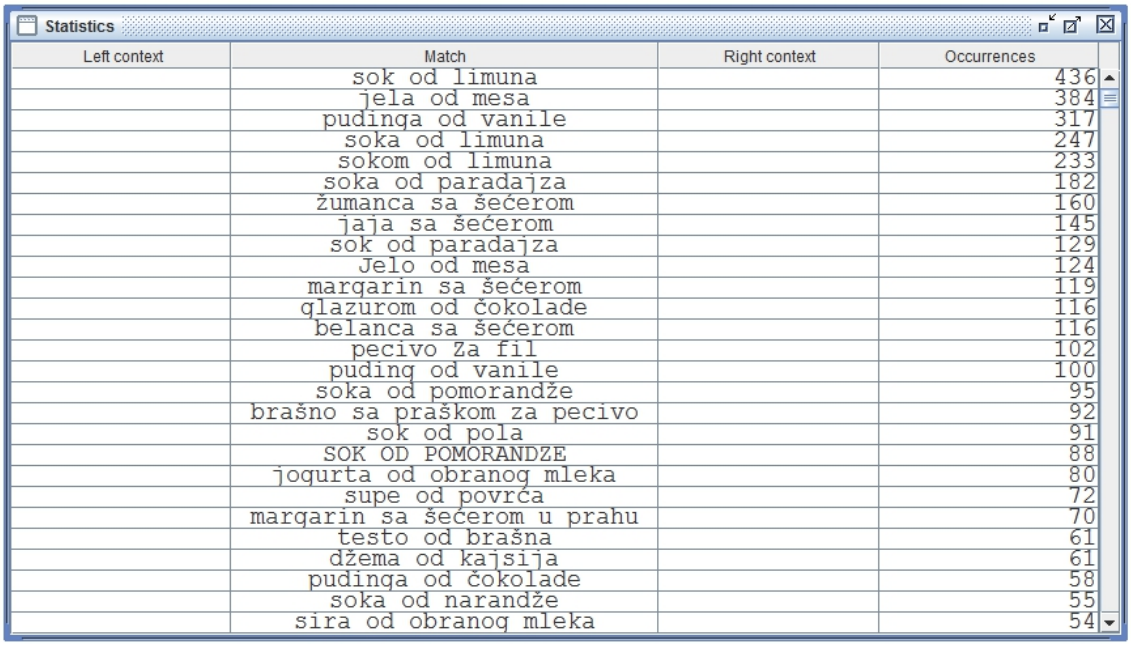
Анализом ових кандидата установљено је да је екстрахован 3.101 коректан облик кандидата, од чега су 54 различити. Пет најфреквентнијих су *маслиново уље*, *лимунов сок*, *кокосово брашно*, *јабуково сирће* и *хељдино брашно*.

Екстракцијом структура „именица означена семантичким маркером +FOOD за којом следи предлог, па поново именица означена семантичким маркером +FOOD“, којима одговара регуларни израз `<N+Food><PREP><N+Food>`, добија се укупно 20.595 екстрахованих ниски (слика 25), од чега су 9.640 различитих облика.



Слика 25. Део конкорданци добијених претрагом кулинарског корпуса регуларним изразом `<N+Food><PREP><N+Food>`.

Анализом првих 1.000 облика сортираних по фреквенцији (слика 26) установљено је да њих 818 представљају добре кандидате за допуну морфолошких речника.



Statistics

Left context	Match	Right context	Occurrences
	sok od limuna		436
	jela od mesa		384
	pudinga od vanile		317
	soka od limuna		247
	sokom od limuna		233
	soka od paradajza		182
	žumanca sa šećerom		160
	jaja sa šećerom		145
	sok od paradajza		129
	Jelo od mesa		124
	margarin sa šećerom		119
	glazurom od čokolade		116
	belanca sa šećerom		116
	pecivo Za fil		102
	puding od vanile		100
	soka od pomorandže		95
	brašno sa praškom za pecivo		92
	sok od pola		91
	SOK OD POMORANDZE		88
	jogurta od obranog mleka		80
	supe od povrća		72
	margarin sa šećerom u prahu		70
	testo od brašna		61
	džema od kajsija		61
	pudinga od čokolade		58
	soka od narandže		55
	sira od obranog mleka		54

Слика 26. Статистички приказ фреквенција структура <N+Food><PREP><N+Food> у кулинарском домену.

Процес доградње морфолошких електронских речника доменском лексиком је дуг и итеративан процес који изискује да се дефинисањем различитих структура и низом експеримената екстрахују кандидати, који се потом лингвистички обрађују, допуњују и уграђују у речнике за коришћење у следећим итерацијама, где могу да учествују у изградњи нових кандидата.

4.4.2 Издвајање онтологије хране

За потребе развијања система за екстракцију информација из текстова кулинарског домена, издвојена је из WordNet-а онтологија хране као онтологија нижег нивоа, поступком објашњеним у поглављу 3.3.3.

Пример 25 приказује SPARQL упит који се користи за креирање класа онтологије хране на основу вредности *Domains* етикете. Овим упитом се формира подскуп синсетова код којих својство типова података *hasDomain* има вредност „gastronomy“. После тога се из овог подскупа издвајају само они синсетови који имају различите вредности својства типова података *hasSumo*. Они су изабрани као класе онтологије хране. На тај начин је издвојена укупно 161 класа из српског WordNet-а.

ПРИМЕР 25.

```
PREFIX rdf: <http://www.w3.org/1999/02/2rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swn30: http://sm.jerteh.rs/swn30#

SELECT distinct ?sumoGroup
       WHERE {?subject swn30:hasDomain ?domainGroup.
              ?subject swn30:hasSumo ?sumoGroup.
              FILTER (?domainGroup="gastronomy" )
              }
       ORDER BY ?sumoGroup
```

Пример 26 приказује SPARQL упит којим се у другом кораку креирају инстанце. Свака инстанца која је члан класе *Synset* у онтологији српског WordNet-а пресликана је у инстанцу оне класе онтологије хране која је одређена на основу вредности својства *hasSumo* те инстанце у српском WordNet-у. На овај начин је преузета 1.091 инстанца.

ПРИМЕР 26.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swn30: <http:// sm.jerteh.rs/swn30#>

