УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЛОЛОШКИ ФАКУЛТЕТ

Јелена Д. Митровић

ЕЛЕКТРОНСКИ ЈЕЗИЧКИ РЕСУРСИ И АЛАТИ ЗА ОБРАДУ СРПСКОГ ЈЕЗИКА И ЊИХОВО УНАПРЕЂИВАЊЕ ПУТЕМ МОДЕЛА ГРУПНЕ РАСПОДЕЛЕ РАДА

Докторска дисертација

Београд, 2018

UNIVERZITET U BEOGRADU
FILOLOŠKI FAKULTET

Jelena D. Mitrović

ELEKTRONSKI JEZIČKI RESURSI I ALATI ZA OBRADU SRPSKOG JEZIKA I NJIHOVO UNAPREĐIVANJE PUTEM MODELA GRUPNE RASPODELE RADA

Doktorska disertacija

Beograd, 2018

UNIVERSITY OF BELGRADE
FACULTY OF PHILOLOGY

Jelena D. Mitrović

ELECTRONIC LEXICAL RESOURCES AND TOOLS FOR NATURAL LANGUAGE PROCESSING OF SERBIAN AND THEIR IMPROVEMENT VIA CROWDSOURCING

Doctoral dissertation

Belgrade, 2018

Ментор:

Проф. др Цветана Крстев, редовни професор, Универзитет у Београду, Филолошки факултет

Чланови комисије:

Проф. др Александра Вранеш, редовни професор, Универзитет у Београду, Филолошки факултет

Проф. др Ранка Станковић, ванредни професор, Универзитет у Београду, Рударско-геолошки факултет

Датум одбране: _ _ _ _ _ _ _ _ _ _

Изјаве захвалности

Највећу захвалност дугујем својој менторки, проф. др Цветани Крстев, захваљујући којој сам се заинтересовала за област обраде природног језика и која ме је својом безрезервном подршком и стрпљењем увек инспирисала и охрабривала да истрајем. Велико хвала проф. др Ранки Станковић и проф. др Александри Вранеш на конструктивним критикама и коментарима захваљујући којима је финална верзија ове тезе знатно побољшана.

Хвала мами, тати и сестри, који се мојим успесима највише радују и Миљану, који ме је одувек безрезервно подржавао и веровао у мене, често стављајући моје потребе испред сопствених. Бољег партнера од њега не могу да замислим.

За Константина и Ану

Наслов дисертације: Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада

Резиме: Овај докторски рад бави се истраживањем најбољег начина да се унапреде језички ресурси и алати за обраду српског језика, са нагласком на обради фигуративног језика, то јест језика који је богат реторичким фигурама. У раду су представљени језички ресурси и алати за обраду српског језика, нарочито Ворднет и Српски ворднет, чија доградња је централни део овог истраживања. Та доградња је спроведена путем модела такозване групне расподеле рада, како смо за потребе израде овог рада превели енглески термин *crowdsourcing*.

Предложени метод користи анотирани корпус савременог српског језика за истраживање семантичког знања садржаног у лингвистичким конструкцијама које имају улогу реторичке фигуре поређење. На основу фреквенције појављивања поређења у корпусу, предложили смо додавање пара нових релација, које повезују синсет именице и синсет придева који представља карактеристични атрибут те именице. Овај приступ је проверен методом провере оцењивача, говорника српског језика, путем модела групне расподеле рада како би валидност аутоматске методе била процењена. Из Корпуса савременог српског језика добили смо кандидате реторичке фигуре поређење, у облику „ПРИДЕВ као ИМЕНИЦА“ које смо кроз пројекте групне расподеле рада понудили говорницима српског језика како бисмо добили одговор на питање: „Која поређења се највише користе у савременом српском говорном језику?“. Резултатима тог истраживање допунили смо Српски ворднет, полуаутоматском методом која се заснива на познавању структуре ове лексичко-семантичке мреже која је веома важан ресурс у обради српског језика. Структура Српског ворднета је тако обогаћена паром нових, инверзних релација које смо назвали *specificOf/specifiedBy,* које повезују именичке и придевске синсетове у Српском ворднету чији литерали повезани везником *као* представљају поређења која су резултат претходно поменутих пројеката групне расподеле рада.

Овако унапређена структура Српског ворднета искоришћена је у процесу аутоматског препознавања реторичких фигура ироније и сарказма у корпусу кратких порука добијених са друштвене мреже Твитер (енг. Twitter). Методологија коришћена у пројектима групне расподеле рада примењена је у сличним истраживањима на савременом грчком језику.

У првом поглављу овог докторског рада говоримо о предмету истраживања, циљевима и очекиваним резултатима. Друго поглавље посвећено је језичким ресурсима и алатима, њиховом опису и примени, нарочито Ворднету и Српском ворднету. У трећем поглављу дајемо опис модела групне расподеле рада, жанрова у оквиру овог модела, његове примене у области обраде природног језика и система провере квалитета резултата добијених у пројектима заснованим на групној расподели рада. Четврто поглавље је посвећено главном истраживању спорведеном у склопу овог докторског рада, реторичким фигурама и њиховој улози у обради природног језика и процесу проширења Српског ворднета новим семантичким релацијама заснованим на реторичкој фигури поређење. У петом поглављу говоримо о примени унапређеног Српског ворднета и научној дисеминацији резултата овог доктората. Коначно, у шестом поглављу износимо закључке и дајемо предлоге за будући рад.

Кључне речи: језичких ресурси, српски језик, обрада природног језика, ворднет, електронски речници, корпус српског језика, групна расподела рада, реторичке фигуре, поређење

Научна област: Библиотекарство и информатика
Ужа научна област: Библиотечка информатика, Обрада природног језика
УДК број: 026/027:004.62(043.3)

Dissertation title: Electronic Lexical Resources and Tools for Natural Language Processing of Serbian and their Enhancement via Crowdsourcing

Abstract: The research presented in this doctoral thesis deals with finding the best way to enhance lexical resources and tools for natural language processing of Serbian language, with a focus on processing of figurative language, i.e. language in which plenty of rhetorical figures are used. The thesis presents lexical resources and tools for natural language processing of Serbian language, especially WordNet and Serbian wordnet, whose enhancment and building via the crowdsourcing model are the central part of the presented research results.

In the suggested method, we use an annotated corpus of contemporary Serbian language to investigate semantic knowledge contained in the linguistic constructs which have a role of the rhetorical figure called simile. Based on the frequency of occurrence of simile in the corpus, we suggested adding of a pair of new relations which connect a noun synset with a synset of an adjective representing the characteristic attribute of that noun. This approach has been validated via the crowdsourcing model in which native Serbian speakers were involved. From the Corpus of contemporary Serbian language we obtained candidates of the rhetorical figure simile, in the format "ADJECTIVE as NOUN", which we have offered to native Serbian speakers via the crowdsourcing project in order to obtain answer to the question: "Which similes are used the most in the contemporary, spoken Serbian language?". Results of this research were used to enhance the Serbian wordnet, utilizing a semi-automatic method based on the structural knowledge of this lexico-semantic network which is a very important resource in natural language processing of Serbian. The structure of Serbian wordnet has been enhanced with a pair of new, inverse relations we have named *specificOf/specifiedBy*, which connect noun and adjective synsets in Serbian wordnet whose literals are connected with a conjuction *као (ser. as)* in such a way that they represent similes which are the result of the previously mentioned crowdsourcing projects.

The newly enhanced structure of the Serbian wordnet has been used in the process of automatic detection of rhetorical figures irony and sarcasm from the corpus of short messages collected from the social network Twitter. The methodology that was used in the crowdsourcing projects was also used in similar research projects focusing on linguistic investigations of contemporary Greek language.

In the first chapter of this thesis, we discuss the research questions, goals and expected results. The second chapter is dedicated to lexical resources and tools, their description and usage, WordNet and Serbian WordNet in particular. In the third chapter, we describe the

crowdsourcing model, its genres, usage in natural language processing and systems of quality control of the results obtained in crowdsourcing projects. The fourth chapter is dedicated to the main research implemented in this thesis, rhetorical figures and their role in natural language processing, as well as to the process of enhancement of Serbian wordnet with new semantic relations based on the rhetorical figure simile. In the fifth chapter we talk about the usage of the enhanced Serbian wordnet and scientific dissemination of the results presented in this thesis. Finally, the sixth chapter gives conclusions and discusses future research directions.

Keywords: Lexical resources, Serbian language, Natural Language Processing, WordNet, Electronic dictionaries, Corpus of Serbian language, Crowdsourcing, Rhetorical Figures, Simile

Research area: Library Science and Information science
Research subarea: Information science, Natural Language Processing
UDC number 026/027:004.62(043.3)

Название диссертации: Электронные лексические ресурсы и инструменты для обработки естественного языка на примере сербского языка и их совершенствование с помощью краудсорсинга

Аннотация: Цель представленной диссертации – найти наилучший способ улучшить лексические ресурсы и инструменты для обработки естественного языка, в частности сербского языка, с уделением особого внимания обработке фигуративного языка, т. е. языка, в котором используется множество риторических фигур. В диссертации представлены лексические ресурсы и инструменты для обработки естественного сербского языка, особенно WordNet и сербский WordNet, чьи усовершенствование и построение с помощью модели краудсорсинга являются центральной частью представленных результатов исследований.

В предложенном методе мы используем аннотированный корпус современного сербского языка для исследования семантических знаний, содержащихся в лингвистических конструкциях, которые играют роль риторической фигуры, называемой сравнением. Атрибут понятия этого существительного (существительное, женский) объясняется частотой появления сравнения в корпусе. Этот подход был подтвержден с помощью модели краудсорсинга, в которой участвовали родные сербские ораторы. Из Корпуса современных сербских языков, формат «ADJECTIVE as NOUN», который доступен для родных сербских ораторов через проект краудсорсинга: «Какие сравнения использовало большинство в современном, разговорном сербском языке? ». Результаты этого исследования были сделаны с использованием полуавтоматического метода, основанного на структурных знаниях этой лексико-семантической сети, которая является очень важным ресурсом в обработке естественного сербского языка. Структура сербского wordnet была подготовлена с помощью пары новых обратных отношений, которые были названы specificOf / definedBy, которые соединяют существительные и прилагательные synsets в сербском wordnet, литералы которых связаны с конъюнкцией као (sr они представляют собой сравнения, которые являются результатом ранее упомянутых проектов краудсорсинга.

Недавно расширенная структура сербского wordnet была использована в процессе автоматического обнаружения риторических фигур иронии и сарказма из корпуса коротких сообщений, собранных из социальной сети Twitter. Методология, используемая в исследовательских проектах, выполненных на современном греческом языке.

В первой главе этого тезиса мы обсудим вопросы исследования, цели и ожидаемые результаты. Вторая глава посвящена лексическим ресурсам и инструментам, их описанию и использованию, WordNet и сербскому WordNet в частности. В третьей главе мы описываем модель краудсорсинга, ее жанры, использование в обработке естественного языка и системы контроля качества результатов, полученных в проектах краудсорсинга. Четвертая глава посвящена основным исследованиям в этом тезисе, риторическим фигурам и их роли в обработке естественного языка, а также процессу расширения сербского wordnet с новыми семантическими отношениями, основанными на риторическом изображении. В пятой главе мы говорим об использовании усовершенствованного сербского wordnet и научного распространения представленных здесь результатов. Наконец, в шестой главе даются выводы и обсуждаются будущие направления исследований.

Ключевые слова: Лексические ресурсы, Сербский язык, Обработка естественного языка, WordNet, Электронные словари, Корпус сербского языка, Краудсорсинг, Риторические фигуры, Simile

Область исследований: Библиотечная наука и информатика
Исследовательский подрайон: информатика, обработка естественного языка
Номер УДК 026/027: 004.62 (043.3)

Садржај

Слике

Табеле

1 Увод

Viva vox docet – (лат. жива реч подучава)

1.1 Предмет истраживања

Природни језици, као што су српски, енглески, грчки или немачки језик, представљају велики изазов за рачунарске системе зато што су вишезначни, непрестано се мењају и веома су опсежни. Обрада природног језика (енг. natural language processing – NLP) заснива се на употреби рачунарских и информационо-комуникационих технологија за задатке као што су анализа текста, проналажење информација, екстракција информација, анализа осећања и ставова који преовлађују у текстовима, аутоматска класификација текста, аутоматско превођење, аутоматско препознавање говора и многи други. Сви ти задаци подразумевају постојање добро конструисаних, савремених, што потпунијих електронских језичких ресурса и алата.

Проблем обраде природног језика посматран је различито у рачунарским наукама и у лингвистици. Области које су проистекле из тих истраживања зато су добиле другачије називе и њихови циљеви су различити, али унутар њих постоје многе заједничке особине, праксе и настојања. Рачунарска лингвистика (енг. computational linguistics) је настала из истраживања у области лингвистике и њен фокус је више на теорији, док је обрада природног језика област рачунарске науке и њен фокус је више на примени ресурса и алата у обради природног језика. У овом докторском истраживању бавимо се обрадом природног језика, обухватајући и теорије рачунарске лингвистике.

Системи за обраду природног језика обично користе неко знање садржано у језику. Тако је, на пример, за препознавање говора потребно знање о фонологији и фонетици, то јест о правилима и начину изговора. Знање о морфологији, то јест о облику и начину изградње речи, потребно је за један од основних задатака у обради природног језика, означавање врста речи (енг. part-of-speech tagging – POS tagging). Знање о синтакси потребно је за парсирање језика, то јест за добијање информација о лингвистичкој структури језика на основу неког текста који се даје као улаз програму за парсирање. Лексичка семантика даје информације о значењу речи што је неопходно за задатке попут разрешавања вишезначности (енг. WSD – word sense disambiguation) (Jurafsky & Martin, 2009), (Manning & Schütze, 1999).

Сви делови природног језика могу бити вишезначни. Истраживања из области психолингвистике су још осамдесетих година прошлог века показала да људи веома вешто и лако доносе одлуку о одговарајућем значењу речи у односу на контекст у коме се та реч користи (Charles, 1988). Тачан механизам по коме људи врше одабир одговарајућег значења у случају постојања вишезначности није потпуно јасан. Да би рачунарски програми и системи који могу да обраде природни језик могли да обраде и „разумеју“ те вишезначности потребно је да их на неки начин обучимо и научимо. Поред вишезначности, природни језик се одликује и употребом језичких структура које директно утичу на промену значења. Такав језик називамо фигуративним језиком. Стилске или реторичке фигуре знатно обогаћују језик текста у коме се појављују, али њихово коришћење отежава рачунарску обраду тих текстова.

Лексичко-семантичке мреже су језички ресурси који у себи садрже знање о значењу речи и њиховој међусобној повезаности. Принстонски ворднет (енг. Princeton WordNet), или једноставно Ворднет (енг. WordNet), како се у литератури често наводи, је једна од најпознатијих лексичко-семантичких мрежа (Miller, Fellbaum, Gross, & Miller, 1990). Изградња и побољшање Српског ворднета (Krstev C. , 2008), који је настао на основу Ворднета, централна је тема овог докторског рада. Говорићемо о структури ове лексичко-семантичке мреже, могућностима њене употребе у обради природног језика, њеном развоју и доградњи и предлозима за будуће примене и унапређења, нарочито у обради фигуративног језика.

Постоје многи начини да се изгради лексичко-семантичка мрежа узимајући у обзир важне факторе као што су време, цена, квалитет података и квалитет саме мреже. Поред ручних и аутоматских стратегија, у том процесу све популарнији постају приступи у којима допринос волонтера, углавном корисника интернета, доноси значајан помак у количини сакупљених података, као и у квалитету саме лексичко-семантичке мреже. Овакав приступ доноси ниже цене израде, као и ефикасност, ако се систем за прикупљање доприноса волонтера постави на прави начин. Све је популарнији тренд коришћења онлајн игара са сврхом и других жанрова у оквиру модела групне расподеле рада (енг. crowdsourcing). Пример једне лексичко-семантичке мреже која се успешно гради и унапређује на тај начин је француски пројекат JeuxDeMots (fr. игра речима). Ова мрежа садржи речи и њихове могуће синониме, а семантичка мрежа се гради повезивањем тих речи и синонима помоћу више од педесет релација чију валидацију врше парови играча преко скупа онлајн игара које су сличне играма асоцијација (Zarrouk, Lafourcade i Joubert 2013).

Изузетни напори уложени у развој језичких технологија за српски језик пружају неопходну основу за унапређивање постојећих ресурса и алата, као и за развој нових. Употребом модела групне расподеле рада, тај подухват је могуће убрзати, олакшати и укључити велики број људи, волонтера који деле љубав према сопственом језику и жељу за његовим очувањем. У предложеном истраживању, највише пажње биће посвећено развоју и могућностима унапређивања Српског ворднета који се може користити у многим подобластима обраде природних језика и зато је веома важно да буде што комплетнији, доступнији и лакши за коришћење. Ако је квалитетно изграђен и адекватно се одржава, он може послужити и као основа за унапређивање других језичких ресурса. У истраживањима спроведеним у оквиру израде овог докторског рада, користили смо модел групне расподеле рада за обогаћивање Српског ворднета и учинили га погодним за примене у области обраде фигуративног језика. Методологија примењена у овом раду може се искористити и за обогаћивање ворднетова на другим језицима.

1.2 Циљ истраживања и очекивани резултати

Циљ истраживања представљеног у овом докторском раду је унапређивање постојећих ресурса за рачунарску обраду природног језика, нарочито Српског ворднета, путем модела групне расподеле рада, ради њихове свеобухватније употребе, превасходно у анализи фигуративног језика и проналажењу информација. У том смислу неопходно је остваривање следећих задатака:

- критички преглед постојеће, релевантне литературе у вези са језичким ресурсима и алатима за обраду српског језика као и моделима групне расподеле рада и њиховим особинама;
- прикупљање и припрема материјала који ће бити основа предложеног пројекта;
- одређивање, примена и вредновање методологије за унапређивање језичких ресурса, нарочито Српског ворднета;
- утврђивање методологије спровођења пројекта групне расподеле рада која је најпогоднија у сврху доградње Српског ворднета;
- одређивање методе евалуације података добијених у пројекту групне расподеле рада;
- утврђивање ограничења предложене методологије и пружање смерница за даљи рад.

У складу са тиме, очекивани резултати истраживања представљеног у овом докторском раду су:

- нови модел групне расподеле рада за унапређивање језичких ресурса за српски језик, нарочито Српског ворднета;
- унапређени електронски језички ресурси које је могуће користити у најактуелнијим светским применама из области обраде природног језика – нарочито Српски ворднет и Онтологија реторичких фигура за српски језик, РетФиг.

2 Језички ресурси и алати у обради природног језика

Језички ресурси представљају скупове језичких података и описа у машински читљивом облику који служе за обраду података изражених природним језиком и могу се користити за изградњу, унапређивање или евалуацију алата за обраду природног језика. Обрада природног језика заснива се на постојању широко доступних ресурса. Ти ресурси су корпуси, стандардни скупови ознака за задатке попут означавања врста речи, парсирања, обележавање значења речи, електронски речници, семантичке мреже, лексичке базе, терминолошке листе, лингвистичке онтологије итд.

У раду посвећеном обради природног језика и језичким ресурсима аутори Годфри и Замполи (енг. Godfrey и Zampolli) наводе три главне активности, неопходне за даљи развој области:

- промовисање поновне употребе лексичких ресура, што подразумева прилагођавање важећим стандардима, унапређивање ресурса како би достигли неопходан ниво квалитета, те постизање договора за стављање ресурса у јавни домен;
- промовисање развоја нових језичких ресурса за језике и домене у којима још увек не постоје, или су само на нивоу прототипа или пак нису јавно доступни за кориснике;
- стварање кооперативне инфраструктуре за прикупљање, одржавање и дисеминацију језичких ресурса (Godfrey & Zampolli, 1996).

Једна од организација које се баве свим овим задацима носи назив ELRA (European Language Resources Association)[1] и основана је фебруара 1995. године како би пружила основу за својеврсну централизовану координацију стварања и управљања језичким ресурсима у Европи.

Организација ELRA дефинише језичке ресурсе као „скуп језичких података и описа у машински читљивом облику, који се користе за изградњу, унапређивање или евалуацију природног језика, алгоритама или система за говор, или као основни ресурси за индустрију језичких услуга и локализације софтвера, за студије језика, електронско објављивање, међународне послове, стручњаке у посебним областима и крајње кориснике.“[2]

[1] ELRA http://www.elra.info/en/

[2] „The term Language Resource refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or,

Деведесетих година 20. века настале су бројне значајне организације и удружења која се баве језичким ресурсима и алатима и њиховом улогом у обради природног језика.

ELSNET[3] је европска мрежа посвећена језичким технологијама. Настала је 1991. године као једна од првих европских такозваних мрежа изврсности (енг. European networks of exellence) подржана од стране Европске комисије. Првобитни циљеви ове мреже били су:

- повезивање заједница истраживача и корисника заинтересованих за језик и говор.
- повезивање академије и индустрије.
- олакшавање истраживања и развоја у оквиру језичких технологија.

ELSNET је и данас својеврстан форум посвећен језичким технологијама, превасходно у европском оквиру, који пружа платформу за обуку и дисеминацију информација о језичким технологијама и језичким ресурсима кроз радионице, летње школе, публикације и процену неопходних будућих корака у развоју језичких ресурса, њиховом дељењу и коришћењу, са гледишта потреба истраживања и индустрије.

Једна од првих иницијатива одређивања будућих корака (енг. Roadmapping) донесена је 1998. године на Институту за лингвистику Универзитета Утрехт у Холандији, и носи назив BLARK (eng. Basic language resource kit) (Krauwer, 1998). Идеја је била да све земље Европе, укључујући и оне у којима се говори језицима који нису довољно присутни у области обраде природног језика, добију својеврстан пакет минималних захтева за квалитет језичких ресурса које би требало испунити да би неки језик био адекватно представљен. Први захтев је да за сваки језик постоји писани корпус опште врсте како би се за тај језик могла спроводити језичка истраживања, те је као предлог дата величина од око 10 милиона речи из новинских текстова, означених, то јест анотираних у складу са опште прихваћеним стандардима. Други захтев је испуњавање сличних услова за корпусе изговореног језика. Трећи захтев је постојање основних алата за управљање и анализу корпуса, а последњи, четврти захтев се односи на скуп вештина које су неопходне за почетак развоја конкурентних производа и услуга у индустрији језичких технологија.

У склопу активности мреже ELSNET за језике Европе је утврђено у којој мери су испуњени услови који су постављени BLARK иницијативом, као и који елементи

as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users."

[3] ELSNET http://www.elsnet.org/

недостају – на пример, алати, материјали за обуку итд. Покренуте су и многе акције које су за резултат имале успостављање редовних обука и летњих школа, као и покретање пројеката за испуњавање основних предуслова неопходних за постојање и коришћење језичких ресурса и алата за све европске језике.

Удружење ELRA/ELDA[4] (енг. European language resources association), основано је 1995. године на сличним принципима као и мрежа ELSNET. Основни циљ ове непрофитне организације јесте обезбеђивање приступа језичким ресурсима за језичке технологије широкој заједници. Зато се у оквиру ове организације спроводи широки спектар активности као што су производња и валидација језичких ресурса, технолошка валидација, дисеминација информација о језичким ресурсима и технологијама и многе друге активности. ELRA/ELDA је организатор једне од највећих конференција из области обраде природног језика, језичких ресурса и технологија под називом LREC (енг. language resources and evaluation)[5] која промовише дељење језичких ресурса, њихову евалуацију засновану на опште прихваћеним принципима и стварање нових ресурса на основу постојећих. Ова организација на свом веб сајту пружа приступ каталозима језичких ресурса чије је стварање покренула и које одржава. Један од тих каталога је META-SHARE[6], који постоји у оквиру META-NET[7] мреже изврсности која је посвећена изградњи технолошких основа за вишејезично информатичко друштво Европе. То је мрежа дељених репозиторијума језичких ресурса и алата за обраду природног језика, на пример, алата за морфолошку анализу, алата за препознавање говора, алата за означавање врсте речи итд. META-NET окупља 60 истраживачких центара из 34 земље, међу којима су две институције из Србије – Математички факултет, Универзитета у Београду, као институција у оквиру које делује Друштво за језичке ресурсе и технологије[8], и Институт Михајло Пупин у Београду[9].

CLARIN (акроним од речи Common Language Resources and Technology Infrastructure) [10] је истраживачка инфраструктура настала са циљем да сви дигитални језички ресурси и алати широм Европе буду доступни преко јединственог онлајн окружења. Године 2012. основан је CLARIN ERIC ради успостављања и одржавања инфраструктуре за подршку и дељење, коришћење и одржавање језичких података и

[4] ELRA/ELDA http://www.elra.info/en/about/
[5] LREC http://www.elra.info/en/lrec/
[6] META-SHARE http://www.elra.info/en/catalogues/meta-share/
[7] META-NET http://www.meta-net.eu/
[8] Јертех http://jerteh.rs/
[9] Институт Михајло Пупин http://bg.ac.rs/sr/clanice/instituti/IMP.php
[10] CLARIN https://www.clarin.eu/content/

алата за истраживање у области хуманистичких и друштвених наука. Неки језички ресурси и алати за српски језик развијају се у склопу пројекта CLARIN.SI, на институту Јожеф Стефан у Љубљани. Један од њих је паралелизовани енглеско-српски корпус [11].

2.1 Језички ресурси и алати у обради српског језика

Област обраде природног језика почела је да се развија у Србији пре скоро 40 година, захваљујући младим истраживачима и научницима, углавном математичке струке. Од тада су оснивачи и сарадници Групе за језичке ресурсе и технологије Математичког факултета у Београду, у сарадњи са професорима, сарадницима и студентима Филолошког факултета у Београду, као и других факултета Београдског универзитета, развили многе ресурсе и алате који су неопходни за обраду српског језика путем рачунарских технологија. У наставку ћемо говорити о тим ресурсима и алатима, њиховој изградњи и примени.

Учешће Групе за језичке ресурсе и технологије у пројекту CESAR[12] (енг. Central and South-european resources) који је био део мреже META-NET било је веома значајно јер је тако српски језик, кроз језичке ресурсе и алате, потврдио место у европској породици језика. Пројекат је трајао од 1. фебруара 2011. године до 31. јануара 2013. године и њиме су у смислу развоја и дељења језичких ресурса и алата, поред српског језика, били покривени следећи језици: пољски, словачки, хрватски, бугарски и мађарски. Пројекат је за циљ имао да пружи опис националних услова, потреба и могућности за језичке технологије. Други скуп циљева односио се на допринос оквиру за размену језичких ресурса на нивоу Европе, META-SHARE[13], у оквиру кога је омогућено заједничко коришћење и размена ресурса којима се може приступити непосредно и које сви могу претраживати. Претрага и размена језичких ресурса омогућени су кроз отворену и сигурну мрежу репозиторијума за дељење и размену језичких података, алата и веб сервиса. У оквиру META-SHARE корисничке сумеђе за претрагу језичких ресурса, као резултат претраге са кључном речју „Serbian“ (енг. српски) добијамо 35 погодака, дакле српски језик је укључен у 35 различитих језичких ресурса којима је могуће приступити преко META-SHARE портала.[14] У склопу пројекта CESAR циљеви су даље били и да се унапреде, прошире, стандардизују и међусобно

[11] https://www.clarin.si/repository/xmlui/handle/11356/1059
[12] CESAR http://www.meta-net.eu/projects/cesar/
[13] META-SHARE http://www.meta-share.org/
[14] http://metashare.ilsp.gr:8080/repository/search/?q=serbian (приступљено 22.12.2017)

повежу алати и ресурси учесника пројекта, да се подстакне сарадња и да се премости технолошки јаз између овог и других делова Европе.

На скупу под називом „Дан језичких технологија“, који је организован у склопу пројекта CESAR, у Београду је 29. октобра 2012. године у хотелу Hyatt представљена књига „Српски језик у дигиталном добу“ као део серије од 23 Беле књиге које су објављене за различите европске језике[15] (Витас, и др., 2012). Наглашена је потреба за очувањем српског језика у дигиталном окружењу с обзиром да је наш језик један од бројних европских језика чија судбина у том смислу није извесна и којима прети изумирање. Зато је рад на ресурсима који ће помоћи да се српски језик у таквом окружењу очува и користи што делотворније и сврсисходније свакако неопходан.

Остали међународни пројекти који су понудили подстицајно развојно окружење за развој језичких ресурса и алата за српски језик су[16]:

- TELRI-II – Trans-European Language Resources Infrastructure (EC Concerted Action PL977085) 1995-1997 i 1999-2001;
- ELAN – European Language Activity Network (MLIS Project 121) 1998-2001;
- BalkaNet – Design and Development of a Multilingual Balkan WordNet (FP5, IST-2000-29388) 2001-2004;
- Bilateral French-Serbian project (2004-2005) Multilingual Dictionary of Proper Names, (with Unversity Tours, France);
- Bilateral Slovene-Serbian project (2004-2005) Development of Slovene and Serbian Language Resources for Machine Translation (with Institute Jozef Stefan, Ljubljana);
- WISE – An Electronic Marketplace to Support Pairs of Less Widely Studied European Languages (BSEC 009 / 05.2007) 2007-2008;
- SEE-ERA.NET – Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages (ICT 10503 RP) 2007-2008;
- CESAR – Central and South-East European Resources (ICT Policy Support Programme, Grant agreement no.: 271022) 2011-2013;
- Parseme – COST Action IC1207 “PARSEME: PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural

[15] Springer, 2012. http://www.meta-net.eu/whitepapers/e-book/serbian.pdf
[16] http://jerteh.rs/?page_id=55

language processing", Management Committee member, Stiring Committee member, 2013-2017;

- Tempus project: Interdisciplinary Curricula in Computing to Meet Labor Market Needs INCOMING" 641, 2013-2017;
- Tempus project: BAEKTEL – Blending academic and entrepreneurial knowledge in technology enhanced learning, 2014-2017

2.1.1 Корпуси српског језика

Реч корпус потиче од латинске речи *corpus*, основног значења „тело" (што је познато из латинске пословице *Mens sana in corpore sano* (у здравом телу, здрав дух). У општем смислу, у обради природног језика корпус представља неку колекцију текстова на којој можемо вршити анализе. Обично је то репрезентативни примерак текстова за неки природни језик.

Текстуелни корпус се у најпростијем смислу може посматрати као низ речи или токена (енг. token). Појединачно појављивање речи одређује позицију у корпусу. Сваки корпус је често припремљен за даљу обраду, тако што се над њим обави такозвано етикетирање, то јест, лексичке јединице добијају синтаксичке и морфолошке ознаке. Корпуси могу бити неанотирани, анотирани (библиографска, структурна, морфолошка, синтаксичка, семантичка, прагматичка, стилистичка анотација, анотација кореференције). Корпуси могу бити једнојезични, вишејезични, у ком случају су то паралелни и упоредни корпуси.

Обично се у литератури наводе три нивоа анотације:

- позициона анотација односи се на анотацију појединачних позиција у корпусу (то јест појединачних речи или токена), на пример додавањем позиционих атрибута као што су врста речи (енг. part-of-speech), морфосинтаксичке информације, основни облици тј., леме;
- структурна анотација је уметање структурних атрибута у корпус, на пример одредница за почетак и крај параграфа или реченице;
- анотација на више нивоа се остварује када је један корпус поравнан са другим корпусом, на пример са верзијом истог корпуса на другом језику (Christ, 1994).

Неки корпуси постоје већ дуже време и могуће им је приступити преко NLTK (енг. natural language tool kit) библиотеке програмског језика Python који је последњих година постао *de facto* програмски језик у области обраде природног језика. NLTK

библиотека садржи мали избор текстова на енглеском језику, али и на другим светским језицима, на пример на урду језику (једном од званичних језика Пакистана, Индије, Бангладеша и Непала). Ти текстови потичу из великог пројекта дигитализације књижевних дела под називом Пројекат Гутенберг (енг. Project Gutenberg)[17]. Тако на пример корпус под називом udhr садржи текстове Универзалне декларације о људским правима (енг. universal declaration of human rights) на преко 300 језика. Остатак електронских текстова из архиве Пројекта Гутенберг која броји преко 25.000 електронских књига, налази се на адреси овог пројекта. Још један значајан корпус коме се може приступити преко NLTK библиотеке је Браунов корпус (енг. Brown Corpus). То је електронски корпус енглеског језика који је први достигао величину од милион речи. Настао је 1961. године на Универзитету Браун (енг. Brown University) и садржи текстове из 500 извора који су категоризовани према жанру, на пример спорт, новински текстови, романтични текстови итд.

У обради српског језика, електронске колекције текстова и корпуси представљају веома важне ресурсе. Када говоримо о колекцијама текстова у односу на корпусе, то су колекције које су сачињене без јасних лингвистичких критеријума, као што је, на пример колекција текстова у оквиру пројекта Растко[18]. Када говоримо о корпусима као колекцијама текстова у чијем формирању су праћени јасни лингвистички критеријуми, за српски језик имамо *Дијахрони корпус српског и српско-хрватског језика*, Ђорђа Костића и *Корпус савременог српског језика* развијен на Математичком факултету Универзитета у Београду (Vitas, Krstev, Obradović, Popović, & Pavlović-Lažetić, 2003).

Дијахрони корпус српског језика формиран је на Институту за експерименталну фонетику и патологију говора у Београду под вођатвом професора Ђорђа Костића. Пројекат изградње корпуса трајао је од 1957. до 1962. године и у њему је учествовало око 400 сарадника, углавном лингвиста, стручњака из других сродних области и техничког особља. Овај удружени рад се донекле може сматрати групном расподелом рада, али с обзиром да није било отвореног позива за допринос изградњи корпуса, и сарадници су били запослени на пројекту изградње корпуса, дакле нису добровољно доприносили изради овог ресурса, групна расподела рада у правом смислу није коришћена као модел изградње овог језичког ресурса. Више о особинама пројеката који се спроводе у складу са правилима модела групне расподеле рада говорићемо у трећем

[17] Пројекат Гутенберг http://www.gutenberg.org/
[18] Пројекат Растко http://www.rastko.rs/

поглављу овог докторског рада. Материјал прикупљен током рада на Дијахроном корпусу српског језика је 1996. године пребачен у дигитални облик. Величина овог корпуса је 11 милиона речи и највише се користи за статистичке анализе, на пример за процену вероватноће појављивања појединих речи и њихових граматичких облика, те процене вероватноћа гласовних и слоговних структура, што су значајни подаци у психолошким истраживањима (Ševa & Kostić , 2003).

Идеја о *Корпусу савременог српског језика* постојала је од осамдесетих година двадесетог века, када је почео процес сакупљања текстова и дигитализације, док је процес изградње започео 2002. године, на иницијативу професора Љубомира Поповића (Popović & Vitas, 2003). Развој корпуса савременог српског језика индиректно је подржаван средствима многих пројеката, али ниједан од њих није био намењен искључиво његовој изградњи. Неки од тих пројеката били су:

- Рачунарство са применама, РЗН СРС, 1986-1990;
- Језичке индустрије (енг. Language Industries), Европска унија, 1989-1991. године;
- Транс-европска инфраструктура језичких ресурса I-II (енг. Trans European Language Resources Infrastructure I-II, скр. TELRI I-II), 1995-2001. године;
- Интеракције између текста и речника, Министарство за науку републике Србије, 2002-2004;
- 148021 Теоријско-методолошки оквир за модернизацију описа српског језика, Министарство за науку Републике Србије и САНУ, 2005-2010. године;
- Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages, SEE-ERA.NET (ICT 10503 RP), 2007-2008;
- Српски језик и његови ресурси: теорија, опис и примене, Министарство за образовање и науку Републике Србије, 2011-2015. године;
- III 47003 инфраструктура за електронски подржано учење у Србији, Министарство за образовање и науку Републике Србије 2011-2015. године;
- Ресурси Средње и Југоисточне Европе (енг. Central and South-East European Resources, skr. CESAR), 2011-2013. године (ICT Policy Support Programme, Grant agreement no.: 271022) (Utvić, 2013).

Прва верзија *Корпуса савременог српског језика* створена је и објављена као производ рада чланова Групе за језичке ресурсе и технологије, Математичког факултета.

Та верзија је била колекција необележених тј. неанотираних текстова, без библиографских података о текстовима и названа је *Неетикетирани корпус српског језика (NETK)*. У тој првој фази радило се на прикупљању текстова за корпус, тестирању алата за креирање корпуса и његову претрагу, изради базе података и веб странице преко које би корисници приступали корпусу, одакле би се вршила администрација корпуса и омогућила једноставна претрага. Корпус је било могуће претраживати преко веб-сучеља које омогућава претрагу засновану на CQP[19]-регуларним изразима, а резултат претраге су биле конкорданце, тј. списак свих појављивања неке речи у корпусу, у оквиру контекста у коме се та реч појављује. Тај првобитни корпус се састојао из 22, 2 милиона корпусних речи. Неетикетирани корпус српског језика доступан је онлајн од 2003. године (Utvić, 2013).

Касније су корпусним текстовима додате библиографске информације и тако је настао *SrpKor 2003*. Библиографска анотација, поред библиографског описа, то јест података о наслову дела, аутору, издавачу, години издања итд., пружа и информације о функционалном стилу текста, као и о томе да ли је текст изворно написан на српском језику или је превод са неког другог језика (Utvić, 2011).

Неетикетирани корпус српског језика (NETK) и *SrpKor 2003* представљају, дакле, истоветну колекцију текстова величине 22,2 речи. Оба корпуса су кодирана коришћењем ASCII карактерског скупа и кодне схеме Аурора (пододељак 2.4, стр. 67). У корпусу *SrpKor2003* сваком тексту је придружен одговарајући библиографски опис, али остали типови анотације, као што су морфолошка или структурна анотација, нису примењени (Krstev & Vitas, 2005).

Применом регуларних израза могуће је, на пример, формирати следеће упите:

plima	{plima}
plim[a-z]	{plima, plime, plimu, plimi,…}
plim[a-z]+	{plima, pliimom, plimski,…}
[a-z]+ski	{plimski, zimski, rimski,…}

Додавањем нових текстова настала је текућа верзија SrpKor корпуса, под називом *SrpKor2013*, која је 2014. године, према подацима изнесеним на семинару Друштва младих лингвиста[20] садржавала 4.889 текстова. Највише текстова потиче из новинских чланака (66%), затим следе административни текстови (19%), књижевно-уметнички

[19] CQP – Corpus Query Processor http://corpora.dslo.unibo.it/TCORIS/cqpman.pdf

[20] Предавање М. Утвића о Корпусу Савременог српског језика http://www.dml.rs/index.php/lat/seminari-lat/2014-seminari-lat/306-korpusna-lingvistika-lat

(7%), научни и научнопопуларни текстови (4%). Већина текстова који су уврштени у овај корпус (87%) објављена је после 2000. године, те се може рећи да је корпус синхронијски и садржи текстове на савременом српском језику. Софтвер којим је индексиран корпус омогућава претрагу помоћу језика CQP. IMS Corpus Workbench[21] чини скуп алата за управљање обимним, анотираним корпусима. Један од алата у оквиру овог скупа је IMS Corpus Query Processor, CQP (Christ, 1994), претраживач корпуса намењен лингвистичким истраживањима.

SrpKor2013 садржи књижевно-уметничке текстове српских писаца у 20. и 21. веку, као и научне и научно-популарне текстове из различитих домена (природне и друштвене науке), административне текстове и опште текстове. Општи текстови представљају чланке дневних новина „Политика“ из периода 2000-2002. и 2005-2010. године, текстове из часописа и магазина објављене у периоду 1991-2002. године („Даница“, „Ебит“, „Економист“, „Гласник“, „НИН“, „Илустрована политика“, „Калибар“, „Моје срце“, „Мостови“, „Православље“, „Свет“, „Теолошки погледи“, „Трн“, „Вива“, „Република“), текстове са интернет портала објављене током 2011. и 2012. године (Пешчаник), вести агенције ТАНЈУГ током 1995. и 1996. године, фељтоне објављене у новинама „Политика“ (2001-2003), „Вечерње новости“ (2008-2011) и „Данас“ (2002-2006). Један део текстова представљају преводи чију већину чине књижевно-уметнички текстови, док мањи део представљају преводи општих текстова (Utvić, 2013).

SrpKor2013 садржи више од 122 милиона корпусних речи, анотиран је библиографски и морфолошки. Под морфолошком анотацијом се у овом корпусу подразумева лематизација и означавање врста речи. Иако је непотпуна, оваква анотација омогућава претраживање корпуса помоћу упита као што су:

[pos = "V" & pos = "N"] – за проналажење глагола и именица који се у корпусу појављују заједно, овим редоследом

[pos = "A" & lemma = ".*ski"] – за проналажење придева који се завршавају наставком –ski

SrpKor2013 је аутоматски анотиран уз помоћ прерађеног подскупа електронског морфолошког речника српског језика у формату LADL/DELA. За разлику од полазног речника који садржи вредности свих морфолошких категорија, прерађени подскуп за

[21] The IMS Corpus Workbench: Corpus Query Processor Manual http://corpora.dslo.unibo.it/TCORIS/cqpman.pdf

сваку одредницу бележи само информацију о леми и врсти речи (Утвић, 2011) . *SrpKor 2013* је доступан уз претходну бесплатну регистрацију на адреси: http://www.korpus.matf.bg.ac.rs.

Лематизирани корпус савременог српског језика (SrpLemKor)[22] је подскуп корпуса SrpKor величине 3,7 милиона корпусних речи који се може преузети и дистрибуирати у складу са лиценцом CC-BY-NC (Утвић, 2011).

Означени корпус српских новинских текстова SETimes.SR је корпус прикупљен из паралелног корпуса енглеског језика и језика Југоисточне Европе, SETimes[23] који објављује вести и ставове из Југоисточне Европе на девет језика: бугарском, босанском, грчком, енглеском, хрватском, македонском, румунском, албанском и српском. Садржи 86.765 одредница и ручно је означен на нивоу леме и морфосинтаксичке ознаке, у складу са смерницама MULTEXT-East V5[24] скупа ознака за српски и хрватски језик. Настао је у склопу такозване RELDI (Regional linguistics data initiative) иницијативе[25].

Нетикетирани корпус Вукових народних пословица[26] за основу има електронско издање књиге „Вукове народне пословице с регистром кључних речи“, библиотека „Одреднице“, уредник Слободан Ђорђевић, Нолит, Београд, 1996. Текст се састоји од пословица и Вукових коментара уз пословице. (Krstev C. , 1997) .

Корпус за евалуацију именованих ентитета (SrpNEval)[27] састоји се из 2.000 кратких вести које су објавиле српске новинске агенције и дневни листови у 2005. и 2006. години. Величина корпуса је 3.343 реченице, 89.425 речи, 7.122 етикета именованих ентитета. Именовани ентитети су аутоматски препознати и ручно кориговани, а корпус је доступан на оба званична писма српског језика – ћирилици и латиници.

Корпус који је саставио Хенинг Моерк (eng.Henning Moerk) са универзитета у Аархусу познат је под називом *Хенингов корпус српскохрватског*. Корпус се састоји од прозних текстова[28] на српскохрватском објављених између 1955. и 1990. године. Програме за конверзију полазних текстова на формат који се интерно користи под CQP-ом и њихово етикетирање обавио је Саша Стевановић.

[22] http://www.korpus.matf.bg.ac.rs/SrpLemKor/
[23] SETimes http://nlp.ffzg.hr/resources/corpora/setimes/
[24] MULTEXT-East V5 http://nl.ijs.si/ME/V5/msd/html/
[25] https://reldi.spur.uzh.ch/hr-sr/2016/05/02/oznaceni-korpus-setimes-sr/
[26] http://alas.matf.bg.ac.rs/~cvetana/proverb/
[27] http://www.korpus.matf.bg.ac.rs/SrpNEval
[28] http://www.korpus.matf.bg.ac.rs/prezentacija/yu-index.html

Изборна криза 2000. године је корпус који се се састоји од комплетних веб-издања дневног листа Политика у периоду од 10. септембра до 20. октобра 2000. године[29].

Етикетирани корпус српског језика се састоји од текстова са минималним скупом структурних етикета (, <head>, <p>, <seg>). Претрага по структурним етикетама је за сада могућа само из командне линије CQP -a[30].

Најважнији вишејезични корпуси српског језика су:

SELFEH (Serbian-English Law Finance Education and Health)[31] паралелни српско-енглески корпус који садржи документе везане за финансије, здравство, права и образовање. Корпус је развијен у склопу учешћа Групе за језичке ресурсе и технологије у пројекту INTERA. SELFEH је двојезични паралелизован корпус који садржи око један милион речи за сваки језик представљен у TMX формату. Корпус садржи преко 150 докумената у XML формату, а српски део корпуса је анотиран на нивоу врсте речи и леме.

Енглеско-српски поравнати корпус (SrpEngKor)[32] садржи дела списатељице Џејн Остин (енг. Jane Austen, тачније њених шест романа на енглеском и њихове превода на српски језик (Krstev & Vitas, 2009). За израду и процесирање овог корпуса коришћен је XML формат и софтверски пакет алата за обраду корпуса Unitex.

Француско-српски поравнати корпус[33] величине од једног милион речи на француском језику и више од једног милиона речи на српском језику. Овај корпус се састоји из књижевних и новинских текстова. Текстови су поравнати на нивоу реченице. (Vitas & Krstev, 2006).

Вишејезично електронско издање романа Жила Верна Пут око света за 80 дана, је паралелизовано на 26 светских језика (Krstev, Vitas , & Erjavec, 2004) (Vitas, Koeva, Krstev, & Obradović, 2008). У склопу независног пројекта на Математичком факултету Универзитета у Београду, верзије Верновог романа на 18 језика (оригинална верзија је на француском) су паралелизоване са српском верзијом. Сви преводи су паралелизовани са неком од верзија на француском, српском или енглеском језику. Српску верзију романа Пут око свет за 80 дана морфосинтаксички је анотирала проф. др Цветана Крстев

[29] http://www.korpus.matf.bg.ac.rs/prezentacija/korpusi.html
[30] http://www.korpus.matf.bg.ac.rs/prezentacija/korpusi.html
[31] http://www.korpus.matf.bg.ac.rs/prezentacija/selfeh.html
[32] http://www.korpus.matf.bg.ac.rs/SrpEngKor/korpus/index1.php
[33] http://www.korpus.matf.bg.ac.rs/SrpFranKor/korpus/index1.php

у формату LADL/DELA. Овим пројектом је руководио Проф. др Душко Витас (Vitas, Koeva, Krstev, & Obradović, 2008) .

Електронска верзија романа Џорџа Орвела 1984 је морфосинтаксички анотирана у два формата: LADL/DELA (видети пододељак 2.1.2) и MULTEXT-East за који су развијене спецификације за српски језик у оквиру истоименог пројекта (Vitas & Krstev, 1998), (Krstev, Vitas , & Erjavec, 2004). MULTEXT-East је вишејезични скуп података за развој језичких ресурса и алата који покрива велики број језика централне и источне Европе.

2.1.2 Електронски морфолошки речници српског језика

Електронски речници су речници који се користе у обради природног језика и садрже информације које их чине кориснима за задатке сегментације и морфолошке обраде текста. (Vitas & Krstev, 2012). Електронски речници се користе за креирање језичких алата као и за евалуацију језичких ресурса (Krstev & Vitas, 2005).

Морфолошке речнике српског језика развили су проф. др Цветана Крстев и проф. др Душко Витас уз помоћ чланова Групе за језичке ресурсе и технологије Математичког факултета у Београду. Њихов рад је био под великим утицајем француског лингвисте Мориса Гроса (фр. Maurice Gross) и његове школе. Грос је 1968. године основао лабораторију за лингвистичка истраживања LADL (фр. Laboratoire d'Automatique Documentaire et Linguistique). Електронски морфолошки речници српског језика развијени су у формату LADL/DELA, у складу са стандардом за електронске морфолошке речнике лабораторије LADL, у оквиру мреже RELEX[34] (Laporte, 2003), а њихова примена је најефикаснија уз помоћа алата заснованих на коначним аутоматима. У оквиру ове мреже развијани су и речници за бугарски и грчки језик и многе друге (Krstev C. , 2008). Систем електронских морфолошких речника у LADL формату је аутоматски конвертован у MULTEXT-East формат (Krstev, Vitas , & Erjavec, 2004) .

Основне компоненте система морфолошких речника за српски језик, SrpMD[35] су:

- DELAS – речник једночланих лексичких јединица (лема);
- DELAC – речник вишечланих лексичких јединица (лема);
- DELAF – речник свих флективних облика речи одредница DELAS речника;

[34] RELEX Network http://ladl.univ-mlv.fr/Relex/Relex.html
[35] SrpMD http://korpus.matf.bg.ac.rs/prezentacija/recnici.html

- DELACF – речник флективних облика вишечланих лексичких јединица (DELAC);
- морфолошке граматике и локалне граматике у облику скупа коначних трансдуктора којима се дефинишу и генеришу сви флективни облици у речницима (Krstev, Stanković, Obradović, Vitas, & Utvić, 2010) (Krstev, Vitas , & Pavlović-Lažetić, 2003).

Електронски морфолошки речници су настали на основу традиционалних речника, али и екстракцијом информација из процесираних текстова коришћењем ресурса и алата за обраду српског језика. За развој електронских морфолошких речника српског језика користи се алат LeXimir, а за коришћење речника и изградњу локалних граматика користи се алат Unitex[36] лабораторије LADL, уз вођење рачуна о посебностима српског језика, као што су:

- употреба два писма – ћирилица и латиница су у српском језику у равноправној употреби и текстови се појављују на оба ова писма;
- правопис је заснован на фонологији – употреба различитих варијанти српског језика (екавски и ијекавски изговор) и постојање дублета због те употребе;
- богат морфолошки систем који се огледа на нивоима флексије и деривације;
- слободан редослед речи у погледу субјекта, предиката, објекта и осталих делова реченице, као и употреба енклитика. (Vitas & Krstev, 2005).

Основна јединица морфолошког електронског речника српског језика је лема. Свакој леми су преко синтаксних и семантичких маркера придружени морфолошки описи – опис флексије, деривације, природног броја и природног рода. Одредници у DELAS и DELAC речнику придружена је ознака врсте речи (једно или више великих слова), веза са подређеним облицима (нумеричка или алфанумеричка ознака која заједно са врстом речи омогућава аутоматско генерисање свих подређених облика за DELAF речник), синтаксни, семантички, употребни, дијалекатски и слични маркери (слободни маркери, алфанумеричке ознаке). Одредници у DELAF и DELACF речнику придружен је канонски облик (лема), ознака врсте речи која је преузета од леме, затим синтаксни, семантички, дијалекатски и слични маркери који су такође преузети од леме, као и скуп кодова који су везани за вредност граматичких категорија које се односе на облик (кодови састављени од великог или малог слова или цифре) (Тртовац, 2016).

[36] UNITEX http://unitexgramlab.org/

Извод из речника простих речи где се могу видети облици са граматичким категоријама, као и извод из речника вишечланих речи дати су у прилогу 2, у табелама 35 и 36.

У структури речника DELAS сваки унос је повезан са кодом који ближе описује врсту речи, синтактичке, семантичке особине и правила промене речи. (Vitas & Krstev, 2012). Неке од ознака су ознаке за род: m – мушки род, f – женски род, n – средњи род; ознаке за број: s – једнина, p – множина, w – паукал; ознаке за падеже су 1 – номинатив, 2 – генитив итд.; ознаке за аниматност су v – живо, q – неживо, g – без значаја. Неке од семантичких ознака које се користе у систему електронских морфолошких речника дате су у табели 30 у прилогу 2 овог докторског рада. Ти семантички маркери су делимично интегрисани у систем електронских речника уз пренос информација из лексичко-семантичке мреже Српски ворднет (пододељак 2.3.5).

У речнику простих речи, на следећем примеру: generaciju,generacija.N600:fs4q, видимо да је ниски *generaciju* додељена лема *generacija.* Ова лема припада флективној класи N600. Код fs4q нам говори да је облик речи *generaciju* у акузативу (4) једнине (s) женског рода (f) неживе аниматности (q) леме *generacija*. После кода флективне класе, леми се могу додати синтаксички и семантички кодови. Тако у примеру: smejali,smejati,V516+Imperf+It+Ref+Ek:Gpm видимо да је реч *smejali* у множини (p) мушког рода (m) радног глаголског придева (G) глагола *smejati* који припада класи флексије глагола V516, а који је несвршени (Imperf), непрелазми (It) и повратни (Ref) глагол екавског изговора (Ek). У примеру: plavo,plav.A17+Col:aens1g:aens4g:aens5g, можемо видети примену семантичких маркера, дакле *plav* је придев из класе A17 са особином боје (Col) (Vitas, Krstev, Obradović, Popović, & Pavlović-Lažetić, 2003). Оваква структура електронског речника омогућава устаљену примену теорије коначних аутомата у процесу тагирања (енг. tagging) и лематизације (енг. lemmatization) корпуса.

У речнику DELAF унос „prozora,prozor.N1:ms2q:mp2q:mw2q:mw4q“ тумачимо на следећи начин: облик „prozora“ је или у генитиву (једнине или множине) или у паукалу уноса „прозор“ који је означен као именица мушког рода, чија је ознака аниматности неживо. На основу садржаја речника DELAF могуће је спровести аутоматску сегментацију текста на речи и спровести морфолошку анализу уз примену метода лексичког препознавања (Silberztein , 1993)

У верзији 2.0 из 2014. године у односу на претходну верзију српских морфолошких речника из 2012. године постоје неке новине и додаци[37]:

- У речнику географских имена (за просте и сложене речи) додат је велики број одредница прикупљен са Википедије (уз додатке на основу анализе разних текстова);
- У речнику енглеских личних имена додата су имена која почињу словима P-З – укупно 1189;
- У речнику простих речи додата је ознака +NVAL= свим бројевима и именицама изведеним из бројева чија је вредност бројчана вредност, на пример jedanaestoro,.NUM06+HumColl+NVAL=11;
- Домени су претворени у пар (ознака, вредност) , ознака је +DOM, на пример за спорт је ознака +DOM=Sport;
- Додате су и ознаке многих нових домена, као и нових семантичких маркера;

Пуни српски електронски речник доступан је у научне сврхе, уз одобрење креатора и главног редактора речника, проф. др. Цветане Крстев.

2.2 Онтологије

2.2.1 Онтологије у обради природног језика

Онтологија је модел којим се представља знање о свету – неко опште знање или знање о одређеном домену. Свака онтологија представља скуп концепата и односа који постоје међу тим концептима. Формална онтологија је онтологија дата на неком формалном језику. Према (Devedžić, 2010), главна сврха онтологија је дељење и вишеструка употреба знања од стране различитих интелигентних агената и апликација. У зависности од тога који део стварности описују, онтологије могу бити: *онтологије највишег нивоа* (енг. top level ontologies) – које представљају опште концепте и знање које је свеобухватно, систематизовано и применљиво у великом броју апликација; *доменске онтологије* (енг. domain ontologies) – када се знања која представљају тичу једног домена или класе проблема, као у случају онтологије CHEMINF[38] која је намењена домену хемије; *онтологије задатака* (енг. task ontologies) или *апликацијске*

[37] Српски морфолошки речници-верзија 2.0, аутори Цветана Крстев и Душко Витас, извештај.
[38] CHEMINF – https://bioportal.bioontology.org/ontologies/CHEMINF

онтологије (енг. application ontologies) – садрже само она знања која су неопходна за извршавање одређених задатака.

Структуру сваке онтологије чине: концепти или класе (енг. concepts, classes), инстанце класа или индивидуе (енг. instances, individuals), релације између класа (енг. relations, properties), атрибути (енг. attributes) и формална правила (енг. axioms). Класе онтологије формирају таксономију ступајући међусобно у релације. Индивидуе представљају примере класа. Релације могу постојати између класа, између индивидуа или између индивидуа и класа. Атрибутима се могу описати и концепти и индивидуе, док се формалним правилима исказују знања која нису дата експлицитно, односно релације међу концептима које се могу извести на основу знања о стварности коју описују.

Прва формална онтологија поравната са Ворднетом (о коме ће бити више речи у одељку 2.3) је онтологија под називом SUMO[39] коју је као пројекат израде новог стандарда 2000. године почео да развија IEEE[40]. SUMO је такозвана горња онтологија јер представља концепте који су довољно општи, апстрактни или по природи генерички да могу да покрију широки опсег подручја на вишем нивоу. То значи да концепти карактеристични за неки одређени домен нису укључени у SUMO, нити у друге онтологије вишег нивоа. У називу SUMO налази се и термин *merged* јер је ова онтологија настала повезивањем више јавно доступних садржаја у јединствену кохерентну структуру (Pease, 2011). Ова онтологија садржи релативно мали број концепата, око 1.000, али велики број формалних тврђења, тачније 4.000, и око 800 правила, што омогућава њено лако разумевање и примену. Неке од општих тема које покрива су структурни концепти, општи типови објеката и процеса, бројеви и мере, временски концепти, делови и целине, основне семиотске релације.

Онтологије су саставни део семантичког веба (енг. semantic web), то јест пројекта израде универзалног медијума за размену информација постављањем докумената са значењем које рачунар може да обради на веб. Главни циљ овакве мреже је постизање узајамног функционисања веб извора на семантичком нивоу, то јест постојање инфраструктуре за машинску интерпретацију и закључивање о садржајима на вебу.

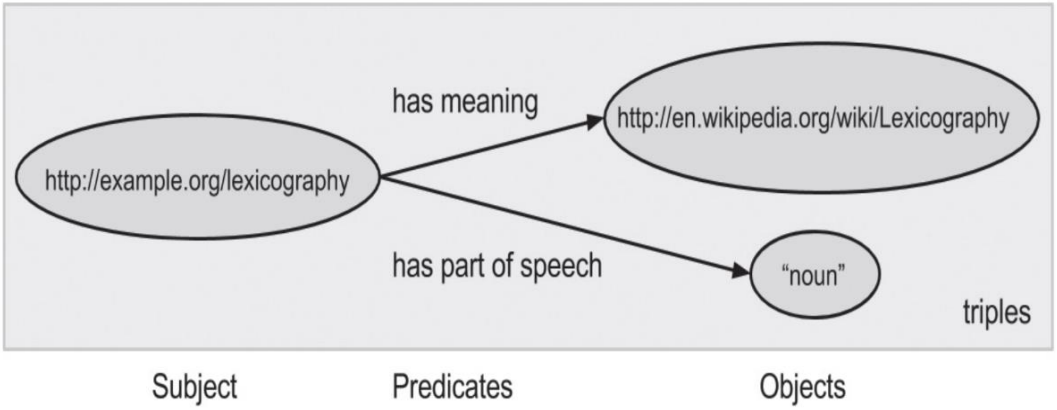
RDF[41] је један од стандарда семантичког веба. То је општи метод за концептуално описивање информација и семантичких веза између електронских извора. Чине га

[39] SUMO – Suggested Upper Merged Ontology

[40] IEEE – Institute of Electrical and Electronics Engineers

[41] RDF – resource description format

уређене тројке (енг. triples) облика: Субјекат – Предикат – Објекат, као што је приказано на слици 1. Још један стандард семантичког веба је SPARQL[42], протокол и језик за семантичке упите над RDF базама података.



Слика 1 RDF тројка[43]

Препорука W3C (World Wide Web Consocrtium) [44] за формални стандардни језик представљања онтологија на семантичком вебу је OWL[45]. Овај језик се користи за објављивање и дељење онтологија, а карактерише га већа изражајност и могућност боље машинске интероперабилности веб садржаја у односу на XML[46] или RDF. Управо због тога је OWL коришћен за грађење онтологије реторичких фигура на српском језику коју смо назвали *РетФиг*, а која ће бити описана у наредном поглављу.

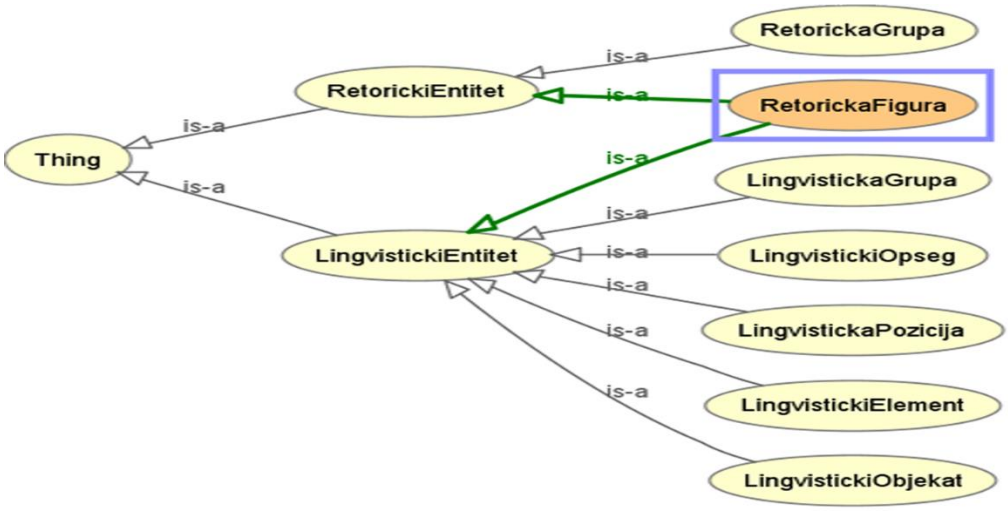
2.2.2 Онтологија реторичких фигура за српски језик – *РетФиг*

Реторичке фигуре, или стилске фигуре су утврђени начин изражавања који се разликује од свакодневног говора, којим се постиже лепота и сугестивност исказа. То су лингвистичка средства за која важе посебна когнитивна правила, која имају функционалну, меморијску и естетску сврху (Ruan, Di Marco, & Harris, 2016). Реторичке фигуре су креативан начин употребе језика ради постизања неког ефекта, било да се ради о мењању редоследа речи или о промени значења. У обради природног језика важно је узети у обзир постојање реторичких фигура, јер оне утичу на значење текста у коме се појављују. Један од начина да уврстимо реторичке фигуре у процесе обраде природног језика јесте да их представимо у виду онтологије.

[42] SPARQL – рекурзивна скраћеница од SPARQL Protocol and RDF Query Language
[43] Пример RDF тројке са предавања Џона Мекреа (енг. John McCree) на летњој школи EUROLAN одржаној 13-25. 07 2015.
[44] W3C http://www.w3.org/
[45] OWL – ontology web language
[46] XML – extensible markup language

Да бисмо остваривали добре резултате у обради природног језика који садржи фигуративни говор, потребно је да имамо корпусе анотиране информацијама о реторичким фигурама. Анотацију корпуса обично врше стручњаци (или учесници у неком пројекту групне расподеле рада); међутим, реторичке фигуре је често веома тешко препознати. Неке фигуре граде се над речима, неке над реченицама или читавим пасусима, а неретко долази до преклапања, те у истом делу текста наилазимо на више реторичких фигура. *РетФиг* онтологија настала је ради превазилажења ових проблема.

РетФиг је прва онтологија реторичких фигура за српски језик (Mladenović & Mitrović, 2013). Она је доменска, формална онтологија која је настала из потребе да се на формалан начин представи структура и начин изградње реторичких фигура у српском језику, те тако описује 98 различитих реторичких фигура. За сваку фигуру је дефинисано којој реторичкој и лингвистичкој групи припада, дефинисани су лингвистички опсег, објекти и елементи који учествују у креирању дате реторичке фигуре, међусобни однос објеката и елемената, као и лингвистичке операције које учествују у процесу креирања реторичке фигуре. Основа таксономије онтологије *РетФиг* дата је на слици 2. На њој се види да је класа *Реторичка фигура* у овој онтологији представљена као лингвистички и као реторички ентитет. Као реторички ентитет, свака фигура припада једној од четири реторичке групе. Као лингвистички ентитет, свака реторичка фигура има дефинисану лингвистичку групу и опсег којима припада, те лингвистичку позицију, лингвистички елемент и лингвистички објекат који учествују у њеном формирању.



Слика 2 – Таксономија онтологије РетФиг

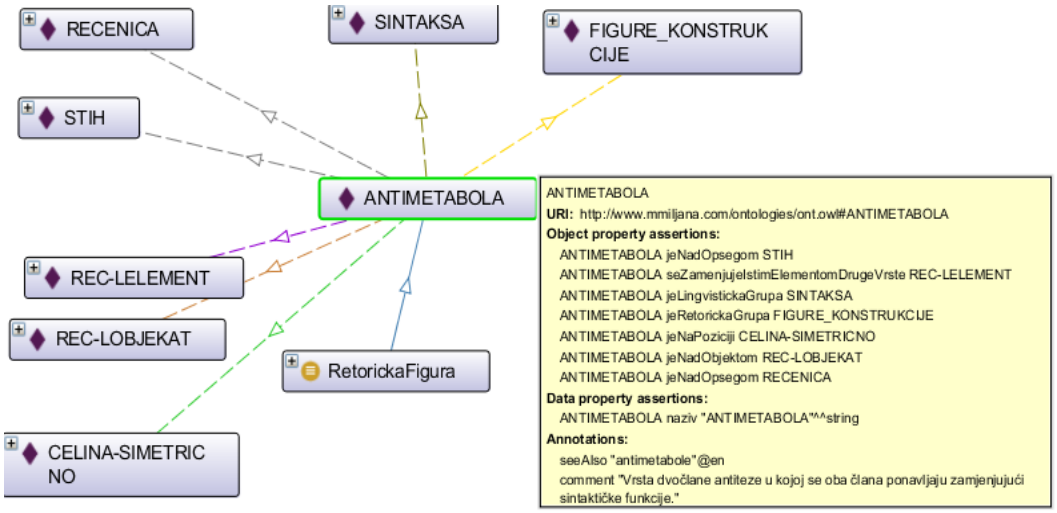
У *РетФиг* онтологији све фигуре подељене су у четири групе:

- фигуре наглашавања (дикције)
- фигуре конструкције (схеме)
- фигуре проширења мисли
- фигуре речи (тропи)

Фигуре наглашавања се заснивају на дејству одређених гласова у говору, односно слова у тексту. Понављање одређених гласова или група гласова у говору, односно слова или група слова у тексту, њихово изостављање или уметање на неочекиваним местима, опонашање одређених звукова и шумова из природе утиче на појачавање или смањење значаја језичких структура над којима се изводе. Ова врста фигура нема утицаја на значење структура над којима се граде и не мења их, већ само наглашава њихово основно значење.

Фигуре конструкције (схеме) настају мењањем распореда речи у реченици или у некој другој већој целини (одломку, стиху и сл.) у односу на уобичајени тј. подразумевани, па се могу посматрати као фигуре поретка или распореда. И ова врста фигура не мења основно значење језичких структура над којима се граде, али утичу на промену значења у ширем контексту тих структура, као и на појачавање значења.

Фигурама мисли мења се основно значење лингвистичке структуре која је комплекснија од речи, док фигуре речи (тропи) имају задатак да промене основно значење речи. Начин представљања реторичких фигура у онтологији *РетФиг* приказан је на слици 3, на примеру реторичке фигуре антиметабола, која припада фигурама конструкције, односно схемама.



Слика 3 – Антиметабола у онтологији РетФиг

Ова реторичка фигура заснива се на замени синтактичких функција њених делова и игра веома важну улогу у политичким говорима, те јој се приписује функција наглашавања убедљивости текста у коме се појављује. Примери ове фигуре који се најчешће наводе у литератури су латинска фраза *Unus pro omnibus, omnes pro uno* (лат. Сви за једнога, један за све), као и чувена реченица из инаугурационог говора америчког председника Џона Ф. Кенедија (енг. John F. Kennedy) *Ask not what your country can do for you; ask what you can do for your country* (енг. Не питајте шта ваша земља може да уради за вас, питајте шта ви можете да урадите за своју земљу).

Поред новог начина концептуалног представљања реторичких фигура, онтологија *РетФиг* такође служи и као свеједрсна база знања о 98 фигура које су њоме представљене. Називи свих реторичких фигура дати су на српском и на енглеском језику, како би овај важан ресурс био погодан за међународну примену. Ова онтологија се може преузети са адресе на којој се налазе и остали семантички ресурси и алати за српски језик, о којима ћемо више говорити у наредном поглављу овог рада. Онтологија *РетФиг* се на овом веб сајту[47] може преузети у облику .xml или .owl датотеке, након неопходне аутентикације.

2.3 Ворднет

„Ворднет је рачунарски речник синонима, тезаурус, лексичка база података, таксономија појмова – списак се може наставити.“ – (Piasecki, Szpakowicz, & Broda, 2009).[48]

Ворднет представља информатичку лексичко-семантичку мрежу, лингвистички ресурс који налази вишеструку примену у обради природног језика и као такав је постао *de facto* стандард у тој области. Развој првог ворднета започео је на Принстонском универзитету (енг. University of Princeton) средином осамдесетих година прошлог века у лабораторији за когнитивне науке. Група психолингвиста и лингвиста, на чијем челу је био професор Џорџ Милер (енг. George Miller), 1985. године је на том универзитету почела да развија лексичку базу података засновану на водећим психолингвистичким теоријама. Тај први ворднет зове се Принстонски ворднет (енг. Princeton WordNet)[49] (у овом раду, Ворднет). Замишљено је да Ворднет буде решење којим би се снага рачунара

[47] Семнтички ресурси српског језика http://resursi.mmiljana.com/RetFigS.aspx

[48] “A WordNet is a computerized dictionary of synonyms, thesaurus, lexical database, taxonomy of concepts – the list can go on.”