SELECT distinct ?subject
       WHERE { ?subject swn30:hasDomain ?domainGroup.
              ?subject swn30:hasSumo ?sumoGroup.
              FILTER (?domainGroup="gastronomy" )
       }
       ORDER BY ?sumoGroup
```

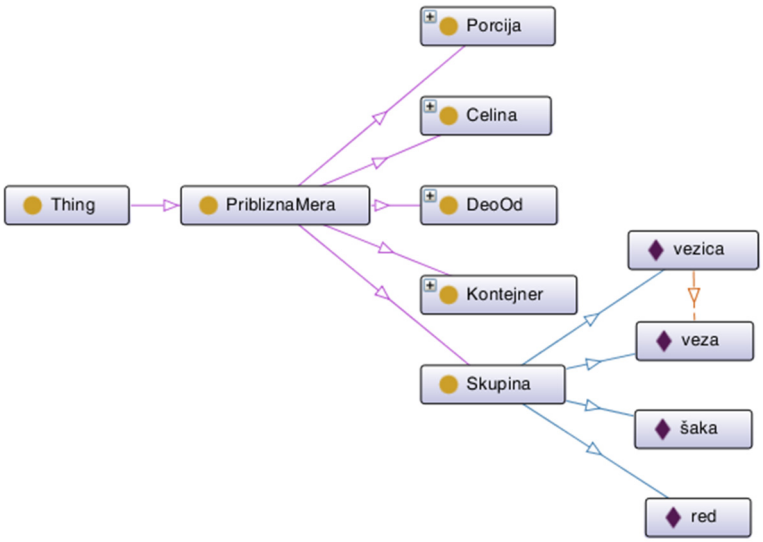
Резултујуће онтологије могу да се прегледају и мењају у онтолошким едиторима. Екстраховане инстанце класе *Paradajz* су приказане на слици 4 (страна 41) употребом онтолошког едитора *Protégé 4.3*[71]. Поред тога, ове онтологије могу да се користе у изградњи различитих апликација.

[71] Protégé 4.3: http://protege.stanford.edu/.

4.4.3 Креирање доменске онтологије приближних мера у кулинарском домену

Екстракција приближних мера које су карактеристичне за кулинарске текстове описана је у поглављу 2.5.2.3. На основу добијених израза, развијена је онтологија за неформалне мере специфичне за кулинарски домен у OWL језику употребом алата Protégé 4.3. Развој онтологије укључује дефинисање класа онтологије, уређење класа у таксономију, дефинисање својстава, опис ограничења која важе за њих (опис дозвољених вредности) и додавање инстанци класама, заједно са вредностима својстава која за њих важе.

Према раније описаној анализи приближне мерне јединице распоређене су у 5 подкласа првог нивоа вршне класе *PriblizнaMera*: *Kontejner*, *Porcija*, *DeoOd*, *Celina* и *Skupina*. Уведено је и својство *jeManja* и његово инверзно својство *jeVeca* којима се означава да је приближна мера мања или већа јединица од друге приближне мере. На слици 27 су приказане наведене класе, поједине инстанце класе *Skupina* (*vezica*, *veza*, *šaka*, *red*) и својство *jeManja* (*vezica jeManja veza*).

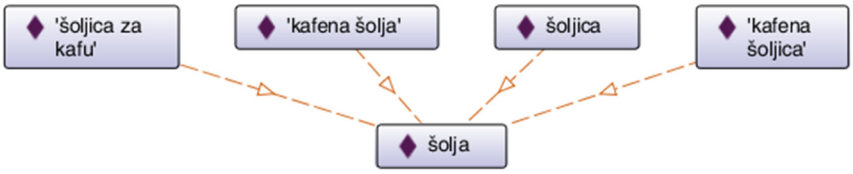


Слика 27. Таксономија онтологије приближних мера, поједине инстанце и њихова својства.

У онтологију су укључена и знања о томе које јединице приближних мера су везане искључиво са одређеном врстом или ограниченим скупом врста хране (на пример, *чен белог лука*, *плоча лиснатог теста*/*плоча лазањи*), као и

да неке инстанце могу да припадају већем броју класа (инстанца *зрнце* у смислу малог *зрна* припада класи *DeoOd* када се односи на *бибер* или *мак* и класи *Porcija* у смислу мале сферне количине када указује на *путер*).

Још један аспект у развоју онтологије јесте могућност да се назначи да се две или више инстанци односе на исту ствар. На пример, инстанце *чен* и *чешањ* треба третирати као исте јединице у кулинарским рецептима. У српском језику су *шољица за кафу*, *кафена шољица*, *шољица* и *кафена шоља* различити изрази за *шољу за кафу*. У онтологији је довољно да се назначи та информација и за један од ових случајева да се наведе да важи релација *jeManja šolja* (на пример, *šoljica jeManja šolja*) да би се закључивањем дошло до закључка да исто важи и за остале три инстанце (слика 28).



Слика 28. Исте инстанце којима је придружена особина *jeManja*.

Класе онтологије (*PriblizнaMera*, *Kontejner*, *Porcija*, *DeoOd*, *Celina* и *Skupina*) су употребљене у развоју доменски специфичног електронског речника приближних мерних јединица. Нови семантички маркери и број инстанци њима одговарајућих класа приказани су у табели 6.

Табела 6. Преглед семантичких маркера предложених за приближне мере у кулинарском домену.

Семантички маркер	**Опис**	**Број инстанци**
+MesApp	приближна мера	108
+Cont	контејнер	33
+Por	порција	33
+Part	део од	30
+Wh	целина	7
+Set	скупина	5

Такође је проверено да ли у српском WordNet-у постоје све издвојене јединице приближних мера и додате су оне јединице за које се испоставило да нису биле раније укључене. Током овог процеса неке јединице приближних мера су премештене из једне класе у другу, која боље одговара позицији у принстонском WordNet-у. Тако је првобитно инстанца *шака* стављена у класу *Skupina*, али је касније премештена у класу *Kontejner* јер је шака хипоним синсета *containerful* у принстонском WordNet-у.

Организација мерних јединица у онтологије погодна је за резоновање и интеграцију у друге онтологије. Једна од таквих примена је приказана у поглављу 4.5.2, где је омогућено да један рецепт који користи приближне мере буде презаписан тако да се приближне мере конвертују у стандардне.

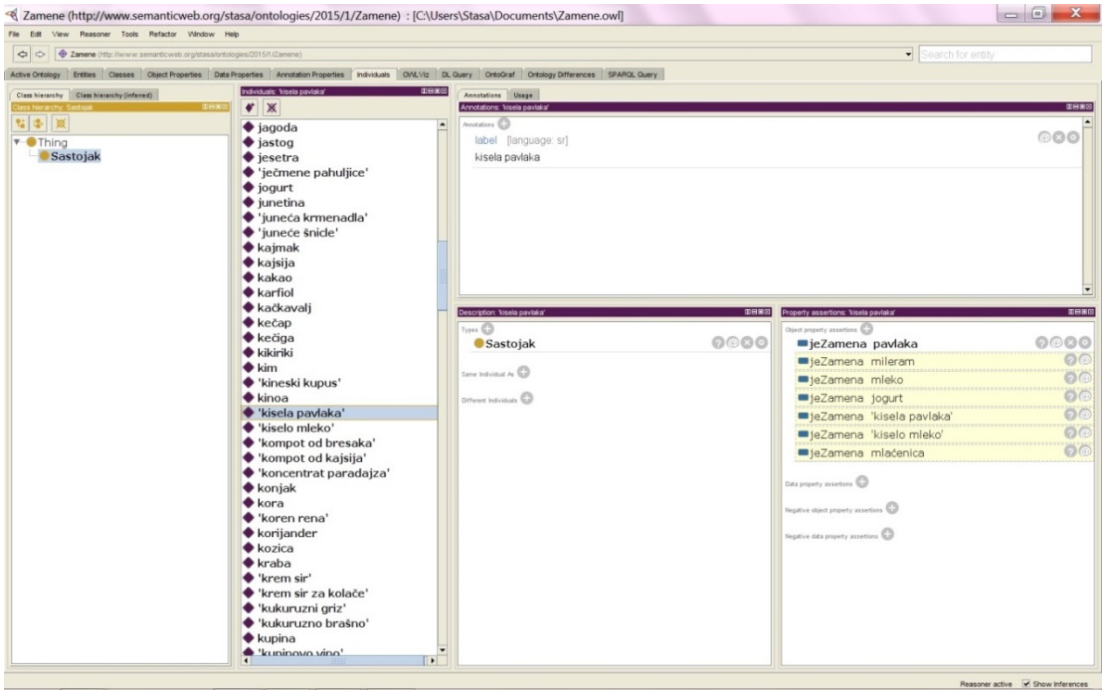
4.4.4 Креирање доменске онтологије замена састојака у кулинарском домену

Анализом састојака који се употребљавају приликом припреме рецепата уочено је да се поједини састојци појављују као међусобне замене (пример 27). Под заменама у овом случају нису сматрани састојци чији су називи синоними, већ они састојци који припадају различитим концептима али је у самом тексту рецепта наведено да се може употребити или један или други. За поједине парове састојака важи да могу увек да буду међусобне замене (на пример, *Нутела* и *Еурокрем*), док за друге важи да некад могу да буду међусобне замене, али некад и не (на пример, *расол* и *лимунов сок* могу да буду међусобне замене приликом припреме киселе чорбе, али би додавање расола у торту од лимуна уместо сока од лимуна, покварило торту). Овакви састојци, који су у тексту били записани као *састојак1 или састојак2*, односно *састојак1 (или састојак2),* полуаутоматски су организовани у доменску онтологију замена.

Прво су у систему Unitex развијени коначни трансдуктори за препознавање описа састојака за припрему облика *састојак1 или састојак2* и *састојак1 (или састојак2)*. Употребом ових трансдуктора из кулинарског корпуса је издвојена листа од 1.279 кандидата за замене. Прегледањем листе је издвојено 266 састојака који граде 183 пара замена.

Онтологија чија вршна класа *Sastojak* има 266 инстанци направљена је у OWL језику употребом алата Protégé 4.3. У њој је уведена особина *jeZamena* и повезани су парови инстанци који су међусобне замене.

ПРИМЕР 27. Састојак *кисела павлака* је у онтологији замена означен као замена за састојке *павлака*, *милерам*, *млеко*, *јогурт*, *кисело млеко*, *млађеница* (слика 29).



Слика 29. Приказ инстанце састојка кисела павлака и његових замена.

Употреба креиране онтологије замена састојака за процењивање сличности већ написаних рецепата биће приказана у наставку у поглављу 4.5.4. Поред тога, постоји могућност њене употребе у системима за саветовање корисника о потенцијалним заменама састојака.

4.5 Примене модела на систем за претрагу рецепата

4.5.1 Проширивање упита употребом онтологије

Аутоматска обрада садржаја на вебу везаних за кулинарски домен, поред представљања знања, треба да обухвата и различите врсте резоновања које би омогућиле ефикасно претраживање рецепата и постављање комплексних упита. Тако би осим могућности претраге рецепата на основу састојака које садрже, категорији јела, начина припреме или времена припреме (што је и омогућено на већини сајтова са којих су рецепти прикупљани), било корисно омогућити и проналажење сличних рецепата, успостављање веза између њих или претрагу не само на основу кључних речи, већ и њихових синонима, различитих облика речи и слично.

Са тим циљем развијен је систем за претраживање кулинарских рецепата, *ReceptiX*, у коме су унапређене и допуњене наведене могућности претраге кулинарских садржаја применом предложеног модела екстракције информација вођене онтологијама. Почетна страна система ReceptiX за напредно претраживање рецепата приказана је на слици 30. Корисник може да унесе појам за претрагу и да пронађе све оне рецепте у којима се тај појам појављује у наслову рецепта.

Слика 30. Почетна страна система *ReceptiX* за напредно претраживање кулинарских садржаја.

Када корисник на почетној страни изабере опцију *Претрага по намирницама* пружа му се неколико различитих могућности за напредну претрагу по намирницама које се користе приликом припреме рецепта (слика 31).

Слика 31. Претрага рецепата по намирницама.

Прва могућност је да корисник у горње текст поље унесе неколико намирница које жели да рецепт садржи и да кликне на дугме *Тражи*.

Као излаз се очекује да буду излистани наслови свих рецепата који одговарају задатим критеријумима претраге, у форми веза ка комплетним текстовима рецепата (слика 32), како би корисник кликом на неки од њих могао да прегледа цео текст рецепта (слика 33).

Систем тада врши проширење наведених појмова за претрагу тако што их прво раздвоји и онда сваки појединачно обрађује кроз неколико фаза.

У првој фази се сваки појам за претрагу лематизује употребом развијених лексичких ресурса и коначних трансдуктора креираних у систему Unitex. Потом се консултовањем онтологије хране сваком од њих придружују одговарајући синоними.

У другој фази систем комуницира са веб сервисом VeBran[72] (Stanković, Obradović, Krstev i Vitas, 2011), који на основу морфолошких електронских речника и флективних граматика скупу речи који је формиран у претходној фази придружује све њихове флективне облике.

Овако формиран проширени скуп појмова за претрагу се користи да се пронађу сви они рецепти у чијој припреми се користе све наведене намирнице (пример 28).

[72] VeBran: http://hlt.rgf.bg.ac.rs/VeBran.

ПРИМЕР 28. За кључне речи *јаја*, *бутер*, *брашно* као резултат претраге се добија 579 резултата (слика 32). Међу резултатима се јављају рецепти у којима се ниска *бутер* не појављује, већ се као елемент експанзије упита јављају њени синоними из синсета у WordNet-у *путер*, *маслац* или *масло*.

На слици 33 је приказан један од пронађених рецепата. У овом рецепту се међу потребним састојцима дословно јавља појам за претрагу *јаја*, али се појам за претрагу *брашно* не појављује дословно, већ се јавља његов флективни облик *брашна*. Поред тога се међу састојцима не појављује *бутер*, већ се наводи да је за припрему потребно *440гр маслаца*, што је флективни облик синонима *маслац*.

Слика 32. Резултати претраге рецепата за кључне речи *јаја*, *бутер*, *брашно*.

ReceptiX

Mamino orahova torta

Kategorija: Deserti

Težina pripreme: srednje

Sastojci:
12 kom. jaja
500 gr šećera
500 gr mlevenih oraha
4 kašike brašna
2 kesice vanilin šećera
6 kom. jaja
300 gr šećera
2 kesice vanilin šećera
6 rebara čokolade za kuvanje
440 gr maslaca (ili 500 gr margarina)
malo ruma

Priprema:
Ispeći dve kore od po: 6 žumanaca umutiti sa 250 gr šećera, dodati sneg od 6 belanaca, 250 gr mlevenih oraha, 2 kašike brašna i jedan vanilin šećer. Kore peći na četvrtastom plehu obloženom papirom za pečenje. Kada se kore ohlade, odseći od svake trećinu po dužini čime se dobija treća kora koju treba staviti u sredinu torte. Za fil skuvati na pari 6 celih jaja sa 300 gr šećera i dva vanilin šećera. Kad se masa zgusne dodati čokoladu i na kraju malo ruma, po ukusu. U ohlađeno umešati prethodno umućen maslac. Ja sam fil kuvala u šerpi sa debelim dnom direktno na ringli, stalno mešajući i time značajno skratila vreme kuvanja. Tortu premazati iznutra i spolja istim filom i obilno posuti mlevenim orasima. Ukrasiti polovinama oraha.

Слика 33. Пример рецепта добијеног претрагом *јаја, бутер, брашно*. Кључна реч *бутер* је у тексту рецепта замењена синонимом маслац.

Када корисник обележи опцију *Појединачно* поред горњег текст поља, фазе проширења појмова за претрагу се одвијају као у раније описаном случају, али се не траже они текстови рецепата који задовољавају да се у њима јављају сви наведени појмови за претрагу, већ је услов ослабљен и траже се они рецепти у којима се јавља бар један од њих.

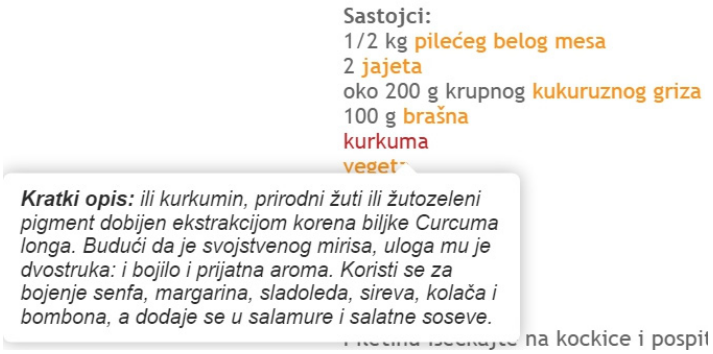
Следећа могућност је да корисник додатно у доњем текст пољу наведе које намирнице не жели да користи у припреми. У овом случају се као излаз очекује да буду излистани наслови свих рецепата који одговарају задатим критеријумима претраге, као везе ка комплетним текстовима рецепата, али се осим намирница које рецепти који буду враћени као одговарајући треба да садрже узимају у обзир и оне намирнице које рецепти не треба да садрже. Појмови за претрагу наведени у овом текст пољу обрађују се на раније описан

начин и користе за елиминацију резултујућих рецепата у којима се јављају кључне речи из првог текст поља (пример 29).

> ПРИМЕР 29. Резултат претраге за кључне речи *јаја*, *бутер*, *брашно* (са обележеном опцијом *Појединачно*) је 6.400 рецепата. У сваком од ових рецепата се јавља најмање један од наведених појмова за претрагу.
>
> Ако корисник наведе да жели да употреби *јаја*, *бутер*, *брашно*, али нема *чоколаду*, добиће као резултат претраге 457 рецепата.

На слици 33 се види да је пронађени рецепт означен категоријом *Десерти*, а да је тежина припреме у овом случају *средње*. Додатне информације овог типа су преузете приликом формирања корпуса. Корисницима је омогућено да кликом на категорију тренутног рецепта добију преглед свих оних рецепата који су означени том категоријом.

Сваки од састојака који се јављају у рецептима повезан је са одговарајућим описима. Када корисник на страни рецепта задржи показивач на састојку чији опис жели да види, приказује се кратак опис (слика 34), а када кликне на састојак отвара се целокупан опис који се састоји из неколико делова (слика 35).



Слика 34. Кратак опис састојка рецепта.

У првом делу је за опис састојка употребљено знање из WordNet-а. Прво су у WordNet-у тражени синсетови који одговарају састојку који се описује. С обзиром да за поједине састојке постоји неколико синсетова код којих се као литерал синсета дословно или као део литерала појављује тражени састојак,

али припадају доменима различитим од гастрономије (на пример, домену хемије или домену ботанике), предност су имали они синсетови који припадају домену гастрономије. За опис појма је из одабраног синсета преузета дефиниција (*Објашњење*). Везе ка осталим пронађеним синсетовима, који садрже сродне појмове или припадају другим доменима, наведене су у наставку овог дела (*Погледајте још*). За поједине састојке нису пронађени синсетови који припадају домену гастрономије, па су они означени поруком „*У WordNet-у не постоји објашњење за тражени појам у домену гастрономије*“ и дата је могућност да корисник може да допуни објашњење за тај појам преусмеравањем на страну „Семантички ресурси српског језика“[73] (Mladenović, Mitrović i Krstev, 2014).

У другом делу је за опис састојка употребљено знање из кулинарског речника. Кулинарски речник је још један од доменски специфичних ресурса чија је изградња започета током рада на овој дисертацији. У њему су прикупљени описи 2.669 гастрономских појмова из различитих извора какав је портал Coolinarika или гастрономска терминологија представљена у (Portić, 2011a). Током његовог даљег развоја биће укључени стручњаци из области гастрономије. У овом случају се кориснику поред описа приказују сродни појмови из кулинарског речника, као и преводи на стране језике уколико постоје. Када у кулинарском речнику не постоји објашњење за разматрани појам кориснику се приказује одговарајућа порука.

Трећи део описа обухвата везе ка чланцима на порталу *Wikipedia*[74] о појму који се описује. Поред српског језика, приказане су везе ка чланцима на француском, енглеском, немачком, руском и хрватском језику, као и одговарајући преводи појма на те језике у случајевима када чланци постоје. У случају да не постоји тражени чланак на српском језику, кориснику се отвара страница за прављење новог чланка.

[73] Семантички ресурси српског језика: http://sm.jerteh.rs/.
[74] Wikipedia: http://www.wikipedia.org/.

Слика 35. Опис састојка рецепта.

4.5.2 Примена онтологије приближних мера за конвертовање

У систему ReceptiX за напредно претраживање рецепата применом онтологије приближних мера омогућено је да приближне мере буду конвертоване у стандардне мере у тексту рецепта где се појављују. Тако би количина изражена у кашичицама или шољицама (слика 36) била замењена одговарајућом количином намирнице изражене у грамима или мililитрима (у зависности да ли се ради о течности или не).

Слика 36. Текст рецепта пре конверзије приближних кулинарских мера.

Оваква врста конвертовања није увек једнозначна нити могућа, а ни потребна. Примера ради, кесица пудинга од ваниле различитих произвођача има тежину од 34 до 42 грама. За неке приближне мере, попут *струк*, *веза*, *чешањ*, *главица* и сличне, тешко је установити на коју се тачно количину мисли (изражену у некој од стандардних мера). Због тога је у систем за напредно претраживање имплементирано конвертовање само најчешћих приближних мера којима се описују количине основних намирница (као што су *шећер*, *брашно*, *уље* и сл.). Ове мере, попут *шоље*, *шољице*, *кашике*, *кашичице*, *чаше* и сличних, су инстанце класе *Kontejner* у оквиру онтологије приближних мера. Све могу бити изражене у јединицама запремине.

За потребе овог система, онтологија приближних мера је додатно проширена тако што је свакој инстанци додато својство типа података *Vrednost* које приказује колико јединичних запремина (у cm^3) има дата приближна мера. Уколико се мере користе за течности (попут *уља*, *млека* и сл.) ова информација је довољна за превођење приближне мере у мililитре. Међутим, за намирнице попут *брашна*, *шећера* или *соли*, чије се количине

стандардно изражавају у јединицама за тежину, потребно је у онтологије уврстити и информацију о тежини јединичне запремине појединих намирница, како би било омогућено превођење. То је учињено проширивањем онтологије хране, тако што је инстанцама које се односе на основне градивне намирнице додељено својство *STezina* које садржи податак о специфичној тежини намирнице по јединици запремине израженој у cm^3.

На рецепт у коме се врши конверзија прво се примењује коначни трансдуктор за препознавање приближне мере, који обележава препознати део текста описан у поглављу 2.5.2.3 (слика 12 на страни 84). Додатно је креиран модул за прерачунавање мера који на основу обележених података и података садржаних у онтологијама (*Vrednost* и *STezina*) прерачунава количину намирнице изражену у некој од стандардних мера. Уколико намирница садржи вредност особине *STezina* сматра се да је у питању чврста супстанца, а у противном се сматра да се ради о течности. Израчуната вредност бива уписана у текст рецепта поред препознате приближне мере (слика 37).

Слика 37. Део текста рецепта после извршене конверзије приближних кулинарских мера у стандардне мере.

Подаци о специфичној тежини намирница по јединици запремине израженој у cm^3, који су употребљени приликом почетне имплементације, преузети су са различитих веб страна, али за целокупну и валидну допуну података морају да буду консултовани стручњаци који се баве овим

проблемом. Поред тога, увек мора да се узме у обзир и природа намирница које се мере с обзиром да у зависности од типа исте врсте намирнице може да дође до варијација у измереној количини (на пример, треба узети у обзир разлике приликом мерења шољице шећера у праху или шољице кристал шећера различите крупноће). На крају, треба нагласити да различите шоље, шољице, кашике итд. које се користе код нас у рецептима не морају да буду једнаке запремине јер се не користе стандардизоване посуде, па самим тим и поменуте конверзије треба узети са резервом.

4.5.3 Успостављање веза између рецепата

Веза између рецепата се успоставља у случајевима када корисник прегледа текст рецепта и уочи да се у његовој припреми појављује део за чију припрему потенцијално постоји рецепт са додатним или другачијим објашњењем. У систему се испод текста рецепта налази дугме *ReceptiX* којим се омогућава повезивање рецепата овог типа. Када корисник притисне ово дугме у тексту припреме се обележавају ниске које одговарају насловима рецепата из кулинарског корпуса. Ове ниске постају везе ка другим рецептима чиме се омогућава њихово прегледање (пример 30).

ПРИМЕР 30. На слици 38 је приказан пример рецепта. Када корисник притисне дугме *ReceptiX*, у тексту припреме се обележавају ниске које одговарају насловима рецепата из кулинарског корпуса. Ове ниске постају везе ка другим рецептима, као што је случај са нискама *бешамел сос* и *мусака* на слици 39, којима се приступа другим рецептима за припрему бешамел соса и мусаке.

Слика 38. Текст рецепта пре означавања веза ка другим рецептима.

Слика 39. Текст рецепта са означеним везама ка другим рецептима.

Да би било постигнуто ово повезивање прво је у систему Unitex направљен коначни трансдуктор за лематизацију текстова описан у поглављу 2.5.2.2 (слика 9) који сваку препознату реч текста који се обрађује замењује њеном лемом, док непрепознате речи преписује у изворном облику (какав је случај, на пример, са погрешно записаном речју *растопјеном*). С обзиром да у морфолошким електронским речницима једном облику речи може да одговара већи број лема, ова замена није једнозначна (на пример, реч *додати* препозната као придев замењује се лемом *додат*, а препозната као глагол лемом *додати*). Међутим, како се у таквим случајевима реч увек замењује првом лемом која је наведена у електронским речницима, замена ће сваки пут бити иста иако не нужно тачна. За даље кораке у повезивању рецепата ова апроксимација решења је у реду.

Применом коначног трансдуктора за лематизацију се за рецепт који се обрађује прави одговарајући лематизован текст припреме, а за све рецепте лематизовани наслови. Добијени лематизовани текст и сваки од добијених лематизованих наслова додатно се обрађују тако што се уклањају појаве карактера „-“ и вишеструких белина насталих обрадом у систему Unitex. Ови карактери се уклањају и из оригиналног текста припреме и оригиналних наслова.

У следећем кораку се обрађује један по један наслов рецепта тако што се проверава да ли се његов лематизовани облик јавља као подниска лематизованог текста припреме рецепта који се обрађује. Ако се лематизовани облик појављује, проверава се на којој позицији се подниска појављује, а онда се на основу броја размака пре те позиције израчунава који редни број речи у лематизованом тексту припреме има прва реч пронађеног наслова. С обзиром да оригинални и лематизовани облици текстова припреме и наслови у општем случају немају једнак број карактера, овај податак се користи да би се у оригиналном тексту одредио редни број речи где се појављује прва реч наслова. На основу овог податка и податка о дужини пронађеног наслова у оригиналном тексту припреме се појава наслова означава као веза ка одговарајућем рецепту коме тај наслов одговара (пример 31).

ПРИМЕР 31. За пример са слике 38 лематизовани наслов рецепта је *бешамел мусака.* Лематизовани текст припреме је:

> *макарони обарити. лук издинстати са млевено месо и зачинити додатак за јело. бешамел-сос направити на следећи начин: упржити 2 кашика брашно на растопјеном путер, додат 2 дл млеко и кад проврети и добити одређен густина, додат парче сир за топљење, жуманце и мало сенф. качкаваљ изрендати. у поуљен плести ставити бешамел, затим половина припремљен макарони, месо, рендан качкаваљ, преостали макарони, прелити*

бешамел и преостали рендан качкаваљ. мусака ставити у рерна и пећи на 200 степен око 20 минут. кад се прохладити мусака исећи на коцка и послужити топао јести се моћи јести и хладно.

а после накнадне обраде лематизовани текст је:

макарони обарити. лук издинстати са млевено месо и зачинити додатак за јело. бешамел сос направити на следећи начин: упржити 2 кашика брашно на растопјеном путер, додат 2 дл млеко и кад проврети и добити одређен густина, додат парче сир за топљење, жуманце и мало сенф. качкаваљ изрендати. у поуљен плести ставити бешамел, затим половина припремљен макарони, месо, рендан качкаваљ, преостали макарони, прелити бешамел и преостали рендан качкаваљ. мусака ставити у рерна и пећи на 200 степен око 20 минут. кад се прохладити мусака исећи на коцка и послужити топао јести се моћи јести и хладно.