[49] https://wordnet.princeton.edu/

ефикасно удружила са традиционалним лексикографским начином представљања информација (Miller, Fellbaum, Gross, & Miller, 1990). Ворднет се састоји из синсетова, то јест скупова речи синонимних значења, који су међусобно повезани основним семантичким релацијама као што су хипонимија и меронимија.

У Ворднету речнички појмови нису представљени алфабетским редом, као у традиционалним речницима, већ концептуално – на основу семантичке меморије и поимања. Дотадашњи речници су често били грађени на основу историјских принципа, то јест истраживањем употребе речи кроз историју, али стандардизовани речници су занемаривали питања савременог начина организовања лексичког знања – ворднет је настао у жељи да се та пракса превазиђе (Miller G. A., 1995). Ипак, Ворднет је првобитно развијен да би послужио као помоћно средство за концептуално претраживање речника, те је замишљено да се користи у блиској вези са конвенционалним онлајн речником.

Милер и његов тим су инспирацију о основном принципу организације Ворднета пронашли у истраживањима о начину организације људске семантичке меморије. Једно такво истраживање бавило се откривањем сличности речи на основу асоцијација, где су испитаници имали задатак да кажу прву реч које се сете када чују неку од опште познатих речи из различитих синтаксичких категорија. Филенбаум и Џоунс (енг. Fillenbaum and Jones) (Fillenbaum & Jones, 1965) открили су да у 79% случајева именица асоцира на именицу, у 65% случајева придев асоцира на неки други придев, док у 43% случајева глагол асоцира на други глагол. На основу ових и других, сличних резултата, центар разумевања у ворднету постао је скуп синонимних речи, то јест речи сличног значења, које припадају истој синтаксичкој категорији – синсет (енг. synset, synonymous set). Синсет је, дакле, скуп когнитивно сличних синонима, то јест речи различитог облика, а истог или сличног значења које представљају исти концепт. Синоним може бити проста реч, сложена реч, вишечлана реч (енг. multi-word unit), фразални глагол, идиоматска фраза или лична именица. Употреба синсетова за представљање значења речи је, дакле, у складу са психолингвистичким доказима да су именице, глаголи, придеви и прилози организовани независно у семантичкој меморији. Ако су концепти представљени синсетовима и ако речи које чине један синсет, тј. синоними морају бити међусобно заменљиви (енг. interchangeable) у одговарајућем концепту, онда речи које припадају различитим синтаксичким категоријама не могу бити синоними нити формирати синсетове јер нису међусобно заменљиви (Miller, Fellbaum, Gross, & Miller, 1990).

Свака од синтаксичких категорија представљена је у Ворднету на другачији начин, јер су аутори сматрали да би исти организациони принцип примењен на све врсте речи онемогућио правилно представљање психолошке комплексности лексичког знања. Ворднет је, дакле, организован као база података која својом структуром осликава начин на који људски ум складишти и користи језичке информације (Fellbaum C. , 1998). Актуелна верзија Ворднета може се претраживати преко корисничког сучеља[50].

Развој Ворднета и других лексичко-семантичких мрежа типа Ворднет које се развијају у свету могуће је пратити у склопу активности професионалне организације Global Wordnet Association која координира рад на развоју Ворднета и свих осталих ворднетова у свету. Сваке две године, почевши од 2002. године, ова организација приређује међународну конференцију под називом Global WordNet Conference, скраћено GWC, на којој су 2014. и 2016. године представљени резултати у развоју Српског ворднета (Mladenović, Mitrović, & Krstev, 2014) и (Mladenović, Mitrović, & Krstev, 2016).

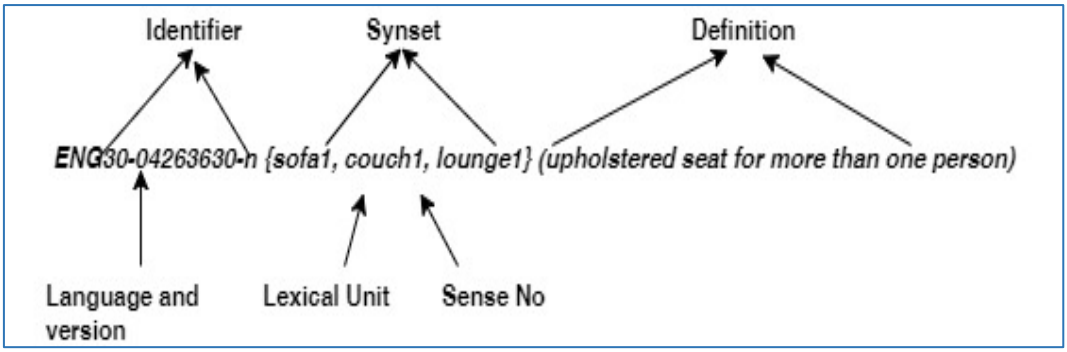
2.3.1 Структура ворднета и релације

У овом одељку објаснићемо основне концепте структуре и начина изградње Ворднета, на основу кога су, у већој или мањој мери, настали сви други ворднетови на свету. Због тога се Ворднет понекад назива „*мајком свих ворднетова*“ и постао је синоним за одређену врсту речничког дизајна. Данас у свету постоји више од 70 ворднетова на око 50 језика, а према подацима на веб страници удружења Global Wordnet[51], свих 70 ворднетова повезано је са Ворднетом, директно или, пак, индиректно са неким другим ворднетом који је повезан са њим.

Централна јединица и главни део сваког ворднета јесте синсет (енг. synset) тј. скуп (енг. set) синонима. Синсетови су лексикализовани концепти који се састоје из речи које имају слично или исто значење у когнитивном смислу – сваки синсет представља неки концепт. Саставни део синсета је нека лексичка јединица (Fellbaum C. , 1998). Индекс значења (енг. sense index) је редни број помоћу кога се разликују лексичке јединице са више значења, где се прво значење означава са 1 (што важи за Ворднет, али није обавезно за остале ворднет мреже).

Пример основне структуре једног синсета из Српског ворднета дат је на слици 4.

[50] PWN корисничко сучеље http://bit.ly/2o5xtfY
[51] Global Wordnet Association http://globalwordnet.org/



Слика 4 – Основна структура синсета Ворднета[52]

Релације у Ворднету успостављају се између лексичких јединица и између синсетова. Најчешће релације између лексичких јединица, то јест лексичке релације, су синонимија и антонимија.

Синонимија је основна лексичка релација за све врсте речи у Ворднету. Према дефиницији која се обично приписује Лајбницу (нем. Leibniz), немачком математичару, филозофу, историчару, политичару, два израза су синонимна ако замена једног од тих израза другим изразом не доводи до промене истинитости изјаве у којој је до те замене дошло. У складу са том дефиницијом, прави синоними су веома ретки. Блажа варијанта ове дефиниције у обзир узима контекст, те су тако два израза синонимна у лингвистичком контексту *C* ако замена једног другим унутар *C* не мења истинитосну вредност. Обично се претпоставља да је релација синонимије симетрична: Ако је *x* слично *y*, онда је *y* слично *x*. (Miller G. A., 1995). Дефиниција синонимије је, између осталог, довела до поделе Ворднета на именице, глаголе, придеве и прилоге. Именице представљају номиналне концепте, глаголи представљају глаголске концепте, а придеви и прилози пружају начине модификовања тих концепата (Miller, Becwith, Fellbaum, Gross, & Miller, 1990).

Антонимија је симетрична релација између два облика речи са супротним значењем или приближно супротним значењем. Тако на пример, имамо релацију антонимије између следећих именских синсетова:

- {female:2, female person:1} (особа која припада полу који може да рађа децу)
- {male:2, male person:1} (особа која припада полу који не може да рађа децу)
- {sorrow:1} (осећање велике жалости повезано са губитком) и
- {joy:1, joyousness:1, joyfulness:1} (осећање велике среће)

[52] Слика нацртана по узору на шему са стране 23 рада (Lohk, 2015)

Антонимија је лексичка релација између облика речи, а не семантичка релација између значења речи. Тако, на пример, значења {rise, ascend} и {fall, descend} су концептуално супротна, али то нису антоними; rise/fall и ascend/descend јесу антоними али већина људи се не би сложила да су rise и descend, или пак ascend и fall антоними (Miller G. A., 1995).

Релација антонимије у Ворднету постоји између:

- *именица*: синсет:{sadness, unhappiness}(осећање које нас обузима када нисмо у добром расположењу) – антоним: {joy, joyousness, joyfulness}(осећање велике среће);
- *придева*: синсет:{ugly}(непријатан за чула) – антоним: {beautiful}(пријатан за чула или интелектуално узбудљив или који изазива емоционално дивљење);
- *глагола*: синсет: {open, open_up}(изазвати отварање или постати отворен) – {close, shut} (померити тако да неки отвор или пролаз постане блокиран; учинити затвореним).

Семантичке релације у Ворднету се успостављају између концепата, то јест између синсетова Ворднета. Релација која је најчешће кодирана између синсетова у Ворднету јесте релација подређености, или *хипонимија* (енг. hyponymy) и релација надређености, или *хиперонимија* (енг. hyperonymy). Овим релацијама је, на пример, повезан синсет општег значења {furniture, piece_of_furniture} (намештај, комад намештаја) са синсетом специфичнијег значења {bed} (кревет). Такође, {maple} (јавор) је хипоним од {tree} (дрво), док је {tree} (дрво) хипоним од {plant} (биљка). Хипоним наслеђује све особине општијих концепата, уз додавање бар једне особине која га издваја од свог надређеног и других хипонима истог надређеног концепта. Тако јавор наслеђује особине свог надређеног концепта, што је дрво, али је другачија од другог дрвећа због неких својих особина, на пример облика лишћа или зато што се од њега добија сируп. (Крстев, и др., 2008). Релација хипонимије је транзитивна: ако је столица за љуљање врста столице, и ако је столица врста намештаја, онда је и столица за љуљање врста намештаја. Због овог принципа наслеђивања Ворднет се назива и системом лексичког наслеђивања (енг. lexical inheritance system) (Miller, Fellbaum, Gross, & Miller, 1990).

Релације хипонимије и хиперонимије су централни организациони принцип за именице у ворднету, као и за глаголе. Због информација о семантичким везама хипонимије и хиперонимије Ворднет се може посматрати и као тезаурус, односно

хијерархијски уређен речник синонима и асоцијативних појмова (Miller, Becwith, Fellbaum, Gross, & Miller, 1990).

Синсетови именица и глагола су различито организовани релацијама хиперонимије и хипонимије. Све глаголе није могуће окупити испод једног чвора на врху (Fellbaum, Gross, & Miller, 1993). Глаголи који се налазе ближе дну хијерархије, који се зову *тропоними*, изражавају све специфичнији начин карактерисања неког догађања или радње, на пример: {communicate} (комуницирати) - {talk} (причати) - {whisper} (шапутати), тако да можемо рећи да је релација *тропонимије* (енг. (manner-name)) за глаголе исто што је и релација хипонимије за именице.

Импликација (енг. entailment) је још једна релација између глагола. Релација импликације постоји између два глагола Г1 и Г2 када реченица *Неко Г1* логички подразумева и реченицу *Неко Г2*. Тако {abort}(прекинути трудноћу извођењем абортуса, подразумева{conceive}(остати трудан, проћи кроз зачеће) или {snore, saw_wood, saw_logs}(дисати гласно током сна) подразумева {sleep, kip, slumber, log_Z's, catch_some_Z's} (спавати). Посебан случај релације импликације је релација *каузалности* (causes) између два глаголска синсета од којих други означава радњу које се дешава само уколико се десила радња коју означава први синсет, као у примеру односа између синсета {feed, give} (дати некоме храну); употреба: „Нахранити гладну децу у Индији“ и синсета {eat}; (уносити чврсту храну); употреба: „Она једе банану“.

Све хијерархије именица крећу се према кореном чвору {entity}. Релацијама хипонимије и хиперонимије генеришу се структуре стабала које у процесу генерализације воде до синсетова који немају надређених хиперонимских синетова и зову се почетни синсетови (енг. unique beginners). У тренутно актуелној верзији PWN-а постоји 25 именских почетних синсетова којима се описују најопштији концепти (прилог 0). Џорџ Милер у свом истраживању (Miller G. A., 1998) наводи да су управо ти почетни именски синсетови одабрани због могућих комбинација облика придев-именица до којих би на основу њих могло доћи.[53] У овом раду се бавимо истраживањем односа именица и придева у функцији коју имају у реторичкој фигури поређење, као и проналажењем начина да ти односи буду представљени у лексичко-семантичкој мрежи каква је ворднет, те је чињеница да је именски део ворднета изграђен узимајући у обзир односе између именица и придева од нарочитог значаја.

[53] “They were selected after considering the possible adjective-noun combinations that could be expected to occur.“

Релацијама *меронимије* и */холонимије* описује се релација дела и целине, члана и групе, састојка и структуре. *Меронимија* (енг. part-name) и њена супротност, *холонимија* (енг. whole-name) помажу да се направи разлика између саставних делова који чине неки концепт. Релација меронимије успоставља се између два именска синсета од којих други означава концепт који је део првог концепта. Меронимија се, на пример, остварује између синсетова {chair} (столица) и {seat} (седиште) или {chair} (столица) и {leg} (нога). Делови се наслеђују од надређених: ако столица има ноге, онда и столица са наслоном за руке (енг. armchair) такође има ноге. Делови се не наслеђују „нагоре“ јер могу бити специфични само за одређене врсте ствари уместо за целу класу, на пример столице и неке врсте столица имају ноге, али неке врсте намештаја немају ноге.

У Ворднету постоји подела на две класе придева – *релационе* и *описне* придеве, који су организовани у складу са релацијом антонимије. Постоје такозвани директни антоними, као што су придеви *мокар* и *сув,* или *млад* и *стар.* Овакви придеви су повезани са већим бројем „семантички сличних“ придева, на пример, у Ворднету, придев *dry* је повезан са *parched, arid, dessicated* и *bone-dry*, а придев *wet* са *soggy, waterlogged,* итд. (Miller, Fellbaum, Gross, & Miller, 1990). Семантички слични придеви се још називају и индиректним антонимима.

Синсетови придева чији литерали немају директне антониме чине посебну групу тзв. придева-сателита (енг. *satellite adjectives*). Такви синсетови се групишу у организоване скупове синонимних синсетова тако што се сваки од њих релацијом типа *similar* повеже са главним синонимним синсетом који, с друге стране, релацијом антонимије остварује везу са антонимним организованим скупом. Литерали синсетова описних придева повезују се релацијом *attribute* са литералима именских синсетова, на пример између придева {perfect}(савршено) и именице {perfection, flawlessness, ne_plus_ultra}. Релациони придеви (енг. relational adjectives или pertainyms) указују на именицу из које су изведени, на пример, придев *злочиначки* изведен је из именице *злочинац*, а придев *љубавни* изведен је из именице *љубав*. Релација *pertains_to* остварује се између релационог придева и именице из које је изведен, на пример *злочиначки* је *pertainym* од *злочинац*. Прилози се повезују са придевима релацијом деривације, а међусобно релацијом антонимије (Miller, Fellbaum, Gross, & Miller, 1990). (Mendes, 2006) (Fellbaum, Gross, & Miller, 1993). У оквиру пројекта Еуроворднет, уведене су још неке релације, о којима ћемо говорити у пододељку 2.3.3.

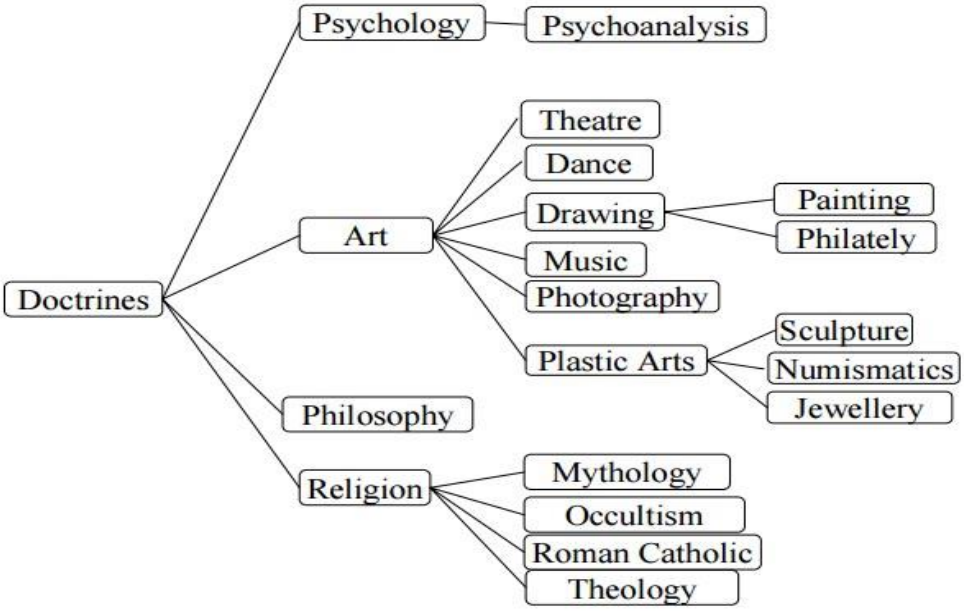
2.3.2 Проширења Ворднета

Структура Ворднета је неколико пута проширивана информацијама које побољшавају његову применљивост у задацима обраде природног језика. Једно од тих проширења је проширење семантичким доменима. Ови домени су природан начин да се успоставе семантичке релације између значења речи у сврху коришћења у разним задацима обраде природног језика. Семантички домени односе се на поља људског интересовања, на пример политика, архитектура, медицина, спорт, култура, економија, од којих свако има одређену, специфичну терминологију и лексичку кохерентност. Коришћење домена за означавање семантичких поља је уобичајено у лингвистици (Крстев, и др., 2008).

WordNet Domains Hierarchy (WDH) [54] је језички независтан ресурс који се састоји од 164 хијерархијски организованих ознака домена. Развијен је на италијанском Институту за научна и технолошка истраживања (ит. Istituto per la Ricerca Scientifica e Technologica – ITCirst). Домени су обележени помоћу двеста обележја домена позајмљених из Џуијеве децималне класификације (енг. Dewey Decimal classification). Та обележја су организована хијерархијски, у структуру дрвета. Лексички ресурс WordNet Domains настао је полуаутоматском методом, додавањем ознака домена синсетовима Ворднета. Сваки синсет Ворднета означен је бар једном ознаком домена, из скупа од око двеста ознака структурираних у складу са WordNet Domain Hierarchy (Magnini & Cavaglià. , 2000), (Bentivogli, Forner, Magnini, & Pianta, 2004).

На слици 5 приказано је једно од пет главних дрвета у WordNet Domains хијерархији.

[54] http://wndomains.fbk.eu/index.html



Слика 5 Једно дрво у WordNet Domains хијерархији[55]

Преостала четири дрвета носе називе free_time, applied_science, pure_science и social_science. Ознака FACTOTUM додељује се у случајевима када ниједна друга ознака није одговарајућа.

У Ворднету је, на пример, синсет {mouse:1} значења „неки од бројних малих глодара који обично личе на умањене пацове пошто имају шиљате њушке и мале уши на издуженим телима са мршавим, обично глатким реповима“ обележен доменом *zoology „зоологија“*, док је синсет {mouse:2; computer mouse:1} значења „електронски уређај који контролише координате курсора на екрану; помера се по равној подлози“ обележен доменом *computer science „рачунарство“*. У Ворднету један домен може да обухвата синсетове различитих врста речи и различитих хијерархија.

Друго значајно проширење Ворднета јесте његово повезивање са SUMO онтологијом (пододељак 2.2). Повезивање је остварено са верзијом Ворднета 1.6, а остварене везе су прослеђене и у касније верзије. За ово повезивање коришћене су три врсте релација: синонимија, хиперонимија и примерак. Ове врсте релација и повезивање са SUMO онтологијом у наставку ћемо приказати на примерима који су делом преузети из рада (Крстев, и др., 2008).

Синсет из Ворднета {battle:1, conflict:3, fight:4, engagement:1} (непријатељски сусрет супротстављених војних снага у току рата) је синониман са концептом „Battle” из

[55] Слика преузета из рада (Bentivogli, Forner, Magnini, & Pianta, 2004)

SUMO онтологије, те се у синсет додаје информација „= Battle“. Синсет који је подређен овом синсету, то јест његов хипоним јесте синсет {naval battle:1} (битка између поморских флота) за који не постоји синонимни концепт у SUMO онтологији. У оваквом случају се синсет повезује са надређеним концептом, тако да се и у овај синсет додаје информација „+ Battle“. Даље, синсет {Iwo:1, Iwo Jima:2, invasion of Iwo:1} (крвава и дугачка операција на острву Иво Џима у којој су се амерички маринци искрцали на острву и поразили јапанске бранитеље у току фебруара и марта 1945. године) представља један случај битке, па се на овакве синсетове примењује трећа врста релације која указује да концепт означен ворднетом представља један члан класе коју означава SUMO концепт. Тада се синсету додаје ознака „@ Battle“.

Понекад се више синсетова из ворднета повезује релацијом синонимије са истим концептом из SUMO онтологије. Тако је ознака „= Battle“ додата и синсетовима {invasion:1} (акт којим једна армија напада противничку територију с циљем да је освоји или опљачка) и {combat:1, armed combat:1} (битка између две војне силе). Хијерархијска грана којој припада онтолошки концепт „Battle” приказана је на слици 6.

```
entity→physical→process→intentional process→
        social interaction→contest→violent contest→battle
```

Слика 6 Хијерархијска грана концепта Battle

Врх хијерархијског дрвета подкласа у SUMO онтологији приказан је на слици 7.

```
entity→
        physical→
                object→
                process→
        abstract→
                quantity→
                attribute→
                set or class→
                relation→
                proposition→
                graph→
                graph element→
```

Слика 7 Врх хијерархијског дрвета подкласа концепта Battle

Остваривање везе између SUMO онтогије и Српског ворднета приказано је у пододељку 2.5.2 овог рада, док је повезивање са доменима описано у пододељку 2.3.6.

2.3.3 Пројекат Еуроворднет

Ворднет је подстакао пројекте изградње сличних ресурса за друге светске језике. Један од првих таквих пројеката био је пројекат *Еуроворднет* (енг. EuroWordNet – EWN) који је од марта 1996. године до јуна 1999. године финансирала Европска заједница (Vossen P. , 1998).

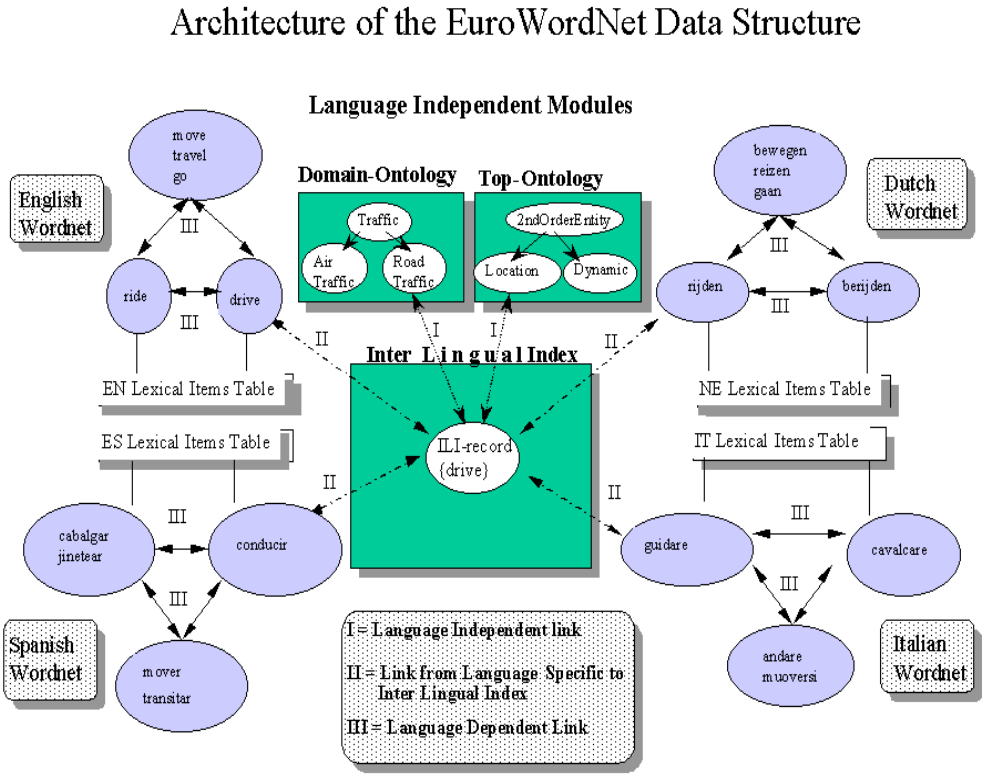
Циљ овог пројекта био је да се изгради вишејезична лексичка база података која би садржала ворднетове на осам светских језика – холандском, италијанском, шпанском, француском, енглеском, немачком, чешком и естонском језику. Структура ворднетова насталих у овиру Еуроворднет пројекта заснивала се на структури Ворднета у верзији 1.5 (Miller, Becwith, Fellbaum, Gross, & Miller, 1990). Сваки ворднет је приликом изградње ипак морао да прати специфичности појединачних језика. Зато сваки ворднет у оквиру Еуроворднет пројекта садржи скуп концепата који су у вези са особеностима лексикализације сваког од представљених језика.

Значајан допринос Еуроворднет пројекта је увођење такозваног интерлингвалног индекса (енг. Inter-Lingual-Index (ILI)). Овај индекс осмишљен је у сврху ефикасног пресликавања између структура ворднетова појединачних језика – ILI се зато може посматрати као универзални индекс значења. (Vossen P. , 1998).

ILI се састоји из синсетова Ворднета у верзији 1.5, без семантичких релација. Такозване релације еквиваленције (енг. equivalence relations) повезују синсетове на језицима који није енглески са ILI записима и тако пружају везе између еквивалентних концепата у различитим језицима. Захваљујући релацијама еквиваленције које се постижу преко ILI индекса, могуће је поредити синсетове и релације између синсетова у различитим језицима. Поред релације еквиваленције, постоји још 15 релација које пружају могућност флексибилног мапирања између концепата различитих језика (Vossen P. et al., 1997).

У сврху обезбеђивања приближно једнаке покривености концептуалних домена у свим језицима, из такозваног скупа заједничких основних концепата (енг. Common Base Concepts) развијене су различите хијерархије концепата по принципу одозго-надоле (енг. top-down). Скуп заједничких основних концепата се састојао из 1.310 синсетова из Ворднета у верзији 1.5 који су одабрани као скуп најважнијих, фундаменталних концепата. Енглески основни концепти су примењени на све језике и проширени локалним основним концептима (енг. Local Base Concepts), то јест концептима који су суштински важни за сваки појединачни језик чије су лексичко-

семантичке мреже грађене у оквиру пројекта Еуроворднет. Надаље су локални ворднетови развијани проширивањем ових основних концепата хипонимима који су били повезани преко ILI записа. Захваљујући овој архитектури, чији је делимичан приказ дат на слици 8, различити ворднетови се заснивају на заједничкој основи, али омогућен је и истовремени развој концептуализација својствених појединачним језицима. ILI запис омогућава и да се додатно знање које је уведено у Ворднет, обележја домена и SUMO концепата (пододељак 2.3.2) , користи у ворднетовима других језика.



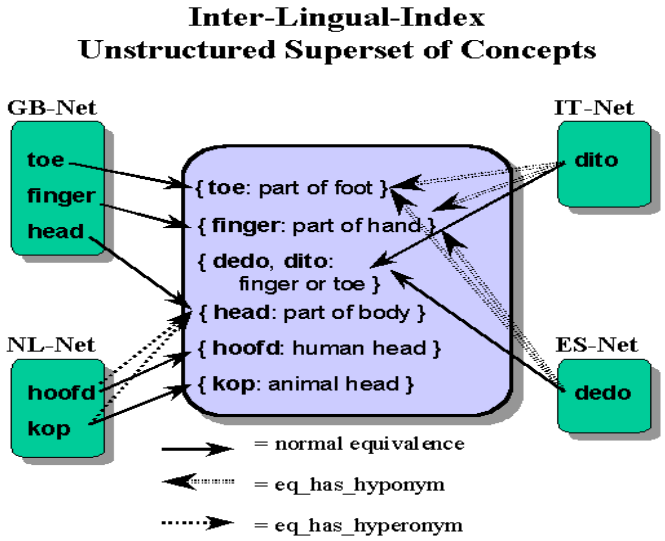
Слика 8 Приказ структуре мреже Еуроворднет[56]

Поред ILI-а, који представља неструктурирана листу синсетова из Ворднета у верзији 1.5, развијена је хијерархијска структура, независна од језика, такозвана Онтологија вишег нивоа (енг. Top Ontology) којој се може приступити преко ILI-а. Ова онтологија се састоји из 63 врховна концепта (енг. Top Concepts), који осликавају основне дистинкције из савремених семантичких и онтолошких теорија. Онтологија вишег нивоа је повезана са скупом заједничких основних концепата као скуп својстава (eng. features), тако што чвор из скупа заједничких основних концепата може бити

[56] Слика преузета са адресе http://projects.illc.uva.nl/EuroWordNet/objectives-ewn.html

повезан са неколико својстава врховних концепата, док својства врховних концепата могу бити наслеђена у концептима својственим појединачним језицима, преко скупа заједничких основних концепата или преко ILI записа (Vossen P., et al., 1997). Онтологија вишег нивоа у пројекту Еуроворднет пружа дељени семантички оквир за све језике на којима су грађени нови ворднетови, док су јединствене особености појединачних језика одржаване у сваком засебном ворднету. Тако је на основу интерлингвалног индекса и дељене врховне онтологије остварено поравнавање једнојезичних ворднетова и Еуроворднет је из вишејезичне лексичко-семантичке мреже претворен у вишејезичну лексичку онтологију (Vossen P., et al., 1997).

Дизајн Еуроворднет базе омогућава прецизно описивање лексикализације у неком језику у односу на концептуални простор. На примеру датом на слици 9 може се видети тај процес. Речи које описују делове тела у свим ворднетовима повезане су са концептима из врховне онтологије, *Part* и *Living*, али лексикализације за делове тела се разликују међу језицима. На слици видимо да у енглеском језику речи head (енг. глава) и leg (енг. нога) могу означавати делове тела људи и животиња, док у холандском језику постоје различите речи за делове тела људи и животиња – kop (хол. глава) и poot (хол. нога) за све животиње осим за коње, и hoofd (хол. глава) за људе и been (хол. нога) за коње. Слично томе, у енглеском и шпанском језику постоје различите речи за прст руке и прст ноге, док се у италијанском и шпанском језику користи иста реч за оба ова дела тела.



Слика 9 Пример примене ILI-а[57]

[57] Слика преузета са адресе http://projects.illc.uva.nl/EuroWordNet/objectives-ewn.html

У оквиру пројекта Еуроворднет дефинисане су две методологије за изградњу локалних ворднетова:

- *Модел проширења* (енг. expand model) или преводилачки модел (енг. translation-based model) подразумева изградњу ворднета на језику који није енглески превођењем основног скупа синсетова (енг. core set of synsets), уз вођење рачуна о очувању семантичких релација из Ворднета у што већој мери. То се постиже изградњом нових синсетова у складу са синсетовима Ворднета, где год је то могуће, и додавањем одговарајућих семантичких релација које већ постоје између тих синсетова. Овај модел коришћен је у пројекту Балканет, те стога и за изградњу Српског ворднета, који је због тога под јаким утицајем Ворднета.
- *Модел спајања* (енг. merge model) подразумева изградњу ворднета независно од Ворднета и других ворднетова на основу доступних једнојезичних језичких ресурса, што олакшава очување особености језика. Нови ворднет се онда поравнава са Ворднетом преко релација еквиваленције. Овај модел коришћен је, на пример, у изградњи пољског ворднета plWordNet 2.0 (Maziarz, Szpakowicz, & Pia, 2012).

Еуроворднет уводи низ нових релација у структуру Ворднета: xpos_near_synonym, xpos_antonym, role, involved, be_in_state, near_antonym, и др. (Vossen P. , 1998). Ове релације су значајне јер повезују појмове који се лексикализују различитим врстама речи.

Важна релација која повезује именичке и придевске синсетове јесте релација „бити у стању нечега“ (енг. be_in_state), као у примеру следећих синсетова повезаних овом релацијом: синсета {cleanness:1} (стање некога или нечега што је чисто) и синсета {clean:1} (који је без прљавштине или нечистоће или има навику да буде чист).

Уведене су и релације јаке и слабе синонимије и антонимије, па тако имамо релацију „скоро супротан“ (near_antonym) којом је синсет {clean:1}(који је без прљавштине или нечистоће или има навику да буде чист) повезан са синсетом {dirty:1, soiled:1, unclean:1} (који на себи има прљавштину или нечистоћу). Још неке релације које се остварују између различитих врста речи су xpos_near_synonym, xpos_antonym, xpos_near_antonym, xpos_fuzzynym, has_xpos_ hyperonym, has_xpos_hyponym. (Pazienza, Stellato , & Tudorache, 2008)

Детаљан приказ принципа, методологије и резултата пројекта Еуроворднет дат је у (Vossen P. , 1999) и на веб презентацији овог пројекта[58].

2.3.4 Пројекат Балканет

Пројекат Балканет (енг. BalkaNet Multilingual Balkan Wordnet) је за циљ имао проширивање вишејезичне базе која је успостављена у оквиру пројекта Еуроворднет балканским језицима (Tufis, Cristea, & Stamou, 2004). Балканет је од септембра 2001. године до августа 2004. године финансирала Европска комисија, а био је под вођством конзорцијума састављеног од представника 13 институција из Грчке, Турске, Србије, Бугарске, Румуније, Француске, Холандије и Чешке[59].

Као наставак и проширење пројекта Еуроворднет, Балканет је за циљ имао развој поравнатих ворднетова за балканске језике (бугарски, грчки, румунски, турски и српски, као и проширење мреже за чешки језик чија је изградња започета у оквиру пројекта Еуроворднет) и развој савремених ресурса за језике Балкана како би био омогућен приступ информацијама на тим језицима (Stamou, et al., 2002).

Главне активности у пројекту Балканет су, дакле, биле развој лексичко-семантичких мрежа по узору на Ворднет, односно Еуроворднет, за балканске језике, као и повезивање ових мрежа са Еуроворднет базом. Ове главне активности су планиране и спроведене синхронизовано, јер су једнојезичке мреже изграђене над заједнички договореним основним скуповима од 8.516 концепата који су већ били присутни у Ворднету. Изван ових основних скупова, за сваки појединачни језик мрежа се развијала независно, али у оквирима које је постављао Ворднет (Christodoulakis, 2004). Због оваквог приступа развоју мреже Балканет током рада су се постављала питања: да ли су концепти језички зависни или не, да ли су обрасци за лексикализацију концепата универзални, да ли је структура мреже Ворднета погодна и за друге језике, и да ли је скуп семантичких релација које су у њега уграђене довољан за све језике? Зато су се сви партнери на пројекту договорили да се као један од резултата рада на овом пројекту угради и скуп концепата који су специфични за балканске језике (Krstev C. , 2006).

Да би се приступило развоју овог скупа, названог Balkan Specific Concepts, требало је, пре свега, утврдити шта ће се сматрати концептом који је специфичан за

[58] Завршни извештај пројекта Еуроворднет http://projects.illc.uva.nl/EuroWordNet/

[59] Завршни извештај пројекта Балканет http://www.dblab.upatras.gr/balkanet/deliverables/finalreport_sub.pdf

Балкан, јер је у међусобној комуникацији установљено да су могући различити приступи:

- Под специфичним концептима се могу подразумевати они концепти који су специфични за појединачне језике (на пример, кајмак и стара штедња за српски);
- Као специфични концепти се могу посматрати они концепти који потичу из једног балканског језика али су се проширили и на друге балканске, па и шире, европске језике (такви су, на пример, концепти Атентат у Сарајеву и шљивовица);
- Специфични концепти Балкана су и они концепти који нису специфични само за Балкан али се препознају као заједнички, док у Ворднету нису забележени (на пример, пирамидална банка или транзиција). (Krstev, Obradović, & Vitas, 2006).

Прва од понуђених дефиниција специфичних концепата је одбачена на нивоу конзорцијума јер је закључено да овако уско одређење не би било корисно. Премда је било заговорника идеје да понуђени скуп треба да садржи само концепте заиста специфичне за Балкан, на крају је преовладало мишљење да је за будуће примене корисније да се у бази података Ворднет нађе што више концепата који су на подручју Балкана препознатљиви као значајни, независно од њиховог порекла и раширености.Договорено је да се поступак утврђивања овако дефинисаних специфичних балканских концепата обави у следећа четири корака:

1) Сваки партнер је требало да припреми листу специфичних концепата за свој језик. Током овог корака је било значајно да сваки учесник темељно провери да ли изабрани концепт већ постоји у Ворднету, како би се избегло непотребно умножавање истих концепата и њихово размимоилажење у разним језицима. Тако је, на пример, *баклава* била природни кандидат за балкански специфични концепт, али је он као концепт већ постојао у Ворднету (енгл. baklava – rich Middle Eastern cake made of thin layers of flaky pastry filled with nuts and honey).

2) Сваки партнер је упоредио своју листу са листама концепата које су понуђене за остале језике да би свој концепт повезао са истим или сличним концептима у другим балканским језицима. Тако је формирано вишејезичко језгро специфичних балканских концепата.

3) Сваки партнер је пронашао у листама осталих партнера концепте који су препознатљиви и у његовом језику и додао их у ворднет свог језика.

4) Коначно, сваки партнер је допунио новим концептима свој ворднет. (Krstev, Obradović, & Vitas, 2006).

У првом кораку је за српски језик било дефинисано 316 концепата којих нема у Ворднету, и то 259 именица, 9 глагола и 47 придева. Највећи број ових концепата се односи на храну (ајвар), породичне односе (јетрва), друштво, а највише на социјалистичко наслеђе и период транзиције (ударник, пирамидална банка), кућанство (куварица), религију (Свети Сава), обичаје (слава), митологију (баук) и историју (Косовска битка). Међу придевима су претежно унети присвојни придеви изведени из именица које се налазе у скупу српских специфичних концепата, (ћавабџијски и ћевабџијин од ћевабџија), а слично је и са глаголима (на пример партизановати од партизан) (Krstev C. , 2006) .

Сваки синсет једног језика у бази Балнканет садржи дефиницију, а специфични балкански синсетови садрже и дефиницију на енглеском да би утврђивање заједничких концепата било једноставније. Осим тога, за српски језик база Балканет садржи и примере употребе који су преузети из Корпуса савременог српског језика[60] (Vitas, Krstev, Obradović, Popović, & Pavlović-Lažetić, 2003) (пододељак 2.1.1) (Vitas 2003). Од 316 концепата који су уврштени у овај први скуп њих 54 нема потврду употребе јер корпус не бележи синониме којима се ти концепти реализују. Међу њима је највише присвојних придева, на пример деверов нема пример употребе за разлику од девер („Помажу нас моје сестре из Ужица и девери из Београда, пре свега, када је у питање одећа и обућа за децу“).

За бугарски је одређено 336 концепата, за грчки 309, за румунски 545, за турски 332 и за чешки 226. С обзиром да је чешки партнер у пројекту касно дао опис својих специфичних концепата, они нису упоређени са концептима осталих учесника. Многи понуђени концепти за друге балканске језика припадају истим доменима као и српски специфични концепти, али има и оних који се односе на биљни и животињски свет, или старе занате и занимања, традиционалну музику и плес, архитектуру, јединице мере, и тако даље.

У следећем кораку је установљено да се од 316 понуђених српских концепата њих 109 појављује бар у још једном од осталих језика. Највише заједничких концепата је пронађено међу бугарским специфичним концептима, њих 67, затим међу грчким 37, па

[60] У питању је био корпус *(NETK)* од 22 милиона речи (видети одељак 2.1.1).

румунским 29 и турским 21 (Krstev C. , 2006). Истовремено су се истим задатком бавили и остали учесници што је као коначан резултат дало скуп од 1.562 концепта.

Једина два заједничка концепта која су свих пет учесника понудила за своје језике су кадаиф и алва (табелаТабела 1). С обзиром да оба представљају посластицу, може изгледати чудно да се у овом скупу не налазе и неке друге посластице са подручја Балкана, познатије од ових. То је зато што су се концепти баклава (енгл. baklava) и ратлук (енгл. Turkish Delight) већ налазили у Ворднету, пре почетка пројекта Балканет.

Табела 1 Заједнички концепти у пет балканских језика

бугарски	**грчки**	**румунски**	**српски**	**турски**
Кадаиф	καντα΄ϊφι	cataif	кадаиф	kadayıf
Халва	αλβάς	halva	алва	kağıt helva

У сврху додавања нових, балканских концепата у Српски ворднет, вршена је провера у Корпусу савременог српског језика, како је то раније наведено, а за остале концепте, који се нису појављивали у овом корпусу, провера је вршена у речницима Матице Српске и Шкаљићевом речнику турцизама (Škaljić, 1989), као референтним речницима за српски језик и ово истраживање. Примећено је да се велики број концепата који немају потврду у корпусу може сматрати застарелим, са турском етимологијом, на пример концепти кајмакан, тахан-халва и рахле. Ипак, има и оних, као што су зуце, таратор-салата и ибришим за које се чини да су, бар у говорном језику присутни и данас, али их корпус не бележи. Уочава се, такође, да знатан број концепата везаних за ислам нема потврду у овом корпусу, на пример, ашкам-намаз, абдест и мувекит (Krstev C. , 2006). Као што показују наведени примери, концепти заједнички за више балканских језика имају често исто порекло, најчешће из турског језика. У тренутној верзији Српског ворднета[61] синсетова са ознаком BILI им 528, док синсетова са ознаком SRP, дакле специфичних концепата за српски језик, има 434.

Након што је сваки партнер у Балканет пројекту развио скуп концепата карактеристичних за сопствени језик, сваки тим је проверио карактеристичне синсетове других тимова како би се утврдило да ли има преклапања, а ти концепти су део базе података под именом BILI (Balkanet Inter-Lingual Index). У том процесу су, дакле, синсетови који одражавају концепте карактеристичне за поједине балканске језике који

[61] Стање забележено 23.12.2017.

нису лексикализовани на енглеском језику, на пример за неке врсте хране, ручно додати у ILI, са префиксом BILI, одакле је омогућено њихово повезивање са синсетовима других језика у којима постоји слично лексикализовано значење. Према завршном извештају пројекта Балканет[62] првобитна верзија ILI-а у овом пројекту била је у складу са верзијом 1.5 Ворднета, док је касније урађено ажурирање на верзију 1.7.1 и коначно на верзију 2.0 Ворднета.

2.3.5 Српски ворднет

Развој Српског ворднета започео је у оквиру пројекта Балканет (пододељак 2.3.4), као што је то био случај и са бугарским, румунским, грчким и турским ворднетом (Stamou, et al., 2002). Српски ворднет се заснива на структури Ворднета и изграђен је на принципу такозваног модела проширивања (који је предложен у оквиру пројекта Еуроворднет (пододељак 2.3.3)), у складу с правилима која је налагао пројекат Балканет, то јест копирањем синсетова из Ворднета у Српски ворднет и њиховим прилагођавањем, уз очување хијерархијске структуре Ворднета (пододељак 2.3.1).

По завршетку пројекта Балканет, Српски ворднет је садржао 8.059 синсетова од којих је 7.736 настало преузимањем из Ворднета, док је 117 припадало скупу балканских специфичних концепата, а 206 скупу концепата специфичних за српски језик (Крстев, и др., 2008). Рад на развоју Српског ворднета је после пројекта Балканет настављен неформално тј. без подршке неког за то намењеног пројекта, већ више на основу волонтерског рада. У првој години после завршетка пројекта Балканет додато је највише синсетова из домена биологије, који представљају биљне и животињске врсте као и више класификационе групе којима те врсте припадају. Избор домена био је усклађен са доградњом електронског морфолошког речника српског језика одредницама из биологије. Исте године Српски ворднет је допуњен и додатним концептима специфичним за балканске језике и концептима специфичним за српски језик (Krstev, 2006).

После те прве фазе, нови синсетови су додавани захваљујући удруженом раду професора, сарадника и студената Катедре за библиотекарство и информатику, Филолошког факултета у Београду и Групе за језичке технологије Математичког факултета у Београду. Допуна је текла у складу са потребама научника и истраживача који су ворднет користили у различитим областима, те је тако овај ресурс допуњаван

[62] BalkaNet Final Report http://www.dblab.upatras.gr/balkanet/deliverables/finalreport_sub.pdf

синсетовима из области лингвистике, биомедицине, микробиологије, генетике, права, гастрономије, рачунарства и информатике, осећања, итд. (Крстев, и др., 2008), (Antonić & Krstev, 2008) (пододељак 2.3.6). С обзиром на такав начин доградње, неки делови ове мреже су ненамерно запостављени, то јест, број неких елемената је мањи у односу на остале. Такав је случај са синсетовима придева у Српском ворднету, којих је у тренутку спровођења истраживања у овом докторском раду било 1.590 у односу на укупан број од 21.234 свих представљених врста речи[63], док је марта 2017. године тај однос био 1.622 од 21.877 (табела 31 у прилогу 4) , а децембра 2017. године 1.907 од 22.530 (табела 32 у прилогу 4).

Развој Српског ворднета се ослањао на постојеће ресурсе српског језика, највише на Речник Матице Српске. Због ограничених и неразјашњених ауторских права у вези са машински читљивом верзијом тог речника, његова употреба је ограничена (Vitas, Krstev, Obradović, Popović, & Pavlović-Lažetić, 2003). У одсуству електронског енглеско-српског речника, синсетови Ворднета су превођени ручно, уз очување семантичке структуре Ворднета. Због непостојања српског речника синонима тај процес је био додатно отежан. Значења литерала су преузимана из Речника Матице Српске, а с обзиром да је овај речник у шест томова објављен 1967. године, многа значења су недостајала (Krstev, Pavlović-Lažetić, & Vitas, 2004). Ипак, где год је то било могуће, ознаке значења литерала у Српском ворднету, које се налазе у склопу етикете <LITERAL> одговарају значењима која су дата у речнику српског језика. Ради одређивања морфолошких, синтаксичких и семантичких особина тих литерала, кодови њихових флективних класа, синтаксички и семантички маркери су уведени из електронског морфолошког речника простих речи, у склопу етикете <LNOTE>. Валидација синсетова је вршена на Корпусу савременог српског језика (Krstev, Pavlović-Lažetić, Obradović, & Vitas, 2003), (Obradović, Krstev, Pavlović-Lažetić, & Vitas, 2004), а као резултат те валидације примери употребе литерала се додају у склопу етикете <USAGE>.

Разлика у структури српског језика у односу на енглески језик, налагала је увођење нових информација у Српски ворднет. Исто важи за бугарски језик, који има сличну структуру, дељену са структуром српског језика и других словенских језика, те су нове информације додате и у бугарски ворднет[64] (Koeva, Krstev, & Vitas, 2008). Информације које су додате тичу се правила флекције и извођења речи. Деривациони

[63] Подаци из јуна 2015. године.
[64] Бугарски ворднет http://dcl.bas.bg/resursi/wordnet/

механизми између енглеског језик и словенских језика су системски различити. У морфо-семантичком погледу, неке од најпродуктивнијих деривационих релација у словенским језицима су аспект глагола, парови родова и деминутиви или умањенице (Krstev, Obradović, & Vitas, 2006), тако да су у Српском и Бугарском ворднету у односу на то уведене релације *derived-pos, derived-vn* и *derived-gender*, редом за присвојне и релационе придеве, глаголске именице и женске или мушке дублете, тј. именице које се у овим и другим словенским језицима користе и у женском роду (изведене су из именица мушког рода моцијом рода), премда је и данас уобичајено коришћење тих именица у мушком роду (на пример за имена занимања), што је феномен који не постоји у енглеском језику – именица teacher означава и мушку и женску особу чији је посао да држи предавања.

Када говоримо о родно равноправним терминима, што је у последње време веома актуелна тема, у Српском ворднету тако имамо, на пример, синсет {predavač:1} који одговара енглеском синсету {teacher:1, instructor:1} (особа чије је занимање да предаје) и има варијанту у женском роду {predavačica} уз дефиницију „женска особа чије је занимање да предаје“.

У Српском ворднету су аспектни парови уврштени у оквиру истог синсета, где год је то имало смисла, на пример, у синсету {zamišljati:2x, zamisliti:2x, dočaravati:2x, dočarati:2x, predočavati:1, predočiti:1} који одговара синсету Ворднета {visualize:1, visualise:3, envision:1, project:9, fancy:1, see:4, figure:3, picture:1, image:1}, док етикета LNOTE која одговара сваком литералу у синсету описује флективне и деривационе особине сваког глагола – на пример, за глагол „замишљати“, садржина LNOTE етикете је V1+Imperf+Tr+Iref+Ref, док је за глагол „замислити“ садржина LNOTE етикете V162+Perf+Tr+Iref+Ref.

Синсет Српског ворднета {lutka:1} који одговара синсету Ворднета {doll:1, dolly:3} (реплика особе која се користи као играчка) стоји у вези *diminutive* са синсетом {lutkica}. Деминутиви тј. умањенице су у Српском ворднету уврштени само у случајевима када и у Ворднету постоји слична лексикализација,најчешће у случајевима када је деминутив добио неко специфично значење, премда би, у духу српског језика, умањенице скоро свих именица могле да буду укључене релативно једноставно, за разлику од енглеског језика, у коме се умањенице обично изражавају фразама (Koeva, Krstev, & Vitas, 2008). Можда ће управо то бити правац неког будућег истраживања са циљем доградње Српског ворднета.

Све наведене релације су у Српском ворднету искоришћене углавном за концепте специфичне за Балкан и за концепте специфичне за српски језик.

Српски ворднет је као лексички ресурс примењен у истраживањима вишечланих лексичких јединица (енг. multi-word units) (Krstev, Stanković, Obradović, Vitas, & Utvić, 2010), (Mitrović, 2014) (Mitrović, Mladenović, & Krstev, 2015), претрази вишејезичних дигиталних база података (Stanković, Krstev, Obradović, & Utvić, 2012), препознавању реторичких фигура (Mladenović M. , 2016), анализи ставова и осећања изражених у тексту (Mladenović M. , Mitrović, Krstev, & Vitas, 2015) моделовању система за детекцију аргумената и аргументације (Mitrović, O'Reilly, Mladenović, & Handschuh, 2017), за побољшавање претраге дигиталног речника дијалеката (Mladenović, Stanković, & Krstev, 2017), за детекцију реторичких фигура иронија и сарказам (Mladenović, Krstev, Mitrović, & Stanković, 2017).

2.3.6 Повезивање Српског ворднета са доменима

Од почетка 2006. године отпочео је кооперативни рад на наставку доградње Српског ворднета. С обзиром да постдипломске студије на Филолошком факултету Универзитета у Београду на Групи за библиотекарство и информатику често уписују студенти са других катедри Филолошког факултета, или са неких других факултета, створени су услови да управо ти студенти допринесу изградњи Српског ворднета својим знањем из различитих домена, у оквиру обавезног семинарског рада који је прописан правилником постдипломских студија (Крстев, и др., 2008).

За одабир подскупа синсетова који највише одговарају појединачним студентима и њиховом стручном знању, коришћен је софтверски алат *WS4LR* (Obradović & Stanković, 2008) (пододељак 2.4). С обзиром да Српски ворднет, као и Ворднет, користи XML формат, овај алат је омогућаваоио да се формулишу XML Path изрази којима могу да се одаберу синсетове из изабраног ворднета и домена. За одабир је коришћена комбинована информација о припадности домену и онтолошкој категорији јер информација о домену често производи превелики подскуп синсетова. Ворднет је, у време кооперативног рада на доградњи Српског ворднета, тако имао 1.181 синсет из домена права, што би било превише за један семинарски рад.

За домен лингвистике, на пример, била је задужена Бојана Ђорђевић, студенткиња докторских студија Језик, књижевност, култура на Филолошком факултету у Београду, која је завршила основне студије из Опште лингвистике. У оквиру пројекта доградње Српског ворднета доменима из области лингвистике требало је преузети

синсетове Ворднета из тог домена и прилагодити их за Српски ворднет. Обрађени скуп синсетова из лингвистичког домена обухватао је категорије: морфеме (16 синсетова), граматику (238 синсетова), карактере (87 синсетова) и природне језике (595 синсетова). Касније су додати још неки синсетови, ради обезбеђивања правилних хијерархијских веза са остатком Српског ворднета, те је укупан број синсетова из домена лингвистике који су прилагођени 946. На слици 10 видимо пример једног од преведених синсетова који је преузет из категорије природних језика.

SYNSET>	<SYNSET>
<ID>ENG20-06480396-n</ID>	<ID>ENG20-06480396-n</ID>
<POS>n</POS>	<POS>n</POS>
<SYNONYM>	<SYNONYM>
<LITERAL>mother tongue	<LITERAL>maternji jezik
<SENSE>1</SENSE></LITERAL>	<SENSE>1</SENSE></LITERAL>
<LITERAL>maternal language	<LITERAL>prvi jezik
<SENSE>1</SENSE></LITERAL>	<SENSE>1</SENSE></LITERAL>
<LITERAL>first language	<LITERAL>rođeni jezik
<SENSE>1</SENSE></LITERAL>	<SENSE>1</SENSE></LITERAL>1
</SYNONYM>	</SYNONYM>
<ILR>	<ILR>
<TYPE>hypernym</TYPE>	<TYPE>hypernym</TYPE>
ENG20-06479855-n</ILR>	ENG20-06479855-n</ILR>
<DEF>one's native language; the language learned by	<DEF>jezik koji je najpre usvojen u
children and passed from one generation to the	detinjstvu ili onaj kome se daje prednost u
next</DEF>	višejezičnoj situaciji</DEF>
<DOMAIN>linguistics</DOMAIN>	<SNOTE>Uradila B. Đorđević,
<SUMO>NaturalLanguage<TYPE>+</TYPE></SUMO>	postdiplomac C. Krstev</SNOTE>
</SYNSET>	</SYNSET>

Слика 10 Повезивање домена природног језика[65]

Други домени који су повезивани на сличан начин су домен биомедицине, домен религије, домен литературе, домен права, домен библиотекарства и издаваштва, што је детаљно описано у раду (Крстев, и др., 2008). Методологија коришћена у пројекту повезивања са доменима била је инспирација за утврђивање методологије спровођења пројекта групне расподеле рада која би била најпогоднија у сврху доградње Српског ворднета новим везама заснованим на реторичкој фигури поређење. Тај процес је био знатно олакшан новим алатима које имамо за доградњу и одржавање овог важног језичког ресурса за српски језик, које ћемо описати у пододељку 2.4.3 овог рада.

У каснијим фазама кооперативног рада на доградњи Српског ворднета, ауторка овог докторског рада је 2011. године, у склопу Самосталног истраживачког рада на

[65] Слика преузета из рада (Крстев, и др., 2008)

докторским студијама Филолошког факултета у Београду допунила Српски ворднет синсетовима придева (више од 460). Потреба за овом врстом доградње се јавила због планова да Српски ворднет користимо у задацима анализе ставова и осећања (резултати тог истраживања дати су у раду (Mladenović M. , Mitrović, Krstev, & Vitas, 2015), али и због планираних истраживања представљених у овом докторском раду.

2.4 Алати за обраду језика

Језички ресурси се могу адекватно користити и унапређивати само ако за њих постоје добро развијени алати којима се ресурси могу обрађивати, евалуирати, допуњавати, унапређивати и повезивати са другим језичким ресурсима и алатима.

2.4.1 Алати за обраду енглеског језика

У свету су јавно доступни многи алати за обраду природног језика које развијају одељења познатих универзитета. Тако Група за обраду природног језика Универзитета Станфорд (енг. University of Stanford) развија и одржава алате за обраду природног језика, највише из области статистичке обраде језика (енг. statistical NLP), али и неке алате потребне за обраду језика засновану на правилима (енг. rule-based NLP). Неки од тих алата су Stanford CoreNLP, Stanford Parser, Stanford POS Tagger, Stanford Named Entity Recognizer, Stanford RegexNER, Stanford Coreference Resolution, Stanford Word Segmenter, Stanford Classifier, Stanford EnglishTokenizer, Stanford TokensRegex, Stanford Temporal Tagger (SUTime), Stanford Pattern-based Information Extraction and Diagnostics (SPIED), Stanford Relation Extractor. Чланови ове групе редовно објављују научне радове на главним конференцијама из области обраде природног језика.[66]

Пројекат Apache OpenNLP[67] је пројекат у оквиру кога се развијају многи корисни алати за обраду природног језика, а функционише по принципу доприноса волонтера, те би се овај пројекат могао посматрати као пројекат групне расподеле рада када би било потпуно јасно како се врши евалуација доприноса учесника у овом пројекту (о чему ћемо више говорити у пододељку 3.5 овог докторског рада).

Ипак, иако су у свету доступни многи алати за обраду природног језика, највише као резултат рада великих светских универзитета чија се одељења баве овом облашћу, највећи број тих алата изграђен је за потребе обраде енглеског језика и других великих светских језика, те је мали број тих алата могуће успешно користити за српски језик.

[66] Радови су доступни на адреси https://nlp.stanford.edu/pubs/.
[67] https://opennlp.apache.org/

2.4.2 Алати за обраду српског језика

На изради неких од првих алата за обраду српског језика радио је професор др. Душко Витас крајем седамдесетих година прошлог века. Први од ових алата носи назив *Аурора* и представља оригиналан и комплексан систем за обраду текста. Основни модул овог система заснива се на могућности да се за улазни текст, означен ознакама логичке структуре, генерише интерна репрезентација тог текста таква да указује на релевантне елементе изворног текста. (Vitas, 1979), (Vitas, Krstev, Pavlović-Lažetić, & Nenadić, 1998). У овом програмском систему први пут је примењена веома важна кодна шема за обраду српског језика, такође названа Аурора. Ова кодна шема се реализује коришћењем карактерског скупа ASCII[68] који не садржи карактере специфичне за српски језик, то јест слова са дијакритицима и латиничне диграфе, и увођењем новог начина записивања слова који омогућава да се постигне разликовање диграфа од консонантских група, на пример, латинично слово „lj“ и ћирилично слово „љ“ су у кодној шеми Аурора представљени као „lx“, док су „nj“ и „њ“ представљени као „nx“. Аурора је најпогоднији интерни код за језичке ресурсе српског језика јер неутралише разлику између латиничног и ћириличног писма, и зато што се на текст кодиран Аурором увек може применити ASCII кодна шема, што је веома значајно зато што многи корисни алати за обраду текста подржавају само ASCII. (Utvić, 2013). Аурора се користи и данас као интерни код у електронским морфолошким речницима српског језика (пододељак 2.1.2), у Српском ворднету (пододељак 2.3.5), у семантичкој мрежи за властита имена Prolex (Maurel, 2008), као и у једнојезичним и вишејезичним корпусима (пододељак 2.1.1).

Систем MORPH за генерацију морфолошких облика именица и придева развијен је почетком осамдесетих година прошлог века. Овај систем за основни облик речи генерише све могуће облике, у складу са правилима промене именица или придева. Систем се заснива на анализи којом се дата реч дели на непроменљиви и променљиви део, на пример, у речи *котао*, систем дели реч на *кот-* и *-ао*. Систем додељује такозване морфографемске дефиниције којима дефинише све промене и њихов редослед које се дешавају приликом генерисања морфолошких облика, на пример, реч *котао* ће у генитиву једнине гласити *котла*, што је феномен који је описан помоћу морфографемске дефиниције (Vitas, 1981).

У оквиру рада Групе за језичке ресурсе и технологије користе се и развијени су многи алати чији је развој инспирисан радом на разноврсним језичким ресурсима, кроз

[68] ASCII American Standard Code for Information Interchange

различите пројекте и методолошке оквире. Алати су развијани како би био олакшан даљи развој и одржавање тих ресурса, као и њихова интеграција са постојећим, светским алатима и ресурсима.

Алат *WS4LR* (енг. акроним настао од речи workstation for lexical resources) (пододељак 2.4.2) је развијен током рада на докторској дисертацији проф. др Ранке Станковић (Stanković, 2009), чији је ментор био проф. др Душко Витас, док је софтверско решење настало под руководством проф. др Цветане Крстев. Овај алат је детаљно описан у радовима (Obradović & Stanković, 2008) и (Krstev, Stanković, Vitas, & Obradović, 2006), а његов развој је настављен је у оквиру Групе за језичке ресурсе и технологије под називом *LeXimir*.

LeXimir[69] је алат за израду, одржавање и експлоатацију језичких ресурса (Stanković, Obradović, Krstev, & Vitas, 2011), (Stanković, Obradović, & Trtovac , 2012). Овај алат омогућава синхронизовано коришћење разнородних ресурса и садржи неколико компоненти које извршавају различите функције:

- Омогућене су различите врсте трансформација ресурса (једне датотеке или скупа датотека) које могу садржати текст, локалне граматике, електронске речнике формата DELAS и DELAF;
- Подсистем за одржавање система морфолошких речника омогућава управљање скупом одабраних речника у DELA формату који садрже просте или сложене речи. Одабрани речници могу да буду дистрибуирани у више датотека. Главна снага алата је могућност ефикасног претраживања и издвајања подскупа лема на основу услова поређења лема, врсте речи, кода флективне класе, и синтаксичких и семантичких ознака;
- Развој и унапређење ворднета је компонента која подржава рад са појединачним ворднетом али и синхронизован рад два ворднета који се остварује преко ILI-а (в. пододељак 2.3.3). Осим тога, синсетови се могу одабрати коришћењем различитих метода, од једноставног сравњивања ниски до комплексних XPath израза[70] за које су припремљени обрасци који одговарају често постављаним захтевима;
- Подсистем за интеракције система електронских речника и онтологија омогућава размену информација између ворднета и морфолошких речника тако

[69] Leximir http://korpus.matf.bg.ac.rs/soft/LeXimir.html
[70] XML Path Language

што се морфо-синтаксичке информације из морфолошких речника могу придружити литералима у синсету, а семантичке информације из синсетова се могу придружити лемама у речницима.

Као надградња овог алата, развијена је веб апликација за проширивање упита под називом *VebRanka* (Stanković, Krstev, Obradović, & Utvić, 2012); , чији је циљ да омогући коришћење и развој српских језичких ресурса на вебу. У том смислу је развијен и веб сервис *wsQueryExpand* који се може користити и независно, као засебна компонента. *VebRanka и LeXimir*, дају кориснику могућност да прошире упит морфолошки, семантички и на још један језик, што зависи од расположивих ресурса.

Acide је окружење за генерисање и рад са паралелним текстовима (Utvić, Stanković, & Obradović, 2008);

NERosetta се користи за претраживање паралелних текстова обележених именованим ентитетима (Krstev, Zečević, Vitas, & Kyriakopoulou, 2013);

Bibliša[71] је алат за претраживање поравнатих текстова из електронских часописа (Stanković, Obradović, & Trtovac , 2012); настао је на основу компоненти алата *LeXimir* и *Вебранка. Bibliša* користи различите језичке ресурсе као што су мреже типа Ворднет, електронски речници и листе термина за морфолошко и семантичко проширење упита, као и превођење упита на друге језике. Развој Библише је започет на корпусу текстова *Инфотека – часописа за дигиталну хуманистику*[72], који се објављује двојезично, те за сваки објављени научни рад или приказ постоји верзија на српском језику и верзија на енглеском језику. Даљи развој је обухватио и друге текстуелне колекције: све радове часописа Менаџмент и Подземни радови који су публиковани двојезично, део радова часописа Стоматолошки гласник и Архитектура и урбанизам, текстове везане за активности на пројектима BAEKTEL, INTERA и CESAR. Библиша користи српске електронске морфолошке речнике, Ворднет и Српски ворднет повезане преко ILI-а и терминолошке речнике међу којима су: двојезични речник библиотечких термина, геолошки и рударски речници. (Stanković, Krstev, Vitas, Vulović, & Kitanović, 2017)

NERanka[73] се користи за обележавање текстова именованим ентитетима (Krstev C. , Vitas, Obradović, & Utvić, 2011). Под именованим ентитетима подразумевамо имена особа, геополитичких назива, организација, производа, итд. Ова веб апликација се

[71] Bibliša http://hlt.rgf.bg.ac.rs/Biblisha/Default.aspx

[72] Инфотека http://infoteka.bg.ac.rs/index.php/sr

[73]http://metashare.elda.org/repository/browse/neranka-named-entity-recognition-and-annotation-tool/0d8e850a8be211e280ed0015171445925f9139c1f5f54dd8aaa7529d90c0a025/

заснива на компонентама алата *LeXimir* и *VebRanka*, док основни део система чине каскаде коначних трансдуктора и може се користити на различитим језичким ресурсима. Обележавање етикетама именованих ентитета у систему *NERanka* је могуће за XML документе и тектуелне документе. XML етикетама се описује врста и подврста сваког именованог ентитета.

Већ постојећи системи за обраду природно-језичких података *Nooj* и *Unitex* користе се такође веома успешно уз прилагођавања карактеристична за српски језик. Nooj модул за српски језик (SrpNooj), настао је у оквиру пројекта CESAR и представља лингвистичко развојно окружење које кроз примену електронских морфолошких речника и граматика анализира корпусе. Овај модул може се користити у задацима анотације, морфосинтаксичког обележавања, проналажења информација или екстракције информација (Stanković, Vitas, & Krstev, 2007).

Unitex је систем отвореног кода (енг. open source) за обраду текста. Користи се за развој и примену коначних аутомата (енг. finite-state automata) и трансдуктора (енг. transducers) у процесу обраде текста на природном језику коришћењем електронских речника у DELA формату и локалних граматика. Систем Unitex је првобитно развијен за француски језик, да би касније био развијен и за енглески, италијански, грчки (класични и модерни), шпански, немачки, корејски, руски, пољски, румунски, фински, норвешки и друге језика (Paumier, Nakamura, & Voyatzi, 2009). Најновија дистрибуција овог система садржи ресурсе за 22 језика, међу којима је и српски језик и његови ресурси на оба званична писма, ћирилици и латиници[74] .

Недавно су развијени нови алати за изградњу и одржавање Српског ворднета, које користимо и за друге примене у оквиру активности Друштва за језичке ресурсе и технологије – Јертех. Више о овим алатима говорићемо у пододељку 2.4.3 овог рада.

2.4.3 Нови алати и побољшања за Српски ворднет

Развој ворднета је захтеван задатак који најчешће подразумева удружени рад великог броја стручњака. Ако желимо да један комплексан ресурс као што је ворднет буде квалитетан, нарочито ако га стварамо од почетка и без коришћења аутоматских метода, за његов развој ће бити потребно неколико година. Тако се може десити да вредан ресурс врло лако постане застарео, јер ворднет у основи има речи, а речи могу лако да изађу из употребе и да више не буду актуелне. У данашње време

[74] http://unitexgramlab.org/language-resources

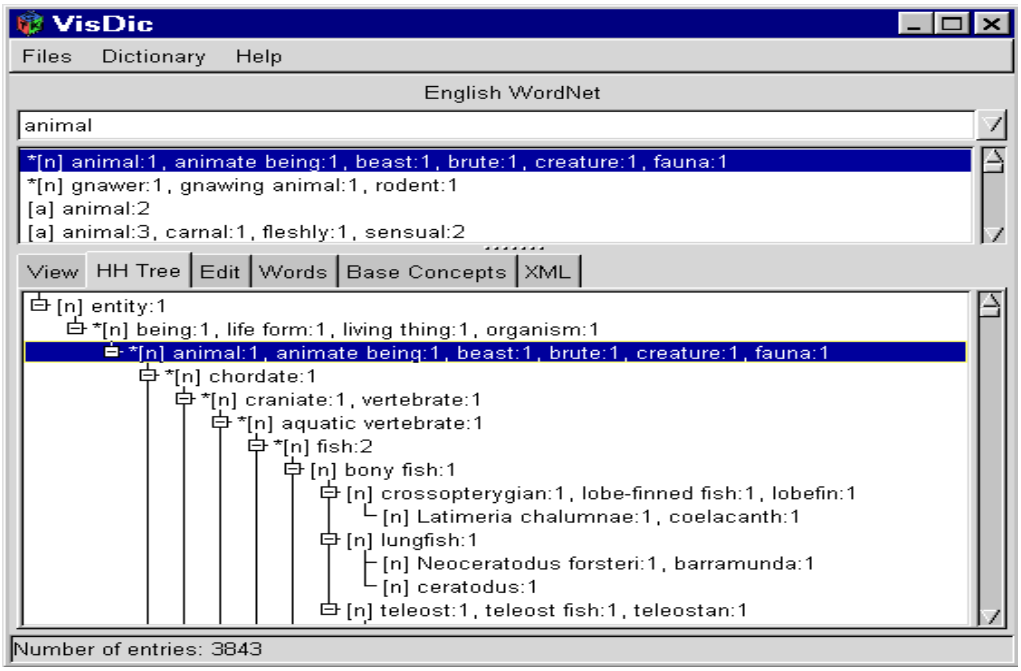
експоненцијалног развоја друштвених мрежа, интернет технологија и сталног откривања нових начина комуникације, у дигиталном свету стално настају нове речи које прате управо та кретања. Тако је за 2009. годину, судећи по истраживању Оксфордских речника (портала који обједињује сва електронска издања ових речника), реч године била реч коју бисмо у слободном преводу на српски језик, у недостатку адекватног превода, морали да опишемо као „прекид пријатељства на некој друштвеној мрежи, најчешће Фејсбуку“ (енг. unfriend), док је за 2015. годину то била реч селфи (енг. selfie), то јест резултат новог тренда фотографисања мобилним телефонима. Реч године за 2016. годину била је реч за коју на српском језику такође још увек немамо усвојени еквивалент, а један од предлога јесте „постчињенични“ (енг. post-truth). Овај израз постоји већ читаву деценију, али у протеклој години је примећен нагли пораст учесталости његовог коришћења у контексту референдума одржаних у земљама Европске уније и председничких избора у Сједињеним америчким државама.[75] Имајући у виду уплив оваквих речи у говорни језик, веома је важно изнаћи начин да се такви трендови прате и да се, у складу са потребама, језички ресурси унапреде овим и сличним неологизмима.