где су појаве подниски лематизованих наслова рецепата *бешамел сос* и *мусака* добијене накнадном обрадом приказане подвучено. Редни бројеви почетних речи лематизованих наслова у лематизованом тексту припреме су 13, 68 и 83. На истим позицијама се налазе означени наслови са придруженим везама у оригиналном тексту припреме (слика 39).

4.5.4 Сличност рецепата

У овом раду се сличност рецепата дефинише као сличност која постоји у описима начина припреме различитих јела у рецептима кулинарског корпуса, при чему се раздваја алгоритам, односно начин припреме рецепата, од података, односно састојака који учествују у припреми. За израчунавање сличности рецепата кулинарског корпуса коришћене су различите мере какве су Еуклидова, Косинусна, Жакарова или Канбера (Lance i Williams, 1967;

Manning, Raghavan i Schütze, 2008). У наставку ће бити описани резултати добијени применом Жакарове мере сличности и Канбера мере удаљености.

За израчунавање Жакарове мере сличности написан је Java програм који обухвата и претходну обраду докумената корпуса и генерисање „индекса“. Претходна обрада докумената рецепата из кулинарског корпуса обухвата обраду у систему Unitex посебно креираним коначним трансдукторима којима се из текстова припреме уклањају све именице и функционалне речи какве су заменице, прилози, предлози или везници и задржавају само глаголи који се уједно лематизују. На овај начин је сваки текст рецепта из кулинарског корпуса замењен одговарајућим текстуалним документом који садржи одговарајућу *врећу речи* (енгл. *bag of words*) – листу термина који представљају текст документа. У овом случају сви термини који представљају текстове су лематизовани глаголи.

Нека важи да корпус садржи p докумената $d_1, d_2, \dots, d_p$ у којима се појављује q термина $t_1, t_2, \dots, t_q$. Следећи корак је генерисање „индекса“ A – матрице чији је сваки елемент $a_{i,j}$ вредност тежине термина t_i у документу d_j, где важи да редови одговарају терминима тј. $i \in \{1, \dots, q\}$, а колоне документима тј. $j \in \{1, \dots, p\}$. Постоје различити приступи за израчунавање тежина термина, као што је комбиновање мера TF – *функција фреквенције термина* (енгл. *term frequency*) и IDF – *инверзна функција фреквенције докумената у којима се тај термин појављује* (енгл. *inverse document frequency*) (Konchady, 2006; Wu, Luk, Wong i Kwok, 2008; Leskovec, Rajaraman i Ullman, 2014) за генерисање матрице A као:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i$$

где сваки елемент $a_{i,j}$ матрице A добија вредност $(tf - idf)_{i,j}$.

Функција фреквенције термина се обично нормализује како би се избегло давање предности дужим документима у којима неки термин може да има већу фреквенцију без обзира на његов значај. Поред тога, при генерисању матрице A, то јест израчунавању тежина термина, треба водити рачуна да се ретким терминима не придаје мањи значај него честим, као и да се терминима

који се вишеструко појављују не придаје мањи значај него онима који се појављују један пут. Отуд се јављају различите варијације за њихово израчунавање од којих су неке уобичајене приказане у табели 7 (Manning, Raghavan i Schütze, 2008).

Табела 7. Варијанте $tf - idf$.

TF – $tf_{i,j}$	**IDF –** idf_i	**Нормализација**
$n_{i,j}$ * (природна)	1 (не)	1 (без)
$1 + \log n_{i,j}$ (логаритамска)	$\log \frac{\lvert D \rvert}{\lvert\{d : t_i \in d\}\rvert}$ ** (инверзна фреквенција докумената)	$\frac{1}{\sum_k n_{k,j}}$ ***
$0{,}5 + \frac{0{,}5 \times n_{i,j}}{\max_i n_{i,j}}$ (проширена)	$\max\{0, \log \frac{\lvert D \rvert - \lvert\{d : t_i \in d\}\rvert}{\lvert\{d : t_i \in d\}\rvert}\}$ (вероватносни инверз фреквенција докумената)	$\frac{1}{\sqrt{\sum_j n_{k,j}^2}}$
$\begin{cases} 1, \text{ако је } n_{i,j} > 0 \\ 0, \text{иначе} \end{cases}$ (бинарна)		

* $n_{i,j}$ означава број појављивања термина t_i у документу d_j

** $|D|$ је укупан број докумената у корпусу,
а $|\{d : t_i \in d\}|$ укупан број докумената у којима се термин t_i појављује

*** сума броја појављивања свих термина у документу d_j

У Java имплементацији у овом раду се $tf_{i,j}$ израчунава као природна функција, тј. број појављивања термина t_i у документу d_j нормализована сумом броја појављивања свих термина у документу d_j, док се idf_i израчунава као инверзна фреквенција докумената, тј. $\log \frac{|D|}{|\{d : t_i \in d\}|}$.

За израчунавање сличности између два документа d_i и d_j употребљена је генерисана матрица A и *Жакарова мера сличности* скупова D_i и D_j која се рачуна по формули:

$$J(d_i, d_j) = J(D_i, D_j) = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}$$

као однос величине пресека скупова D_i и D_j и величине њихове уније, где D_i и D_j одговарају редом колонама докумената d_i и d_j матрице A. Сличност се израчунава за све парове докумената, односно колона матрице, при чему важи да је за две колоне d_i и d_j: $J(d_i, d_j) = J(d_j, d_i)$ и $J(d_i, d_i) = 1$.

Важи да је:

$$J(d_i, d_j) = J(D_i, D_j) = \frac{|D_i \cap D_j|}{|D_i| + |D_j| - |D_i \cap D_j|}$$

где је $|D_i \cap D_j| = \sum_{s=1}^{q} \min(a_{si}, a_{sj})$, $|D_i| = \sum_{s=1}^{q} \min(a_{si})$, $|D_j| = \sum_{s=1}^{q} \min(a_{sj})$, q је број редова (термина), а $i, j \in \{1, \dots, p\}$, при чему је p број колона (докумената) у генерисаној матрици A.

Број текстова p у кулинарском корпусу који су коришћени за мерење сличности је 11.670, а број термина q који се у њима јављају је 6.915.

Анализом добијених резултата направљена је хијерархија парова сличних рецепата од три нивоа. Један рецепт може да учествује у већем броју парова, али један пар рецепата припада тачно једном нивоу хијерархије. На првом нивоу су парови рецепата код којих је мера сличности $0 \leq J(d_i, d_j) < 0{,}75$. Такви рецепти су у попуности различити. На трећем нивоу су парови рецепата код којих је мера сличности $0{,}9 < J(d_i, d_j) \leq 1$. То су рецепти чији текстови се понављају на истом сајту са различитим насловима или се понављају на различитим сајтовима са истим или различитим насловима. Из корпуса је уклоњено укупно 173 рецепта који су дупликати овог типа. У појединим случајевима се међу овим рецептима јављају минималне измене у текстовима припреме где аутор на другачији начин формулише један или два корака припреме.

Средњи ниво хијерархије сличности рецепата након уклањања дупликата чини 529 парова рецепата сличне припреме код којих важи да је $0{,}75 \leq J(d_i, d_j) < 0{,}9$. Међу овим рецептима важи да један рецепт може да буде

сличан већем броју рецепата и сви такви парови се броје. Овај ниво хијерархије може додатно да се профини поделом у две подгрупе – подгрупу у којој се рецепти слично припремају употребом различитих састојака и подгрупу у којој је припрема описана на сличан начин, али су састојци пописани у једном рецепту синоними или замене састојака који се јављају у другом рецепту. Додатно поређење се прави на основу састојака ових рецепата употребом лексичких ресурса тако што се проверава да ли се у паровима рецепата користе састојци који су истоветни, синоними су или су међусобне замене.

У додатном поређењу се прво екстракцијом састојака за парове рецепата праве два скупа састојака S_1 и S_2 који се употребљевају у њиховој припреми. Састојци у скуповима се при екстракцији лематизују. У овом кораку се примењују лексички ресурси, електронски морфолошки речници и коначни трансдуктори. Од њиховог степена развијености зависи да ли ће сви састојци исправно бити екстраховани. Тако је у примеру 32 екстрахован састојак *сода*, уместо *сода бикарбона*, у првом скупу, и *шећер*, уместо *ванил шећер*, у другом. Препознате грешке се користе за поновну допуну лексичких ресурса.

Да би се извршио следећи корак у утврђивању сличности састојака који се употребљавају у припреми рецепата користи се онтологија замена састојака описана у поглављу 4.4.4.

На састојке који припадају скупу $S_2 \setminus S_1$ примењују се онтологија замена и WordNet и проверава се да ли међу њиховим синонимима и састојцима који су њихове замене има састојака који припадају скупу $S_1 \setminus S_2$. Ако се за синоним или замену неког састојка из скупа $S_2 \setminus S_1$ установи да припада скупу $S_1 \setminus S_2$, тај састојак се замењује пронађеним синонимом или заменом. На тај начин се формира скуп S_2'. Тако су у примеру 32 састојци *кисела павлака* и *путер* из другог скупа замењени редом састојцима *јогурт* и *уље*.

На крају се рачуна Жакарова мера сличности скупова S_1 и S_2'. Скупови су мање слични што је мера ближа 0 и у том случају се рецепти сврставају у прву подгрупу, а код мера које су ближе 1 се рецепти сврставају у другу подгрупу.

За рачунање *Канбера раздаљине* (енгл. *Canberra distance*) описа припреме јела у рецептима почетног кулинарског корпуса од 11.670 рецепата

употребљен је пакет `stylo` (Eder, Kestemont i Rybicki, 2013) софтверског алата *R*[75]. Канбера раздаљина се дефинише као:

$$C(d_i, d_j) = \sum_{k=1}^{q} \frac{|tf_{k,i} - tf_{k,j}|}{|tf_{k,i}| + |tf_{k,j}|}$$

где је $tf_{k,i}$ фреквенција термина t_k у тексту d_i, а $tf_{k,j}$ фреквенција термина t_k у тексту d_j.

Анализом добијених резултата направљена је хијерархија сличности парова рецепата од три нивоа на сличан начин како је описано код Жакарове мере, поштујући границе својствене Канбера мери раздаљине. Према овим резултатима препознато је 157 рецепата који су дупликати других рецепата. Сви такви рецепти су пронађени и употребом Жакарове мере сличности. Након уклањања дуплих рецепата, пронађено је 456 парова сличних рецепата, где такође важи да један рецепт може да буде сличан већем броју рецепата и сви такви парови се броје.

У систему ReceptiX су испод сваког рецепта приказане везе ка сличним рецептима пронађене употребом описаних мера поређане опадајуће према сличности (пример 32).

ПРИМЕР 32. На сликама 40 и 41 су приказани рецепти који припадају трећем нивоу сличности, јер се њихови описи припреме разликују само у последњем кораку. Припрема рецепта са слике 41 допуњена је у односу на припрему рецепта са слике 40 реченицом *Мафине прелити отопљеном чоколадом или посути шећером у праху*, док су сви остали кораци (и састојци) истоветни.

***

Рецепти приказани на сликама 40 и 42 припадају другом нивоу у хијерархији сличности, јер се њихови начини припреме разликују. На њих је додатно примењено мерење различитости састојака који се

[75] The R Project for Statistical Computing: http://www.r-project.org/.

употребљавају у њиховој припреми (слика 43). $S_1 = \{$*јаје*, *шећер*, *брашно*, *прашак за пециво*, *вишња*, *јогурт*, *сода*, *уље*$\}$, $S_2 = \{$*јаје*, *шећер*, *прашак за пециво*, *кисела павлака*, *путер*, *брашно*, *вишња*, *компот*$\}$. Састојци *кисела павлака* и *путер* из другог скупа замењени су састојцима *јогурт* и *уље*, па је $S_2' = \{$*јаје*, *шећер*, *прашак за пециво*, *јогурт*, *уље*, *брашно*, *вишња*, *компот*$\}$.

У овом случају је $J(S_1, S_2') = 0{,}875$, односно у рецептима се приликом припреме употребљавају слични састојци.

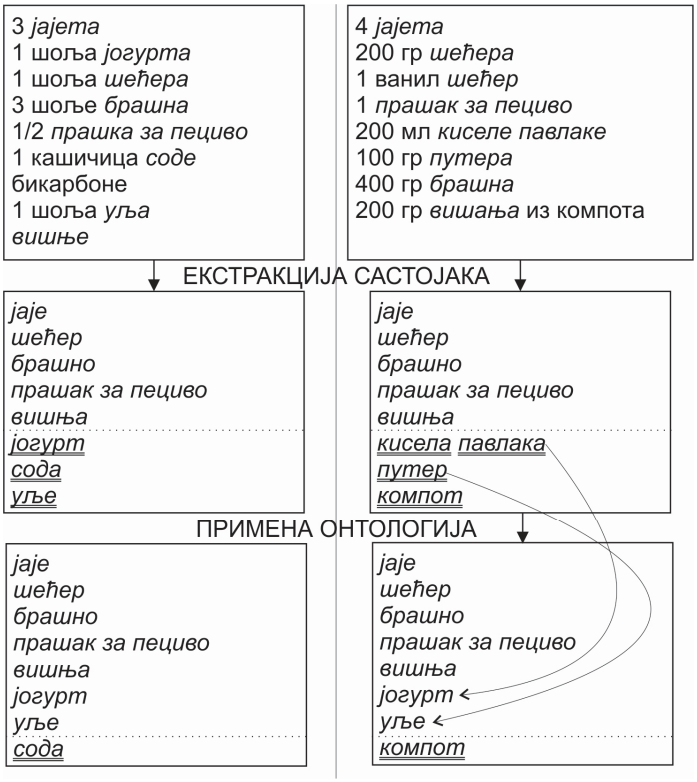
***

Рецепти приказани на сликама 40 и 44 припадају другом нивоу у хијерархији сличности мереној у односу на припрему и за њих је $S_1 = \{$*јаје*, *шећер*, *брашно*, *прашак за пециво*, *вишња*, *јогурт*, *сода*, *уље*$\}$, $S_2 = \{$*јаје*, *уље*, *јогурт*, *бело кукурузно брашно*, *пшенично брашно*, *прашак за пециво*, *сир*$\}$, а $S_2' = \{$*јаје*, *уље*, *јогурт*, *брашно*, *прашак за пециво*, *сир*$\}$. Сличност је у овом случају једнака $J(S_1, S_2') = 0{,}556$, па произилази да је због веће различитости између састојака који се употребљавају већа мера различитости између ових рецепата него што је било у претходном случају.

Слика 40. Пример рецепта са приказаним везама ка сличним рецептима.

Слика 41. Пример рецепта који се разликује од рецепта са слике 40 у последњем делу описа припреме.

Слика 42. Пример рецепта који је сличан рецепту приказаном на слици 40.



Слика 43. Сличност састојака у рецептима приказаним на сликама 41 и 42.

Слика 44. Пример рецепта који је сличан рецепту приказаном на слици 40.

Резултати мерења сличности рецепата зависе од покривености доменски специфичне лексике морфолошким речницима и онтологијама, као и начина на који су глаголи употребљавани приликом описа припреме рецепата.

5

Закључак и даљи рад

5.1 Закључак

Област екстракције информација, као подобласт области природних језика, зависи у великој мери од описа одређеног природног језика. Док је ова област веома развијена за језике попут енглеског, за словенске језике, а посебно за српски језик, још увек је у повоју.

Ова дисертација предлаже поступак екстракције информација за српски језик вођен правилима, а кроз развој одговарајућих лексичких ресурса у облику морфолошких и семантичких речника. Поред општег модела за екстракцију информација, рад се фокусирао на посебан домен – на домен кулинарства – који је занемарен у литератури из ове области. У том смислу је развијен информатички модел за екстракцију информација из комплексних описа кулинарских феномена, које одликује специфична доменска семантика и посебне класе именованих ентитета. У овом светлу, предложени модел омогућава екстракције релевантних информација из доменског корпуса и њихово опремање маркерима који дефинишу њихово значење. Полазни проблем је проширен увођењем онтологија које дефинишу доменске концепте, њихове особине и релације између њих, што је омогућило да се препознатим лексичким елементима доделе додатне информације које нису експлицитно садржане у анализираном тексту. Систем за екстракцију информација, структуриран на овај начин, захтевао је опис и формализацију лексичких и корпусних ресурса неопходних за изградњу адекватне репрезентације изучаваног домена.

Стога је као главни резултат дисертације предложени модел за решавање проблема екстракције информација вођене онтологијама, који предвиђа унапређење резултата екстракције информација интегрисањем језичких ресурса и алата, организован кроз следеће модуле – модул за формирање корпуса текстова, модул за дефинисање задатака екстракције информација, модул за изградњу коначних модела за екстракцију информација, модул за примену развијених коначних модела, модул за доградњу морфолошких електронских речника, модул за проширење WordNet-а и модул за изградњу нових онтологија. Оваква организација доприноси итеративном креирању и доградњи језичких ресурса и алата.

За развој информатичке подршке за обраду текстова кулинарског домена модел је имплементиран кроз посебан систем за екстракцију доменских информација и напредно претраживање рецепата. Модули система су реализовани формирањем кулинарског корпуса текстова на српском језику који је послужио за препознавања кулинарске лексике којом су проширени и допуњени WordNet и морфолошки електронски речници, као и за развој доменских онтологија. У систему корисник на основу корпуса кулинарских текстова добија одговоре на постављене упите о рецептима, наводећи називе јела или намирнице које жели да се у њима користе или не користе, не водећи рачуна о специфичностима српског језика на коме поставља упит, попут падежа или синонимије. Такође је имплементиран метод за успостављање веза између рецепата. Везе се успостављају када се у опису поступка припреме једног рецепта јави назив јела за чију припрему постоји посебан рецепт. Тада се кориснику омогућава да погледа додатни опис. Даље је дефинисана сличност рецепата као сличност текстова који описују поступак припреме рецепата и искоришћене развијене онтологије утврђивање оваквих сличности. Поред тога су онтологије примењене за решавање задатака конвертовања приближних кулинарских мера у стандардне мере.

5.2 Правци даљег рада

Описани резултати који су постигнути у кулинарском домену представљају тек почетак истраживања која се налазе на граници наука о обради природних језика, математике, лингвистике и гастрономије. Како је већ поменуто, област *кулинарска лингвистика,* која се налази на граници између гастрономије и лингвистике, у самом је зачетку. Такође, позната је употреба различитих врста израчунавања у гастрономији. Отуда из тачке гранања ове четири науке потиче нова научна област којој су професори Душко Витас и Дени Морел наденули назив *гастроматика*.

Темељи ове области постављени су радом на овој дисертацији, а треба наставити са њеним развојем даљом исцрпном систематизацијом и описом терминологије која се користи у непрофесионалним рецептима, али и у професионалним текстовима. Поред састојака и прибора потребно је систематизовати и описати терминологију која се односи на различите начине припреме. Даље, могуће је развијати различите апликације попут лексички контролисаних едитора за унос рецепата којима се анализира комплетност рецепта. Ови едитори ће се заснивати на моделима реченица које прецизно описују рецепте како не би били изостављени параметри попут атрибута којима се описује како треба припремити нешто (на пример, на колико степени) или коју тачно врсту намирнице треба употребити уколико у терминолошком систему намирница има подређене појмове (на пример, ако *зелена салата* има подређене појмове *ајзберг салата*, *ендивија* и *путерица* није довољно рећи зелена салата, него је неопходно тражити да буде прецизно одређено који од три подређена термина треба да буде употребљен).

Различите обраде и анализе текстова из кулинарског домена употребом развијених ресурса и алата могу да се употребе за развој система за означавање рецепата који су погодни за посебну исхрану људи каква је исхрана дијабетичара или вегетеријанаца. Могу да послуже стручњацима који се баве гастрономијом за различита културолошка испитивања или за медицинска истраживања. Резултати могу да укажу на неке специфичности наше кухиње или да се употребе за поређење са кухињама других народа.

Пример анализа које развијени ресурси и алати добијени применом представљеног модела омогућавају представљен је у прилогу Ђ.

Представљени модел за екстракцију информација вођену онтологијама могао би да се примени за развој области гастроматике и у језицима са већим бројем говорника, какви су енглески или француски, где ова област још нема адекватне темеље, што указује на значај рада у његовом развоју. Поред тога, правци даљег развоја обухватају и проширење на вишејезично окружење, имплементације у различитим доменима, али и реализације тих имплементација за различите врсте уређаја и њихову интеграцију у друге апликације.

ЛИТЕРАТУРА

Adelberg, B. (1998). NoDoSE – a Tool for Semi-Automatically Extracting Structured and Semistructured Data from Text Documents. *Proceedings of the 1998 ACM SIGMOD international Conference on Management of Data* (str. 283-294). New York, NY, USA: Association for Computing Machinery Press.

Aho, A. i Ullman, J. (1972). *The Theory of Parsing, Translation, and Compiling.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

AIMS Agricultural Information Management Standards. (2014). URL: http://aims.fao.org/standards/agrovoc/concept-scheme.

Anthropology of Food. (2015). URL: https://aof.revues.org/.

Appelt, D. i Israel, D. (1999). Introduction to Information Extraction Technology. *Tutorial at the Sixteenth International Joint Conference on Artificial Intelligence* (str. 1-41). Artificial Intelligence Center: Menlo Park, CA, USA. URL: http://www.ai.sri.com/~appelt/ie-tutorial/.

Appelt, D., Hobbs, J., Bear, J., Israel, D. i Tyson, M. (1993). FASTUS: A Finite-State Processor for Information Extraction from Real-World Text. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (str. 1172-1178). San Mateo, CA, USA: Morgan Kaufmann.

Ayuso, D., Boisen, S., Fox, H., Gish, H., Ingria, R. i Weischedel, R. (1992). BBN: Description of the PLUM System as Used for MUC-4. *Proceedings Fourth Message Understanding Conference (MUC-4)* (str. 177-185). San Fransisco, CA, USA: Morgan Kaufmann Publishers.

Baldwin, T. i Lui, M. (2010). Language Identification: the Long and the Short of the Matter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (str. 229-237). Stroudsburg, PA, USA: Association for Computational Linguistics.

Beisswanger, E., Lee, V., Kim, J.-J., Rebholz-Schuhmann, D., Splendiani, A. i Dameron, O. (2008). Gene Regulation Ontology (GRO): Design Principles and Use Cases.

EHealth Beyond the Horizon - Get IT There: Proceedings of MIE2008, the XXIst International Congress of the European Federation for Medical Informatics. 136, str. 9-14. Amsterdam: IOS Press.

Bentivogli, L. i Pianta, E. (2000). Looking for Lexical Gaps. *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000* (str. 663-669). Stuttgart: Institut für Maschinelle Sprachverarbeitung.

Bentivogli, L., Forner, P., Magnini, B. i Pianta, E. (2004). Revising the WordNet Domains Hierarchy: Semantics, Coverage and Balancing. *Proceedings of the Workshop on Multilingual Linguistic Resources* (str. 101-108). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bergamaschi, S., Guerra, F. i Vincini, M. (2005). *Critical Analysis of the Emerging Ontology Languages and Standards.* Modena, Italy: University of Modena and Reggio Emilia. URL: http://www.dbgroup.unimo.it/wisdom/deliverables/fase_1/d1r1.pdf

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biodiversity Heritage Library. (2015). URL: http://www.biodiversitylibrary.org.

Bliksen, K. (2013). *Babetina gozba.* (R. Kosović, Prev.) Beograd: Geopoetika.

Blocker, L. i Hill, J. (2007). *Culinary Math.* New York, USA: John Wiley & Sons.

Bober, P. P. (1999). *Art, Culture, and Cuisine.* Chicago, USA: University of Chicago Press.

Borgo, S. (2007). How Formal Ontology Can Help Civil Engineers. *Ontologies for Urban Development*, *61*, 37-45.

Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse.* Enschede, The Netherlands: Centre for Telematica and Information Technology, University of Twenty.

Bon Appétit. (2015). URL: http://www.bonappetit.com.

Boscarino, C., Koenderink, N., Nedović, V. i Top, J. (2014). Automatic Extraction of Ingredient's Substitutes. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (str. 559-564). New York City, NY, USA: Association for Computing Machinery Press.

Brickley, D. (1999). *Message to RDF Interest Group: "WordNet in RDF/XML: 50,000+ RDF class vocabulary".* URL: http://lists.w3.org/Archives/Public/www-rdf-interest/1999Dec/0002.

Brill, E. (1992). A Simple Rule-Based Part of Speech Tagger. *Proceedings of the Workshop on Speech and Natural Language* (str. 112-116). Stroudsburg, PA, USA: Association for Computational Linguistics.

Brillat-Savarin, J. A. (1848). *La Physiologie du Goût.* (G. d. Gonet, Ur.) URL: http://gallica.bnf.fr/ark:/12148/bpt6k1063697.

Brin, S. (1999). Extracting Patterns and Relations from the World Wide Web. *Selected papers from the International Workshop on The World Wide Web and Databases* (str. 172-183). London, UK: Springer-Verlag.

Califf, M. E. i Mooney, R. (1999). Relational Learning of Pattern-Match Rules for Information Extraction. *Proceedings of the Sixteenth National Conference on Artificial intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence* (str. 328-334). Orlando, FL, USA: AAAI Press.