Алат *VisDic* је годинама коришћен за развој и одржавање Српског ворднета, од самог почетка пројекта Балканет (пододељак 2.3.4), и показао се као поуздан и лак за коришћење (Horák & Smrž, 2004). Поред овог алата, Српски ворднет је могуће користити и преко алата *LeXimir*, *VeBranka* и *Bibliša*, описаних у пододељку 2.4.2. *VisDic* је био нарочито користан за рад на неколико ворднетова идентичне XML структуре истовремено. Претрага се спроводила у изворном ворднету, а захваљујући ILI-ју резултате је било могуће приказати и за све остале ворднетове чије су датотеке биле учитане у VisDic уређивач. Веза између тих докумената остваривала се на два начина – преко функције *AutoLookUp*, која је повезивала синсетове различитих ворднет докумената са истим идентификационим ознакама (ILI), где је резултат било њихово представљање једног поред другог, као и преко функције *CopyEntryTo* која је омогућавала копирање садржаја неког синсета из једног ворднет документа у други. Захваљујући функцији *CopyEntryTo* овог алата, синсетови Ворднета копирани су у Српски ворднет и дорађивани, уз вођење рачуна о очувању хијерархијске структуре.

Функционалност претраживања у овом алату ослањала се на представљање синсетова путем структуре дрвета у оба смера (према корену и према листовима). У том

[75] https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016

смислу су примењиване две операције: *TopmostEntries* и *FullExpansion*. Прва операција је пружала добијање листе свих синсетова који су представљали корен релационе хијерархије. Друга операција пружала је увид у све синсетове који су представљали делове поддрвета у датој претрази. На слици 11 се може видети начин на који су синсетови вордента и њихови хијерархијски односи били приказани овим алатом.



Слика 11 Алат VisDic

Помоћу VisDic алата било је могуће да се донекле контролише доследност података, на пример, било је могуће добити информацију о неким неусклађеностима – синсетовима са идентичним идентификационим ознакама, дупликатима парова *Literal/Sense*, или о дуплим везама између синсетова.

У првих неколико година развоја Српског ворднета, овај бесплатни алат је значајно допринео том процесу. Ипак, због чињенице да је алат ограничен на такозвану једнокорисничку употребу, тимски рад је био отежан. С обзиром да је развоју српског ворднета веома често истовремено доприносило више волонтера, одржавање је било отежано (Крстев, et al., 2008). С обзиром да је све документе које су корисници градили требало увек изнова спајати у један, главни документ, често су се поткрадале грешке и настајале недоследности. Због тога је било неопходно да се унапреди или промени алат за одржавање Српског ворднета. XML документи који су били коришћени у алату VisDic нису имали корени елемент, те није постојала ни функција за проверу валидности и

добре формираности XML докумената у односу на неки DTD[76] или XSD[77]. Резултат такве употребе алата била је разноврсна структура синсетова у Српском ворднету. С обзиром да није постојала контрола валидности, корисници алата су могли да уносе и неочекиване и нежељене вредности етикета. Ограничен систем морфолошког обележавања у алату VisDic такође није био погодан за уношење синсетова на морфолошки богатом српском језику. Зато су морфолошке информације касније додате ручно, уз помоћ Српских електронских морфолошких речника (пододељак 2.1.2), а за то је била задужена главна уредница Српског ворднета, професор др. Цветана Крстев. Морфолошке информације су додаване унутар елемента LNOTE који испрва није био намењен за ову врсту информација (пододељак 2.5.1).

Овакав начин рада био је подложан грешкама и знатно је успоравао процес додавања нових уноса. Исти проблем се појавио и приликом додавања SUMO ознака (пододељак 2.5.2) за синсетове који су специфични само за Српски ворднет, те зато нису пренесени из Ворднета, као што је то био случај са синсетовима који носе BILI ознаке – синсетови који су додати у оквиру пројекта Балканет 2.3.4).

Примећено је, такође, да алат VisDic нема неке функције које су у раду често потребне, као што је, на пример, могућност провере да ли неки синсет „виси“, то јест, да ли се негде поткрала грешка и постоји случај да неки синсет нема свој надређени синсет који је с њим у хипернимном односу. Недостајали су и основни алати за статистичку анализу – за одређивање броја синсетова и литерала у односу на врсту речи, број полилексичких литерала у односу на једноставне литерале, број литерала са највећим бројем значења, број синсетова са највећим бројем литерала, итд.

Немогућност повезивања Српског ворднета са SUMO онтологијом и другим онтологијама вишег нивоа (eng. upper level ontologies), као и са доменским онтологијама, успоравала је развој алата за онтолошко закључивање заснованих на нашем ворднету. Чињеница да је било немогуће трансформисати XML документ у друге формате докумената, нарочито у RDF[78] и OWL[79] још више је отежавала развој база знања заснованих на онтологијама које су повезане са Српским ворднетом.

Систем претраживања алата VisDic ослањао се на елементарне упите над садржајем, без могућности подешавања логичких филтера или могућности паметне

[76] DTD – document type definition
[77] XSD – XML scheme definition
[78] RDF – resource description format
[79] OWL – ontology web language

претраге, на пример коришћењем XPath синтаксе. Тако је, на пример, помоћу нових алата сада могуће генерисати упит „//SYNSET[DOMAIN='geology']“ за издвајање домена или „//SYNSET[ILR/TYPE/child::text()='derived']“ за добијање литерала насталих деривацијом.

Узевши у обзир све предности и недостатке постојећег софтверског решења, развијен је скуп алата који за циљ имају побољшање процеса изградње Српског ворднета, као и осталих семантичких ресурса за рачунарску обраду српског језика.

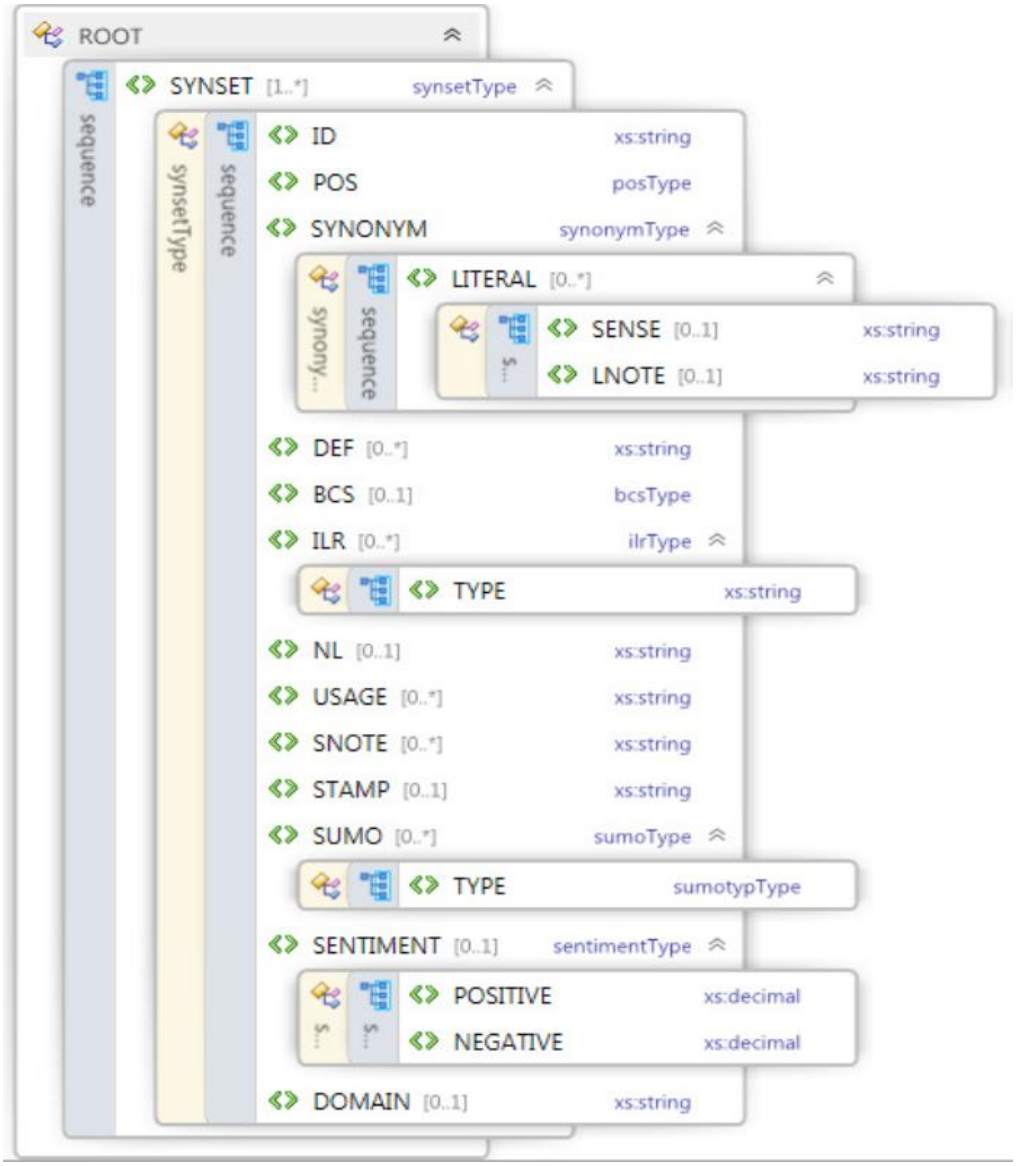
Веб апликација коју смо назвали Serbian WordNet Edition (SWNE)[80] последњих неколико година се користи за изградњу, одржавање и унапређивање Српског ворднета (Mladenović & Mitrović, 2014). Ова апликација обезбеђује:

- побољшани систем морфолошког означавања синсетова, кроз бољу повезаност Српског ворднета (пододељак 2.3.5) и електронских морфолошких речника српског језика (пододељак 2.1.2) која је постигнута побољшавањем система контроле постојећих веза, контроле над новим везама које настају приликом додавања нових синсетова, као и ограничавањем списка могућих вредности етикете LNOTE;
- побољшани систем синхронизованог приказивања синсетова и њихових семантичких релација у Српском ворднету и Ворднету као и синхронизована претрага синсетова у оба ова ресурса;
- повезивање синсетова Српског ворднета са концептима из SUMO онтологије (пододељци 2.2 и 2.5.2) уз омогућавање контролисања исправности постојећих SUMO етикета и проверу успостављања веза са SUMO концептима током додавања нових синсетова и вредности SUMO етикета, у зависности од врсте речи, то јест POS етикете синсета;
- Нове семантичке етикете у Српском ворднету, на основу семантичког ресурса SentiWordNet (Esuli & Sebastiani, 2006);
- Комплекснији систем претраживања захваљујући додавању логичких филтера и нових параметара приликом формирања упита;
- Побољшана контрола над XML документом ворднет, на основу XSD шеме;
- Требало је омогућити и благовремено ажурирање Српског ворднета у складу са новом верзијом Ворднета, с обзиром да је по завршетку пројекта Балканет, Српски ворднет био повезан са верзијом 2.0 Ворднета.

[80] SWNE апликација http://sm.jerteh.rs/

Основна предност алата SWNE лежи у омогућавању и олакшавању колаборативног рада на изградњи Српског ворднета. Овај алат омогућава надгледање процеса развоја ворднета захваљујући статистичким подацима које је могуће добити, на пример, укупан број синсетова, број семантичких релација, релације у вези с другим семантичким ресурсима, квалитет и брзина додавања синсетова кроз списак аутора који су додавали нове синсетове и датума када су ти додаци извршени (Mladenović, Mitrović, & Krstev, 2014) .

На слици 12 дата је графичка репрезентација актуелне верзије XSD схеме која се користи за валидацију XML структуре документа Српског ворднета, што значајно доприноси процесу валидације овог важног ресурса и омогућава бржи кооперативни рад на његовој доградњи.



Слика 12 XSD схема XML документа Српског ворднета[81]

2.5 Везе Српског ворднета

2.5.1 Веза између Српског ворднета и електронских морфолошких речника српског језика

Ова два важна ресурса за рачунарску обраду српског језика – Српски ворднет и електронски морфолошки речници функционишу у интегрисаном окружењу од када је у оквиру Групе за језичке технологије Математичког факултета у Београду развијена „радна станица“ за језичке ресурсе (енг. *WS4LR: A Workstation for Lexical Resources*) (Krstev, Stanković, Vitas, & Obradović, 2006). Сваки од ова два ресурса користи специфичан скуп информација и података које други ресурс не садржи. По узору на рад

[81] Слика преузета из (Младеновић, 2016)

овог алата, омогућено је да SWNE веб апликација (пододељак 2.4.3) такође може да користи податке из електронских морфолошких речника српског језика, те тако сваки синсет Српског ворднета који носи етикету <LNOTE> може да буде повезан са скупом морфолошких правила лексеме која је дефинисана етикетом <LITERAL> на коју се етикета < LNOTE > односи. Такође, није могуће одабрати морфосинтактичку етикету за < LNOTE > ако она не припада скупу дозвољених етикета. Српски морфолошки речници садрже различите морфолошке речнике у LADL формату (пододељак 2.1.2) али за потребе повезивања са SWNE апликацијом, коришћен је речник простих лема – DELAS, чији је основни унос облика

lemma.Knnn [+SinSem]

где је Knnn јединствена ознака класе флективног правила, чији је део K ознака врсте речи. У SWNE алату, све Knnn ознаке из DELAS скупа могуће је повезати са литералима синсетова Српског ворднета, при чему се везује само ознаке чија је врста речи у складу с врстом речи литерала.

2.5.2 Веза између Српског ворднета и SUMO онтологије

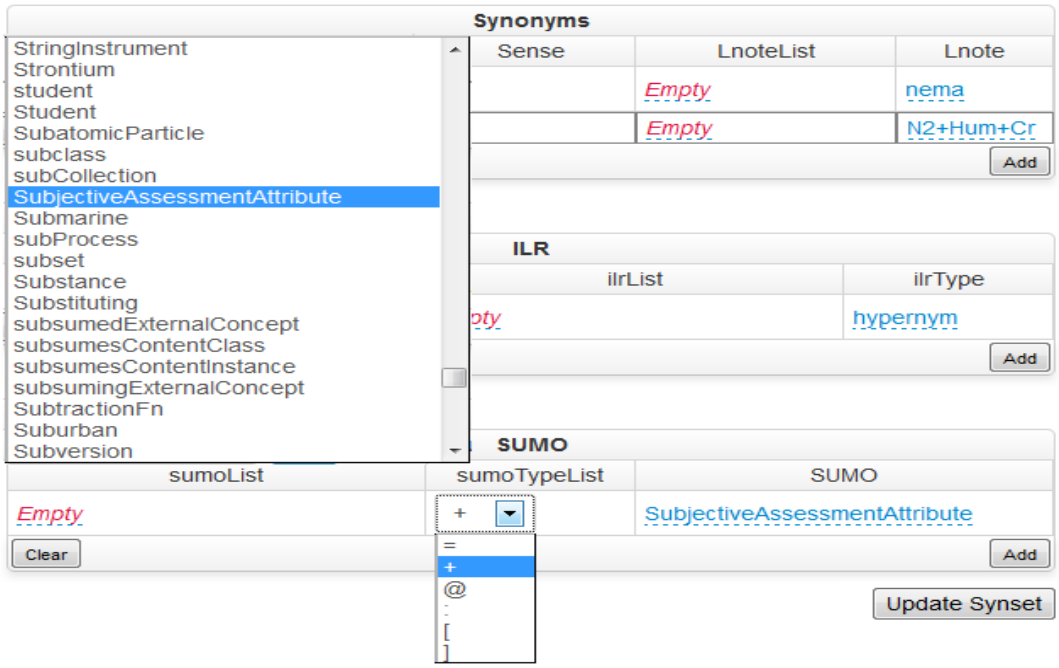
Онтологија представља концептуалну репрезентацију знања на формалан начин. SUMO онтологија (акроним од речи Suggested Upper Merged Ontology) развијена је 2000. године на иницијативу Адама Пизија (енг. Adam Pease) (Pease, 2011), као прва формална онтологија која остварује мапирање са свим синсетовима Ворднета. Првобитна верзија Ворднета за коју је урађено ово мапирање је била верзија 1.6 и тај процес је био ручни, док је од верзије 3.0 мапирање између ова два важна ресурса аутоматско и редовно се ажурира.

SUMO садржи релативно мали број концепата (око 1.000), уз око 400 тврдњи (енг. assertions) и око 800 правила, што омогућава процес аутоматског закључивања, проналажења информација и може се користити у многим задацима из области обраде природног језика. SUMO хијерархија се заснива на концептима и релацијама између тих концепата. На њеном врху се налази општи концепт – *entity* (енг. ентитет), а на најнижим нивоима се налазе веома специфични концепти, на пример *atoll* (енг. врста коралног острва). Ова онтологија је универзална у смислу описивања уобичајених концепата и односа међу њима, без обзира на било који филозофски систем или културне утицаје.

База података ворднета се може обогатити додавањем етикета сваком синсету у складу са одговарајућим SUMO концептом. SUMO и Ворднет дефинишу неке

концептуализације, то јест поједностављивање начина на који схватамо свет. Циљ ворднета јесте да те концептуализације представи природним језиком, док је циљ SUMO онтологије да их организује у логичку структуру.

Захваљујући јавно доступним датотекама које садрже недвосмислене парове за мапирање између синсетова Ворднета и SUMO концепата, било је могуће да у Српски ворднет додамо SUMO етикете. Тако су синсетови Српског ворднета добили етикете SUMO концепата кроз процес хоризонталне паралелизације (између енглеског синсета и српског синсета) и вертикалног наслеђивања (српски неетикетирани синсет је наслеђивао SUMO етикету родитељског синсета, ако такав постоји, или је додаван ручно) (слика 13).



Слика 13 Повезивање синсета Српског ворднета са SUMO онтологијом [82]

Описано повезивање обављено је помоћу алата SWNE (пододељак 2.4.3).

Додавањем етикета SUMO концепата синсетовима Српског ворднета створени су повољни услови за интеграцију *РетФиг* онтологије реторичких фигура за српски језик (пододељак 2.2.2) са SUMO онтологијом у циљу омогућавања аутоматског закључивања у процесу препознавања реторичких фигура (Mladenović, 2016).

[82] Слика преузета из рада (Mladenović & Mitrović, 2014)

2.5.3 Повезивање Српског ворднета са SentiWordNet ресурсом

Нови алати за изградњу и одржавање Српског ворднета омогућили су унапређивање овог важног ресурса у смеру који га чини повољним за коришћење у задацима анализе ставова и осећања (енг. sentiment analysis).

У последње време, изузетна количина информација која се сваког тренутка ствара на блоговима, форумима, сајтовима друштвених мрежа, комерцијалним сајтовима као што је Амазон, на којима је могуће оставити коментаре за производе које смо преко тог сајта купили, отвара могућност и ствара потребу за изналажењем аутоматских метода за анализирањем ставова и осећања која преовлађују у таквим порукама. Анализа ставова и осећања је подобласт обраде природног језика која се заснива на постојању семантичких ресурса који могу да подрже нове, све сложеније захтеве који се јављају уз свеприсутну експанзију информација на интернету. Ти ресурси су обично интегрисани са Ворднетом или на неки начин остварују комуникацију са Ворднетом. Такав ресурс, који се веома често користи у овим задацима јесте SentiWordNe (Esuli & Sebastiani, 2006), јавно доступан лексички ресурс за анализу ставова и осећања, заснован на Ворднету. Овај ресурс додељује три ознаке интензитета и поларитета осећања (енг. sentiment) сваком запису који је у вези са Ворднет синсетом – позитиван, негативан и објективан став и осећање. Тако се одређују позитивни и негативни концепти представљени синсетовима Ворднета, у складу са неким ставом или осећањем које одражавају. За сваки синсет у Ворднету постоји један запис у бази SentiWordNet коме се додељују нумеричке ознаке Pos – за позитиван и Neg – за негативан став и осећање, који описује која је мера објективности, позитивних ставова и осећања, или негативних ставова и осећања у вези са концептом који је описан датим синсетом. На пример, један запис у SentiWordNet бази који одговара Ворднет синсету који дефинише придев *gladsome* представљен је на следећи начин:

a 01361705 0.75 0 gladsome#1

где су вредности 0.75 и 0 степен нумеричке јачине за позитиван и негативан став и осећање, „a 01361705“ значи да је горе приказаним записом представљен придев у Ворднет датотеци са ознаком ID=01361705 и ознака #1 представља број значења. Све три мере су бројеви од 0 до 1, где 0 означава одсуство неког поларитета, а 1 присуство у највећој мери. Збир ове три мере је 1, па зато у ворднету мера за објективност није приказана јер се може израчунати на основу две дате мере, нпр. за пример на слици 14, мера за објективност ће бити 0.25.

У процесу обогаћивања Српског ворднета етикетама интензитета и поларитета осећања из SentiWordNet базе, увели смо две етикете за сваки синсет Српског ворднета и остварили везу са одговарајућим записом у SentiWordNet ресурсу. Апликација у оквиру алата SWNE је развијена посебно у ту сврху. Пример једног синсета из Српског ворднета након додавања ознаке интензитета и поларитета осећања приказан је на слици 14.

```
<SRPWN>
<SYNSET>
<ID>ENG30-01361705-a</ID>
<POS>a</POS>
<SYNONYM>
<LITERAL>drag<SENSE>1</SENSE><LNOTE></LNOTE></ LITERAL >
</ SYNONYM >
<DEF>koji oseća ili izražava radost i veselje</DEF>
<BCS></BCS>
<ILR>ENG30-01361414-a<TYPE>similar_to</ TYPE ></ILR>
<NL>yes</NL>
<USAGE></USAGE>
<SNOTE></SNOTE>
<STAMP>08/10/2012 00:00:00 jeca</STAMP>
<SUMO>EmotionalState<TYPE>+</TYPE></SUMO>
<SENTIMENT>
<POSITIVE>0.75000</POSITIVE>
<NEGATIVE>0.00000</NEGATIVE>
</SENTIMENT>
</SYNSET>
</SRPWN>
```

Слика 14 Синсет Српског ворднета са етикетама интензитета и поларитета осећања

Као и у случају остваривања веза са SUMO онтологијом (пододељак 2.5.2), праћен је принцип хоризонталне паралелизације (енглески синсет – српски синсет) и вертикалног наслеђивања (српски синсет без етикете наслеђује етикету SENTIMENT свог родитељског синсета, ако такав синсет постоји, у супротном се етикета додаје ручно). Расподела класа сентимената у Српском ворднету приказана је у оквиру прилога **Error! Reference source not found.**, у табелама Табела 35 и Табела 36. Овако унапређени Српски ворднет коришћен је у оквиру хибридног система за анализу ставова и осећања званог САФОС који је изграђен у оквиру израде докторске дисертације др. Миљане Младеновић (Младеновић, 2016), а резултати коришћења овог система представљени су у раду (Mladenović, Mitrović, Krstev, & Vitas, 2015).

3 Групна расподела рада

3.1 Одређење појма

Групна расподела рада (енг. crowdsourcing) је модел управљања и пословни модел који се све више примењује у науци и култури. С обзиром на чињеницу да је овај термин настао као комбинација речи crowd (енг. група људи) и outsourcing (што описно значи препуштање дела пословних задатака особама које нису запослене у некој компанији, обично особама које ће бити плаћене мање него сопствени запослени), у српском језику је веома тешко пронаћи одговарајућу сложеницу, те смо се, у сврху писања овог рада, одлучили за описни назив „*групна расподела рада*“, који ћемо надаље користити за опис овог модела.

Термини за означавање групне расподеле рада настали на основу енглеског термина усвојени су, пак, у неким другим језицима. Тако у грчком језику постоји адекватна кованица – *Πληθοπορισμός* (што значи групни напор, труд, настојање). Страница на Википедији посвећена овом феномену [83] написана је на 39 светских језика. На хрватском језику наслов те странице носи назив „*Набава из мноштва*“, а написана је само једна, уводна реченица која описује овај модел рада. На бугарском и руском језику написани су дужи, детаљни чланци, а наслов је само транскрибован, енглески назив, то јест „*Краудсорсинг*“. На француском језику, користи се термин „*Externalisation ouverte*“ (у слободном преводу, отворено ангажовање спољних сарадника) и „*production participative*“ (у слободном преводу, производња захваљујући сарадњи). Страница на Википедији посвећена групној расподели рада на немачком језику носи исти назив као и страница на енглеском језику, то јест *Crowdsourcing*. Страница на Википедији на српском језику настала је недавно као део рада на овој тези, те је употребљен термин који овде користимо – групна расподела рада[84].

Овај модел рада је релативно нов и као такав још увек није добио коначну и потпуну дефиницију. Предузећа и појединци који га користе увек изнова проналазе нове начине његове употребе, па тако настају и разне дефиниције. Сматра се да је термин *групна расподела рада* увео Џеф Хауи и то у чланку под називом Успон групне расподеле рада (енг. The rise of crowdsourcing) који је објављен у часопису Wired. Дајући примере из сфере бизниса, Хауи је показао како се знање и вештине неке групе

[83] Wikipedia Crowdsourcing https://en.wikipedia.org/wiki/Crowdsourcing
[84] Grupna raspodela rada https://sr.wikipedia.org/wiki/Grupna_raspodela_rada

појединаца (масе) могу веома добро искористити, чак иако они нису запослени у компанији којој је нека услуга потребна, и да је најбоље давати задатке преко отвореног позива. Још тада је било јасно да интернет игра веома важну улогу у овом, тада још увек недовољно истраженом феномену, премда се овај модел рада користио и пре настанка интернета.

3.1.1 Неке дефиниције групне расподеле рада

Један од првих покушаја да се дефинише модел групне расподеле рада долази од творца овог термина, Џефа Хауија, који је овај модел назвао пословном праксом која дословно значи расподелити задатке (енг. термин који је употребљен је „outsource“) у неком пројекту групи људи (Howe, 2006). Хау је касније, у другом раду, указао на значај овог начина расподеле рада и назвао га механизмом по коме се таленат и знање једне групе људи могу упарити са потребама других (Howe, 2008). Касније су други аутори давали сопствене дефиниције, те је Дејвид Алан Грир (енг. David Alan Grier) описао групну расподелу рада као начин да се интернет употреби за обједињавање доприноса великог броја људи који другачије не би имали прилику да сарађују на истим пројектима (Grier, 2011). Казаи (енг. Kazai) се такође бавио истом проблематиком те ову врсту сарадње сматра отвореним позивом групи људи да реше неки проблем или да обаве неки задатак за који је неопходна људска интелигенција, дакле за задатак који рачунари не могу да обаве, обично у замену за малу новчану надокнаду, стицање статуса у друштву или ради забаве (Kazai, 2011).

Дарен Брабам (енг. Daren Brabham) групну расподелу рада дефинише као модел дистрибуиран преко интернета, који се користи за решавање проблема и производњу, а користе га профитне организације попут Threadless[85] (портал на коме свако може да предложи дизајнерско решење за изглед одевних предмета, за које остали учесници онда гласају) и iStock[86] (репозиторијум најразличитијих фотографија високог квалитета, по ниским ценама, које могу користити новинске агенције, аутори блогова, итд.). Исти аутор у другом раду (Brabham D. C., 2008) дефинише групну расподелу рада као стратешки модел који привлачи заинтересовану, мотивисану групу људи који су способни да пруже решења која квалитетом и квантитетом превазилазе решења која се могу добити традиционалним облицима пословања.

[85] Threadless https://www.threadless.com/
[86] iStock http://www.istockphoto.com/

У књизи индикативног назива, *Crowdsourcing*, (Brabham D. C., 2013) наилазимо на још једну, овога пута свеобухватну, дефиницију овог модела, чији ћемо превод овде навести у целости – „Групна расподела рада је онлајн, дистрибуирани модел решавања проблема и производње који користи моћ колективне интелигенције онлајн заједница за специфичне организационе циљеве.“[87] Брабам наглашава да је групна расподела рада јединствена по томе што комбинује креативни, отворени процес који долази „одоздо“ са организационим циљевима који долазе „одозго“. Групна расподела рада није производња софтвера отвореног кода, коме недостаје компонента одозго-надоле; то није ни истраживање тржишта у коме се учесницима нуди кратак списак могућих избора; то је квалитативно другачија активност у односу на пројекте отворених иновација и производне процесе засноване на сарадњи, који су постојали пре дигиталног доба, јер су таквим пројектима недостајали брзина, доступност, велике могућности, као и све предности које нам је донео интернет. Брабам надаље описује интелектуалне корене идеје групне расподеле рада кроз концепте као што су колективна интелигенција, групна мудрост и дистрибуирано рачунарство. Он говори и о најважнијим проблемима у пројектима групне расподеле рада као што су мотивација учесника, погрешна представа о оваквим врстама пројеката коју имају учесници аматери, прикупљање новчаних средстава (енг. crowdfunding) и опасност од експлоатације волонтера (енг. crowdsploitation). Брабам сматра да ће модел групне расподеле рада тек показати да може да игра веома важну улогу у новинарству, управљачким системима, националној безбедности, науци и здравству.

Ванг, Хоанг и Кан (енг. Wang, Hoang and Kan) дају општу дефиницију по којој је групна расподела рада стратегија на основу које се уједињеним напором јавности решава неки проблем, или настаје неки ресурс (Wang, Hoang, & Kan, 2013), док Бучелер (енг. Buecheler) и његови сарадници (Buecheler, Sieg, Füchslin Rudolf M, & Pfeifer, 2010) описују групну расподелу рада као посебну врсту колективне интелигенције (енг. collective intelligence).

Саксон, Оу и Кишори (енг. Saxton, Oh and Kishore) (Saxton, Oh, & Kishore, 2013) сматрају да је модел групне расподеле рада могућ ако су остварена три услова: 1) постојање „гомиле“[88], то јест разнолике групе учесника, 2) расподела рада изван неке

[87] “Crowdsourcing as an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals. Online communities, also called crowds, are given the opportunity to respond to crowdsourcing activities promoted by the organization, and they are motivated to respond for a variety of reasons.”

[88] енг. Crowd

компаније и 3) коришћење напредних интернет технологија. Исти аутори тако и дефинишу групну расподелу рада као „модел у коме организације користе напредне интернет технологије како би искористиле напоре неке групе људи у покушају да спроведу специфичне организационе задатке.“

Једна од најсвеобухватнијих дефиниција групне расподеле рада дата је у раду (Estellés-Arolas & González-Ladrón-de-Guevara, 2002) и овде ћемо је навести у целости – „Групна расподела рада је врста учествовања у некој онлајн активности у којој особа, институција, непрофитна организација или компанија нуде групи особа различитих нивоа знања, хетерогености и броја, преко флексибилног отвореног позива, да се добровољно посвете неком задатку. Посвећивање задатку, који је разнолике комплексности и модуларности и у коме група људи учествује својим знањем или искуством, новцем, радом, увек укључује узајамну корист. Учеснику ће бити задовољена нека потреба, било да је то економска, друштвена, лична потреба, или развијање сопствених вештина, док ће особа, институција или организација која нуди задатак, оно чиме су учесници допринели, а чији облик ће зависити од врсте задатка, искористити за сопствене потребе.”[89]

3.1.2 Преглед термина сличних групној расподели рада

Концепт групне расподеле рада проширио се на многе области људског деловања, те тако постоје и многи термини који у својој основи имају енглеску реч *crowd*, у значењу група људи, гомила, маса. Надаље ћемо користити реч група због могућег пежоративног значења речи маса и гомила у српском језику. Такви концепти и термини дати су у раду ауторке Весне Ињац-Малбаше (Ињац-Малбаша, 2012/2013), а овде ћемо навести и ближе објаснити неке од њих.

Crowdfunding је финансирање које потиче од групе људи, процес финансирања пројекта у коме велики број појединаца улаже мале износе средстава како би се реализовао неки групни циљ. Овако су покренути многи значајни пројекти, а у новије време се чак и велике филмске продукције финансирају управо на основу овог модела.

[89] “Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.”

Учесници у једном оваквом пројекту донирају одређену суму новца у замену за неке услуге или добра, која ће им бити на располагању само у случају да се планирана сума новца сакупи до унапред одређеног датума. Обично се одреди време за које треба прикупити потребна средства (најчешће шест месеци), и ако овакав начин сакупљања новчаних средстава не успе, свим учесницима се уложени новац враћа. У случају да се уједињеним напором људи ипак дође до тражене суме, награде се крећу од карата за премијеру филма за који су се средства прикупљала, до излиставања имена финансијера као продуцената филма, или чак присуства на снимању филма и упознавања са главним протагонистима. Овакав приступ најчешће се користи у пројектима из области уметности, а платформе које га подржавају су Kickstarter[90], Crowdrise[91] и SeedUps[92]. Threadless[93] је једна од првих креативних платформи које су покренуте на основу групне расподеле рада, а основана је управо кроз crowdfunding модел. На платформи Threadless свако може да постави своје креативно решење за дизајн мајице кратких рукава, и однедавно још неких одевних предмета, а остали учесници гласају за најбоља решења која се на крају пуштају у продају као готови производи.

Crowdslapping се односи на ситуацију када узнемирена група људи „узвраћа ударац“ и уместо подршке промотивној кампањи реализује антипропагандни садржај.

Crowddamping је процес у коме велики број произвођача засипа тржиште робом различитог квалитета коју група појединаца ипак уме да процени.

Crowdcomputing је свеобухватни термин који подразумева велики број алатки које омогућавају размену идеја, нехијерахијско одлучивање и пуно коришћење когнитивног вишка групе појединаца.

Crowdblanking је отворен формулар или веб-сајт за нове идеје или речи; на пример, група појединаца може да смишља и додаје своје неологизме у неки речник.

Crowdsharing је назив за размену идеја и мишљења у оквиру групе људи.

Crowdmotivation се односи на мотивацију групе појединаца који учествују у активностима групне расподеле рада.

Crowdslang или жаргон групе људи, је речник свих неологизама који потичу од корена речи crowd.

[90] Kickstarter https://www.kickstarter.com/
[91] Crowdrise https://www.crowdrise.com/
[92] SeedUps https://www.seedups.com/
[93] Threadless https://www.threadless.com/

Crowdstorming је такмичење у идејама, „олуја“ идеја, при чему група појединаца непрестано развија и додаје нове идеје на неку тему.

Crowdvoting се појављује када неки веб сајт прикупи мишљење и процену великог броја појединаца о некој теми.

Неки термини се односе на концепт групне расподеле рада, али у свом називу немају реч crowd.