Califf, M. E. i Mooney, R. (2003). Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction. *Journal of Machine Learning Research, 4*, 177-210.

Chang, C. H., Kayed, M., Girgis, M. i Shaalan, K. (2006). A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering, 18*(10), 1411-1428.

Chang, C.-H. i Kuo, S.-C. (2004). OLERA: Semisupervised Web-Data Extraction with Visual Support. *IEEE Intelligent Systems, 19*(6), 56-64.

Chang, C.-H. i Lui, S.-C. (2001). IEPAD: Information Extraction Based on Pattern Discovery. *Proceedings of the 10th International Conference on World Wide Web* (str. 681-688). New York City, NY, USA: Association for Computing Machinery Press.

Chevalier, M., Dansereau, J. i Poulin, G. (1978). *TAUM-METEO: Description du Système.* Montréal: Université de Montréal.

Chinchor, N. (1997). *MUC-7 Named Entity Task Definition*. URL: http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html.

Chinchor, N. i Marsh, E. (1998). *MUC-7 Information Extraction Task Definition (version 5.1).* URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.* New York, USA: Springer Science & Business Media.

Cocks, C. (1850). *Bordeaux ses Environs et ses Vins.* Bordeaux: Feret Fils.

Corcho, O., Fernández-López, M. i Gómez-Pérez, A. (2003). Methodologies, Tools and Languages for Building Ontologies. Where is their Meeting Point? *Data & Knowledge Engineering, 46*(1), 41-64.

Courtois, B. (1989). *Dictionnaire Electronique du LADL pour les Mots Simples du Français.* Paris: LADL, Université Paris 7.

Cowie, J. i Lehnert, W. (1996). Information Extraction. *Communication ACM, 39*(1), 80-91.

Crane, G. (Ur.). (2015). *Perseus Digital Library*. Tufts University. URL: http://www.perseus.tufts.edu.

Crescenzi, V., Mecca, G. i Merialdo, P. (2001). ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites. *Proceedings of the 27th Very Large Data Base Endowment Conference* (str. 109-118). San Fransisco, CA, USA: Morgan Kaufmann Publishers.

Cunningham, H., Maynard, D. i Bontcheva, K. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (str. 168-175). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I. i sar. (2011). *Text Processing with GATE (Version 6).* Sheffield, UK: The University of Sheffield, Department of Computer Science.
URL: http://tinyurl.com/gatebook.

De Luca, E. W., Eul, M. i Nurnberger, A. (2007). Converting EuroWordNet in OWL and Extending it With Domain Ontologies. *Proceedings of the Workshop on*

Lexical-Semantic and Ontological Resources. In conjunction with GLDV 2007 (str. 39-48). Stroudsburg, PA, USA: Association of Computational Linguistic.

Diemer, S., i Frobenius, M. (2013). When making pie, all ingredients must be chilled. Including you: Lexical, syntactic and interactive features in online discourse - a synchronic study of food blogs. U C. Gerhardt, M. Frobenius, i S. Ley, *Culinary Linguistics* (T. 10, str. 53-82). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Ding, Y., Lonsdale, D., Embley, D., Hepp, M. i Xu, L. (2007). Generating Ontologies via Language Components and Ontology Reuse. U *Natural Language Processing and Information Systems* (str. 131-142). Berlin Heidelberg: Springer.

DuCharme, B. (2013). *Learning SPARQL.* Sebastopol, CA, USA: O'Reilly Media, Inc.

Đurin, T. (2012). Rableov homo bibens i prevođenje pijanke. *Godišnjak Filozofskog fakulteta u Novom Sadu, Knjiga XXXVII*, 47-60.

Edelstein, S. (2008). *Managing Food and Nutrition Services: For the Culinary, Hospitality, and Nutrition Professions.* Sudbury, MA, USA: Jones & Bartlett Learning.

Eder, M., Kestemont, M. i Rybicki, J. (2013). Stylometry with R: a Suite of Tools. *Digital Humanities 2013: Conference Abstracts* (str. 487-489). Lincoln: University of Nebraska-Lincoln.

Epstein, B. J. (2009). What's Cooking: Translating Food. *Translation Journal, 13(3)*. Digital Online Journal. URL: http://translationjournal.net/journal/49cooking.htm.

Erjavec, T. i Fišer, D. (2006). Building Slovene WordNet. *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC* (str. 1678-1683). Paris: ELRA/ELDA. URL: http://www.lrec-conf.org/proceedings/lrec2006/.

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M. i Vitas, D. (2003). The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. *Proceedings of the Workshop on Morphological Processing of Slavic Languages* (str. 25-32). Budapest: 10th Conference of the European Chapter, EACL 2003.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Yates, A. (2005). Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence, 165*(1), 91-134.

Europeana. (2015). URL: http://www.europeana.eu/.

Fellbaum, C. (1998a). *WordNet.* Hoboken, NJ, USA: Wiley Online Library.

Fellbaum, C. (Ur.). (1998b). *WordNet: An Electronic Lexical Database.* Cambridge, MA, USA: MIT Press.

Fellbaum, C. (2010). WordNet. U R. Poli, M. Healy i A. Kameas (Urednici), *Theory and Applications of Ontology: Computer Applications* (str. 231-243). Netherlands: Springer.

Fensel, D., van Harmelen, F. i Horrocks, I. (2003). OIL & DAML+OIL: Ontology Languages for the Semantic Web. U J. Davies, D. Fensel i F. van Harmelen (Urednici), *Towards the Semantic Web: Ontology-Driven Knowledge Management* (str. 11-31). West Sussex, UK: John Wiley & Sons.

Fišer, D. (2009). *Izdelava slovenskega semantičnega leksikona z uporabo eno- in večjezičnih jezikovnih virov, doktorska disertacija.* Ljubljana: Univerza v Ljubljani.

Freitag, D. (1998). Information Extraction from HTML: Application of a General Machine Learning Approach. *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence* (str. 517-523). Wisconsin, USA: Madison.

Freyne, J. i Berkovsky, S. (2010). Intelligent Food Planning: Personalized Recipe Recommendation. *Proceedings of the 15th International Conference on Intelligent User Interfaces* (str. 321-324). New York City, NY, USA: Association for Computing Machinery Press.

Friburger, N. i Maurel, D. (2004). Finite-State Transducer Cascades to Extract Named Entities in Texts. *Theoretical Computer Science, 313*(1), 93-104.

Friedman, C., Kra, P. i Rzhetsky, A. (2002). Two Biomedical Sublanguages: a Description Based on the Theories of Zellig Harris. *Journal of Biomedical Informatics, 35*(4), 222-235.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. i Schneider, L. (2002). Sweetening Ontologies with DOLCE. *Proceedings of the 13th International Conference on*

Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web. (str. 166-181). London, UK: Springer-Verlag.

Gastronomica. (2015). URL: http://www.gastronomica.org.

Gerhardt, C., Frobenius, M., i Ley, S. (Urednici). (2013). *Culinary Linguistics: the Chef's Special* (T. 10). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Goldfarb, C. (1980). *Document Composition Facility Generalized Markup Language: Concepts and Design Guide, Form No. SH20-9188-0.* Colorado, USA: IBM Corporation, White Plains.

Gómez Pérez, A. i Corcho, O. (2002). Ontology Languages for the Semantic Web. *IEEE Intelligent Systems, 17*(1), 54-60.

Gómez Pérez, A., Fernández López, M. i Corcho Garcia, O. (2004a). Ontological Engineering. *Computing Reviews, 45*(8), 478-479.

Gómez Pérez, A., Fernández López, M. i Corcho Garcia, O. (2004b). *Ontological Engineering: with Examples from the Areas of Knowledge Management, e-Commerce and Semantic Web.* London, UK: Springer-Verlag.

Graovac, J. (2013). Wordnet-Based Serbian Text Categorization. *INFOtheca: Journal of Informatics and Librarianship, 14*(2), 2a-17a.

Grau, B. C., Horrocks, I., Kazakov, Y. i Sattler, U. (2007). Just the Right Amount: Extracting Modules from Ontologies. *Proceedings of the 16th International Conference on World Wide Web* (str. 717-726). New York City, NY, USA: Association for Computing Machinery Press.

Graves, A. i Gutierrez, C. (2006). Data Representations for WordNet: A Case for RDF. U P. Sojka, K.-S. Choi, C. Fellbaum i P. Vossen (Ur.), *Proceedings of the 3rd International WordNet Conference* (str. 165-169). Brno: Masaryk University.

Grishman, R. i Kittredge, R. (Urednici). (1986). *Analyzing Language in Restricted Domains: Sublanguage Description & Processing.* NJ, USA: Lawrence Erlbaum Associates, Inc., Publishers.

Grishman, R. i Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. *COLING, 96*, 466-471.

Gross, M. (1975). *Méthodes en Syntaxe: Régime des Constructions Complétives.* Paris: Hermann.

Gross, M. (1988). Methods and Tactics in the Construction of a Lexicon-Grammar. U T. L. Korea (Ur.), *Linguistics in the Morning Calm 2* (str. 177-197). Seoul: Hanshin Publishing Co.

Gross, M. (1989). The Use of Finite Automata in the Lexical Representation of Natural Language. U M. Gross i D. Perrin (Urednici), *Electronic Dictionaries and Automata in Computational Linguistics* (str. 34-50). Berlin Heidelberg: Springer.

Gross, M. (1993). Local Grammars and their Representation by Finite Automata. U M. Hoey (Ur.), *Data, Description, Discourse. Papers on the English Language in Honour of John McH Sinclair* (str. 26-38). London, UK: Harper-Collins Publishers.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition, 5*(2), 199-220.

Gruber, T. R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies, 43*(5), 907-928.

Guarino, N. (1998). Formal Ontology in Information Systems. *Proceedings of the First International Conference (Volume 46)* (str. 3-15). Amsterdam: IOS press.

Gucul-Milojević, S. (2010). Personal Names in Information Extraction. *INFOtheca: Journal of Informatics and Librarianship, 11*(1), 53a-63a.

Harabagiu, S. i Maiorano, S. (2000). Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC).* Paris: ELRA/ELDA. URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/347.pdf.

Harris, Z. S. (1960). *Structural Linguistics.* Chicago: University of Chicago Press.

Harris, Z. S. (1968). *Mathematical Structures of Language.* New York, USA: Willey.

Harris, Z. S. (1982). *A Grammar of English on Mathematical Principles.* New York, USA: Willey.

Harris, Z. S. (1991). *A Theory of Language and Information: a Mathematical Approach.* New York, USA: Clarendon Press, Oxford University Press.

Hašek, J. (2003). *Doživljaji dobrog vojnika Švejka u prvom svetskom ratu.* (S. Vinaver, Prev.) Beograd: Dereta.

Hemingvej, E. (1952). *Starac i more.* (S. Vinaver, Prev.) Beograd: Duga.

Hepple, M. (2000). Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based POS Taggers. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (str. 278-277). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hilbert, M. i López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science, 332*(6025), 60-65.

Hobbs, J., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M. i Tyson, M. (1997). FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *Finite-State Language Processing* (str. 383-406). Cambridge, MA, USA: The MIT Press.

Horridge, M., Knublauch, H., Rector, A., Stevens, R. i Wroe, C. (2004). *A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0.* Manchester, UK: University of Manchester.

Hu, H., Du, X., Liu, D. i Ouyang, J. H. (2006). Ontology Learning using WordNet Lexicon. U *Computational Methods* (str. 1249-1253). Dordrecht, Netherlands: Springer Science & Business Media.

Huang, X.-x. i Zhou, C.-l. (2007). An OWL-Based WordNet Lexical Ontology. *Journal of Zhejiang University, 8*(6), 864-870.

Huffman, S. (1995). Learning Information Extraction Patterns from Examples. U *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, Lecture Notes in Computer Science* (T. 1040, str. 246-260). Berlin Heidelberg: Springer.

IBM. (2014). *IBM's Watson Cognitive Cooking*. URL: http://www.ibm.com/smarterplanet/us/en/cognitivecooking/.

International Journal of Gastronomy and Food Science. (2015). URL: http://www.journals.elsevier.com/international-journal-of-gastronomy-and-food-science/.

ISO 24610-2:200. (2007). *ISO 24610-2:200, Language Resource Management – Feature Structures – Part 2: Feature System Declaration.* ISO/TC 37/SC 4.

ISO 8879:1986. (1986). *ISO 8879:1986, Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML).* ISO.

ISO/IEC JTC1/SC 18. (1990). *ISO/IEC JTC1/SC 18, Document Processing and Related Communication.* ISO.

Jacobs, P. i Rau, L. (1990). The GE NLToolset: a Software Foundation for Intelligent Text Processing. *Proceedings of the 13th Conference on Computational Linguistics (Volume 3)* (str. 373-375). Stroudsburg, PA: Association for Computational Linguistics.

Jones, T. (2008). *Culinary Calculations: Simplified Math for Culinary Professionals.* New York, USA: John Wiley & Sons.

Jurafsky, D. (2014). *The Language of Food: A Linguist Reads the Menu.* New York City, NY, USA: W.W. Norton & Company.

Jurafsky, D. i Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Cambridge, MA, USA: MIT Press.

JUS ISO 1087-2. (2005). *Terminološki rad – rečnik: Primene računara.* Beograd, Srbija: Institut za standardizaciju Srbije.

Kaplan, R. i Kay, M. (1981). Phonological Rules and Finite-State Transducers. *Proceeding of Linguistic Society of America Meeting Handbook, Fifty-Sixth Annual Meeting*, 27-30.

Karttunen, L. (1983). KIMMO: a General Morphological Processor. *Texas Linguistic Forum, 22*, 163-186.

Kim, J.-J. i Rebholz-Schuhmann, D. (2011). Improving the Extraction of Complex Regulatory Events from Scientific Text by Using Ontology-Based Inference. *Journal of Biomedical Semantics, Proceedings of the Fourth International Symposium on Semantic Mining in Biomedicine (Volume 2)* (str. 1-13). London: BioMed Central.

Kim, J.-T. i Moldovan, D. (1995). Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. *IEEE Transactions on Knowledge and Data Engineering*, 713-724.

Kleene, S. C. (1956). Representation of Events in Nerve Nets and Finite Automata. U C. E. Shannon i J. McCarthy (Urednici), *Automata Studies* (str. 3-42). Princeton, NJ, USA: Princeton University Press.

Koeva, S., Mihov, S. i Tinchev, T. (2004). Bulgarian Wordnet – Structure and Validation. *Romanian Journal of Information Science and Technology, 7*(1-2), 61-78.

Konchady, M. (2006). *Text Mining Application Programming.* Boston, MA, USA: Charles River Media.

Koskenniemi, K. (1983). Two-Level Model for Morphological Analysis. *Proceedings of the 8th International Joint Conference on Artificial Intelligence (Volume 2)* (str. 683-685). CA, USA: International Joint Conference on Artificial Intelligence. URL: http://ijcai.org/Past%20Proceedings/IJCAI-83-VOL-2/CONTENT/content.htm.

Krstev, C. (2008). *Processing of Serbian. Automata, Texts and Electronic.* Belgrade: University of Belgrade, Faculty of Philology.

Krstev, C. i Lazić, B. (2015). Glagoli u kuhinji i za stolom. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opisi primene 44/3*, 117-135.

Krstev, C., Đorđević, B., Antonić, S., Berček-Ivković, N., Zorica, Z., Crnogorac, V. i Macura, L. (2008). Kooperativni rad na dogradnji srpskog WordNeta. *INFOteka, časopis za digitalnu humanistiku, 9*(1-2), 57-75.

Krstev, C., Obradović, I. i Vitas, D. (2008). An Approach to the Development. *Southern Journal of Linguistics, Special Theme: South Slavic and Balkan Languages, 29*(1/2), 106-118.

Krstev, C., Obradović, I., Utvić, M. i Vitas, D. (2013). A System for Named Entity Recognition based on Local Grammars. *Journal of Logic and Computation, 24*(2), 473-489. doi:10.1093/logcom/exs079.

Krstev, C., Pavlović-Lažetić, G., Vitas, D. i Obradović, I. (2004). Using Textual and Lexical Resources in Developing Serbian Wordnet. *Romanian Journal of Information Science and Technology, 7*(1-2), 147-161.

Krstev, C., Stanković, R. i Vitas, D. (2010). A Description of Morphological Features of Serbian: a Revision using Feature System Declaration. *Proceedings of the*

7th International Conference on Language Resources and Evaluation (LREC 2010) (str. 816-819). Paris: ELRA/ELDA.

Krstev, C., Stanković, R., Obradović, I., Vitas, D. i Utvić, M. (2010). Automatic Construction of a Morphological Dictionary of Multi-Word Units. U H. Loftsson, E. Rögnvaldsson i S. Helgadóttir (Urednici), *Advances in Natural Language Processing, Lecture Notes in Computer Science* (T. 6233, str. 226-237). Berlin-Heidelberg: Springer-Verlag.

Krstev, C., Stanković, R., Vitas, D. i Obradović, I. (2006). WS4LR - a Worksation for Lexical Resources. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (str. 1692-1697). Paris: ELRA/ELDA. URL: http://www.lrec-conf.org/proceedings/lrec2006/.

Krstev, C., Vitas, D., Maurel, D. i Tran, M. (2005). Multilingual Ontology of Proper Names. U Z. Vetulani (Ur.), *Proceedings of the Second Language and Technology Conference* (str. 116-119). Poznań: Faculty of Mathematics and Computer Science of Adam Mickiewicz University.

Krstev, C., Vitas, D., Obradović, I. i Utvić, M. (2011). E-dictionaries and Finite-State Automata for the Recognition of Named Entities. *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing* (str. 48-56). Stroudsburg, PA, USA: Association for Computational Linguistics.

Krstev, C., Vujičić Stanković, S., & Vitas, D. (2014). Approximate Measures in the Culinary Domain: Ontology and Lexical Resources. U T. Erjavec, & J. Žganec Gros (Ur.), *Proceedings of the 9th Language Technologies Conference IS-LT 2014* (str. 38-43). Ljubljana, Slovenia: Institut "Jožef Stefan".

Krupka, G., Jacobs, P., Rau, L., Childs, L. i Sider, I. (1992). GE NLTOOLSET: Description of the System as Used for MUC-4. *Proceedings Fourth Message Understanding Conference (MUC-4)* (str. 177-185). San Fransisco, CA, USA: Morgan Kaufmann Publishers.

Kuhlins, S. i Tredwell, R. (2002). Toolkits for Generating Wrappers. U M. Aksit, M. Mezini i R. Unland (Urednici), *Objects, Components, Architectures, Services, and Applications for a Networked World, Lecture Notes in Computer Science* (T. 2591, str. 184-198). Berlin Heidelberg: Springer-Verlag.

Laender, A., Ribeiro-Neto, B. i da Silva, A. (2002). DEByE - Data Extraction by Example. *Data & Knowledge Engineering, 40*(2), 121-154.

Laender, A., Ribeiro-Neto, B., da Silva, A. i Teixeira, J. (2002). A Brief Survey of Web Data Extraction Tools. *ACM Sigmod Record*, 84-93.

Lance, G. i Williams, W. (1967). Mixed-Data Classificatory Programs I - Agglomerative Systems. *Australian Computer Journal, 1*(1), 15-20.

Langendoen, T. (1981). The Generative Capacity of Word-Formation Components. *Linguistic Inquiry, 12*(2), 320-322.

Lassila, O. i McGuinness, D. (2001). *The Role of Frame-Based Representation on the Semantic Web, Knowledge Systems Laboratory Report KSL-01-02.* Stanford, USA: Stanford University.

Leech, G. N. (1991). The State of the Art in Corpus Linguistics. U K. Aijmer i B. Altenberg (Urednici), *English Corpus Linguistics* (str. 8-29). London, UK: Longman.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E. i Soderland, S. (1994). Evaluating an Information Extraction System. *Journal of Integrated Computer-Aided Engineering, 1*(6), 453-472.

Lehrer, A. (2009). *Wine & Conversation.* New York: Oxford University Press.

Leskovec, J., Rajaraman, A. i Ullman, J. D. (2014). *Mining of Massive Datasets.* Cambridge, UK: Cambridge University Press.

Levi-Stros, K. (2008). *Mitologike knjiga 1 - Presno i pečeno.* (D. Udovički, Prev.) Novi Sad: Prometej.

Li, P. (2013). A Survey of Machine Translation Methods. *TELKOMNIKA Indonesian Journal of Electrical Engineering, 11*(12), 7125-7130.

Liddy, E., Jorgensen, C., Sibert, E. i Edmund, Y. (1993). A Sublanguage Approach to Natural Language Processing for an Expert System. *Information Processing & Management, 29*(5), 633-645.

Liddy, E., Symonenko, S. i Rowe, S. (2006). Sublanguage Analysis Applied to Trouble Tickets. *FLAIRS Conference* (str. 752-757). Orlando, FL, USA: AAAI Press.

Liu, L., Pu, C. i Han, W. (2000). XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources. *Proceedings of 16th International*

Conference on Data Engineering (str. 611-621). San Diego, CA: The Institute of Electrical and Electronics Engineers.

Liu, X., Zhang, S., Wei, F. i Zhou, M. (2011). Recognizing Named Entities in Tweets. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1*, 359-367.

Lonsdale, D., Embley, D., Ding, Y., Xu, L. i Hepp, M. (2010). Reusing Ontologies and Language Components for Ontology Generation. *Data & Knowledge Engineering, 69*(4), 318-330.

Magnini, B. i Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. *Proceedings of the 2nd Conference on Language Resources and Evaluation* (str. 1413-1418). Athens, Greece: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2000/.

Magnini, B., Strapparava, C., Pezzulo, G. i Gliozzo, A. (2002). Comparing Ontology-Based and Corpus-Based Domain Annotations in WordNet. *Proceedings of the First International WordNet Conference*, (str. 146-154). Mysore, India.

Manning, C. i Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA, USA: MIT press.

Manning, C., Raghavan, P. i Schütze, H. (2008). *Introduction to Information Retrieval (Volume 1)*. Cambridge, MA, USA: Cambridge University Press.

Marcus, J. (2013). *Culinary Nutrition: The Science and Practice of Healthy Cooking.* Waltham: Academic Press.

Marković, S. P. (1939). *Moj kuvar.* Beograd: Politika a.d.

Marković, S. P. (1956). *Veliki narodni kuvar.* Beograd: Narodna knjiga.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H. i Wilks, Y. (2001). Named Entity Recognition from Diverse Text Types. *Recent Advances in Natural Language Processing 2001 Conference*, (str. 257-274). Tzigov Chark, Bulgaria.

McCuinness, D. (2005). Ontologies Come of Age. U D. Fensel, J. Hendler, H. Lieberman, W. Wahlster, & T. Berners-Lee (Urednici), *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential* (str. 171-188). Cambridge, MA, USA: The MIT Press.

McEnery, T., Xiao, R. i Tono, Y. (2006). *Corpus-Based Language Studies.* London: Routledge.

Mijo, K. (2012). *Rečnik zaljubljenika u gastronomiju.* (O. Stefanović, Prev.) Beograd: Službeni glasnik.

Miller, G. (1995). WordNet: a Lexical Database for English. *Communications of the ACM, 38*(11), 39-41.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. i Miller, K. (1990). Five Papers on WordNet. *Special Issue in International Journal of Lexicography, 3*(4).

Milne, R. M., O'Keefe, R. i Trotman, A. (2012). A Study in Language Identification. *Proceedings of the Seventeenth Australasian Document Computing Symposium* (str. 88-95). New York City, NY, USA: Association for Computing Machinery Press.

Mladenović, M. i Mitrović, J. (2013). Ontology of Rhetorical Figures for Serbian. U *Text, Speech, and Dialogue, Lecture Notes in Computer Science* (T. 8082, str. 386-393). Berlin Heidelberg: Springer-Verlag.

Mladenović, M. i Mitrović, J. (2014). Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool. *Natural Language Processing for Serbian: Resources and Applications* (str. 1-11). Belgrade: University of Belgrade.

Mladenović, M., Mitrović, J. i Krstev, C. (2014). Developing and Maintaining a WordNet: Procedures and Tools. U H. Orav, C. Fellbaume i P. Vossan (Ur.), *The Proceedings of Seventh Global WordNet Conference* (str. 55-62). Tartu, Estonia: University of Tartu.

Moens, M.-F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Dordrecht, Netherlands: Springer Science & Business Media.

Mori, S., Sasada, T., Yamakata, Y. i Yoshino, K. (2012). A Machine Learning Approach to Recipe Text Processing. *Proceedings of the 1st Cooking with Computer Workshop*, (str. 29-34). URL: http://liris.cnrs.fr/cwc/cwc2012/.

Muslea, I. (1999). Extraction Patterns for Information Extraction Tasks: A Survey. *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction* (str. 1-6). Orlando, FL, USA: AAAI Press.

Muslea, I., Minton, S. i Knoblock, C. (1998). STALKER: Learning Extraction Rules for Semistructured, Web-Based Information Sources. *Proceedings of AAAI-98*

Workshop on AI and Information Integration (str. 74-81). Orlando, FL, USA: AAAI Press.

Nadeau, D., Turney, P. i Matwin, S. (2006). Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. *Advances in Artificial Intelligence, 4013*, 266-277.

Navigli, R. i Velardi, P. (2002). Automatic Adaptation of WordNet to Domains. *Proceedings of 3rd International Conference on Language Resources and Evaluation* (str. 1023-1027). Las Palmas: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/47.pdf.

Nedović, V. (2013). Learning Recipe Ingredient Space Using Generative Probabilistic Models. U A. Cordier, E. Nauer i M. Wiegand (Ur.), *Cooking with Computers 2013 Workshop Proceedings* (str. 13-18). URL: http://liris.cnrs.fr/cwc/papers/CwC2013-Proceedings.pdf.

Niles, I. i Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proceedings of the IEEE International Conference on Information and Knowledge Engineering* (str. 412-416). San Diego, CA, USA: Institute of Electrical and Electronics Engineers.

Nirenburg, S. i Raskin, V. (2004). *Ontological Semantics.* Cambridge, MA, USA: MIT Press.

Onfre, M. (2002). *Gurmanski um: Filozofija ukusa.* Čačak: Umetničko društvo Gradac.

Padró, M. i Padró, L. (2004). Comparing Methods for Language Identification. *Procesamiento del lenguaje natural, 33*, 155-162.

Pajić, V. (2012). *Modeli konačnih stanja u ekstrakciji informacija, doktorska teza.* Beograd: Univerzitet u Beogradu, Matematički fakultet.

Pascal ISO/IEC 7185:1990. (1990). *Pascal ISO/IEC 7185:1990.*

Paumier, S. (2014). Unitex 3.1 User Manual. URL: http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf.

Pavlović-Lažetić, G. i Graovac, J. (2010). Ontology-Driven Conceptual Document Classification. *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (str. 383-386). Setúbal, Portugal: Science and Technology Publications.

Pelaprat, A. P. (1969). *Veliki Pelaprat: Prvi kuvar sveta.* Beograd: Prosveta.

Petrov, S. (2009). The Cevapcici Guide To The Balkans. URL: http://balkancevapcici.blogspot.rs/.

Pinard, Y. (2010). *Food in the Louvre (Musee Du Louvre).* Paris: Flammarion.

Poibeau, T. (2000). A Corpus-Based Approach to Information Extraction. *Journal of Applied Systems Studies, 1*(2), 254-267.

Popović, Lj. (2000). Bivalentni kontrolori kongruencije : problem leksikografskog opisa konkurencije gramatičkog i semantičkog slaganja. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opisi primene, 29/1*, 65-80.

Popović, Lj. i Vitas, D. (2003). Konspekt za izgradnju referentnog korpusa standardnog srpskog jezika. *Naučni sastanak slavista u Vukove dane*, 221-227.

Popović, Z. (2008). *Evaluacija programa za obeležavanje (etiketiranje) teksta, magistarska teza.* Beograd: Univerzitet u Beogradu, Matematički fakultet.

Popović, Z. (2010). Taggers applied on texts in Serbian. *INFOtheca: Journal of Informatics and Librarianship, 11*(2), 19-36.

Portić, M. (2011a). *Gastronomija.* Beograd: Univerzitet Singidunum.

Portić, M. (2011b). *Gastronomski proizvodi.* Novi Sad: Univerzitet u Novom Sadu, Prirodno-matematički fakultet.

Rable, F. (1989). *Gargantua i Pantagruel.* (S. Vinaver, Prev.) Beograd: Samostalno prevodilačko izdanje K. Vinavera i dr.

Riloff, E. (1993). Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceedings of the 11th National conference on Artificial Intelligence* (str. 811-816). Palo Alto, CA, USA: AAAI Press/MIT Press.

Riloff, E. i Lorenzen, J. (1999). Extraction-based Text Categorization: Generating Domain-specific Role Relationships Automatically. U *Natural Language Information Retrieval, Text, Speech and Language Technology* (T. 7, str. 167-196). Dordrecht, Netherlands: Springer Science & Business Media.

Ritchie, G. D. (1992). *Computational Morphology: Practical Mechanisms for the English Lexicon.* Cambridge, MA, USA: The MIT press.

RMSMH. (1967). *Rečnik srpskohrvatskoga književnog jezika* (*Tom 1-6*). Beograd-Zagreb: Matica Srpska, Matica Hrvatska.

RMSMH. (1976). *Речник српскохрватскога књижевног језика* (*Том 6*). Нови Сад: Матица српска.

Sager, N., Friedman, C. i Lyman, M. (1987). *Medical Language Processing: Computer Management of Narrative Data.* MA, USA: Addison-Wesley, Reading.

Sarawagi, S. (2008). Information Extraction. *Foundations and Trends in Databases, 1*(3), 261-377.

Segaran, T., Evans, C. i Taylor, J. (2009). *Programming the Semantic Web.* Sebastopol, CA, USA: O'Reilly Media, Inc.

Sekine, S. i Isahara, H. (2000). IREX: IR and IE Evaluation Project in Japanese. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2000)* (str. 1975-1980). Paris: ELRA.

Sekine, S. i Nobata, C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)* (str. 1977-1980). Paris: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2004/.

Sekine, S., Sudo, K. i Nobata, C. (2002). Extended Named Entity Hierarchy. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2002).* Las Palmas, Canary Islands, Spain. URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/120.pdf.

Silberztein, M. D. (1989). The Lexical Analysis of French. *Electronic Dictionaries and Automata in Computational Linguistics* (str. 93-110). Berlin: Springer.

Slocum, J. (1985). A Survey of Machine Translation: its History, Current Status, and Future Prospects. *Computational Linguistics, 11*(1), 1-17.

Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning, 34*(1-3), 233-272.

Soderland, S., Fisher, D., Aseltine, J. i Lehnert, W. (1995). CRYSTAL: Inducing a Conceptual Dictionary. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)* (str. 1314-1319). Menlo Park, CA, USA: AAAI Press.

Stanković, R. (2009). *Modeli ekspanzije upita nad tekstuelnim resursima, doktorska teza.* Beograd: Univerzitet u Beogradu, Matematički fakultet.

Stanković, R., Obradović, I. i Trtovac, A. (2012). An Approach to Development of Bilingual Lexical Resources. *Proceedings of the Fifth Balkan Conference in Informatics* (str. 101-104). Novi Sad: Faculty of Sciences, University of Novi Sad.

Stanković, R., Obradović, I., Krstev, C. i Vitas, D. (2011). Production of Morphological Dictionaries of Multi-Word Units Using a Multipurpose Tool. *Proceedings of the Computational Linguistics-Applications Conference (Selected papers)* (str. 77-84). Katowice: Polskie Towarzystwo Informatyczne, Oddział Górnośląski.

Stecher, R., Niedere, C., Nejdl, W. i Bouquet, P. (2008). Adaptive Ontology Re-Use: Finding and Re-Using Sub-Ontologies. *International Journal of Web Information Systems, 4*(2), 198-214.

Studer, R., Benjamins, R. V. i Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering, 25*(1), 161-197.

Szatrowski, P. (2014). *Language and Food: Verbal and Nonverbal Experiences* (Pragmatics & Beyond New Series, *Volume 238*). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Tatar, S. i Cicekli, I. (2011). Automatic Rule Learning Exploiting Morphologicalcal Features for Named Entity Recognition in Turkish. *Journal of Information Science, 37*(2), 137-151.

Teng, C.-Y., Lin, Y.-R. i Adamic, L. (2012). Recipe Recommendation Using Ingredient Networks. *Proceedings of the 4th Annual ACM Web Science Conference* (str. 298-307). New York City, NY, USA: Association for Computing Machinery Press.

The New York Public Library. (2015). URL: http://digitalcollections.nypl.org/.

Tomić, B. (2014). Kategorije broja i brojivosti u nazivima za jela u srpskom jeziku. *Naš jezik, 45*(3/4), 39-49.

Tufiş, D., Cristea, D. i Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information Science and Technology, 7*(1-2), 9-43.

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A. i Ştefănescu, D. (2008). Romanian WordNet: Current State, New Applications and Prospects. *Proceedings of 4th Global*

WordNet Conference (str. 441-452). Szeged: University of Szeged, Department of Informatics.

Tufiş, D., Verginica, M. B., Ştefănescu, D. i Ion, R. (2013). The Romanian Wordnet in a Nutshell. *Language Resources and Evaluation, 47*, 1305-1314.

Utvić, M. (2014). Izgradnja referentnog korpusa savremenog srpskog jezika, doktorska teza. Beograd: Univerzitet u Beogradu, Filološki fakultet.

van Assem, M., Gangemi, A. i Schreiber, G. (2006). RDF/OWL Representation of WordNet. *W3C Public Working Draft*, *19.*

van Assem, M., Menken, M. R., Schreiber, G., Wielemaker, J. i Wielinga, B. (2004). A Method for Converting Thesauri to RDF/OWL. *The Semantic Web - ISWC 2004* (str. 17-31). Berlin Heidelberg: Springer.

Vasiljević, N. (2014). *Automatska obrada pravnih tekstova na srpskom jeziku, doktorska disertacija.* Beograd: Univerzitet u Beogradu, Filološki fakultet.

Vetulani, Z. (2000). Electronic Language Resources for Polish: POLEX, CEGLEX and GRAMLEX. *Second International Conference on Linguistic Ressources and Evaluation (LREC 2000)* (str. 367-374). Paris: ELRA.

Vintar, Š. i Fišer, D. (2011). Enriching Slovene WordNet with Domain-Specific Terms. *Translation: Computation, Corpora, Cognition, 1*(1), 29-44.

Vitas, D. (1993). *Matematički model morfologije srpskohrvatskog jezika (imenska fleksija), doktorska disertacija.* Beograd: Univerzitet u Beogradu, Matematički fakultet.

Vitas, D. (1997). O elementarnoj morfografemskoj klasi. *Naučni sastanak slavista u Vukove dane, "Međuodnos gramatike i rečnika u srpskom jeziku", 26/2* (str. 195-206). Beograd: Međunarodni slavistički centar.

Vitas, D. (2006). *Prevodioci i interpretatori: uvod u teoriju i metode kompilacije.* Beograd: Univerzitet u Beogradu, Matematicki fakultet.

Vitas, D. (2007). O problemu ne(pre)poznate reči u obradi tekstova na srpskom jeziku. U *Zbornik Matice srpske za filologiju i lingvistiku* (str. 111-120). Novi Sad: Matica srpska.

Vitas, D. (2014). O različitosti sličnog. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opisi primene, 43/3*, 31-49.

Vitas, D., Krstev, C. i Sabo, O. (2003). O predstavljanju morfološke informacije u elektronskim rečnicama slovenskih jezika. *Književnost i jezik - časopis Društva za srpskohrvatski jezik i književnost, 50*(1-3), 97-106.

Vitas, D., Vasiljević, N. i Krstev, C. (2014). Informatički pogled na korpus zakona Republike Srbije. *Srpski jezik - studije srpske i slovenske, sv. 19*, 377-394.

Vossen, P. (1998). *A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

Vujičić Stanković, S. (2012). Named Entity Recognition in the System for Information Extraction. U S. Halupka-Rešetar, M. Marković, T. Milićev i N. Milićević (Ur.), *Selected Papers from SinFonIJA 3* (str. 206-223). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Vujičić Stanković, S. (2013). Model sistema za ekstrakciju informacija iz tekstova pisanih na srpskom jeziku. *Info M*, 47/2013, 4–9.

Vujičić Stanković, S. i Pajić, V. (2012). Information Extraction from the Weather Reports in Serbian. *Proceedings of the Fifth Balkan Conference in Informatics* (str. 105-108). Novi Sad: Faculty of Sciences, University of Novi Sad.

Vujičić Stanković, S. i Pajić, V. (2013). Formiranje domenskog korpusa – kulinarska leksika. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opisi primene, 43*(3), 51-59.

Vujičić Stanković, S. i Pajić, V. (2015). Upotreba vlastitih imena u kulinarskom domenu. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opisi primene, 44/3*, 137-142.

Vujičić Stanković, S., Kojić, N., Rakočević, G., Vitas, D. i Milutinović, V. (2012). A Classification of Data Mining Algorithms for Wireless Sensor Networks, and Classification Extension to Concept Modeling in System of Wireless Sensor Networks Based on Natural Language Processing. *Advances in Computers: Connected Computing Environment, 90*, 223-283.

Vujičić Stanković, S., Krstev, C. i Vitas, D. (2014). Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain. *The Proceedings of Seventh Global WordNet Conference* (str. 127-132). Tartu, Estonia: University of Tartu.

Vujičić Stanković, S., Rakočević, G. i Milutinović, V. (2011). A Metadata-Supported Distributed Approach for Data Mining Based Prediction in Wireless Sensor Networks. *10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS, Volume 1)* (str. 181-185). Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Vukić, M. i Portić, M. (2009). *Kuvarstvo sa praktičnom nastavom.* Beograd: Zavod za udžbenike.

Weischedel, R., Boisen, S., Bikel, D., Bobrow, R., Crystal, M., Ferguson, W. i Wechsler, A. (1996). Progress in Information Extraction. *TIPSTER '96 Proceedings of a Workshop* (str. 127-138). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wimalasuriya, D. i Dou, D. (2009). Using Multiple Ontologies in Information Extraction. *Proceedings of the 18th ACM conference on Information and knowledge management* (str. 235-244). New York: ACM press.

Wimalasuriya, D. i Dou, D. (2010). Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches. *Journal of Information Science, 36*(3), 306-323.

Woods, W. (1970). Transition Network Grammars for Natural Language Analysis. *Communications of the ACM, 13*(10), 591-606.

Wu, H. C., Luk, R. W., Wong, K. F. i Kwok, K. L. (2008). Interpreting tf-idf Term Weights as Making Relevance Decisions. *ACM Transactions on Information Systems (TOIS), 26*(3), 13:1-13:37.

Yangarber, R. i Grishman, R. (1998). NYU: Description of the Proteus/PET System as Used for MUC-7 ST. *Proceedings of the 7th Message Understanding Conference (MUC-7).* San Fransisco, CA, USA: Morgan Kaufmann Publishers.

Yu, L. (2011). *A Developer's Guide to the Semantic Web.* Berlin Heidelberg: Springer-Verlag.

Zaid, A., Hughes, H., Porceddu, E. i Nicholas, F. (2007). Biotehnološki rečnik za hranu i poljoprivredu. (G. Rehm, H. Uszkoreit, Urednici i T. Č. S. Stojanović, Prev.) FAO. URL: http://www.fao.org/biotech/docs/serbglos.pdf

Zečević, A. i Vujičić Stanković, S. (2013). The Mysterious Letter J. *Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages*

and Language Variants (str. 40-44). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA. URL: http://www.aclweb.org/anthology/W13-53.

Zgusta, L. (1971). *Manual of Lexicography.* Prague: Publishing House of the Czechoslovak Academy of Sciences.

Zhai, Y. i Liu, B. (2005). Web Data Extraction Based on Partial Tree Alignment. *Proceedings of the 14th International Conference on World Wide Web* (str. 76-85). New York City, NY, USA: Association for Computing Machinery.

Zhang, Q., Hu, R., Mac Namee, B. i Delany, S. J. (2008). *Back to the Future: Knowledge Light Case Base Cookery.* Dublin: Dublin Institute of Technology.

ZTIN. (1932). *Zoološka terminologija i nomenklatura.* Beograd: Ministarstvo prosvete Kraljevine Jugoslavije.

ПРИЛОГ А

Сатоши Секине и сарадници су направили проширену хијерархију именованих ентитета који се најчешће јављају у текстовима новинских чланака (Sekine, Sudo i Nobata, 2002; Sekine i Nobata, 2004). Она је 2002. године обухватала око 150 типова именованих ентитета, као и бројне подтипове, да би је касније проширили на 200 типова и њихове подтипове. Хијерархија је кроз експерименте на развоју упитничких система и система за екстракцију информација упоредо развијана за јапански и енглески језик. Овде је приказан извод из те хијерархије за подтипове проширеног именованог ентитета производ (*Product*),[76] међу којима се као подтип јавља и храна (*Food*):

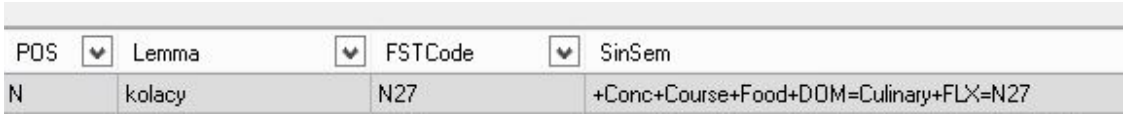
Подтип проширеног именованог ентитета		**Пример**
Product_Other		државне обвезнице
Material		керозин
Clothing		кимоно, фармерке
Money_Form		новчић, кованица
Drug		Аспирин, HCI
Weapon		нуклеарно оружје
Stock		акције Микрософта
Award		Нобелова награда
Decoration		Легија части
Offence		убиство, превара
Service		Air Serbia #333
Class		прва класа, фелери
Character		Мери Попинс
ID_Number		IE 1234-5678, 1605
	Vehicle_Other	Harley-Davidson

[76] Секинеова дефиниција проширене хијерархије именованих ентитета дата је на адреси: http://nlp.cs.nyu.edu/ene/version7_1_0Beng.html.

Vehicle	**Car**	Фиат 500
	Train	TGV
	Aircraft	Конкорд, Боинг 737
	Spaceship	Аполо 1, Мир
	Ship	Титаник, Наутилус
Food	**Food_Other**	Соса Cola, вода
	Dish	пиво, лазања
Art	**Art_Other**	Давид, Победник
	Picture	Ноћна стража
	Broadcast_Program	Београдска хроника
	Movie	Грк Зорба
	Show	Мадам Батерфлај
	Music	Мокрањчеве руковети
	Book	Гаргантуа и Пантагруел
Printing	**Printing_Other**	Устав Републике Србије
	Newspaper	Политика
	Magazine	Инфотека
Doctrine_Method	**Doctrine_Method_Other**	арапски бројеви
	Culture	ренесанса
	Religion	хришћанство
	Academic	лингвистика
	Sport	пливање
	Style	импресионизам
	Theory	теорија релативитета
	Plan	Маршалов план
Language	**Language_Other**	врањски дијалект
	National_Language	српски
Unit	**Unit_Other**	кг, минут
	Currency	$, динар

ПРИЛОГ Б

У овом прилогу је дата XML репрезентација лексичких информација која укључује морфолошке, синтаксичке и семантичке категорије које чине пример речника у LMF формату за одредницу *колач* приказану на слици 45.



POS	Lemma	FSTCode	SinSem
N	kolacy	N27	+Conc+Course+Food+DOM=Culinary+FLX=N27

Слика 45. Пример одреднице речника.

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<!--Lexical markup framework (LMF): ISO/TC 37/SC 4 N130 Rev.9-->
<Lexicon version="LMF: ISO/TC 37/SC 4 N130 Rev.9">
 <LexiconInformation language="srp" />
 <LexicalEntry>
  <feat att="ID" val="14590" />
  <feat att="partOfSpeech" val="N" />

  <Lemma>
   <feat att="writtenForm" val="kolacy" />
   <feat att="inflectionalParadigm" val="N27" />
  </Lemma>

  <WordForm>
   <DelafRed>kolacyima,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mp7q</DelafRed>
   <feat att="writtenForm" val="kolacyima" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="plural" />
   <feat att="grammaticalCase" val="locative" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>

  <WordForm>
   <DelafRed>kolacyima,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mp6q</DelafRed>
   <feat att="writtenForm" val="kolacyima" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="plural" />
   <feat att="grammaticalCase" val="instrumental" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>

  <WordForm>
   <DelafRed>kolacyima,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mp3q</DelafRed>
   <feat att="writtenForm" val="kolacyima" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="plural" />
   <feat att="grammaticalCase" val="dative" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>

  <WordForm>
   <DelafRed>kolacyi,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mp5q</DelafRed>
   <feat att="writtenForm" val="kolacyi" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="plural" />
   <feat att="grammaticalCase" val="vocative" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>
```

```
<WordForm>
 <DelafRed>kolacyi,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mp1q</DelafRed>
 <feat att="writtenForm" val="kolacyi" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="plural" />
 <feat att="grammaticalCase" val="nominative" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>

<WordForm>
 <DelafRed>kolacyem,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:ms6q</DelafRed>
 <feat att="writtenForm" val="kolacyem" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="singular" />
 <feat att="grammaticalCase" val="instrumental" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>

<WordForm>
 <DelafRed>kolacye,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mp4q</DelafRed>
 <feat att="writtenForm" val="kolacye" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="plural" />
 <feat att="grammaticalCase" val="accusative" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>

<WordForm>
 <DelafRed>kolacyu,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:ms5q</DelafRed>
 <feat att="writtenForm" val="kolacyu" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="singular" />
 <feat att="grammaticalCase" val="vocative" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>

<WordForm>
 <DelafRed>kolacyu,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:ms7q</DelafRed>
 <feat att="writtenForm" val="kolacyu" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="singular" />
 <feat att="grammaticalCase" val="locative" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>

<WordForm>
 <DelafRed>kolacyu,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:ms3q</DelafRed>
 <feat att="writtenForm" val="kolacyu" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="singular" />
 <feat att="grammaticalCase" val="dative" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>

<WordForm>
 <DelafRed>kolacya,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mw4q</DelafRed>
 <feat att="writtenForm" val="kolacya" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="paukal" />
 <feat att="grammaticalCase" val="accusative" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>

<WordForm>
 <DelafRed>kolacya,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mw2q</DelafRed>
 <feat att="writtenForm" val="kolacya" />
 <feat att="grammaticalGender" val="masculine" />
 <feat att="grammaticalNumber" val="paukal" />
 <feat att="grammaticalCase" val="genitive" />
 <feat att="grammaticalAnimateness" val="non-animate" />
</WordForm>
```

```
  <WordForm>
   <DelafRed>kolacya,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:mp2q</DelafRed>
   <feat att="writtenForm" val="kolacya" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="plural" />
   <feat att="grammaticalCase" val="genitive" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>

  <WordForm>
   <DelafRed>kolacya,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:ms2q</DelafRed>
   <feat att="writtenForm" val="kolacya" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="singular" />
   <feat att="grammaticalCase" val="genitive" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>

  <WordForm>
   <DelafRed>kolacy,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:ms4q</DelafRed>
   <feat att="writtenForm" val="kolacy" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="singular" />
   <feat att="grammaticalCase" val="accusative" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>

  <WordForm>
   <DelafRed>kolacy,kolacy.N+Conc+Course+Food+DOM=Culinary+FLX=N27:ms1q</DelafRed>
   <feat att="writtenForm" val="kolacy" />
   <feat att="grammaticalGender" val="masculine" />
   <feat att="grammaticalNumber" val="singular" />
   <feat att="grammaticalCase" val="nominative" />
   <feat att="grammaticalAnimateness" val="non-animate" />
  </WordForm>

  <sense>
   <nasxSinSem>+Conc+Course+Food+DOM=Culinary+FLX=N27</nasxSinSem>
   <nasxKomentar>
   </nasxKomentar>
   <synSet>
   </synSet>
   <nasxDict>delas-im.dic</nasxDict>
   <syntacticBehavior>
   </syntacticBehavior>
  </sense>
 </LexicalEntry>

</Lexicon>
```