Citizen science или наука грађана је научно истраживање које воде делом или у целини појединци који нису стручњаци ни научници, а обично се спроводи путем активности групне расподеле рада.

Collective intelligence или колективна интелигенција је способност групе људи да пронађе више решења и боља решења неких проблема него што би их пронашао сваки њен појединачни члан.

Open innovation – отворена иновација означава нову парадигму која подразумева да организација може и треба да користи и интерне и екстерне идеје и моделе са постојећег тржишта како би унапредила своју технологију.

Social collaboration – друштвена сарадња, односи се на интеракцију између великог броја појединаца који размењују идеје како би остварили неки заједнички циљ; природно окружење друштвене сарадње је интернет који омогућава и чини лакшим проток нових идеја, сарадњу и размену информација.

Virtual volunteering – виртуелно волонтирање, термин који означава волонтера који извршава задатке, делом или у целини, изван матичне организације, користећи интернет из своје куће, школе, канцеларије, користећи и друге уређаје осим рачунара. Познато је и као онлајн волонтирање, сајбер волонтирање, телементоринг, итд.

3.1.3 Модел групне расподеле рада пре настанка интернета

Модел групне расподеле рада свакако није настао захваљујући модерним технологијама, али је захваљујући њима доживео убрзан развој. Премда се данас подразумева да у једном пројекту групне расподеле рада кључну улогу имају интернет технологије, сам концепт овакве врсте организовања пројеката постојао је много пре настанка интернета.

Далеке 1714. године, Британска влада је потражила решење такозваног „Проблема географске дужине“ (енг. the longitude problem). Одређивање географске ширине у то време није представљало тако велики проблем јер ју је било могуће израчунати на основу надморске висине и положаја сунца у подне. Проблем одређивања

географске дужине је био мало тежи, и од његовог решења је зависило одређивање тачне позиције бродова на отвореном мору. Зато је Британска влада понудила 20.000 фунти (што је данас близу 5 милиона америчких долара) обичним људима који би понудили решење за овај проблем. Тако је решење и пронашао човек из радничке класе Џон Харисон (John Harrison), ниског нивоа формалног образовања, а награђени су и многи други појединци који су пружили занимљива решења овог проблема (O'Donnell, 2002).

Сличан пројекат описан је у раду (Proctor, 2013). Наиме, кустос Џозеф Хенри (енг. Joseph Henry), први секретар Смитсонијан института је, користећи тадашњу нову технологију, телеграф, подстицао грађане да му шаљу временске извештаје из целе земље и тако је, 1856. године, направио прву мапу временских услова за Сједињене Америчке Државе.

Један од раних пројеката групне расподеле рада из области лексикографије је пројекат изградње Оксфордског речника енглеског језика (енг. Oxford English Dictionary – OED). Већина историјских и лексичких информација које се налазе у овом речнику заснована је на милионима цитата који су сакупљени из енглеских текстова кроз такозвани Читалачки програм (енг. Reading programme) како би се што боље представила употреба речи. Читалачки програм започет је 1857. године, када су волонтери почели да прикупљају цитате за Нови енглески речник у издању Британског филолошког друштва. У овом пројекту је коришћен папир и традиционалне поштанске услуге захваљујући којима су прикупљане вредне информације. Две деценије касније, покренут је шири читалачки програм, објављен путем јавног позива волонтерима у Британији, али и у Америци и Британским колонијама.[94]

Пројекат изградње Сиднејске опере се такође може сматрати пројектом групне расподеле рада. Идеја је била да се удруженом снагом великог броја људи изнађе архитектонско решење за овај објекат. Тако је 1955. године објављен јавни позив за пројекат изградње зграде опере у Сиднеју. Услови су били да то буде зграда са салом од 3.000 места и са мањом салом од 1.200 места, те да свака од њих буде изграђена за различите намене. На овај позив пријавило се 233 архитеката из 32 земље. Пошто је пројекат одабран, јавност се побунила јер је одбачен онај пројекат за који је она сматрала да је најбољи. Влада је на крају попустила и после две године је објављено да је ипак победио дански архитекта Јерн Уцон (Jørn Utzon), чији се пројекат испрва налазио међу

[94] Oxford dictionaries blog http://blog.oxforddictionaries.com/2014/02/can-world-englishes-benefit-crowdsourcing/

одбаченима. Његова грађевина налик на барку са једрима данас је симбол Сиднеја (Stuart, 2011).

Пројекат Математичке табеле је установљен за време велике економске кризе 1938. године у Америци. На њему је ангажовано 450 незапослених стручњака и то као део програма за олакшавање проблема насталих услед економске рецесије када је много људи остало без посла. Они су израдили Математичке табеле које је објавио Колумбија универзитет. Стручњаци су имали задатак да направе математичке табеле експоненцијалних функција, логаритамских таблица, тригонометријских функција, итд. Пројекат је био један од највећих и најзначајнијих у области математике пре проналаска рачунара. *Приручник математичких функција са формулама, графиконима и математичким табелама* објављен је 1949. године и постао је једна од најчешће цитираних математичких и научних референци. (Grier, 1998).

Загат водичи су још један резултат групне расподеле рада пре доба модерних технологија. Године 1979. брачни пар Загат, Тим и Нина (енг. Tim and Nina Zagat), почели су да прикупљају информације о светким ресторанима како би направили својеврсну ранг листу на основу њиховог квалитета. У почетку се истраживање односило само на град Њујорк и испитаници су били пријатељи и познаници брачног пара Загат, а касније су одговори стизали и од грађана из целог света. Водичи Загат у 2005. години покривају преко 70 градова, захваљујући одговорима преко 250.000 појединаца. У овим водичима се рангирају и хотели, ноћни живот, продавнице, музичка дешавања, филмови, позоришта, авионски превоз, итд. Компанија Гугл је 2011. откупила права на Загат водиче, и они су сада јавно доступни и интегрисани у Гугл сервисе. [95]

3.1.4 Мотивација учесника у пројектима групне расподеле рада

Питање мотивације учесника је једно од најважнијих питања која се постављају приликом организовања једног пројекта групне расподеле рада. Аутори (Estellés-Arolas & González-Ladrón-de-Guevara, 2002) повезују мотивацију за учешће са Масловљевом пирамидом потреба (Maslow , 1943). Тако мотивација, поред новчане надокнаде, може бити и прилика да се развију креативне способности и друге вештине, да се забавимо, да поделимо знање с другима, љубав према заједници, привржености задатку који треба обавити, жеља да у друштву коме припадамо будемо препознати захваљујући својим доприносима у оквиру неког пројекту.

[95] Zagat survey, https://en.wikipedia.org/wiki/Zagat

Истраживање аутора (Chamberlain, Kruschwitz, & Poesio, 2012) указује на три главна мотива која покрећу учеснике неког пројекта групне расподеле рада: лични, друштвени и финансијски мотив.

1) Лични мотиви огледају се у жељи учесника да једноставно буду део неког пројекта. Уопштено речено, лични мотив је присутан увек када је некоме câмо учествовање у пројекту довољна награда и када неко жели да допринесе пројекту за који сматра да је вредан. Добар пример мотивације учесника види се у пројекту Duolingo (детаљније у одељку 3.3.2), у коме учесници преводе делове веба, а оно што добијају заузврат јесте прилика да бесплатно уче стране језике, и то на веома добро структуриран и ефикасан начин (Vesselinov & Grego, 2012).

2) Друштвени мотив и начин подстицања односи се на награђивање побољшавањем друштвеног статуса учесника у контексту друштвених интернет мрежа и у односу на друге учеснике у пројекту групне расподеле рада. Показало се да је коришћење система бодовања и нивоа ефикасно, јер учесници често испуњавају више задатака прописаних пројектом управо да би достигли већи број бодова и прешли на следећи ниво (von Ahn & Dabbish, 2008).

3) Када говоримо о новчаној надокнади, то јест подстицању учесника пројеката групне расподеле рада на финансијски начин, Amazon Mechanical Turk (описан у пододељку 3.2.3) се намеће као платформа која се у те сврхе најчешће користи. За сваки задатак који учесник обави преко ове платформе, одређена је сума новца коју ће добити од организације која је тај пројекат осмислила и понудила. Amazon.com, као власник ове платформе, за сваку појединачну исплату коју учесник у пројекту добије, узима додатних десет посто од особе или организације која је осмислила дати пројекат. Најнижа цена која се може понудити за појединачни задатак на овој платформи јесте $0.005 по задатку, тако да компанија Amazon, у том случају, од сваког исплаћеног хонорара добија $0.0005.

3.2 Поделе и жанрови у оквиру модела групне расподеле рада

Модел групне расподеле рада користи се у најразличитије сврхе, те у оквиру њега постоје многе поделе. Неки пројекти се могу сврстати у више од једне категорије, али углавном на основу мотивације учесника можемо закључити која је примарна група којој неки пројекат припада.

Једна од првих подела у оквиру пројеката групне расподеле рада настала је на основу *врсте задатака* које је потребно обавити. С обзиром да сваки пројекат зснован

на овом моделу треба да привуче велики број учесника, обично волонтера, задатке је потребно пажљиво структурирати имајући у виду циљ самог пројекта.

Микрозадаци (енг. microtasks) су мали, добро дефинисани задаци које преко одређених онлајн платформи, као што је платформа Mechanical Turk (одељак 3.2.3) обавља група људи. Сваки учесник испуњава више малих делова целокупног задатка, да би се на крају сва решења „склопила“ у коначно решење до кога треба доћи у датом пројекту.

Проблеми који се могу решити преко микрозадатака су они проблеми који се лако могу поделити на велики број једноставних задатака, како би радници, то јест учесници у пројекту групне расподеле рада, могли лако да их реше, за релативно кратко време, без поседовања неких специфичних знања и вештина. Тако се микрозадаци могу користити за означавање и разврставање садржаја (Maji, 2011), за проналажење специфичних садржаја на интернету, и уопште за задатке које људи с лакоћом обављају, а рачунарски програми још увек нису достигли степен развоја који би им омогућио да буду једнако успешни у њиховом извршавању.

Кроз микрозадатке се могу спровести све врсте задатака који су довољно модуларни, то јест који се лако могу изменити, раставити на ситније задатке, и онда поново саставити у смислену целину. На разним платформама преко којих се могу постављати и извршавати микрозадаци, било да су то велике, добро познате платформе, какав је MTurk, или мање платформе, изграђене за потребе мањих пројеката, спроводе се најразличитије врсте задатака. Неки од тих задатака су: означавање слика (енг. tagging images), означавање делова слика (на пример лица, делове одеће), конверзија слика, оптимизација слика, пречишћавање база података, послови транскрипције, преводилачки послови, дигитализација докумената (на пример, укуцавање података са пословне картице у неки веб формулар), конверзија датотека, модерација садржаја (енг. content moderation), верификација садржаја, претрага садржаја (на пример, проналажење одређеног писма електронске поште из скупа контаката), категоризација садржаја (на пример, одабир најбоље категорије за неки елемент, креирање садржаја, прикупљање података, итд).

Да би се микрозадаци пружили на увид потенцијалним учесницима у пројекту неопходно је да буду добро дефинисани, дакле да учесницима буде јасно шта треба да ураде, те да на неки начин буду организовани у групе, било да је то по сличности или по припадности истом делу пројекта. Показало се да се кроз микрозадатке неки послови обављају брже и ефикасније него када би те исте задатке на уобичајен начин обављале

запослене особе у оквиру обавеза које им налаже радно место (Chamberlain, Kruschwitz, & Poesio, 2012).

Макрозадаци (енг. macrotasks) се односе на пружање читавог задатка или проблема на увид групи људи, како би свако могао да одреди који и колики део ће решити и како ће пројекту допринети у складу са својим знањима и компетенцијама. Учесници сами одређују ниво сопствених способности и унапред треба да процене да ли уопште треба да се прикључе неком пројекту. Тако свако може да одреди најбољи начин обављања задатака, као и део пројекта који ће преузети на себе. По томе се модел групне расподеле рада заснован на макрозадацима умногоме разликује од модела заснованог на микрозадацима, јер у случају потоњег често нису потребна никаква посебна знања и вештине – задаци које учесници тако обављају често имају везе са природно усађеним способностима свих људи.

Макрозадаци су добри за развојне и истраживачке пројекте (такозване R&D - Research and development) и за иновацију производа (енг. product innovation). Највеће платформе које користе макрозадатке су Quirky[96], Innocentive[97] и Chaordix[98]. Макрозадаци су у пројектима групне расподеле рада много мање заступљени, па не постоји ни много примера добре праксе, нити много предлога за најбољи начин спровођења таквих пројеката.

У оквиру модела групне расподеле рада постоје многи жанрови. У раду (Braslavski, Mukhin, Ustalov, & Kiselev, 2016) се говори о три врсте жанрова:

1) Групна мудрост (енг. wisdom of the crowds – WOTC)

2) Механизовани рад (енг. mechanized labor – MLab)

3) Игре са сврхом (енг. games with a purpose – GWAP).

Слична подела дата је и у раду (Wang, Hoang, & Kan, 2013), с тим што се поред групне мудрости и игара са сврхом, као трећи жанр наводи Amazon Mechanical Turk (MTurk), као главна платформа на којој се одвијају пројекти такозваног механизованог рада, те је зато и сам жанр назван по њој.

Аутори Квин и Бедерсон групну расподелу рада називају *distributed human computation* (дистрибуирано људско рачунање), уместо уобичајеног енглеског термина c*rowdsourcing*, те предлажу поделу на осам жанрова (Quinn & Bederson, 2009):

1) Игре са сврхом

[96] Quirky https://www.quirky.com/
[97] Innocentive https://www.innocentive.com/
[98] Chaordix https://www.chaordix.com/

2) Механизовани рад (или MTurk)

3) Групна мудрост

4) Групна расподела рада (користе термин crowdsourcing)

5) Рад са двоструком сврхом (енг. dual-purpose work),

6) Велика претрага (енг. grand search)

7) Генетски алгоритми засновани на раду људи (енг. human-based genetic algorithm)

8) Прикупљање знања од волонтера.

Још једна подела на жанрове у оквиру парадигме групне расподеле рада дата је у раду (Yuen, Chen, & King, 2009):

1) Иницијативно људско рачунање (енг. initiatory human computation)

2) Дистрибуирано људско рачунање (енг. distributed human computation)

3) Људско рачунање с волонтерима засновано на друштвеним игрицама (енг. social game-based human computation with volunteers)

4) Плаћени инжењери

5) Играчи онлајн игрица

Неки аутори су анализирали и специфичне теоријске аспекте модела групне расподеле рада. Тако су у раду (von Ahn & Dabbish, 2008) представљени општи принципи дизајна игара са сврхом. Аутори овог рада одређују три класе игара и за сваку дефинишу основна правила игре и услове за победу такве да је у интересу играча да добро изведе задатак који је циљ игре у којој учествује. Исти аутори су и творци једне од најуспешнијих игара са сврхом, о којој ћемо говорити у пододељку 3.3.2.

У наредним одељцима ћемо говорити о жанровима групне расподеле рада који се најчешће користе у области обраде природног језика.

3.2.1 Групна мудрост

На основу до сада спроведених пројеката групне расподеле рада показало се да је за различите врсте ових пројеката неопходно ангажовати различито структуриране групе људи, како би групна мудрост (енг. wisdom of the crowd) била што боље искоришћена.

Групна мудрост односи се на појаву да у неком пројекту група која се састоји од великог броја учесника може бити много успешнија од неколико стручњака. На примеру сајта Википедија (енг. Wikipedia) доказано је да велики број учесника који чине једну хетерогену групу, у смислу нивоа образовања и стручности, побољшава ниво квалитета

чланака на Википедији (Arazy, Morgan, & Patterson, 2006), премда многи чланци на Википедији неретко остају ниског квалитета.

У неким пројектима је боље користити допринос што хетерогеније групе особа (на пример, за пројекте обележавања слика или за пројекте који укључују прикупљање мишљења о новим производима). У другим пројектима је, ипак, неопходно упослити групу људи која је хомогена, то јест сличног нивоа образовања и стручности, како би резултати били бољи, на пример у случају уметничких пројеката, где је неопходно изнаћи решење за дизајн новог производа (Estellés-Arolas & González-Ladrón-de-Guevara, 2002).

Transcribe Bentham[99] је такође веома важан пројекат групне расподеле рада који се заснива на групној мудрости. Настао је на Лондонском универзитетском колеџу (енг. University College London), а циљ му је да преко интернета ангажује велики број људи у циљу транскрипције оригиналних и непроучених рукописа које је на енглеском језику написао велики филозоф и реформатор Џереми Бентам (енг. Jeremy Bentham) (1748-1832). До сада је у оквиру овог пројекта, захваљујући непроцењивој помоћи учесника, транскрибовано 95 посто сачуваних рукописа (Causer, Tonra, & Wallace, 2012).

3.2.2 Игре са сврхом

Многи задаци који су тривијални за људе и даље представљају велики изазов рачунарским програмима. Идеја на којој се заснивају игре са сврхом почива управо на покушају да се снага људског ума искористи за испуњавање неких задатака који су релативно лаки, али за рачунаре су и даље недостижни. Основне принципе дизајна и процене квалитета игара са сврхом поставили су у многим научним радовима и практичним применама аутори Вон Ан и Дабиш (енг.Von Ahn и Dabbish).

Људи који играју неку од игара са сврхом спроводе једноставне задатке које још увек није лако аутоматизовати. Ове игре су једноставне, забавне и обично постоји неки систем награђивања и рангирања учесника. Такав је био и случај са игром ESP (акроним од Extra Sensory Perception), једном од првих широко познатих игара ове врсте (von Ahn & Dabbish, 2004), која је развијена како би људи стварали метаподатке за слике на вебу, све кроз играње онлајн игрице. Тако је слика на којој се налазе човек и пас у овој игри добијала етикете „човек“, „пас“, „кућни љубимац“. Овај пројекат трајао је од 2008. до

[99] Transcribe Bentham http://blogs.ucl.ac.uk/transcribe-bentham/

2011. године и за то време су играчи допринели са више од 50 милиона етикета. Сврха овог пројекта била је унапређивање система претраге слика на вебу.

Научници са Универзитета Вашингтон осмислили су онлајн игрицу Foldit[100] како би омогућили што већем броју волонтера да помогну у решавању старог проблема из области молекуларне биологије. Дуже од једне деценије било је покушаја да се открију детаљи структуре протеина вируса хумане имунодефицијенције (ХИВ) који му помаже да се умножава. Разумевање структуре и облика тог протеина умногоме помаже у процесу развијања лекова против ХИВ вируса. Током три недеље од покретања, у овом пројекту је учествовало више од 57.000 играча, од којих већина није имала никакво претходно знање из области молекуларне биологије. Тако је по први пут један важан научни проблем решен путем онлајн игрице (Savage, 2012).

У неким врстама пројеката заснованих на групној мудрости, рачунари обаве један, припремни део посла, те учесници касније имају улогу побољшавања тих резултата. Тако је на Универзитету МекГил (енг. McGill University) у сврху бољег слагања секвенци ДНК ради разумевања начина на који настају генетске болести развијена игрица Phylo[101]. Phylo корисничко сучеље показује учесницима делове решења која је пронашао рачунарски програм и од њих тражи да побољшају та решења. Ова игрица је лансирана 2010. године и за две године је у њој учествовало 35.000 људи, који су заједничким радом побољшали резултате добијене рачунарским путем за чак 70 посто.

У пројекту под називом *Phrase Detectives* (Chamberlain, Kruschwitz, & Poesio, 2012), снага друштвене мреже Фејсбук (енг. Facebook)[102] искоришћена је за пројекат групне расподеле рада у коме учесници играјући игрице анотирају, то јест обележавају делове текстова. Играчи се на почетку „тренирају“ на текстовима који су одређени као златни стандард, или златни задаци (текстови које су анотирали стручни лингвисти). Када играч заврши анотирање текстова одређених за златни стандард, додељује му се оцена на основу процента тачно анотираних „златних“ текстова[103] којима се унапређује основна функционалност платформи на којима се извршавају задаци групне расподеле рада. Златни задаци су они задаци за које је одговор већ познат, или је тривијалан, те се прецизност одговора учесника у пројекту одређује управо на основу одговора на те

[100] Foldit http://fold.it/portal/
[101] Phylo игрица http://phylo.cs.mcgill.ca/
[102] Facebook www.facebook.com
[103] у литератури се јављају и термини „златни стандард“ и „златна јединица“

задатке. Помоћу њих се може установити да ли се неким учесницима може веровати, јер ако су дали тачне одговоре на питања за која систем већ зна одговор, постоји одређени степен сигурности да ће и одговори који систему нису унапред познати бити тачни.

Ова врста дистрибуираног рада разликује се од других по томе што се не заснива на алтруизму (што је случај са моделом заснованом на групној мудрости) или финансијским потребама учесника (као у механизованом раду, о коме ћемо говорити у следећем одељку) као мотивима за учешће у неком пројекту, већ се заснива на људској потреби за забавом. У пројекту *Phrase Detectives*, као и у многим другим, сличним пројектима, заснованим на моделу групне мудрости, нагласак није на добити, јер само најбољи играчи добијају награду у виду бедževa и похвале коју могу поделити на друштвеним мрежама.

Тако је игра са сврхом, дакле, игра у којој играч обавља неки користан задатак као „пропратни ефекат“ уживања у игрању неке рачунарске игрице. Људи играју због забаве, а не првенствено како би допринели неком пројекту. Ипак, оваква врста рада, може да утиче на погоршање радних услова учесника, нарочито ако се за расподелу задатака користе платформе као што је Mturk (Kittur, et al., 2013).

3.2.3 Механизовани рад (MTurk)

Једна од најпопуларнијих платформи за пројекте такозваног механизованог рада на којој се може спровести пројекат групне расподеле рада јесте Mechanical Turk[104] (MTurk), компаније Amazon. Ова платформа је толико заступљена у пројектима групне расподеле рада да многи аутори један тип овог модела називају управо по њој.

Занимљиво је порекло назива ове платформе. У 18. веку је постојао такозвани „Турчин“ (енг. The Turk), то јест машина за играње шаха, коју је конструисао Волфганг фон Кемпелен (енг. Wolfgang von Kempelen), проналазач и писац мађарског порекла. Ова машина, или аутомат, како су га другачије називали, обилазила је Европу и добијала шаховске партије чак и против чувених људи као што су Наполеон Бонапарта и Бенџамин Френклин. Касније је обелодањено да та машина уопште није била аутомат, већ се у њеној кутији скривао човек, шаховски мајстор, који је управљао покретима лутке налик човеку која је „играла“ шах. Тако и платформа Mechanical Turk омогућује људима да помогну рачунарским програмима (модерним машинама) да изведу задатке за које још увек нису потпуно оспособљени.

[104] Mechanical Turk https://www.mturk.com/

MTurk је врста онлајн тржишта на коме свако може да постави задатке и да одреди надокнаду коју ће радници добити за њихово извршавање. Систем је осмишљен тако да то буду једноставни задаци који би били веома тешки за рачунарске програме, а људи их без муке могу обавити. Задаци које је могуће решавати преко ове платформе обично захтевају веома мало времена и напора, те су и радници плаћени веома мало (реда величине неколико центи, до десетак центи по микрозадатку).

Задаци на MTurk платформи носе назив HIT (енг. human intelligence tasks). Приликом постављања тих задатака, вођа пројекта може да одреди више параметара – колико одговора је потребно за сваки задатак; може се одредити коме је дозвољено да решава неке задатке, рецимо, у зависности од нивоа познавања неког страног језика (у случају пројеката који се баве побољшавањем резултата машинског превођења, или за анотацију корпуса); могуће је ограничити и број учесника који могу одговарати на сваки постављени задатак.

Amazon Mechanical Turk уствари пружа API (енг. application program interface), то јест апликационо програмско сучеље, за веб сервисе[105]. Ово апликационо програмско сучеље се може користити да би се преко Amazon Mechanical Turk веб сајта задаци поставили, да би се они који су испуњени одобрили и да би се одговори уврстили у друге софтверске апликације (Kittur, Chi, & Suh, Crowdsourcing User Studies Using Mechanical Turk, 2008) (Vukovic, Kumara, & Greenshpan, 2010) (Doan, Ramakrishnan, & Halevy, 2011).

Аутори Форт, Ада, Саго и Мар критикују све већу употребу ове платформе јер сматрају да постоје многа етичка питања која треба решити у вези са овом врстом коришћења снаге људи за испуњавање задатака (Fort, Adda, Sagot, & Mar, 2011). Истраживање које су ови аутори спровели даје увид у социо-економске чињенице (земље из које долазе учесници, годиште) као и у начин коришћења саме платформе (број задатака које обаве недељно, упупну зараду преко ове платформе, итд.) и у то како учесници квалификују сопствене активности. Тако 91% учесника користи ову платформу због жеље да заради новац, без обзира на то што су суме које се добијају по појединачним задацима понекад изузетно мале. Чак 60% учесника сматра да је рад на овој платформи уносан начин да се проведе слободно време и успут заради мало новца. Само 30% спомиње интересовање за саме задатке који су понуђени, а 20% њих (од чега 5% чине учесници из Индије) тврди да користе овакав начин рада само да би прекратили

[105] Ово сучеље дефинише начине на које веб сервис може да се користи, те одређује речник и правила позивања која програмер треба да примени да би користио те сервисе.

време, дакле, из досаде. Значајно је споменути да 20% учесника (од којих је 30% из Индије) користи MTurk како би обезбедили основна средства за живот, а за исти проценат учесника MTurk је примарни извор прихода.

Аутори даље наводе да 20% најактивнијих учесника, који проводе више од 15 сати недељно обављајући неке задатке на овој платформи и који су заслужни за 80% активности које се на њој одвијају, можемо сматрати запосленим радницима, који би онда, у складу са тим, требало да имају и нека права. То се посебно односи на учеснике који наводе да им је основни извор финансијских средстава управо све што успеју да зараде преко MTurk платформе, што је неретко мање од 2 долара на сат – зато су и покренуте многе расправе на ову тему, и доводи се у питање етичност коришћења снаге људи за обављање једноставних задатака уз малу новчану надокнаду.

Ова платформа је свакако привукла велику пажњу у академској заједници, те су и многе конференције и радионице које се редовно одржавају посвећене управо размени искустава у вези са прикупљањем података преко ње. Тако је у склопу једне од највећих конференција из области обраде природног језика, Language Resources and Evaluation (LREC), 2013. године организована радионица посвећена управо овој платформи и начинима грађења и унапређивања језичких ресурса преко ње, док су у наредним годинама радови који су за тему имали употребу модела групне расподеле рада за језичке ресурсе били пожељни, што је и наглашавано у позивима за предају дужих или краћих радова и постера који би допринели овој важној конференцији.

Као што постоје различите дефиниције групне расподеле рада, тако постоје и различити приступи решавању проблема путем овог модела, на различитим платформама. Овде ћемо дати кратак опис неких од тих платформи.

Поред MTurk платформе, која је најразвијенија и има највише учесника, како особа које постављају задатке, тако и радника, постоје многе друге, сличне платформе за решавање једноставних задатака. Сваки задатак постављен на некој од тих платформи, обавља више људи, те се резултат за који се определило највише радника узима за валидан. Такве платформе су Clickworker[106], Microtask[107] (која се користи углавном у пројектима дигитализације, као у пројекту Digitalkoot) и CloudCrowd[108] (која пружа функционалност сличну MTurk платформи, уз бољу контролу тока и уграђену контролу квалитета, као и све погодности рада „у облаку“).

[106]Clickworker http://clickworker.com/
[107] Microtask http://microtask.com/
[108] CloudCrowd http://cloudcrowd.com/

Нека решења за пројекте групне расподеле рада користе MTurk као основу на коју надограђују сопствене функционалности ради добијања бољег квалитета понуђених решења и лакшег коришћења. Такве су платформе CrowdFlower[109], Scalableworkforce[110], CrowdGuru[111], Smartsheet[112] и CloudFactory[113].

Механизовани рад се користи и за добијање решења путем такмичења. Група људи обавља неки задатак, а само творци одабраних, најбољих решења, добијају награду у виду новчане компензације. Такмичења се користе за дизајн логоа и осмишљавање назива нових производа и пословних подухвата. Обично постоји јавни позив на који се заинтересовани јављају. Системи награђивања су веома различити, а најчешће је то такозвано већинско гласање (енг. majority voting), те учесници гласају за најбоља решења других учесника. Такве платформе су 99designs[114], crowdSPRING[115] и Squadhelp[116] за разна дизајнерска решења, Mykindacrowd[117] за нове идеје и заједничко решавање проблема, HireTheWorld[118], где постоји могућност запошљавања „1 на 1“ за послове дизајна, административне послове, програмерске послове или писање.

3.3 Примери добре праксе у пројектима групне расподеле рада

3.3.1 Пројекат Digitalkoot

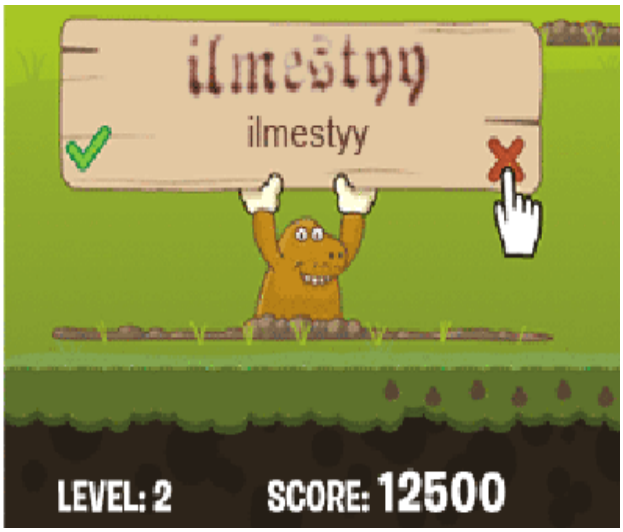
Дигитализација у библиотекама, где год се оне налазиле, обично подразумева поседовање скупе опреме, обучавање запослених за нове задатке које је потребно обавити, неретко и запошљавање нових људи који ће у том процесу учествовати, али и препуштање дела посла особама које нису запослене у библиотеци. Модел групне расподеле рада се у таквим случајевима може користити веома успешно – такав је пример пројекта Digitalkoot[119], Народне библиотеке Финске, покренут у сарадњи са финском компанијом Mikrotask[120] 8. фебруара 2011. године, који је за циљ имао

[109] CrowdFlower http://crowdflower.com/
[110] ScalableWorkforce http://scalableworkforce.com/
[111] CrowdGuru http://www.crowdguru.de
[112] Smartsheet http://smartsheet.com/
[113] CloudFactory http://cloudfactory.com/
[114] 99designs http://99designs.com/
[115] CrowdSPRING https://www.crowdspring.com/
[116] Squadhelp http://www.squadhelp.com/
[117] MyKindaCrowd http://www.mykindacrowd.com /
[118] HireTheWorld http://www.hiretheworld.com
[119] Digitalkoot http://www.digitalkoot.fi/
[120] Microtask http://www.microtask.com/

исправљање грешака насталих у процесу дигитализације финских новина с краја 19. века, Aamulethi, које се чувају у тој библиотеци.

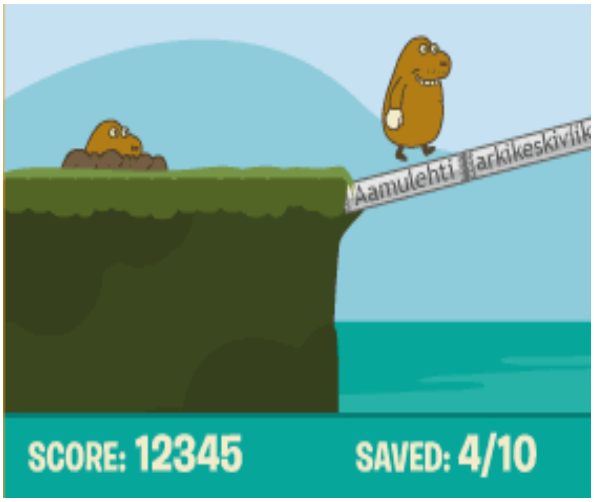
Реч *talkoot* у финском језику користи се за обичај који подразумева окупљање људи како би без новчане надокнаде заједно радили на неком пројекту, на пример, да би нешто изградили или поправили. Зато је овај пројекат, где се волонтери окупљају у дигиталном окружењу како би исправили грешке настале сканирањем новинске архиве, назван Digitalkoot. Као први овакав програм у Европи, он је кроз модел групне расподеле рада покренуо велики број људи с циљем да олакшају дигитализацију милиона страна архивских материјала. Алгоритми система за оптичко препознавање текста добро раде на текстовима писаним савременим фонтовима и врстама слова, али тешкоће настају када треба препознати старе фонтове и карактере. С обзиром на специфичности финског језика, особама које овај језик не познају довољно добро било је тешко да дају тачне одговоре. Као прво, потребно је користити посебне самогласнике, као што су: å, ö и ä, а речи су обично сложене и дуге, с обзиром на аглутинативну природу финског језика.

Пројекат Digitalkoot је искомбиновао забаву и добровољни рад, а спроводио се кроз две онлајн игрице. У „Лову на кртице” (енг. Mole hunt; фин. Myyräjahti), пред играча се постављају две речи, за које, што је то брже могуће, треба утврдити јесу ли исте, при чему се мора обратити пажња на велика слова и на интерпункцијске знаке (слика 15). Једна реч је оригинална реч извучена из скенираног текста, док је друга реч верзија те оригиналне речи добијена технологијом оптичког препознавања текста. После сваког одговора, кртица нестане са екрана, али тек на крају одиграног нивоа играч може да види колико тачних одговора је дао. На тај начин се откривају словне грешке у дигитализованим архивским материјалима.



Слика 15 – Игрица Лов на кртице

Друга игрица носи назив „Мост за кртице” (енг. Mole bridge, фин. Myyräsilta). Од играча се очекује да тачно откуцају речи које се појављују на екрану, при чему се такође мора обратити пажња на велика слова и интерпункцијске знаке. Тачан одговор, односно тачно укуцана реч, помаже кртицама да изграде мост преко реке. Са сваком новом укуцаном речју, мост за кртице добија нов део од дрвета. Систем онда утврђује јесу ли одговори тачни тако што се нови делови који су добијени за тачне одговоре претварају у челичне делове моста, док они делови који су мосту додати након нетачног укуцавања речи експлодирају и нестају носећи са собом и неколико околних делова моста. Када играч успе да изгради мост и спасе кртице, ниво се завршава и израчунавају се добијени поени (слика 16).



Слика 16 – Игрица Мост за кртице

У овој игрици се, као и у претходној, добијају позитивни поени за тачне одговоре, а негативни за нетачне одговоре. У игрици „Мост за кртице” постоји и дугме на које играч може да кликне ако не може да препозна реч, то јест ако не може да је укуца, такозвано Impossible дугме. Притиском на то дугме, у горњем делу екрана се, уместо речи коју играч није могао да препозна, појављује нова реч, а број стечених поена остаје непромењен, то јест играч не губи поене.