ПРИЛОГ В

У овом прилогу је дат опис спољашњих програма система Unitex који су коришћени за екстракцију информација у овом раду.

За нормализацију текста у оквиру претходне обраде користи се спољашњи програм *Normalize*. Синтакса позива програма је:

```
Normalize <tekst> [OPCIJE]
```

где се уместо `<tekst>` наводи комплетна путања до *txt* документа текста који се нормализује. Уколико овај документ има назив *tekst_za_obradu.txt*, тада се након извршења нормализације аутоматски креира директоријум под називом *tekst_za_obradu_snt* у који се смешта резултујући документ са истим називом као почетни и екстензијом *snt* (*tekst_za_obradu.snt*). Овај директоријум ће касније користити и други програми Unitex-а који се позивају да у њега сместе резултате извршавања. Опционо може да се наведе путања до посебно форматираног документа у коме су описана правила нормализације која треба да се примене.

У Unitex систему се разликују две врсте интерних репрезентација графова: графички приказ и одговарајући текстуални формат репрезентације граматика. Unitex графови се чувају у датотекама у текстуалном формату са екстензијом *grf.* У њој се чувају подаци о визуалном приказу стања и прелаза графа, описи њихових позиција и изгледа (боје, величине итд.). Док корисник мења граф у графичкој корисничкој сумеђи подаци се чувају у овој датотеци. Када је потребно да се граф искористи као улазни податак за даљу обраду текста, компилира се и конвертује у *fst2* формат. У овом формату се чувају подаци о стањима и прелазима једног или више графова, али се у потпуности занемарује њихов визуални приказ. *fst2* формат графова разликује се од коначних трансдуктора јер задржава структуру са подграфовима унутар граматика. Због тога се у Unitex-у употребом *Flatten* програма омогућава

конверзија граматике из *fst2* формата у коначни трансдуктор, када је то могуће, или конструкција апроксимативног трансдуктора када није. При компилирању се сваки појединачни позив подграфа замењује тим подграфом. Замене се врше до одређене дубине, а позиви подграфова који се налазе након те дубине замењују се празном речју. Резултат компилирања је и у том случају коначни трансдуктор, али који није еквивалентан полазном графу. Такође, постоји опција да се позиви подграфова задрже, чиме се постиже еквивалентност, али резултат компилирања није коначни трансдуктор. Позив програма *Flatten* има облик:

```
Flatten <fst2> [OPCIJE]
```

где се са `<fst2>` задаје комплетна путања до главног графа са екстензијом *fst2*. Опције које се користе при позиву како би се постигла описана трансформација су: `-r/--rtn` и `-d N/--depth=N`, где је целобројна вредност `N` дубина до које се при трансформацији врши замена позива подграфова конкретним подграфовима. `N` је подразумевано 10.

У оквиру претходне обраде, док још није извршена подела текста на лексичке јединице, позивом програма *Fst2Txt*:

```
Fst2Txt <fst2> [OPCIJE]
```

примењује се трансдуктор који се налази на путањи `<fst2>` на текст који се обрађује. Позивом овог програма улазни текст ће бити измењен. Комплетна путања документа текста са екстензијом *snt*, `TXT`, наводи се у делу [`OPCIJE`] као: `-t TXT/--text=TXT`. Путања `ALPH` до фајла са алфабетом језика текста који се обрађује наводи се са `-a ALPH/--alphabet=ALPH`. Додатне опције за креирање текста излазног документа су `-M/--merge`, којом се бира да се излаз из коначног трансдуктора споји са улазним текстом, `-R/--replace` да се делови улазног текста препознати трансдуктором замене одговарајућим излазним текстом одређеним излазом трансдуктора.

За токенизацију генерисаног текста из *snt* документа позива се програм *Tokenize*:

```
Tokenize <tekst> [OPCIJE]
```

где се уместо `<tekst>` наводи комплетна путања до *snt* документа текста, а као опција се наводи путања до фајла којим се дефинише који карактери припадају алфабету језика текста који се обрађује. Током извршења програм *Tokenize* креира текстуални документ *tokens.txt* који садржи листу токена текста. Сваком токену је додељен јединствени кôд. Ови кодови се потом користе да се цео текст кодира и затим сними у бинарну датотеку *text.cod*. Поред ових докумената креирају се и: *tok_by_freq.txt* који садржи токене сортиране по учесталости појављивања у тексту, *tok_by_alph.txt* који садржи токене сортиране лексикографски, *stats.n* који садржи информације о броју сепаратора реченица, броју токена, броју речи и бројева и *enter.pos* који садржи листу позиција знака за нови ред у тексту и који се користи за усаглашавање конкорданци са оригиналним текстом.

Програм *Dico* се користи да се генеришу речници текста који се обрађује. Након позива програма:

```
Dico <tekst> [OPCIJE]
```

у директоријуму *tekst_za_obradu_snt* се креирају фајлови *dlf* и *dlc* који садрже редом речник простих речи и речник полилексичких јединица у DELAF формату, и речник непознатих речи *err*. Програм *Dico* у директоријуму текста креира фајлове: *dlf* који садржи речник простих речи из текста, *dlc* који садржи речник полилексичких јединица из текста, *err* који садржи речник непрепознатих речи, *tags.ind* који садржи секвенце које је потребно уметнути у аутомат текста и *stat_dic.n* који садржи информације о броју простих речи, полилексичких јединица и непрепознатих речи.

Како би се на текст примениле граматике позива се програм *Locate*:

```
Locate <fst2> [OPCIJE]
```

Под применом граматика се подразумева претрага текста или примена трансдуктора. Након извршења примене граматика програм *Locate* креира датотеке *concord.ind* и *concord.n* у директоријуму *tekst_za_obradu_snt*. Прва

датотека садржи референце до пронађених појављивања израза који одговарају графу, а друга садржи број појављивања и проценат препознатих формалних речи унутар текста. Међу опцијама се поред задавања комплетне путање до текста и одређивања алфабета текста могу задати и различита ограничења: `-l/--all` се поставља како би се пронашли сви изрази који одговарају граматици, `-S/--shortest_matches` се поставља како би се пронашли само најкраћи изрази, `-L/--longest_matches` се поставља како би се пронашли најдужи, `-A/--all_matches` се поставља како би се пронашли сви изрази, а `-n N/--number_of_matches=N` се поставља како би се претрага зауставила након што се пронађе првих `N` израза. Могу се задати и различите опције које ће важити при креирању излаза попут: `-I/--ignore` чиме се у резултатима игноришу излази трансдуктора, `-M/--merge` чиме се излази трансдуктора убацују на одговарајући начин у улазни текст или `-R/--replace` где се препознати изрази улазног текста замењују одговарајућим излазима трансдуктора.

Програм *Locate* у директоријуму текста креира два фајла: *concord.ind* који садржи референце до пронађених појављивања израза који одговарају графу и *concord.n* који садржи информацију о броју појављивања као и о проценту препознатих формалних речи у тексту. Подаци о препознатим изразима који су смештени у *concord.ind* датотеку се користе у даљем процесу екстракције информација за њихово екстраховање, обраду или структурирање.

ПРИЛОГ Г

У овом прилогу дат је Java програмски кôд класа *POSTaggerSR*, *UnitexPart* и *ExternalCommand* које чине део GATE додатка за подршку обради текстова на српском језику, а којима се решава проблем додељивања граматичких категорија.

Прво се позивају класе *UnitexPart* и *ExternalCommand* да се позове програм *UnitexToolLogger.exe* који омогућава покретање свих спољашњих Unitex подпрограма из командне линије чиме се формирају електронски речници за текст који се обрађује. На основу њих се касније у класи *POSTaggerSR* речима придружују одговарајуће граматичке категорије које се придружују тексту као анотације за употребу у следећим фазама обраде.

```
package gate.posserb;

import java.io.File;

public class UnitexPart extends ExternalCommand {
  public static String ressourceDir;
  public static String graphResDir;
  public static String othersResDir;
  public static String CorpusWorkPath;
  public static String delaResDir;

  public static String UnitexToolLoggerPath;
  public static String encodingLine;
  public String docName;
  public String encoding;

  protected void beforeExecute() {
          // Postavljanje parametara za GATE:
          ressourceDir = ".\\plugins\\Lang_Serbian\\resources\\";
          graphResDir = ressourceDir + "Graphs\\";
          othersResDir = ressourceDir + "Others\\";
          CorpusWorkPath = ressourceDir + "Corpus\\";
          delaResDir = ressourceDir + "Dela\\";

// Postavljanje putanje do UnitexToolLoggera:
UnitexToolLoggerPath = ".\\plugins\\Lang_Serbian\\prog\\UnitexToolLogger.exe";

// Podaci o kodnoj semi dokumenta
if (encoding.equalsIgnoreCase("UTF-16")){
  encodingLine = "-qutf16be-bom";
}
          else if (encoding.equalsIgnoreCase("UTF-8")){
                  encodingLine = "-qutf8-no-bom";
          }

          // Pravljenje foldera u kome ce biti smesteni rezultati
          File CorpusDirectory = new File(CorpusWorkPath + docName+ "_snt");
          try {
```

```
                if (CorpusDirectory.mkdir()) {
                        System.out.println("Directory " + CorpusWorkPath
                                        + docName + "_snt created");
                } else {
                        System.out.println("Directory " + CorpusWorkPath
                                        + docName + "_snt is not created");
                }
        } catch (Exception e) {
                e.printStackTrace();
        }

}

// Formiranje komandne linije za prosledjivanje ProcessBuilder-u
protected String[] getCommandArguments() {
        String[] command = { UnitexToolLoggerPath, "{", "Normalize",
                CorpusWorkPath + docName + ".txt", "-r",
                othersResDir + "Norm.txt", encodingLine, "}", "{", "Grf2Fst2",
                graphResDir + "Preprocessing\\Sentence\\Sentence.grf", "-y",
                "--alphabet=" + othersResDir + "Alphabet.txt", encodingLine,
                "}", "{", "Flatten",
                graphResDir + "Preprocessing\\Sentence\\Sentence.fst2",
                "--rtn", "-d5", encodingLine, "}", "{", "Fst2Txt",
                "-t" + CorpusWorkPath + docName + ".snt",
                graphResDir + "Preprocessing\\Sentence\\Sentence.fst2",
                "-a" + othersResDir + "Alphabet.txt", "-M", encodingLine,
                "}", "{", "Grf2Fst2",
                graphResDir + "Preprocessing\\Replace\\Replace.grf", "-y",
                "--alphabet=" + othersResDir + "Alphabet.txt", encodingLine,
                "}", "{", "Fst2Txt", "-t" + CorpusWorkPath + docName + ".snt",
                graphResDir + "Preprocessing\\Replace\\Replace.fst2",
                "-a" + othersResDir + "Alphabet.txt", "-R", encodingLine,
                "}", "{", "Tokenize", CorpusWorkPath + docName + ".snt",
                "-a" + othersResDir + "Alphabet.txt", encodingLine, "}",
                "{", "Dico", "-t" + CorpusWorkPath + docName + ".snt",
                "-a" + othersResDir + "Alphabet.txt",
                "-m" + delaResDir + "latDelacf-im-new.bin",
                "-m" + delaResDir + "latdelacf-Food-.bin",
                "-m" + delaResDir + "latdelacf-imenice.bin",
                "-m" + delaResDir + "latdelaf-Food-.bin",
                "-m" + delaResDir + "latdelaf-Srpski.bin",
                "-m" + delaResDir + "latdelaf-im-nove.bin",
                delaResDir + "latdelaf-Srpski.bin",
                delaResDir + "latdelaf-Licna.bin",
                delaResDir + "latdelaf-Strana.bin",
                delaResDir + "latdelaf-Vlastita.bin",
                delaResDir + "Acr-new+.fst2", delaResDir + "Razno.fst2",
                delaResDir + "NaKraju+.fst2",
                delaResDir + "latdelacf-imenice.bin",
                delaResDir + "latdelacf-neprom.bin",
                delaResDir + "latdelacf-NProp.bin",
                delaResDir + "latdelacf-pridevi.bin",
                delaResDir + "latDelacf-im-new.bin",
                delaResDir + "latdelaf-im-nove.bin",
                delaResDir + "latdelaf-gl-novi.bin",
                delaResDir + "delaf-im-nove.bin",
                delaResDir + "delaf-gl-novi.bin", delaResDir + "Acr+.fst2",
                delaResDir + "NewBrojSlovima.fst2",
                delaResDir + "latdelacf-Food-.bin",
                delaResDir + "latdelaf-Food-.bin",
                delaResDir + "delaf-Food-.bin", delaResDir + "Acr+.fst2",
                delaResDir + "latFilterUmeren-.bin",
                delaResDir + "NaKraju+.fst2",
                delaResDir + "NewBrojSlovima.fst2", delaResDir + "Razno.fst2",
                encodingLine, "}" };

        return command;
}

protected String getCommandLine() {
        String command_line = "CMD line";
        String[] s = getCommandArguments();
        for (String str : s)
                command_line += " " + str;
```

```
        return command_line;
  }

  public UnitexPart(String docName, String encoding) {
        this.docName = docName;
        this.encoding = encoding;
        execute();
  }
```

```
package gate.posserb;

import java.io.BufferedReader;
import java.io.IOException;
import java.io.InputStream;
import java.io.InputStreamReader;

public abstract class ExternalCommand {
  protected abstract void beforeExecute();

  protected abstract String[] getCommandArguments();

  protected abstract String getCommandLine();

  public void execute() {
        execute(new LineWriter() {
                public void writeLine(String line) {
                        System.out.println(line);
                }
        });
  }

  public void execute(LineWriter lw) {
        beforeExecute();
        lw.writeLine("> " + getCommandLine());
        ProcessBuilder pb = new ProcessBuilder(getCommandArguments());

        pb.redirectErrorStream(true);

        try {
        Process p = pb.start();

        InputStream in = p.getInputStream();
        BufferedReader bin = new BufferedReader(new InputStreamReader(in));
        String line;
        while ((line = bin.readLine()) != null)
                lw.writeLine(line);
        } catch (IOException e) {
                throw new RuntimeException(e);
        }
  }
}
```

```
package gate.posserb;

import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.IOException;
import java.util.ArrayList;
import java.util.Iterator;
import java.util.List;
import java.util.Scanner;
import java.util.StringTokenizer;
import java.util.regex.Matcher;
import java.util.regex.Pattern;

public class POSTaggerSR {

private String encoding;
```

```
// Konstruktor POS tagera sa podrazumevanom (platformskom) kodnom semom.
  public POSTaggerSR() throws IOException {
                this("UTF-8");
        }

// Konstruktor POS tagera sa specificnom kodnom semom.
public POSTaggerSR(String encoding) throws IOException{
                this.encoding = encoding;
        }

 /* Pokretanje POS tagera nad listom recenica.
  * Svaka recenica je lista reci koje treba anotirati.
  * Povratna vrednost je lista anotiranih recenica, od kojih je svaka lista parova
  * niski reci i njihovih odgovarajucih gramatickih kategorija
  */
 public List runTagger(List sentences, String docName){

        UnitexPart un = new UnitexPart(docName, encoding);
        List output = new ArrayList();

        List taggedSentence = new ArrayList();
        Iterator sentencesIter = sentences.iterator();
        while (sentencesIter.hasNext())
        {
                List sentence = (List)sentencesIter.next();
                Iterator wordsIter = sentence.iterator();
                while(wordsIter.hasNext())
                {
                        String newWord = (String)wordsIter.next();
                        newWord = newWord + ",";
                        // rec se dopunjava gramatickim kategorijama
                        oneStepMultiCat(newWord, taggedSentence,docName);
                }
                output.add(taggedSentence);
                taggedSentence = new ArrayList();
        }
        return output;
 }

 /*
  * Pridruzuje novoj reci odgovarajuce gramaticke kategorije i pridruzuje te podatke
  * taggedSentence listi.
  */
 protected void oneStepMultiCat(String word, List taggedSentence, String docName){

        List stringList = new ArrayList();

        //Otvara se recnik dlf za citanje
        Scanner sc=null;
        try
        {
                sc=new Scanner(new
                File(".\\plugins\\Lang_Serbian\\resources\\Corpus\\"+docName+"_snt
                \\dlf"), "Unicode");

        }
        catch(FileNotFoundException e)
        {
                e.printStackTrace();
                System.out.println("Nisam uspeo da napravim skener za dlf\n "+
                e.getMessage());
        }

        int ind = 0;
        word = word.toLowerCase();

        // Citaju se podaci iz dlf recnika
        String cat = null;
        String lemma = null;
        String[] strnew;

        while(sc.hasNextLine())
        {
```

```
String line=sc.nextLine();

// Da li linija sadrzi rec?
int pos=line.indexOf(word);

// regularni izraz za prepoznavanje okoline kategorije
String regx = ",(.*)(\\.(.*))";

Pattern pat = Pattern.compile(regx);
Matcher mat = pat.matcher(line);

// Ako linija sadrzi rec
if(pos == 0)
{
	// i ako ta rec nije vec obradjivana
	// onda je vredi obradjivati
	if(ind == 0)
	{
		ind = 1;

		while (mat.find())
		{
			lemma = mat.group(1);
			cat=mat.group(3);

			String podgrupa=mat.group(2);

			// Ako group(1) sadrzi : ili +
			String regx1= "\\.((\\w)*)[:|\\+]";
			Pattern pat1 = Pattern.compile(regx1);
			Matcher mat1 = pat1.matcher(podgrupa);

			if(mat1.find())
			{
				cat = mat1.group(1);
			}

			strnew = new String[]{lemma,cat};
			if (alreadyIn(stringList, strnew))
			{
				stringList.add(strnew);
			}

		}//while (mat.find())
	}
	else
	{
		String cat1;

		while (mat.find())
		{
			lemma = mat.group(1);

			cat1=mat.group(3);
			String podgrupa=mat.group(2);

			// Ako group(1) sadrzi : ili +
			String regx1= "\\.((\\w)*)[:|\\+]";
			Pattern pat1 = Pattern.compile(regx1);
			Matcher mat1 = pat1.matcher(podgrupa);

			if(mat1.find())
			{
				cat1 = mat1.group(1);
			}

			strnew = new String[]{lemma,cat1};
			if (alreadyIn(stringList, strnew))
			{
				stringList.add(strnew);
			}
		}//while (mat.find())
	}
}
```

```
                }

                /*
                * Ako nijedna linija ne sadrzi rec
                * dodeljuje se UNK - nepoznata kategorija
                */
                if (ind == 0)
                {
                        cat = "UNK";
                        lemma = "UNK";
                        //System.out.println("Kategorija: "+cat +"\n" );
                        strnew = new String[]{lemma,cat};
                        if (alreadyIn(stringList, strnew))
                        {
                                stringList.add(strnew);
                        }
                }

                taggedSentence.add(stringList);

        }//oneStepMultiCat(String word, List taggedSentence, String docName)

        /*
         * Ako se par {cat,lemma} vec ukljucen u listu povratna vrednost je false
         * inace true
         */
        protected boolean alreadyIn(List<String[]> list, String[] str)
        {
                for(int i=0; i<list.size(); i++)
                {
                 if( list.get(i)[0].equals(str[0]) && list.get(i)[1].equals(str[1]))
                        {
                                return false;
                        }
                }
                return true;
        }// alreadyIn(List<String[]> list, String[] str)

}
```

ПРИЛОГ Д

Извод са примерима куварских термина из електронског морфолошког речника простих речи:

```
adobo,N1067+Food+Conc+Course+DOM=Culinary+FLX=N1067
akvavit,N1001+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1001
aniset,N1001+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1001
aniseta,N623+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N623
anisonka,N623+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N623
arak,N1001+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1001
armanjak,N9+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N9
asimina,N600+Bot+DOM=Bio+DOM=Bot+Food+Conc+Alim+DOM=Culinary+FLX
=N600
benediktin,N1001+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1001
branč,N28+Food+Meal+DOM=Culinary+FLX=N28
bri,N1063+Food+Conc+Course+DOM=Culinary+FLX=N1063
čili,N1063+Food+Conc+Course+DOM=Culinary+FLX=N1063
čizburger,N1+Food+Conc+Course+DOM=Culinary+FLX=N1
burger,N1+Food+Conc+Course+DOM=Culinary+FLX=N1
bulgur,N1001+Food+Conc+Prod+DOM=Culinary+FLX=N1001
dispenser,N1+Conc+Uten+DOM=Culinary+FLX=N1
drambui,N1063+Food+Conc+Drink+Prod+Erg+DOM=Culinary+FLX=N1063
džambalaja,N600+Food+Conc+Course+DOM=Culinary+FLX=N600
falafel,N1001+Food+Conc+Course+DOM=Culinary+FLX=N1001
fondi,N1063+Food+Conc+Course+DOM=Culinary+FLX=N1063
fondu,N1067+Food+Conc+Course+DOM=Culinary+FLX=N1067
fritata,N600+Food+Conc+Course+DOM=Culinary+FLX=N600
gaspaćo,N1067+Food+Conc+Course+DOM=Culinary+FLX=N1067
hemendeks,N1001+Food+Conc+Course+DOM=Culinary+FLX=N1001
hotdog,N81+Food+Conc+Course+DOM=Culinary+FLX=N81
kebab,N1+Food+Conc+Course+DOM=Culinary+FLX=N1
```

kirš,N1001+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1001
konzome,N1069+Food+Conc+Course+DOM=Culinary+FLX=N1069
kurasao,N1069+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1069
koantro,N1069+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1069
kuantro,N1069+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1069
medovača,N600+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N600
minestrone,N1069+Food+Conc+Course+DOM=Culinary+FLX=N1069
najgvirc,N1+Bot+DOM=Bio+DOM=Bot+Food+Conc+Alim+DOM=Culinary+FLX=N1
rendalica,N650+Conc+Uten+DOM=Culinary+FLX=N650
sake,N1069+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1069
saki,N1063+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1063
sašimi,N1063+Food+Conc+Course+DOM=Culinary+FLX=N1063
skoč,N1027+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1027
složenac,N41+Food+Conc+Course+DOM=Culinary+FLX=N41
suvlaki,N1063+Food+Conc+Course+DOM=Culinary+FLX=N1063
šnaps,N1001+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1001
tekila,N1+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N1
terijaki,N1063+Food+Conc+Course+DOM=Culinary+FLX=N1063
tokajac,N41+Food+Conc+Drink+Prod+DOM=Culinary+FLX=N41
vok,N297+Conc+Uten+DOM=Culinary+FLX=N297
zaimača,N600+Conc+Uten+DOM=Culinary+FLX=N600
zester,N1+Conc+Uten+DOM=Culinary+FLX=N1
bapka,N660+Food+Conc+Prod+DOM=Culinary+FLX=N660
baći,N63+Food+Conc+Course+DOM=Culinary+Erg+FLX=N63

Извод са примерима куварских термина из електронског морфолошког речника полилексичких јединица:

```
aćeto balzamiko(balzamiko.N67:ms1q),
NC_2XN3+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_2XN3

biljni(biljni.A2:adms1g) želatin(želatin.N1:ms1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

azuki pasulj(pasulj.N27:ms1q),
NC_2XN+Bot+DOM=Bot+DOM=Bio+Conc+Food+Alim+DOM=Culinary+Comp+Flj=
NC_2XN

balzamiko sirće(sirće.N314:ns1q),
NC_2XNr+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_2XNr

čokoladna(čokoladni.A2:aefs1g) bananica(bananica.N650:fs1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

krem bananica(bananica.N650:fs1q),
NC_2XN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_2XN

čoko bananica(bananica.N650:fs1q),
NC_2XN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_2XN

basmati riža(riža.N600:fs1q),
NC_2XN3+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_2XN3

agar-agar(agar.N1:ms1q),
NC_2XN3+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_2XN3

bolonjez sos(sos.N297:ms1q),
NC_2XNr+Conc+Food+Course+DOM=Culinary+Comp+Flj=NC_2XNr

špagete(špagete.N624:fp1q) bolonjeze,
NC_N2X2+Conc+Food+Course+DOM=Culinary+Comp+Flj=NC_N2X2

bolonjeze sos(sos.N297:ms1q),
NC_2XNr+Conc+Food+Course+DOM=Culinary+Comp+Flj=NC_2XNr

dezertno(dezertni.A2:aens1g) vino(vino.N300:ns1q),
NC_AXN+Comp+Conc+DER=ZS+Drink+Food+Prod+DOM=Culinary+Flj=NC_AXN

dezertni(dezertni.A2:adms1g) tanjir(tanjir.N2:ms1v),
NC_AXN+Conc+Uten+DER=ZS+DOM=Culinary+Comp+Flj=NC_AXN

desertni(desertni.A2:adms1g) tanjir(tanjir.N2:ms1v),
NC_AXN+Conc+Uten+DER=SZ+DOM=Culinary+Comp+Flj=NC_AXN

ekspandirana(ekspandirani.A2:fs1gae) riža(riža.N600:fs1q),
NC_AXN3+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN3
```

```
gril tava(tava.N600:fs1q),
NC_2XN+Conc+Uten+DOM=Culinary+Comp+Flj=NC_2XN

heljdino(heljdin.A1:aens1g) brašno(brašno.N300:ns1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

jelovna(jelovni.A2:fs1gae) kašika(kašika.N612:fs1q),
NC_AXN+Comp+Conc+Uten+DOM=Culinary+Flj=NC_AXN

lorberov(lorberov.A1:akms1g) list(list.N81:ms1q),
NC_AXN+Conc+Food+Alim+DOM=Culinary+Comp+Flj=NC_AXN

njeguški(njeguški.A2:adms1g) pršut(pršut.N1:ms1q),
NC_AXN+Conc+Food+Prod+Erg+DOM=Culinary+Comp+Flj=NC_AXN

njeguška(njeguški.A2:aefs1g) pršuta(pršuta.N600:fs1q),
NC_AXN+Conc+Food+Prod+Erg+DOM=Culinary+Comp+Flj=NC_AXN

orašast(orašast.A6:akms1g) plod(plod.N81:ms1q),
NC_AXN+Bot+DOM=Bot+DOM=Bio+Conc+Food+Alim+DOM=Culinary+Comp+Flj=
NC_AXN

orašasto(orašast.A6:aens1g) voće(voće.N308:ns1q),
NC_AXN3+Bot+DOM=Bot+DOM=Bio+Conc+Food+Alim+DOM=Culinary+Comp+Flj
=NC_AXN3

ovsene(ovseni.A2:aefp1g) pahuljice(pahuljice.N624:fp1q),
NC_AXN3+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN3

peršunov(peršunov.A1:akms1g) list(list.N81:ms1q),
NC_AXN+Conc+Food+Alim+DOM=Culinary+Comp+Flj=NC_AXN

peršunov(peršunov.A1:akms1g) listić(listić.N1:ms1q),
NC_AXN+Conc+Food+Alim+DOM=Culinary+Comp+Flj=NC_AXN

peršunovo(peršunov.A1:aens1g) lišće(lišće.N308:ns1q),
NC_AXN3+Conc+Food+Alim+DOM=Culinary+Comp+Flj=NC_AXN3

plećna(plećni.A2:aefs1g) slanina(slanina.N600:fs1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

poširano(poširan.A6:aens1g) jaje(jaje.N314:ns1q),
NC_AXN+Conc+Food+Course+DOM=Culinary+Comp+Flj=NC_AXN

povrćna(povrćni.A2:aefs1g) kocka(kocka.N602:fs1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

rikota sir(sir.N83o:ms1q),
NC_2XN3+Conc+Food+Prod+Comp+DOM=Culinary+Flj=NC_2XN3

feta sir(sir.N83o:ms1q),
NC_2XN3+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_2XN3
```

```
roštiljska(roštiljski.A2:aefs1g) kobasica(kobasica.N650:fs1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

salatni(salatni.A6:akms1g) preliv(preliv.N1:ms1q),
NC_AXN+Conc+Food+Prod+Ek+DOM=Culinary+Comp+Flj=NC_AXN

sipino(sipin.A1:aens1g) mastilo(mastilo.N300:ns1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

komova(komov.A1:aefs1g) rakija(rakija.N600:fs1q),
NC_AXN+Conc+Food+Prod+Drink+DOM=Culinary+Comp+Flj=NC_AXN

sirni(sirni.A2:adms1g) namaz(namaz.N1:ms1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

stišnjena(stišnjen.A6:aefs1g) šunka(šunka.N748:fs1q),
NC_AXN+Conc+Food+Prod+DOM=Culinary+Comp+Flj=NC_AXN

supena(supeni.A2:aefs1g) kašika(kašika.N612:fs1q),
NC_AXN+Conc+Uten+DOM=Culinary+Comp+Flj=NC_AXN

šam-rolna(rolna.N660:fs1q),
NC_2XN6+Conc+Food+Course+DOM=Culinary+Comp+Flj=NC_2XN6

šaum-rolna(rolna.N660:fs1q),
NC_2XN6+Conc+Food+Course+DOM=Culinary+Comp+Flj=NC_2XN6

šitaki(šitaki.N3001:mp1q) gljive,
NC_N2X2+Bot+DOM=Bot+DOM=Bio+Conc+Food+Alim+DOM=Culinary+Comp+Flj
=NC_N2X2

šitaki(šitaki.N3001:mp1q) pečurke,
NC_N2X2+Bot+DOM=Bot+DOM=Bio+Conc+Food+Alim+DOM=Culinary+Comp+Flj
=NC_N2X2

štapni(štapni.A2:adms1g) mikser(mikser.N1:ms1q),
NC_AXN+Conc+Uten+DOM=Culinary+Comp+Flj=NC_AXN
```