Игрице „Лов на кртице” и „Мост за кртице” било је могуће играти директно преко Digitalkoot веб странице, пријављивањем помоћу адресе електронске поште играча, али и на друштвеној мрежи Фејсбук. Играње преко Фејсбук налога било је популарније јер су играчи тако могли да поделе свој успех са пријатељима и да се такмиче за неко од првих места на табели победника. Свакодневно се на Фејсбук страници пројекта

Digitalkoot могао видети списак шест најбољих играча обе игрице, као и број поена које су постигли.

На основу ReCAPTCHA технологије (која ће бити описана у пододељку 3.5.3), резултати заједничког рада волонтера из Финске и осталих земаља у којима се говори овим језиком, довели су до значајног побољшања квалитета дигитализованих текстова.

У обраду података добијених захваљујући волонтерима, укључен је и систем за проналажење грешака. Овакви системи су неопходни јер, нажалост, увек има играча који намерно укуцавају погрешне одговоре и уопште не желе да својим радом допринесу дигитализацији. Како би идентификовао злонамерне волонтере, систем на почетку игрице поставља низ *„златних задатака”,* за које су одговори већ утврђени и познати. Када играч, дајући одређени проценат тачних одговора, докаже да заиста жели да игра по правилима и да жели да помогне, проценат ових задатака који служе за проверу постепено се смањује. Овај процес је потпуно неприметан, те чак ни злонамерни играчи који разумеју механизам по ком провера функционише, неће моћи да га заобиђу.

Слика 17 приказује изглед сканиране странице текста из Aamulethi новина (лево), затим изглед те исте странице обрађене технологијом оптичког препознавања текста (у средини), те коначан изглед дигитализованог текста (десно) настао захваљујући Digitalkoot пројекту и компанији Microtask, која је све обављене микрозадатке на крају спојила у комплетан исправљен текст.

Слика 17 - Процес исправке грешака дигитализације у пројекту Digitalkoot

Модел групне расподеле рада који је спроведен у овом пројекту може се сврстати у модел игара са сврхом, јер су игрице осмишљене тако да буду веома забавне. Ипак, алтруизам, то јест жеља да се помогне у очувању сопственог културног наслеђа је оно што је већину учесника на првом месту и довело до ових игрица, док је елемент забаве, донекле и жеља за доказивањем и такмичарски дух, оно што их је задржало и због чега су се изнова враћали.

Пројекат је завршен 29.11.2012. године и био је веома успешан – скоро 110.000 учесника обавило је преко 8 милиона задатака исправљања словних грешака, а све кроз волонтерски рад и играње онлајн игрица.

3.3.2 Пројекат Duolingo

Duolingo је пројекат који је осмислио један од твораца и највећих заговорника CAPTCHA технологије, професор Луис вон Ан са Карнеги Мелон универзитета. Пројекат Duolingo представља новину у учењу страних језика и заснива се на учењу кроз превођење делова текста.

Овај пројекат је започео од идеје да се интернет странице које нису на енглеском језику преведу како би људи који те језике говоре могли да их користе. Велики део интернета је на енглеском језику, тако да људи који тај језик не разумеју не могу да уживају у многим предностима које пружају савремене веб технологије. Зато је основна, покретачка идеја пројекта Duolingo била да се преведу сви веб сајтови на интернету, и то на све, условно речено, важније светске језике. Велики број људи широм света заиста жели да учи стране језике, а Duolingo је пројекат који настоји да искористи ту чињеницу у циљу превођења светом раширене мреже. Данас се преко овог пројекта могу учити 22 светска језика, међу којима су и чешки, мађарски, ирски, корејски итд.

Duolingo се заснива на принципу учења кроз праксу (енг. learning by doing), јер корисници од почетка преводе делове текста, испрва веома кратке и просте реченице с веба, да би касније, стицањем нових знања, могли да преводе и сложеније реченице. Корисничко сучеље које се у овом пројекту користи је веома пријемчиво и лако за коришћење, а својеврсна маскота је једна зелена сова (слика 18).

Слика 18 – Маскота пројекта Duolingo

Учење се заснива на превођењу реченице која се појави с леве стране екрана, с тим да се постављањем курсора миша изнад речи у тим реченицама може видети њихово значење, предлози превода и граматички облици, те је учесников задатак да од тих речи на десној страни екрана склопи реченицу на одговарајућем језику.

Na Duolingo веб сајту или апликацији за мобилне уређаје, корисници из целог света могу да науче основе многих светских језика. Сваки страни језик је представљен својеврсним „дрветом“ вештина, где се свака вештина односи на неку идеју или област знања, на пример „породица“ или „места“. Сваки корисник такође може да остави коментаре и питања, што омогућава другим корисницима да помогну, дају одговор или можда поставе нека сопствена питања на исту тему, чиме се ствара осећај припадања заједници.

Корисници преко ове платформе уче и да читају и да пишу на страном језику који желе да науче. Систем је осмишљен тако да се знање надограђује, те су нивои веома пажљиво осмишљени и ослањају се на претходно стечена знања. Учи се и кроз мале тестове, у којима корисник има четири срца (или живота, у жаргону видео игрица), то јест могућност да направи само четири грешке ако жели да „положи“. Такође се учесници охрабрују да јачају своје језичке способности и да редовно уче, јер се сваки страни језик тако најбоље учи.

Овакав начин учења језика је веома ефикасан, бесплатан је, а доноси и додатно задовољство јер корисник зна да чини добро док преводи делове интернета, са језика који говори, на онај који учи. CAPTCHA технологија омогућава упоређивање одговора корисника, тј. њихових превода, и утврђивање најприближнијих превода, што као резултат даје прецизност равну прецизности превода које су урадили професионални преводиоци. Овакво превођење веба представља и праведан пословни модел за учење

страних језика, с обзиром да су пакети за учење страних језика скупи, па већина људи не може да их приушти (von Ahn & Dabbish, 2008).

Овај пројекат је испрва финансиран захваљујући моделу прикупљања новчаних средстава путем упућивања јавног позива великој групи људи, тј. *Crowdfunding* моделу, о коме смо више говорили у пододељку 3.1.2 овог докторског рада. Уз више од 10 милиона корисника, који на овај начин могу да уче преко 25 страних језика, Duolingo је прави пример моћи модела групне расподеле рада, који функционише управо зато што су мотивација корисника и потребе вођа пројекта компатибилни. Спада у игре са сврхом, у комбинацији са Групном мудрошћу. Ипак, овај пројекат за резултат има стварање нових знања, па ће можда неки будући теоретичари модела групне расподеле рада, неку нову врсту модела назвати управо Duolingo.

Једна од новина у раду ове платформе која је уведена 2014. године јесте то што особе којима је матерњи језик шпански, португалски или француски, могу да сарађују како би превели чланке са веб сајтова BuzzFeed[121] и CNN[122], а све то, наравно, док вежбају своје знање енглеског језика. Сви ти преведени чланци објављују се у међународним верзијама ових агрегатора вести који спадају у најпосећеније на свету.

Duolingo апликација за мобилне телефоне доступна је од 2016. године чиме је учење језика преко овог занимљивог пројекта заснованог на групној расподели рада додатно олакшано.

3.4 Групна расподела рада у обради природног језика

Групна расподела рада се у обради природног језика успешно користи за многе задатке у којима је, на пример, потребно извршити анотацију, валидацију или евалуацију лингвистичких података (Munro, et al., 2010) (Sarasua, Simperl, & Noy, 2012) (Biemann & Nygaard, 2010) (Filatova, 2012) (Gurevych & Kim, 2013).

Анотирани корпуси, то јест колекције текстова чији су лингвистички елементи означени на различите начине, проналазе широку примену у обради природног језика. Такви корпуси се, на пример, користе за „тренирање“ то јест обучавање модела машинског учења за обављање задатака као што су парсирање, машинско превођење, сумаризација. Резултати процеса који се обављају над анотираним корпусом умногоме зависе од квалитета тог ресурса, те се велика пажња посвећује управо његовој изградњи

[121] Buzzfeed.com https://www.buzzfeed.com/
[122] CNN.COM http://edition.cnn.com/

и успостављању стандарда квалитета како би се постигла висока поузданост података и могућност њихове касније репродукције. Веома је важно пронаћи стручно и обучено особље које ће на адекватан начин извршити потребне анотације. Пројекти као што су SemCor[123] (Mihalcea, 1998) и Penn Treebank (Marcus, 1994) су значајни примери таквог модела стварања анотираних корпуса и налазе широку примену у многим задацима обраде природног језика. Настанком веба, а посебно веб 2.0 парадигме, омогућено је лакше и брже креирање корпуса, било да су они једнојезични или паралелни. Садржаји које креирају сами корисници веба и друштвених мрежа постали су изузетно значајни, нарочито садржаји са друштвене мреже Твитер, на којој корисници често размењују мишљења и ставове о актуелним светским темама.

Модел групне расподеле рада под називом игре са сврхом (пододељак 3.2.2) широко је распрострањен у различитим задацима обраде природног језика као што су обележавање значења речи (енг. word sense tagging), парафразирање, прикупљање асоцијација између термина и њихових обележја, прикупљање општег знања о речима, прикупљање мишљења о интензитету и поларитету осећања изражених неком реченицом, за изградњу онтологија, анотирање података, одређивање односа између заменица и елемената реченице на које се те заменице односе (енг. anaphora resolution).

Једна од првих игара са сврхом која је за циљ имала креирање корпуса за обраду природног језика носила је назив *1001 paraphrases.* Игром су прикупљане парафразе које су коришћене за обучавање система машинског превођења који би могао да препозна различите варијанте неких фраза. Такав систем би за машинско превођење са српског на енглески језик реченице „Деца воле да отварају веб странице“ и „Деца воле да сурфују вебом“, превео на исти начин, на пример „Children like to surf the web“.

Парафразе су у овом пројекту прикупљане од волонтера на вебу тако што им је преко веб платформе било омогућено да понуде парафразе за одређене делове текста. Систем валидације се заснивао на додељивању више бодова ако играч понуди решење које је неко већ понудио, док би се за ново решење добијало мање бодова (Chklovski, 2005).

Open Mind Word Expert је једна од раних игара којом су прикупљане ознаке значења речи од учесника на вебу. План је био да се добије велика количина података за обучавање система за обраду природног језика по ниској цени, која би овако свакако била много нижа од традиционалне методе запошљавања лексикографа и лингвиста.

[123] SemCor https://www.gabormelli.com/RKB/SemCor_Corpus

Подаци добијени овом игром искоришћени су за креирање корпуса са ознакама значења речи (Chklovski & Mihalcea, 2002).

Игра са сврхом под називом *PlayCoRef* развијена је за енглески и чешки језик, а слична је пројекту *Phrase detectives* који смо већ описали у пододељку 3.2.2 овог рада, као репрезентативан пример модела групне расподеле рада који се спроводи кроз игре са сврхом. *PlayCoRef* је игра за два играча у којој сваки играч означава парове речи које се у једном тексту појављују заједно. Циљ игре је да се повеже што више таквих парова, а с обзиром да сваки играч може да види колико парова је његов супарник повезао, мотивација да се победи је већа (Hladká, Mírovský, & Schlesing, 2009). *Verbosity* је игра помоћу које је такође прикупљано опште знање о свету, дакле чињенице које су познате већини људи, на пример „млеко је бело“ или „рингла је врућа после кувања кафе“ (von Ahn, Kedia, & Blum, 2006).

Игра са сврхом *JeuxDeMots* (fr. игра речима) која постоји од септембра 2007. године а актуелна је и данас има за циљ изградњу велике лексичко-семантичке мреже сличне ворднету.[124] Ова мрежа садржи термине и њихове могуће синониме. Семантичка мрежа се гради повезивањем термина помоћу више од педесет релација чију валидацију врше парови играча преко скупа онлајн игара које су сличне играма асоцијација. Релације добијају тежинске ознаке на основу броја парова играча који су их одабрали. На почетку овог пројекта, база података ове игре са сврхом попуњена је са око 140.000 термина из неколико речника француског језика. Када играчи понуде нови термин, он се додаје у базу података, а постаје нови чвор у лексичко-семантичкој мрежи само ако се бар један пар играча сложи у одређивању његовог значаја (Zarrouk, Lafourcade, & Joubert, 2013).

OntoGames је скуп игара који помаже у изградњи семантичког веба као пројекта израде универзалног медијума за размену информација постављањем докумената са значењем које рачунар може да обради на вебу. С обзиром да је циљ постизање узајамног функционисања веб извора на семантичком нивоу, допринос људи у процени и провери семантичких релација између вишеструких онтологија, као и у означавању текстуалних и мултимедијалних информација је пресудна. У играма са сврхом ови задаци су представљени на забаван начин и учесници их у слободно време радо обављају (Siorpaes & Martin , 2008).

[124] JeuxDeMots http://www.jeuxdemots.org/jdm-accueil.php

3.5 Провера квалитета у пројектима групне расподеле рада

3.5.1 Системи провере квалитета у пројектима групне расподеле рада

У сваком пројекту групне расподеле рада веома је важно проверити квалитет доприноса учесника или анотатора. Принцип отвореног позива на учешће свакако носи опасност од добијања неквалитетних и непотпуних података. У зависности од мотивације која учеснике покреће да својим радом допринесу неком пројекту, обично је неопходно уврстити и различите системе провере квалитета. Приликом расподеле задатака на платформи каква је Amazon Mechanical Turk и друге сличне платформе, увек постоји опасност од злоупотреба јер се дешава да учесници намерно дају погрешне одговоре или решавају неке задатке само да би испунили минимум прописаних услова и добили одређену компензацију. С обзиром да је мотивација учесника у таквим пројектима финансијске природе, често се дешава да учесници непажљиво испуне веома велики број задатака, само да би испунили одређену квоту која ће им донети новац, не водећи рачуна о квалитету својих одговора. Управо због таквих случајева, користе се разне методе евалуације, од којих је прва провера редундантности (енг. redundancy check), то јест провера кроз модел у коме више учесника обавља исти задатак (Oleson, et al., 2011). Исти подаци се понуде на проверу двема различитим групама учесника, те се на основу тога види да ли су подаци које је прва група учесника означила као тачне, тако означени и од припадника друге групе учесника. На тај начин се, поред утврђивања тачног одговора, може измерити квалитет доприноса учесника у једном пројекту групне расподеле рада.

Често коришћена метода евалуације описана је у (Dawid & Skene, 1979), где се провера поузданости доприноса учесника заснива на статистичком алгоритму под називом Очекивање-максимизација (енг. expectation - maximization – EM)[125] чији је резултат скуп процењених тачних одговора за сваки задатак, као и матрица која, за сваког учесника понаособ, даје вероватноћу прављења грешке. Из те матрице се директно може измерити квалитет рада сваког учесника (Sundberg, 1976). Сличан приступ представљен је у (Ipeirotis, Provost, & Wang, 2010) где се EM алгоритам користи за израчунавање квалитета доприноса за сваког учесника у пројекту, а прилагођен је

[125] EM алгоритам је итеративна процедура за процену параметара на основу критеријума максималне веродостојности (МЛ - енг. Maximum Likelihood) или оцене апостериорног максимума (МАП - енгл. Maximum a posteriori), код којих су присутне недоступне латентне варијабле (нису видљиве у опсервацијама), које имају особину θ коју не можемо да измеримо, бар не директно.

случајевима када може доћи до грешака, те се за сваког учесника добија и стопа грешке, чиме се може израчунати право стање квалитета рада појединачних учесника у неком пројекту групне расподеле рада. Решење које се успешно примењује у многим пројектима групне расподеле рада јесте успостављање система златних задатака (одељак 3.2.2).

3.5.2 Статистичке мере за процену слагања доприноса учесника у пројектима групне расподеле рада

Многе мере које се данас користе за процену валидности доприноса учесника у пројектима групне расподеле рада развијене су за потребе мерења у психолошким истраживањима. Природа мерења у тим истраживањима налагала је да дође до развоја терија о менталним тестовима који узимају у обзир грешке у мерењу и поузданости (Schaer, 2012).

Информације које се прикупљају током истраживања класификују се у зависности од нивоа мерења, од чега зависи њихова статистичка обрада. Различити нивои мерења података садрже различите количине информације без обзира на предмет мерења. У систему класификације који је представљен у (Stevens, 1946) нивои мерења су номинални, уређени, интервални и размерни ниво. Како ниво мерења расте од најнижег (номинални ниво) до највишег (размерни ниво), расте и количина информација коју садрже подаци, као и математичке операције које се могу изводити над тим подацима.[126] Подаци на номиналном нивоу мерења се називају номинални или категорички. Категорички подаци бележе квалитет или карактеристику неке особе, као што су: боја очију, припадност полу, нацији или политичкој партији, мишљење о неком друштвеном проблему, итд. Категорички подаци сврставају индивидуе у групе и ови подаци се обично представљају као број и/или проценат броја или особа који спадају у одређене групе. Уређена скала одређује шта је веће или мање, али разлике између појединих јединица скале нису једнаке. На овај начин се прикупљају следећи подаци: рангови, школске оцене (у случају описног оцењивања; оцена A, B, C или D; оцена на факултетима, нпр. 0-50, 51-60 итд.). и др.

Карактеристика интервалне скале је да одређује шта је веће или мање, а разлике између појединих јединица скале су једнаке на сваком делу скале и у сагласности са мереном особином.

[126] Statistika u društvenim naukama http://www.e-statistika.rs/index.php?pa=56&idTeksta=35

Пример интервалних скала су резултати на психолошким тестовима, иако неки теоретичари сматрају да је примереније податке добијене психолошким тестирањем третирати као уређене.

Размерни ниво мерења има све особине интервалне скале и још има апсолутну 0. То значи да су бројчани односи исти с односима у мереној појави. Мерења у физици су на размерној скали, на пример дужина, тежина, отпор и др.

Примери скала мерења дати су у табели 2.

Табела 2 Скале мерења

Номинална скала мерења	
Пол особе	
1. Женски	2. Мушки
Уређена скала	
Да ли сте задовољни пруженом услугом:	
1. Веома задовољни	
2. Делимично задовољни	
3. Неутрални	
4. Делимично незадовољни	
5. Веома незадовољни	
Интервална скала	
Одаберите термин за разговор са разредним старешином:	
1. 16-17	
2. 17-18	
3. 18-19	
Размерна скала	
Структура неписмених особа у некој општини је:	
1. 10 на 300	
2. 15 на 300	
3. 35 на 300	

За анализу података у оквиру пројеката групне расподеле рада Хајес и Крипендорф (енг. Hayes и Krippendorff) (Hayes & Krippendorff, 2007) су предложили коришћење такозваног Крипендорфовог алфа коефицијента јер, за разлику од других,

сличних мера које се такође користе, ова мера испуњава све услове за успешно приказивање степена слагања учесника у једном таквом пројекту. Исти аутори наводе да свака ефикасна мера за процену такозваног слагања међу анотаторима (енг. inter-annotator agreement) треба да задовољи следеће услове:

1) Треба да процени слагање између два или више посматрача (анотатора, учесника) који описују сваку јединицу анализе независно једни од других. У случају када постоји више посматрача, ова мера не би требало да зависи од броја посматрача нити да се мења у зависности од пермутације и селективног учешћа посматрача. У случају да су та два услова испуњена, на слагање доприноса учесника не би утицали индивидуални идентитети посматрача који стварају или процењују податке у неком пројекту.

2) Треба да буде заснована на расподели категорија или скала бодова које сами посматрачи користе. Тако поузданост доприноса посматрача не зависи од разлике која би настала између тога како су аутори упутства за учешће у неком пројекту замислили да ће подаци изгледати и тога какви су подаци заиста добијени.

3) Треба да произведе нумеричку скалу између најмање две тачке која ће имати разумну интерпретацију поузданости. Савршеним слагањем се, по конвенцији, сматра 1.000 (то јест 100%), или у случају недостатка слагања, типично означеног са 0.000 (што не мора обавезно бити доња тачка скале поузданости).

4) Мера би требало да одговара нивоу мерења података. То значи да приликом примене статистичких мера на неколико врста података, мера мора задржати своју математичку структуру, осим и једино у случају одлика мерног система. То омогућава поређења између различитих мерних система, ако је то потребно за стандард поузданости.

5) Начин прикупљања узорака би требало да буде познат, или да га је бар могуће израчунати.

Надаље ћемо навести кратак опис статистичких мера које се користе за анализу садржаја и доприноса учесника у пројектима групне расподеле рада, као и њихову подобност у смислу испуњавања претходно наведених услова.

Проценат слагања (енг. percent agreement) је пропорција јединица анализе са својим одговарајућим описима, за које се два посматрача слажу (табела 3). Ова мера се веома лако израчунава, тако што се број категорија у којима су се анотатори сложили подели са бројем укупних категорија. У датој табели, у којој су анотатори бирали између

номиналних вредности тј. могућих одговора 2,3,4 или 5, резултат процента слагања је 3 подељено са 5, дакле проценат слагања је 60%.

Табела 3 Израчунавање процента слагања међу анотаторима

Јединица анализе	Анотатор 1	Анотатор 2	Слагање
1	3	4	0
2	2	2	1
3	5	4	0
4	4	4	1
5	3	3	1
			3 /5

Проценат слагања има недостатке у готово свим аспектима који су важни за квалитет једне статистичке мере која се може користити у процени слагања међу анотаторима. Први услов јесте задовољен, али само зато што се мера односи на процену доприноса два посматрача. С обзиром да је слагање на овај начин теже проценити за више посматрача, ова мера не испуњава први услов. Други услов такође није испуњен, зато што 100% јесте недвосмислен индикатор поузданости, али 0% није. Трећи услов такође није испуњен код ове мере јер сва слагања која нису потпуна, то јест 100%, постају бескорисна. Четврти услов јесте испуњен, али само за номиналне податке. Пети услов је испуњен.

Коенова Капа (Cohen's Kappa) (Cohen, 1960) је мера која је донела многа побољшања у односу на меру процента слагања. Коенов капа коефицијент је статистичка мера за одређивање сагласности у анотацији између два оцењивача номиналном скалом мерења. Користи се само када су два посматрача независна један од другог. Посматрајмо најпре тзв. матрицу сагласности у случају два анотатора и три категорије приказану у табели 4.[127]

Табела 4 Матрица сагласности оцена 2 анотатора и 3 категорије

		Анотатор А			
Анотатор В	Категорија	1	2	3	piB
	1	0.25 (0.20)	0.13 (0.15)	0.12 (0.15)	0.50
	2	0.12 (0.12)	0.02 (0.09)	0.16 (0.09)	0.30
	3	0.03 (0.08)	0.15 (0.06)	0.02 (0.06)	0.20
	piA	0.40	0.30	0.30	pi=1.00

[127] Делови објашњења преузети из презентације др Миљане Младеновић са семинара Друштва за језичке ресурсе и технологије JepTex: http://jerteh.rs/wp-content/uploads/2017/01/seminarJerteh_122016.pdf

Анотатор А је 40% својих оцена доделио категорији 1, а анотатор Б је истој категорији доделио 50% својих оцена. То значи да би очекивано слагање оцена анотатора у овој категорији било 0.50*0.40=0.20. На исти начин оцењујемо очекивано слагање анотатора и по осталим категоријама и те вредности се налазе у заградама матрице сагласности.

Уведимо ознаке P_o – релативни опсервирани број анотација у којима постоји слагање (проценат опсервација у којима постоји слагање) и P_e – очекивана вероватноћа слагања анотација.

У примеру из табеле 4 видимо да се матрица сагласности попуњава тако што прва колона, означена бројем 1, представља јединице анализе којима је анотатор А доделио ознаку 1. Од тог броја, у истој колони, у првом реду се налази број који каже колико од тог броја је и анотатор Б дао ознаку 1. У другом реду је број који каже колико је јединица анализе и анотатор Б означио са 2 итд. Уочимо да се ради о независним, случајним догађајима које посматрамо у паровима (догађај да је анотатор А доделио ознаку Y_1 јединици анотације Z и догађај да је анотатор Б доделио ознаку Y_2 јединици анотације Z).

P_o= 0.25 + 0.02 + 0.02 = 0.29

P_e = 0.20 + 0.09 + 0.06 = 0.35

Коенова капа степен сагласности (A_m – agreement measure) изражава се као однос

$$Am = \frac{Po - Pe}{1 - Pe}$$

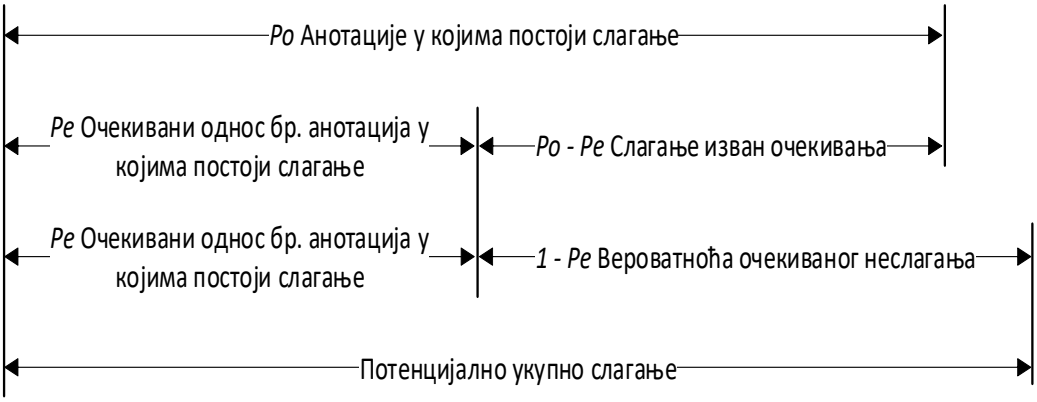
У претходном примеру

$$Am = \frac{0.29 - 0.35}{1 - 0.35} = -0.092$$

$P_o - P_e$ представља оне случајеве који су се појавили као слагање изван очекиваног слагања.

$1 - P_e$ представља вероватноћу очекиваног неслагања.

A_m се тумачи као удео слагања изван очекиваног у делу у коме се очекује неслагање. Графички приказ ових односа дат је на слици 19.



Слика 19 – Коенов коефицијент степена сагласности

Интерпретација резултата степена сагласности на основу Коеновог капа коефицијента дата је у табели 5.

Табела 5 Ниво сагласности на основу Коеновог капа коефицијента

Коенов капа коефицијент	**Ниво сагласности**
< 0.00	слаб
0.00 - 0.20	низак
0.21 - 0.40	коректан
0.41 - 0.60	умерено
0.61 - 0.80	висок
0.81 - 1.00	веома висок

Даља унапређења статистичке оцене из класе капа коефицијената су Флајсов капа (Fleiss' kappa) и Крипендорфов алфа (Krippendorff's alpha) коефицијент.

Флајсова Капа (енг. Fleiss's K) (Fleiss, 1971) је, у односу на Коенову Капу, донела уопштавање на више посматрача, то јест, она узима у разматрање већи број анотатора, уместо само два. Основни приступ израчунавања Флајсове капе и *Крипендорфове алфе* је исти. Обе ове мере користе се за израчунавање степена слагања дефинишући слагање на следећи начин:

$$\mathrm{A} = 1 - \frac{Do}{De}$$

где је слагање означено са A (енг. agreement), степен уоченог, посматраног неслагања је означен са D_o (енг. observed disagreement), а степен очекиваног неслагања са D_e (енг. expected disagreement).

Крипендорфов алфа коефицијент (енг. Krippendorff's alpha coefficient), Крипендорфова алфа (енг. Krippendorff's alpha) или Калфа (енг. Kalpha) је мера која

испуњава све услове за успешност једне статистичке мере у процесу одређивања степена слагања међу анотаторима. Она обједињује предности претходно наведених статистичких мера и доноси сопствена побољшања. Калфа израчунава неслагања, уместо да коригује проценте слагања, и по томе се разликује од осталих мера. Ова мера омогућава оцену сагласности анотације када је:

- број анотатора већи од 2
- број категорија (ознака) већи од 2
- скала мерења може бити произвољна (номинална, уређена, интервална, размерна)
- анотатори не морају свакој јединици за анотацију доделити етикету (енг. missing data)
- скупови за анотацију немају захтевани минимум података.

Крипендорфова алфа је развијена за мерење поклапања доприноса посматрача, оцењивача, или неких инструмената за мерење којима се утврђују разлике између типично неструктурираних феномена, као и у случајевима када су подаци непотпуни. Крипендорфов алфа коефицијент првобитно је коришћен за анализу садржаја, али примену проналази у свим пројектима у којима се на исти скуп објеката или јединица анализе примењују две или више метода генерисања података, а потребно је закључити колико се подацима може веровати (Kripendorff, 2011) и (Hayes & Krippendorff, 2007).

Крипендорфова алфа је мера која је развијена и пре неких других мера које се користе за анализу података у пројектима групне расподеле рада у хуманистичким наукама, али с обзиром да није била уврштена у стандардне статистичке пакете који се користе у анализама тих података, била је запостављена.

У смислу првог услова за погодност употребе у пројектима групне расподеле рада, овом мером се броје парови категорија или тачака на скали које су посматрачи доделили индивидуалним јединицама, третирајући посматраче тако да се њихова позиција може пермутовати и тако да резултати не зависе од броја посматрача. То развејава уврежено мишљење да је поузданост теже одредити како број посматрача расте. Крипендорфова алфа је заснована на подацима које генеришу сви посматрачи, чиме су испуњени други и трећи услов. Ова мера такође испуњава четврти услов и то тако што мери слагања за номиналне, ординалне, интервалне и размерне податке. Пети услов је такође испуњен, јер постоје јасне смернице за прикупљање података.

Када је вредност Крипендорфовог α коефицијента у интервалу [0,1] ниво слагања се креће од потпуног неслагања, када је $\alpha = 0$, до потпуног слагања, у случају када је $\alpha = 1$. Вредност ове мере може да буде и негативна, до -1, што се обично дешава ако дође до грешке у прикупљању података, или до систематског неслагања. Када говоримо о прихватљивом нивоу неслагања, у радовима (Lombard, Snyder-Duchand, & Campanella Bracken, 2002), (Hayes & Krippendorff, 2007) и (Maggetti, 2013) је приказано да су слагања чије су вредности изнад 0,667 поуздана, док се вредности веће од 0,8 могу сматрати веома поузданим – те смернице смо такође користили у смислу анализе података добијених у пројектима групне расподеле рада спроведених у оквиру ове докторске тезе.

Као и у случају израчунавања Флајсове капе, уочено неслагање (енг. observed disagreement - D_o) указује на проценат неслагања анотатора у вредностима јединица у којима се анализа врши. Уочено неслагање се израчунава по следећој формули[128]:

$$D_o = \frac{1}{n}\sum_{c}\sum_{k} o_{ck\ metric}\,\delta^2_{ck}$$

где су c и k из скупа свих јединица анализе то јест иду од 1 до броја јединица анализе (питања) која су анотатори оцењивали, а o_{ck}, n_c, n_k и n односе се на фреквенције вредности у матрицама. Прецизније, у формули:

$$D_o = \frac{1}{n}\sum_{c\in R}\sum_{k\in R}\delta(c,k)\sum_{u\in U} m_u \frac{n_{cku}}{P(m_u,2)}$$

где је δ (метричка) функција разлика, n укупан број елемената који се упарују, m је број оцена јединице анализе u, n_{cku} је број (c , k) пара у јединици анализе u, и P је функција пермутације. Функције разлика $\delta\,(v, v')$ између вредности v и v' одражавају својства мерења варијабле.

У општем случају:

$\delta\,(v , v') \geq 0$

$\delta\,(v , v) = 0$

$\delta\,(v , v') = \delta\,(v' , v)$

специфично за номиналне податке:

$$\delta \text{ nominal } (v , v') = \begin{cases} 0 \text{ za } v = v' \\ 1 \text{ za } v \neq v' \end{cases}$$

где v i v' имају улогу имена (јединица анализе).

[128] Формула преузета из рада (Kripendorff, 2011)

Очекивано неслагање (енг. expected disagreement - D_e) је оно неслагање које би се добило када би оцењивање анотатора било насумично и може се израчунати по формули:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{k\,metric} \delta^2_{ck}$$

где се c и k крећу од 1 до броја јединица анализе, а аргументи o_{ck}, n_c, n_k, n фреквенције појављивања вредности дате у матрици коинциденције приказаној на слици 20. Општа формула за o_{ck} је:

$$o_{ck} = \sum_u \frac{\text{број парова } c-k \text{ у јединици анализе } u}{m_u - 1}$$

У овој матрици је, на пример, o_{11} фреквенција за све јединице анализе анотатора који су се сложили у оцени и ставили ознаку 1.

	1	.	**k**	
1	o_{11}	.	o_{1k}	$n_1 = \sum_k o_{1k}$
.	.	.	.	.
.	.	.	.	.
c	o_{c1}	.	o_{ck}	$n_c = \sum_k o_{ck}$
.	.	.	.	
	n_1	.	n_k	$n = \sum_c \sum_k o_{ck}$

Слика 20 – Матрица коинциденције

У табели 6 можемо видети резултате анотације употребом номиналне скале када је број анотатора већи од 2, број категорија већи од 2, а неке од јединица за анотацију немају етикету (у табели означене тачком (.)).

Табела 6 Пример резултата анотације које Калфа може да обради

	JA1	JA2	JA3	JA4	JA5	JA6	JA7	JA8	JA9	JA10
Анот_А	1	3	3	2	4	4	1	5	.	.
Анот_Б	1	2	3	2	2	4	2	5	.	.
Анот_В	.	2	3	2	3	4	1	.	1	2
Анот_Г	1	2	3	2	1	4	1	5	1	.

За појединачне вредности: [1, 2, 3, 4, 5] је укупан број вредности 33, од тога за појединачне вредности :{n_1=9, n_2=10, n_3=6, n_4=5, n_5=3}. Број анотатора је 4, број јединица анализе је 10, број недостајућих података је 7.

У матрици коинциденције се рачунају o_{ck} за сваки пар вредности оцена, у конкретном случају 1-5, тако да се добија матрица 5x5. Уочимо да свака јединица има $m_u(m_u–1)$ случајева и у матрици коинциденције рачунају се све парови вредности пронађени у јединици анализе u. Прва јединица анализе (JA1) садржи 3(3-1)=6 парова који одговарају оцени 1, а то доприноси 6/(3-1)=3 елементу o11 матрице (један за сваку вредност). Друга јединица анализе JA2 садржи 4(4-1)=12 парова, 6 одговарајућих 2-2 парова, 3 неодговарајућа 2-3 пара, и 3 неодговарајућа 3-2 пара. Тиме додаје вредност од 6/(4-1)=2 у o_{22}, 3/(4-1)=1 у o_{23}, 1 у o_{32}, и 4 укупну вредност n, и тиме обрачунава своје 4 вредности. Пета јединица JA5 садржи 4(4-1)=12 парова вредности које се не поклапају, при чему свака додаје 1/(4-1)=1/3 другом елементу матрице.