Извод из електронског морфолошког речника кулинарских апроксимативних мера:

```
na vrh noža,ADV+MesApp+Unit+DOM=Culinary+Comp+Flj=ADV
kocka,N602+Wh+MesApp+DOM=Culinary+Flj=N602
kockica,N650+Dem+Por+MesApp+DOM=Culinary+Flj=N650
štapić,N27+Wh+MesApp+DOM=Culinary+Flj=N27
zrnce,N330+Part+MesApp+DOM=Culinary+Flj=N330
zrno,N300+Part+MesApp+DOM=Culinary+Flj=N300
prut,N89+Wh+MesApp+DOM=Culinary+Flj=N89
sloj,N83+Por+MesApp+DOM=Culinary+Flj=N83
traka,N612+Part+MesApp+DOM=Culinary+Flj=N612
kora,N600+MesApp+Part+Wh+DOM=Culinary+Flj=N600
deo,N93+Ek+Part+MesApp+DOM=Culinary+Flj=N93
dio,N2093+Ijk+Part+MesApp+DOM=Culinary+Flj=N2093
sredina,N600+Part+MesApp+DOM=Culinary+Flj=N600
kraj,N83+Por+MesApp+DOM=Culinary+Flj=N83
odrezak,N25+Por+MesApp+DOM=Culinary+Flj=N25
parče,N310+Por+MesApp+DOM=Culinary+Flj=N310
polutka,N728+Part+MesApp+DOM=Culinary+Flj=N728
režanj,N105+Por+MesApp+DOM=Culinary+Flj=N105
filet,N1+Por+MesApp+DOM=Culinary+Flj=N1
ploška,N724+Por+MesApp+DOM=Culinary+Flj=N724
grudvica,N650+Sr+Por+MesApp+DOM=Culinary+Flj=N650
kap,N696+Por+MesApp+DOM=Culinary+Flj=N696
kolut,N89+Por+MesApp+DOM=Culinary+Flj=N89
kolutić,N27+Por+MesApp+DOM=Culinary+Flj=N27
kriška,N632+Por+MesApp+DOM=Culinary+Flj=N632
šnicla,N660+Sr+Por+MesApp+DOM=Culinary+Flj=N660
vrh,N292+Por+MesApp+DOM=Culinary+Flj=N292
rolna,N660+Wh+MesApp+DOM=Culinary+Flj=N660
tabla,N660+Wh+MesApp+DOM=Culinary+Flj=N660
vekna,N660+Sr+Wh+MesApp+DOM=Culinary+Flj=N660
cvet,N81+Ek+Part+MesApp+DOM=Culinary+Flj=N81
cvijet,N2081+Ijk+Part+MesApp+DOM=Culinary+Flj=N2081
```

cvetić,N27+Ek+Part+MesApp+DOM=Culinary+Flj=N27
cvjetić,N27+Ijk+Part+MesApp+DOM=Culinary+Flj=N27
čen,N81+Part+MesApp+DOM=Culinary+Flj=N81
češanj,N87+Part+MesApp+DOM=Culinary+Flj=N87
glava,N600+Part+MesApp+DOM=Culinary+Flj=N600
glavica,N650+Part+MesApp+DOM=Culinary+Flj=N650
grančica,N650+Part+MesApp+DOM=Culinary+Flj=N650
iglica,N650+Part+MesApp+DOM=Culinary+Flj=N650
koren,N89+Ek+MesApp+Part+DOM=Culinary+Flj=N89
korijen,N89+Ijk+MesApp+Part+DOM=Culinary+Flj=N89
korica,N650+Part+MesApp+DOM=Culinary+Flj=N650
list,N81+Part+MesApp+DOM=Culinary+Flj=N81
listić,N1+Part+MesApp+DOM=Culinary+Flj=N1
perce,N312+Dem+Part+MesApp+DOM=Culinary+Flj=N312
plod,N81+Part+MesApp+DOM=Culinary+Flj=N81
rebro,N330+Part+MesApp+DOM=Culinary+Flj=N330
stabljika,N612+Part+MesApp+DOM=Culinary+Flj=N612
čaša,N600+Cont+MesApp+DOM=Culinary+Flj=N600
čašica,N600+Cont+MesApp+DOM=Culinary+Flj=N600
čep,N81+Cont+MesApp+DOM=Culinary+Flj=N81
flaša,N600+Cont+MesApp+DOM=Culinary+Flj=N600
flašica,N650+Cont+MesApp+DOM=Culinary+Flj=N650
kašičica,N650+Sr+Cont+MesApp+DOM=Culinary+Flj=N650
kašika,N612+Sr+Cont+MesApp+DOM=Culinary+Flj=N612
kesa,N600+Cont+MesApp+DOM=Culinary+Flj=N600
kesica,N650+Cont+MesApp+DOM=Culinary+Flj=N650
kutija,N600+Cont+MesApp+DOM=Culinary+Flj=N600
kutlača,N600+Cont+MesApp+DOM=Culinary+Flj=N600
paket,N1+Cont+MesApp+DOM=Culinary+Flj=N1
paketić,N27+Cont+MesApp+DOM=Culinary+Flj=N27
šolja,N600+Cont+MesApp+DOM=Culinary+Flj=N600
šoljica,N650+Cont+MesApp+DOM=Culinary+Flj=N650
tanjir,N2+Sr+Cont+MesApp+DOM=Culinary+Flj=N2
tanjur,N2+Cr+Cont+MesApp+DOM=Culinary+Flj=N2
tegla,N660+Cont+MesApp+DOM=Culinary+Flj=N660
teglica,N650+Cont+MesApp+DOM=Culinary+Flj=N650

```
varjača,N600+Cont+MesApp+DOM=Culinary+Flj=N600
vrećica,N650+Cont+MesApp+DOM=Culinary+Flj=N650
žlica,N650+Cr+Cont+MesApp+DOM=Culinary+Flj=N650
disk,N297+Por+MesApp+DOM=Culinary+Flj=N297
komad,N1+Por+MesApp+DOM=Culinary+Flj=N1
komadić,N27+Por+MesApp+DOM=Culinary+Flj=N27
konzerva,N660+Por+MesApp+DOM=Culinary+Flj=N660
limenka,N628+Por+MesApp+DOM=Culinary+Flj=N628
krug,N297+Por+MesApp+DOM=Culinary+Flj=N297
lopta,N660+Por+MesApp+DOM=Culinary+Flj=N660
loptica,N650+Por+MesApp+DOM=Culinary+Flj=N650
pakovanje,N300+Por+MesApp+DOM=Culinary+Flj=N300
ploča,N600+Por+MesApp+DOM=Culinary+Flj=N600
šipka,N632+Por+MesApp+DOM=Culinary+Flj=N632
šnita,N600+Por+MesApp+DOM=Culinary+Flj=N600
štangla,N660+Por+MesApp+DOM=Culinary+Flj=N660
štanglica,N650+Por+MesApp+DOM=Culinary+Flj=N650
red,N81+Set+MesApp+DOM=Culinary+Flj=N81
struk,N297+Part+MesApp+DOM=Culinary+Flj=N297
šaka,N612+Cont+MesApp+DOM=Culinary+Flj=N612
veza,N600+Set+MesApp+DOM=Culinary+Flj=N600
vezica,N650+Set+MesApp+DOM=Culinary+Flj=N650
kriškica,N650+Por+MesApp+DOM=Culinary+Flj=N650
šnitica,N650+Por+MesApp+DOM=Culinary+Flj=N650
režnjić,N27+Por+MesApp+DOM=Culinary+Flj=N27
jufkica,N650+MesApp+Part+DOM=Culinary+Flj=N650
vrč,N83+Cont+MesApp+DOM=Culinary+Flj=N83
jufka,N660+MesApp+Part+DOM=Culinary+Flj=N660

supena(supeni.A2:aefs1g) kašika(kašika.N612:fs1q),
NC_AXN+Cont+MesApp+DOM=Culinary+Comp+Flj=NC_AXN

šoljica(šoljica.N650:fs1q) za kafu,
NC_N4X+Cont+MesApp+DOM=Culinary+Comp+Flj=NC_N4X

jelovna(jelovni.A2:fs1gae) kašika(kašika.N612:fs1q),
NC_AXN+Cont+MesApp+DOM=Culinary+Comp+Flj=NC_AXN

čajna(čajni.A2:aefs1g) kašika(kašika.N612:fs1q),
NC_AXN+DOM=Culinary+Comp+Cont+MesApp+Flj=NC_AXN
```

```
kafena(kafeni.A2:aefs1g) kašičica(kašičica.N650:fs1q),
NC_AXN+DOM=Culinary+Comp+Cont+MesApp+Flj=NC_AXN

kafena(kafeni.A2:aefs1g) kašika(kašika.N612:fs1q),
NC_AXN+DOM=Culinary+Comp+Cont+MesApp+Flj=NC_AXN

kafena(kafeni.A2:aefs1g) šolja(šolja.N600:fs1q),
NC_AXN+DOM=Culinary+Comp+Cont+MesApp+Flj=NC_AXN

kafena(kafeni.A2:aefs1g) šoljica(šoljica.N650:fs1q),
    NC_AXN+DOM=Culinary+Comp+Cont+MesApp+Flj=NC_AXN
```

ПРИЛОГ Ћ

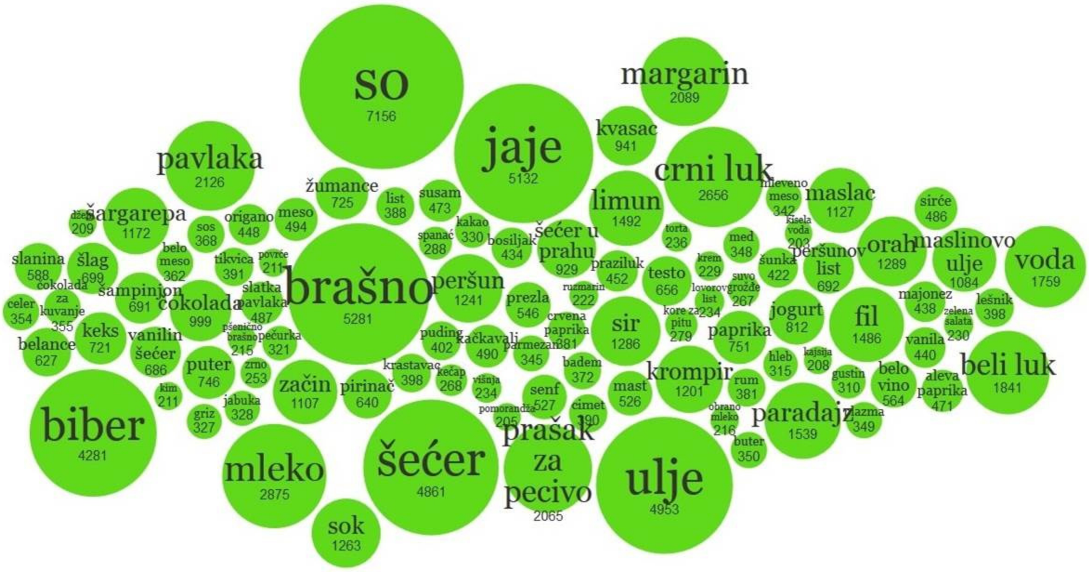
За анализу фреквенција састојака који се јављају у рецептима из кулинарског корпуса прикупљених за потребе овог рада направљени су Java програми који на основу развијених лексичких ресурса израчунавају фреквенције појединачних састојака и фреквенције парова састојака.

На слици 46 приказане су фреквенције сто најфреквентнијих састојака. Може се видети да су *со* (7.156), *брашно* (5.281), *јаје* (5.132), *уље* (4.953) и *шећер* (4.861) састојци који се најчешће појављују приликом набрајања намирница потребних за припрему рецепата.

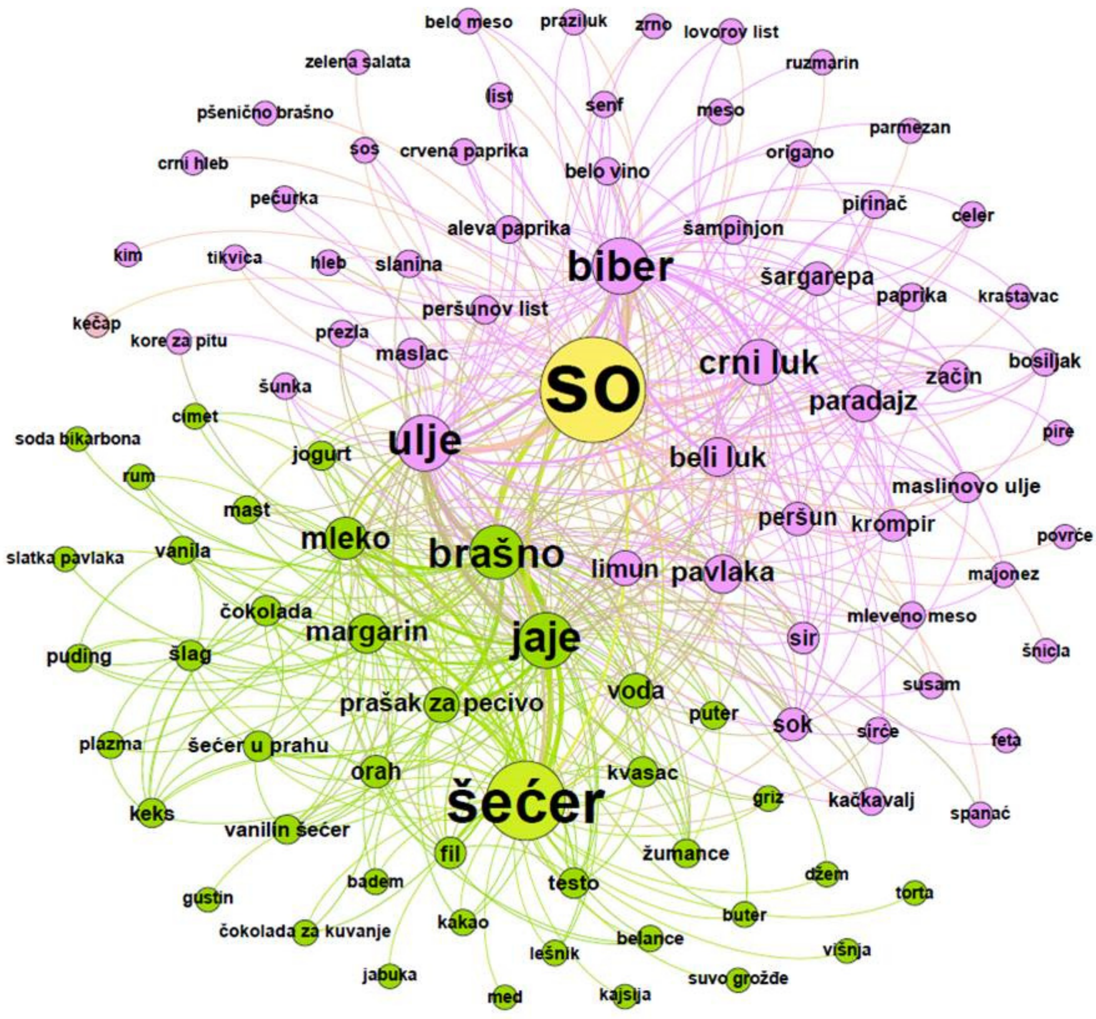
На слици 47 приказано је првих 101 састојака који се појављују у паровима у анализираним рецептима. Ови састојци граде са 504 међусобне релације. Специјално, слика 48 приказује састојке са којима се у рецептима најчешће појављује *лук*.

За визуелизацију фреквенција је коришћен пројекат *Bubble Cloud*[77], а за визуелизацију парова састојака *Gephi*[78] платформа за приказ различитих врста мрежа и графова.

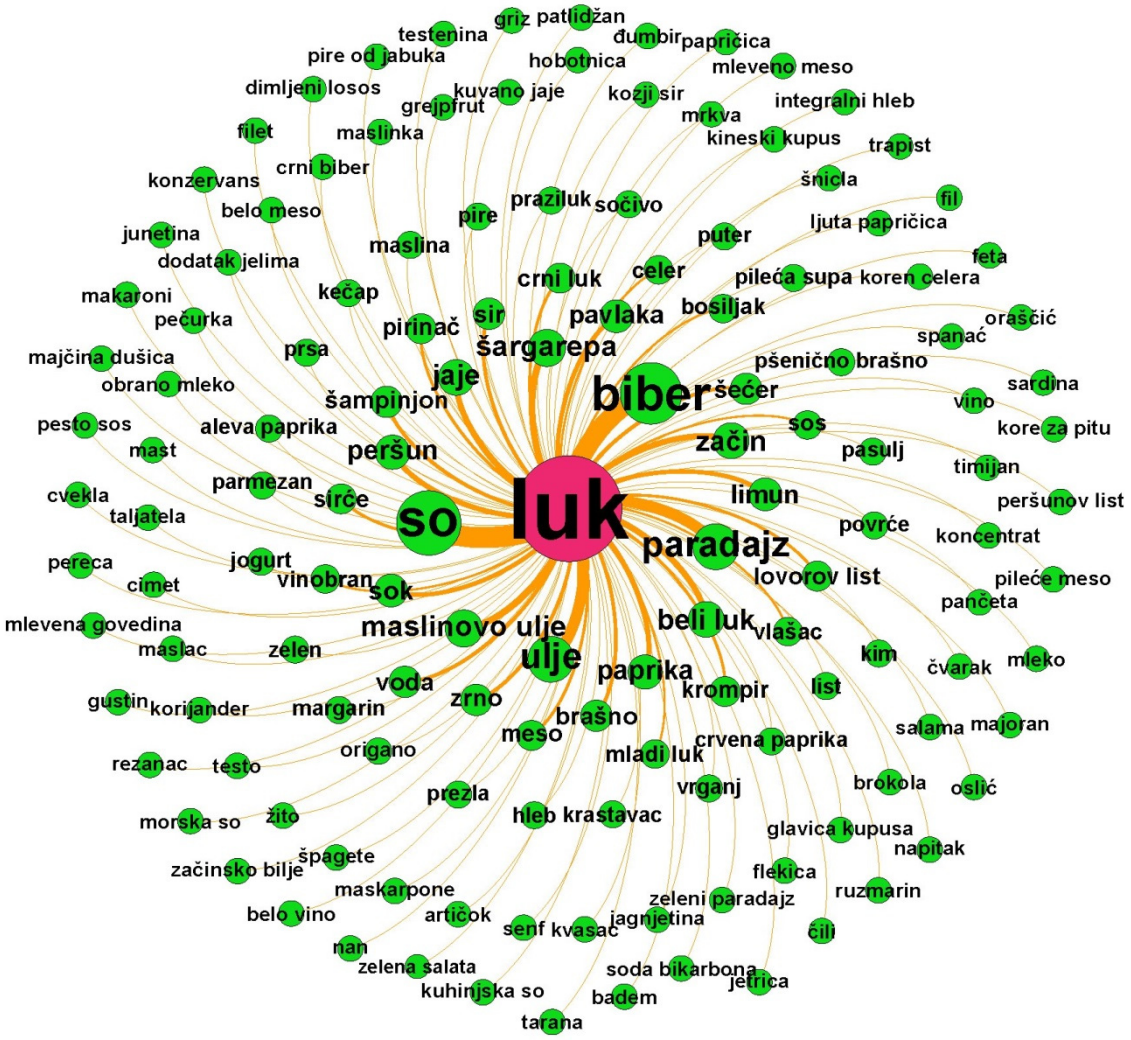
[77] Bubble Cloud пројекат: http://vallandingham.me/building_a_bubble_cloud.html.
[78] Gephi платформа: http://gephi.github.io/.



Слика 46. Фреквенције сто најфреквентнијих састојака у кулинарском корпусу.



Слика 47. Парови састојака који се најчешће јављају заједно у рецептима кулинарског корпуса.



Слика 48. Састојци који се најчешће јављају у пару са састојком лук у рецептима кулинарског корпуса.

* ТЕХНИЧКА НАПОМЕНА: У штампаној верзији текста дисертације стране број 10, 41, 77, 93, 107, 119, 127, 128, 129, 131, 132, 134, 135, 136, 137, 138, 139, 141, 142, 143, 145, 146, 155, 156, 157, 211 и 212 су, за разлику од осталих страна, штампане у боји. Због тога се боја текста на њима разликује у односу на боју текста на осталим странама.

БИОГРАФИЈА АУТОРА

Сташа Вујичић Станковић рођена је у Београду, 1982. године. Завршила је Основну школу „Бранко Ћопић“ као носилац диплома „Вук Караџић“ и „Ђак генерације“. Након тога је завршила Математичку гимназију у Београду и Математички факултет у Београду, где је дипломирала на смеру „Рачунарство и информатика“ са просечном оценом 9,39. Током школовања је била носилац стипендије Фонда за младе таленте Министарства просвете Републике Србије.

Од 2007. године запослена је на Математичком факултету Универзитета у Београду, као Сарадник у настави на Катедри за рачунарство и информатику, а од 2009. године као Асистент у настави. Држала је вежбе из следећих курсева: Програмирање 1, Програмирање 2, Објектно оријентисано програмирање, Информациони системи, Увод у организацију рачунара, Основи управљања и Управљање пројектима у индустрији и науци.

Основне области интересовања су јој обрада природних језика, екстракција информација, базе података, претраживање информација и истраживање веба. Учесник је научних пројеката „Српски језик и његови ресурси: теорија, опис и примене“ и „Инфраструктура за електронски подржано учење у Србији“ које финансира Министарство науке Републике Србије. Као резултат научног рада, објавила је већи број научних радова и учествовала на неколико међународних конференција.

Од 2014. године члан је и један од оснивача Друштва за језичке ресурсе и технологије.

Прилог 1.

Изјава о ауторству

Потписана Сташа И. Вујичић Станковић

број уписа 2003/2007

Изјављујем

да је докторска дисертација под насловом

Екстракција информација вођена онтологијама (Модел за српски језик)

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 01.05. 2016.

Сташа Станковић

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Сташа И. Вујичић Станковић

Број уписа 2003/2007

Студијски програм Рачунарство и информатика

Наслов рада

Екстракција информација вођена онтологијама (Модел за српски језик)

Ментор проф. др Душко Витас

Потписани Сташа И. Вујичић Станковић

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 01.05.2016.

С.Вујичић Станковић

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Екстракција информација вођена онтологијама (Модел за српски језик)

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

(2.) Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 01.05.2016.

Вујичић Станковић

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.