Десета јединица анализе JA10 има једну вредност 2 и нема поређења, тако да има 1(1-1)=0 парова и не додаје ништа. Дакле, маргине матрице коинциденције не узимају све вредности које се појављују у матрици поузданости, само оне које се могу упарити, тако да је овде n=32 поредиве вредности од свих 33 оцењених вредности. Када се уклони појављивање вредности 2 из JA10, тада појединачне вредности постају:{n_1=9, n_2=9, n_3=6, n_4=5, n_5=3}.

Дајемо пример рачунања вредности:

o_{11}= 3(3–1)/(3-1) {JA1}+ 4(4-1)/4 {JA7} + 2(2-1)/2 {JA9} = 3 + 3 + 1= 7

o_{23}= 3/(4-1) {JA2} + 1/(4-1) {JA5} = 1+ 1/3 = 1.33

…

Матрица коинциденције има вредности дате у табели 7.

Табела 7 Вредности матрице коинциденције

o_{ck}	[1]	[2]	[3]	[4]	[5]
[1]	7.00	1.33	0.33	0.33	0.00
[2]	1.33	6.00	1.33	0.33	0.00
[3]	0.33	1.33	4.00	0.33	0.00
[4]	0.33	0.33	0.33	4.00	0.00
[5]	0.00	0.00	0.00	0.00	3.00

Коначно се може израчунати калфа вредност према формули:

$$_{nominal}\alpha = 1 - \frac{D_0}{D_e} = \frac{A_0 - A_e}{1 - A_e} = \frac{\frac{\sum_c O_{cc}}{n} - \frac{\sum_c n_c(n_c-1)}{n(n-1)}}{1 - \frac{\sum_c n_c(n_c-1)}{n(n-1)}} = \frac{(n-1)\sum_c O_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)}$$

Што за претходни пример даје вредност:

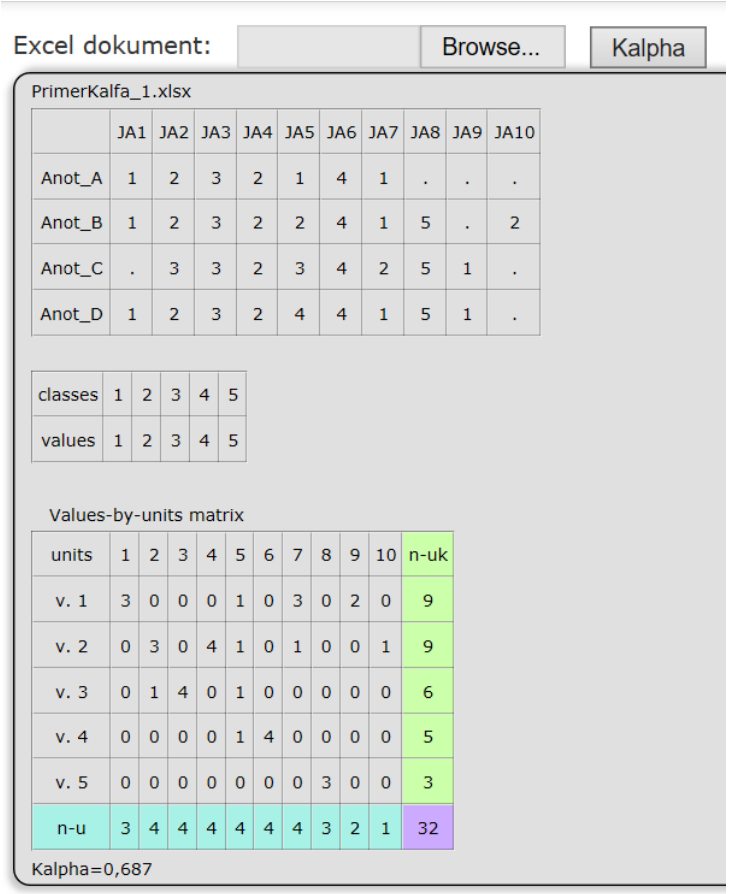
$$\alpha = \frac{(32-1)(7+6+4+4+3) - [9(9-1) + 9(9-1) + 6(6-1) + 5(5-1) + 3(3-1)]}{32(32-1) - [9(9-1) + 9(9-1) + 6(6-1) + 5(5-1) + 3(3-1)]}$$
$$= 0.687$$

Веб алат за оцену сагласности резултата испитаника Крипендорфовим α коефицијентом, као део апликације за семантичке ресурсе за српски језик, SWNE (пододељак 2.4.3) , који смо користили за израчунавање, приказан је на слици 21. Приликом коришћења овог алата, податке за обраду потребно је унети у формату Excel табеле (xlsx), у коме један ред представља све оцене једног посматрача. Недостајуће оцене се означавају тачком (.), а излаз прорачуна се исписује на веб страни алата. Овај алат је успешно искоришћен за израчунавање степена сагласности учесника у пројектима групне расподеле рада који су описани у одељку 4.4 овог рада, а осмишљен је тако да се може користити у свим будућим, сличним пројектима.

Слика 21 – Део апликације за израчунавање вредности Калфа[129]

[129] Ова апликације је изграђена у оквиру израде докторске тезе „Информатички модели у анализи осећања засновани на језичким ресурсима“, ауторке Миљане Младеновић, на Математичком факултету у Београду.

На слици 22 је приказан резултат који добијамо када документ са подацима о резултатима анотације приказан у табели 6 пропустимо кроз апликацију за оцену сагласности резултата испитаника Крипендорфовим α коефицијентом, која је део SWNE апликације за семантичке ресурсе за српски језик.



	JA1	JA2	JA3	JA4	JA5	JA6	JA7	JA8	JA9	JA10
Anot_A	1	2	3	2	1	4	1	.	.	.
Anot_B	1	2	3	2	2	4	1	5	.	2
Anot_C	.	3	3	2	3	4	2	5	1	.
Anot_D	1	2	3	2	4	4	1	5	1	.

classes	1	2	3	4	5
values	1	2	3	4	5

units	1	2	3	4	5	6	7	8	9	10	n-uk
v. 1	3	0	0	0	1	0	3	0	2	0	9
v. 2	0	3	0	4	1	0	1	0	0	1	9
v. 3	0	1	4	0	1	0	0	0	0	0	6
v. 4	0	0	0	0	1	4	0	0	0	0	5
v. 5	0	0	0	0	0	0	0	3	0	0	3
n-u	3	4	4	4	4	4	4	3	2	1	32

Слика 22 Резултат израчунавања Крипендорфовог α коефицијента у SWNE

3.5.3 Евалуација у пројектима обраде природног језика

Ако се постојеће платформе групне расподеле рада користе уз додатне технике управљања и уз проверу квалитета обављених задатака, то јест евалуацију доприноса учесника, комплексни задаци у оквиру обраде природног језика могу се обавити веома успешно. Модел заснован на микрозадацима се често користи у многим пројектима, обично за разне врсте анотације и провере тачности података. С друге стране, у неким пројектима у којима се користи више платформи и више начина сакупљања доприноса учесника, потребно је изнаћи решење за другачији начин евалуације.

У раду (Munro, et al., 2010) је описано неколико пројеката у којима је за евалуацију коришћен Капа коефицијент (Kripendorff, 2011). Коенова капа се користи и

у анализи садржаја[130] у многим пројектима из области обраде природног језика, јер дозвољава упоређивање различитих резултата учесника, а користи се и у случајевима када одређивање поузданости резултата уобичајеним методама (на пример, помоћу златних задатака) није довољно (Carletta, 1996).

У пројекту CROWDMAP (Sarasua, Simperl, & Noy, 2012) поред MTurk платформе коришћена је и платформа CrowdFlower, која је основа имплементације пројекта. Циљ пројекта је да се истражи веома значајна област у обради природног језика, поравнавање онтологија (енг. ontology alignment), што је постигнуто расподелом посла преко микрозадатака. Већ постојећи, аутоматизовани процеси, побољшании су доприносом људи, а CrowdFlower платформа делује као посредник јер истовремено поставља микрозадатке на другим, сличним платформама (укључујући MTurk, Crowd Guru[131], Getpaid[132]). Евалуација је извршена на основу златних задатака, као и помоћу Крипендорфовог алфа коефицијента.

У пројекту генерисања и анализе специјализованог корпуса који се може користити за идентификовање сарказма и ироније у текстовима (Filatova, 2012), такође су коришћене две контролне процедуре. Прва врста провере је просто већинско гласање (енг. majority voting) – учесници су гласали за најбоља решења која су понудили други учесници, док је друга врста провере заснована на Крипендорфовом алфа коефицијенту.

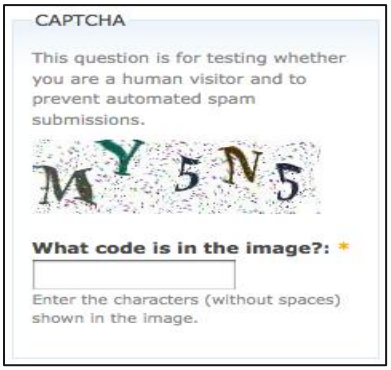
За евалуацију у пројектима заснованим на моделу игара са сврхом врло често се користе технологије CAPTCHA и ReCAPTCHA. У пројектима Duolingo и Digitalkoot (стр. 99), који су, сваки на свој начин, у вези са обрадом природног језика, користе се ове технологије, те ћемо их у наставку ближе описати.

CAPTCHA (изговара се ‘кепча’) је, у најширем смислу, програм који може да утврди да ли је његов корисник људско биће или рачунарски програм. CAPTCHA низови карактера се могу видети на дну многих веб страница, и то најчешће као разнобојне слике текста са искривљеним словима. CAPTCHA је програм који штити веб странице од такозваних ботова (енг. bots, скраћено од robots), тако што генерише и оцењује тестове које људи могу да реше с лакоћом, а рачунарски програми то још увек не успевају (слика 23).

[130] Метода у друштвеним наукама која се користи за квантитативну и квалитативну анализу садржаја штампе, телевизије, филма, интернет медија, итд.

[131] Crowd Guru https://www.crowdguru.de/en/

[132] Getpaid https://www.get-paid.com/



Слика 23 – Изглед CAPTCHA слагалице

Термин CAPTCHA (акроним од речи Completely Automated Public Turing Test To Tell Computers and Humans Apart), осмислили су 2000. године Луис вон Ан (Luis von Ahn), Мануел Блум (Manuel Blum), Николас Хопер (Nicholas Hopper) и Џон Ленгфорд (John Langford) са Универзитета Карнеги Мелон (Carnegie Mellon University).

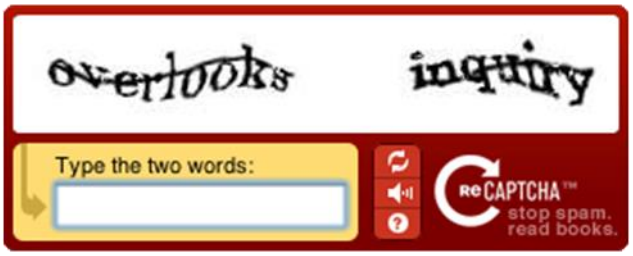
CAPTCHA се углавном користи у сигурносне сврхе, а највише за заштиту од постављања нежељених коментара у блоговима, заштиту приликом регистрације на веб сајтовима, онлајн гласање, заштиту лозинки, заштиту од вируса и спамова тј. нежељених порука.

ReCAPTCHA је бесплатна CAPTCHA технологија која помаже у дигитализацији књига, новина, и старих радио емисија.

У циљу архивирања људског знања и боље доступности информација, данас се у свету спроводе многи пројекти дигитализације књига. Странице књига се обично на неки начин сканирају, те претварају у текстуалне документе техником оптичког препознавања карактера (енг. Optical Character Recognition – OCR). Трансформација у текст је корисна јер се сканирањем књига добијају само слике, које је тешко и скупо складиштити и претраживати. Ипак, OCR технологија не даје увек жељене резултате. Зато ReCAPTCHA побољшава процес дигитализације књига тако што речи које рачунари не могу да протумаче шаље на веб како би их преко CAPTCHA технологије корисници веб сајтова дешифровали. Свака реч коју OCR техника не може да прочита правилно, поставља се као једна од речи у ReCAPTCHA слагалици. Ако корисник правилно реши реч за коју је одговор познат, систем претпоставља да је његов одговор за другу реч такође тачан. Затим се та иста слагалица с две речи нуди другим корисницима, те се статистичким моделима одређује проценат тачности одговора, и на крају се добија тачан одговор за дату непознату реч. ReCAPTCHA технологија помаже дигитализацију старих издања новина The New York Times и књига обухваћених пројектом Google Books.

Процењено је да се око 200 милиона ReCAPTCHA слагалица (слика 24) реши сваког дана (von Ahn & Dabbish, 2008). Сваки пут када нека особа реши ReCAPTCHA задатак, то траје око десет секунди. То индивидуално не представља велики напор, али укупно чини више од 50.000 сати сваког дана. С том чињеницом на уму је и настала технологија ReCAPTCHA, која тај људски напор усмерава у онлајн "читање" књига.

Можемо закључити да је ReCAPTCHA светски пројекат групне расподеле рада заснован на микрозадацима јер корисници интернета широм света свакодневно помажу дигитализацију књига, реч по реч.



Слика 24 – Изглед ReCAPTCHA слагалице

4 Нове семантичке релације у Српском ворднету на основу реторичке фигуре поређење

У овом поглављу ћемо дати опис истраживања које за циљ има унапређивање структуре Српског ворднета новим семантичким релацијама како бисмо тако унапређени ворднет могли да користимо у напредним системима за детекцију реторичких фигура. Надаље ћемо говорити о реторичким фигурама, њиховој функцији у језику, као и њиховој улози у обради природног језика. Даћемо детаљан опис корака који су спроведени у овом истраживању, почевши са претрагом Корпуса савременог српског језика, преко описа два пројекта групне расподеле рада и анализе њихових резултата.

4.1 Реторичке фигуре и њихова улога у језику

Реторичке фигуре су широко распрострањене у тексту, говору, дијалогу и од давнина представљају богате изворе за комуникацију, књижевност и расправе, аргументе, аргументацију. Истраживања реторичких фигура и њихове улог сежу чак два и по миленијума у прошлост (Mitrović, O'Reilly, Mladenović, & Handschuh, 2017). Реторичке фигуре се могу посматрати као језичка средства која су под когнитивним надзором и имају функционалну, меморијску и естетску сврху[133] (Ruan, Di Marco, & Harris, 2016:1).

Реторика је наука која се бави изучавањем начина на који се језик користи за постизање разних ефеката. Према Аристотелу, реторика је вештина која се огледа у способности убеђивања (Аристотел, 1987). Квинтилијан, антички теоретичар и предавач реторике, препознао је да је необично коришћење језичких норми често делотворан начин да се добије пажња публике, као и њено пристајање и сагласност (Квинтилијан, 1985). У класичној реторици све реторичке фигуре називане су тропима, док у савременој теорији књижевности и реторици, као делу стилистике, тропи представљају једну групу фигура. У српској теорији књижевности и стилистици најчешћа подела реторичких фигура, предложена као најближа савременом схватању српског језика, јесте подела на фигуре дикције, фигуре речи (тропе), фигуре конструкције (схеме) и фигуре мисли (Солар , 1981) (Ковачевић , 1998). Тропи и схеме су фигуре које највише утичу на

[133] „Rhetorical figures are cognitively governed linguistic devices that serve functional, mnemonic, and aesthetic purposes.“

промену значења текста у коме се појављују, а у овом докторском раду највише ћемо се бавити улогом тропа.

Тропи су фигуре са неочекиваним заокретом у значењу речи (грч. tropos – „окрет“) и настају променом основног, уобичајеног значења појединих речи. У наставку ћемо навести особине и примере неких тропа.

- *Метафора* се заснива на пренесеном значењу речи, где речи тумачимо на основу неке сличности, као у примеру „Зуби су јој бисери.“
- *Поређење* је фигура која се заснива на поређењу два појма, лица или предмета по сличности, такође по неким заједничким особинама. Има три поредбена члана: предмет који се пореди, заједничку особину, предмет са којим се пореди, као у примеру „Зуби су јој бели као бисери.“
- *Метонимија* се, као и метафора, заснива на пренесеном значењу речи, али се то значење не преноси по сличности, него према неким стварним односима или особинама, као у примеру „Оловка је моћнија од мача.“
- *Зевгма* је фигура која се састоји у томе да се један предикат односи на више осталих делова реченице. Често је коришћена у српској народној поезији. Пример једне зевгме је: „Наизменично је мрцварио свој мозак и свог магарца.“
- *Хипербола* је фигура која се заснива на преувеличавању, као у примеру „Исплакала је море суза.“
- *Литота* је фигура којом се неки исказ ублажава да би се код читаоца изазвао јачи супротан утисак. Овом фигуром се избегава права реч или право вредновање и увек се користи негација, као у примеру „Ајнштајн није лош научник.“
- *Парадокс* се добија стављањем једне поред друге две мисли које су супротне и једна другу искључују, као у примеру „Знам да ништа не знам.“
- *Оксиморон* је врста парадокса којим се нови појам добија повезивањем супротних појмова, на пример „гласна тишина“.
- *Иронија* је фигура којом се речима придаје супротно значење од оног које имају у дословном значењу, на пример „Сунчано је као зими на Аљасци“.

У наставку овог докторског рада говорићемо о две реторичке фигуре које спадају у тропе, а којима се често приписују слична својства – метафора и поређење.

4.2 Метафора и поређење

Метафора је реторичка фигура из групе тропа чијој аутоматској обради је у литератури посвећено много пажње, од неких од првих радова (Black, 1962), преко репрезентативних истраживања Лејкофа и Џонсона (Lakoff & Johnson, 1980), (Lakoff, 1990), (Lakoff, 2012), до најновијих радова посвећених статистичком процесирању ове реторичке фигуре (Mason, 2004), (Hardie, Koller, Rayson, & Semino, 2007) , (Shutova, Teufel , & Korhonen, 2013).

Сматра се да је прву дефиницију метафоре дао Аристотел, рекавши да је то „промена значења речи од њене уобичајене употребе до нове“ (Richards, 1965).

Џорџ Милер, творац Ворднета, заступао је традиционални став о метафори као скраћеном поређењу (Miller G. A., 1979), док су Лејкоф и Џонсон, најпознатији теоретичари метафоре, у предговору новом издању књиге „Метафоре по којима живимо“ (енг. Metaphors we live by), оригинално објављене 1980. године (Lakoff & Johnson, 1980), јасно изнели став да метафора није само скраћено поређење.

Један од највећих светских пројеката посвећен систематизацији и праћењу развоја стилских фигура, Silva Rhetoricae[134] (лат. шума реторике), фигуру *поређење* (енг. simile) дефинише као експлицитно упоређивање уз коришћење везника *као* (енг. „like“ или „as“), уз репрезентативни пример:

You are like a hurricane: there's calm in your eye, but I'm getting blown away – Neil Young (енг. Ти си као ураган: у твом оку је смирај, али бивам однесен ветром – Нил Јанг).

Реторичка фигура *поређење* повезује уобичајене, устаљене концепте са изузетним, необичним, често измишљеним концептима и тако омогућава узајамно деловање норме и креативности, судећи по ауторима Вил и Хао (енг. Veale и Hao), који су се бавили истраживањем фигура метафоре и поређења (Veale & Hao, 2013). Исти аутори запажају и да многа уверења која користимо у расуђивању о свакодневним појавама нису увек тачна, нити су логички доследна. Људи се често ослањају на народну мудрост у виду стереотипа и клишеа, а управо они чине основу према којој поредимо и дефинишемо друге, мање познате феномене. Тако су, на пример, честа поређења особина људи са особинама животиња као што су змија, медвед, ајкула, горила, иако већина људи није лично искусила особине тих животиња. Ипак, сви поседујемо довољно општег знања о овим животињама и зато их користимо као врсту пречице у описивању

[134] Silva Rhetoricae http://rhetoric.byu.edu/

нечијих особина или начина понашања („отровна као змија“, „опасна као змија“, „снажан као медвед“, „здрав као медвед“ су неки од примера из Корпуса савременог српског језика (одељак 2.1.1) који илуструју овакву употребу поређења).

Вили и Хао дефинишу поређење и као средство преношења народне мудрости које користи експлицитна синтактичка средства (за разлику од метафоре) да означи концепте који су најкориснија обележја лингвистичког описа. Поред одређеног броја концепата који су универзалног карактера, сваки народ карактеришу и посебни концепти и само њему својствени односи међу тим концептима, што и ствара основу за национални поглед на свет – што је једно од полазишта за истраживање спроведено у оквиру овог докторског рада.

На истраживању фигуре поређење и њеног значаја у обради природног језика радила је са својим тимом и професорка Кристијан Фелбаум (енг. Christiane Fellbaum), сада главна уредница Ворднета, у оквиру истраживачког програма фондације Александер фон Хумболт (нем. Alexander von Humboldt). Део резултата тог истраживања приказан је у раду Патрика Хенкса (Hanks, 2005), члана тима професорке Фелбаум. У раду је представљено појављивање везника *like* у Британском националном корпусу (енг. British National Corpus), из чега је закључено да се фигура поређење обично заснива на неким културним стереотипима заједнице у којој се користи.

Функција многих поређења је исказивање нечега веома једноставног на директан и живописан начин. Поређења подстичу нашу машту и имају много јачи ефекат неко обичан говор. Када кажемо да је неко „брз као ракета“, обично замишљамо начин на који се ракета креће, то јест како се лансира и полеће увис. У истраживању тима професорке Фелбаум закључено је и да свака именица која често учествује у формирању поређења има неки карактеристичан атрибут (или више атрибута) који представљају основу за формирање поређења. Зато смо у истраживању које је спроведено за потребе овог докторског рада такође пошли од претраживања Корпуса савременог српског језика како бисмо пронашли која поређења су најчешћа, то јест који придеви често стоје уз неке именице.

Ворднет укључује нека фигуративна значења – на пример, за глагол „grasp“, једно од значења јесте “get the meaning of something”; за глагол „skim“, постоји значење које је у вези са употребом као што је „skim through a text“, то јест “read superficially”. Наравно, много фигуративних значења није укључено у тренутну верзију Ворднета. Неки истраживачи указали су на тај пропуст и на потребу да се на систематичан начин укључе ознаке за пренесено значење у ресурсима какав је ворднет (Shutova, Teufel , & Korhonen,

2013) јер би системи за процесирање реторичких фигура могли да користе тако унапређени ворднет за изградњу или за евалуацију. Истраживањем које је спроведено у овом докторском раду желимо да допринесемо и унапређењу структуре ворднета у смислу представљања фигуративног говора.

4.3 Фигура поређење у домаћој литератури

Поређење је семантичка фигура која се састоји у приближавању једног појма другом по некој семантичкој величини и заснива се на синтаксичком функционисању речи (Грицкат, 1967). Поређење, као исконски, универзални начин упознавања, спознавања, тумачења и описивања света, појава и дешавања у њему и око њега, подразумева поређење двају објеката, откривање њихових сличности и разлика, при чему се нови објекат упоређује са већ познатим. Нека особина је означена у односу на остале, представља основу поређења и служи као основна мера за означавање степена особине, карактеристике, стања или радње одређеног појма који се пореди.

Фигура поређење коју истражујемо у овој тези, у домаћој литератури се назива *поредбеним фразеологизмом*. Структурни облик поредбених фразеологизама, то јест њихова такозвана *тематско-рематска* структура подразумева да садрже два елемента повезана поредбеним везником *као*: *основу поређења* (референт или базни део поређења, тема, леви део поређења) са једне стране, и *компаративни део* (или еталон, рема, десни део поређења) са друге стране (Драгићевић, 2010). Таква структура подразумева да се у устаљеним поређењима одвијају семантички процеси посебне природе засновани на зближавању референта и еталона. Еталони су средства којима се сликовито (изражајно) мери свет, тј. стварност и представљају измерена својства и особине предмета, појава, објеката и најчешће се одражавају у језику у устаљеним поређењима, али еталон може бити било која човекова језичка манифестација одмеравања света. Дакле, еталон је својеврсна јединица одмеравања света, која се разликује од језика до језика и зависи од колективног сазнања (Голяк, 2009).

Према Миливоју Солару „поређење или компарација настаје када се нешто с нечим пореди на основу неких заједничких особина које иначе нису непосредно приметне. Поређење упозорава на посебна својства ствари, појава и особа, откривајући сличности и разлике које често измичу непосредном искуству, изненађују и узбуђују читаоца, или му указујуна одређени посебан аспект посматрања“ (Solar, 2005:84).

Говорећи о речима као оквирима или знацима извесног садржаја Белић истиче важност онога од чега тај садржај зависи – друштвене прилике, степен културе,

окружење, обичаји, навике, социјални односи, веровања и друго (Белић, 1998). То се нарочито огледа у јединицама фразеолошког нивоа. Телија истиче да се еталон може дефинисати као карактеролошки сликовита замена особине човека или предмета неком реалијом или особом, објектом из природе, стварју, који постају знак који у њима доминира, с тачке гледишта животно-културног искуства (Телија, 1996).

У (Драгићевић, 2010:182) се наводи: „Конкретне појаве обично се дефинишу тако што се наводе карактеристике њиховог изгледа или функције, док се апстрактни ентитети често дефинишу прототипичном ситуацијом у којој се јављају. То није достигнуће когнитивне лингвистике и концептуалне анализе, јер се појавама овако приступа и интуитивно".

Велики број фразеолошких поређења припада типу код кога се упоређује један конкретан предмет с другим предметом који нам није појединачно познат и чије је лексичко значење „сужено на значење доминантне карактеристике". Међутим, није свако поређење фразеолошко. Углавном се фразеолошким поређењем приближавају неистородни објекти уз квалификовање појма и појачавање његовог значења (често је нпр. референт човек, његово понашање или делатност, а еталон може бити предмет, животиња и др., са сликовитим представама о њиховим особинама или делатностима), јер се нефразеолошким поређењем упоређују истородни објекти (лице с лицем, животиња са животињом) „којима је заједничка нека реална особина" (Мршевић-Радовић, 1982:42).

Код поређења се референт и еталон приближавају, док се код метафоре поистовећују. „Метафора подразумева идентификацију објеката и изражава стабилну, константну особину (а не и тренутну), док поређење само указује на зближавање објеката и може да се користи и при описивању случајне или тренутне сличности" (Голяк, 2009:211).

У складу са претходним ставовима, у овом докторском раду спровели смо пројекат групне расподеле рада у сврху прикупљања најчешће коришћених поређења у српском језику како бисмо, управо на основу тих поређења, доградили лексичко-семантичку мрежу Српски ворднет која има широку примену у задацима обраде природног језика. Циљ је био добијање увида у семантичко знање и национални поглед на свет садржано у српском језику, на основу говорног српског језика.

4.4 Додавање нових семантичких релација у Српски ворднет на основу фигуре поређење

Ворднет је флексибилан ресурс динамичне структуре који се може мењати у складу са потребама које се јављају у неком задатку обраде природног језика. Тако се неке потешкоће које произилазе из особености језика богате морфологије и правила извођења речи, то јест деривације, какав је српски, могу решити увођењем морфо-семантичких релација (Koeva, Krstev, & Vitas, 2008) (пододељак 2.3.5). Додавањем нових семантичких релација у ворднет могу се постићи различити ефекти, од којих неки утичу и на представљање фигуративног значења. Тако Кути (енг. Kuti) и сарадници представљају нову семантичку релацију названу *скаларна средина* помоћу које се релација антонимије између два синсета описних придева трансформише у троструку, градациону структуру (Kuti, Varasdi , Gyarmati, & Vajda, 2008) , док Мациарц, Спаковиц и Пиа (енг. Maziarz, Szpakowicz и Pia) уводе скуп релација које су у вези са степеном компарације придева, на пример, релације *comparative* и *superlative* за описне придеве (Maziarz, Szpakowicz, & Pia, 2012). Деривациона релација *characteristic* дефинише релацију између придева и именице тамо где је садржај или особина објекта који се описује именицом познат на основу придева, нпр. на основу изјаве „Ако је неко познат, окарактерисан је славом“, те ће релација *characteristic* бити успостављена између именице *слава* и придева *познат*. Нова семантичка релација између именица и придева у Португалском ворднету описана у радовима (Marrafa, et al., 2006) и (Mendes, 2006) дата је у облику пара инверзних релација названих *a characteristic of / has as a characteristic*. Сврха ове релације је, према ауторима, да означи значајну карактеристику именице изражену придевом (‘{carnivorous} is a characteristic of {shark}’, то јест {месождерски} је карактеристика {ајкуле}). Аутори истичу да увођење ове релације обогаћује ворднет, да може допринети процесу одређивања семантичког домена коме припада придев и да се може користити у апликацијама за аутоматско расуђивање (енг. reasoning applications). Неки истраживања бавила су се и дефинисањем релација којима се остварује веза са спољним ресурсима, као што је случај са релацијом названом *здрав разум* (енг. common sense) помоћу које се литерал у синсету повезује са линковима на Википедији у којима је описан (Angioni, Demontis, Deriu, & Tuveri, 2008).

Метода коју предлажемо у овом докторском истраживању ослања се, са једне стране, на ова истраживања која се баве односима именица и придева, осим што за разлику од њих ми трагамо за специфичним везама између именица и придева, односно

разматрамо оне описне придеве који су специфични за мали или врло мали скуп именица, или за само једну именицу. Извор генерисања семантичких релација које предлажемо лежи у реторичкој фигури поређење, која има значајну фреквентност у текстовима на природном језику, чији је значај у обради природног језика истакнут у поглављу 4.2 овог рада.

Друга врста истраживања на коју се ослања истраживање у овом докторском раду приказана је у радовима Вилија и Хаоа (Veale & Hao, 2008) и (Hao & Veale, 2010) и бави се развојем аутоматских метода екстракције семантичког знања из примера примене реторичке фигуре поређење. Ови аутори предлажу проширење ворднета на основу семантичког знања садржаног у језичким конструкцијама облика „ПРИДЕВ као ИМЕНИЦА“ (енг. ADJ as a NOUN), које представљају реторичку фигуру поређење. Аутори су у сврху тог истраживања, како би дошли до примера поређења, из Ворднета извукли листу свих антонимних парова придева и тако добили списак кандидата за претрагу. За сваки придев из те листе, формиран је упит облика „<ADJ as a *>“, где <ADJ> означава један од одабраних придева, и послат Гугл претраживачу. Затим је од добијених резултата формирана колекција од око 2000 конструкција „ПРИДЕВ као ИМЕНИЦА“ за именице над којима је извршен задатак разрешавања значења вишезначних речи (енг. word-sense disambiguation). У том процесу је једној именици придружено више придева, на основу различитих семантичких основа. Структура која се тако добија, названа *frame:slot:filler* састоји се из именице (frame), особине те именице (slot) и придева као вредности која је у вези са особином (filler). Тако за једну именицу може постојати више таквих структура. Репрезентативне именице које су коришћене у овом истраживању су оне именице које су у Ворднету лексикализоване једном лексемом, док су сва остала појављивања поређења у којима је други део упита „<ADJ as a *>“ синтаксички комплексан, изостављане. Просечан број придева који је придружен свакој појединачној именици у истраживању ових аутора је 8. На примеру именице паун (енг. peacock), у табели 8 можемо видети како то изгледа. Приказана метода предлаже додавање 7 релација од којих је прва облика {peacock}Has feather{brilliant}.

Табела 8 Нове релације у Ворднету на основу фигуре поређење

{peacock}	Has feather	{brilliant}
{peacock}	Has plumage	{extravagant}
{peacock}	Has strut	{proud}
{peacock}	Has tail	{elegant}
{peacock}	Has display	{colorful}
{peacock}	Has manner	{stately}
{peacock}	Has appearance	{beautiful}

Исти аутори су ради бољег разумевања структуре језика извршили упоредну анализу ироничних поређења на енглеском и кинеском језику. Тако је показано да су поређења и лингвистички и културолошки феномен, јер се скуп поредбених фраза које су изабране за коришћење у овом раду може применити на кинески језик само у 3-4% случајева (Veale & Hao, 2012).

У једном од раних радова на тему структуре Ворднета (Fellbaum, Gross, & Miller, 1993) говорило се о томе да је везу између именице у Ворднету и придева, као њеног модификатора могуће видети као такозвану рестрикцију одабира (енг. selectional restriction) за тај придев. Тај однос није првобитно уврштен у Ворднет заједно са другим релацијама (више о релацијама у Ворднету и осталим ворднетовима, у поглављу 2.3.1) зато што је то једносмерна релација – ако кажемо да је *врабац мали*, та изјава би се могла представити помоћу означеног показивача (енг. labeled pointer) баш као што је именични синсет *врабац* хипоним именичног синсета *птица*. Ипак, не бисмо могли да поставимо повратни показивач са придева *мали* на именицу *врабац*, док имамо релацију хипернимије која повезује синсет *врабац* са синсетом *птица*. Још тада је закључено да би додавањем сличних, али двосмерних, семантичких релација у Ворднет, чија би семантичка улога била јасна, ова лексичко-семантичка мрежа била значајно побољшана и резултати задатака заснованих на семантичким релацијама дали би боље резултате.

Милер и сарадници су приликом првих разматрања о структури Ворднета (Miller, Fellbaum, Gross, & Miller, 1990) одабрали 25 јединствених почетних синсетова (енг. unique beginners) (прилог 0) „након разматрања могућих комбинација именица-придев до којих би могло доћи“[135] (Miller, 1998:29) што додатно оправдава истраживање односа

[135] “…after considering the possible adjective-noun combinations that could be expected to occur.”

ових речи и проналажење начина да се одговарајући синсетови ворднета повежу семантичким релацијама између именица и придева.

У овом докторском раду такође истражујемо поређења која се користе у савременом, говорном, живом српском језику, ослањајући се на Корпус савременог српског језика као на референтну колекцију текстова у којима су та поређења садржана. У том смислу, предлажемо екстракцију лингвистичких конструкција облика облика „ПРИДЕВ као ИМЕНИЦА“ из анотираног дела Корпуса савременог српског језика (Utvić, 2011) како бисмо у Српском ворднету добили кандидате за повезивање семантичким релацијама које су у уској вези са реторичком фигуром поређење.

4.5 Полуаутоматско проширење Српског ворднета паром нових семантичких релација

Процедуру проширења ворднета релацијама које смо назвали *specificOf/specifiedBy* које овде предлажемо, показаћемо, дакле, на примеру Српског ворднета, који је незаменљив ресурс за многе примене у области обраде српског језика. Овим проширењем желимо да побољшамо семантичку структуру ворднета путем модела додавања нових семантичких веза. Тај модел се може сматрати језички независним у смислу поступка којим се долази до кандидата за повезивање новим семантичким везама, те се може користити и за друге светске језике, уз неопходна прилагођавања.

У том смислу, предлажемо екстракцију лингвистичких конструкција облика „ПРИДЕВ као ИМЕНИЦА“ из анотираног дела Корпуса савременог српског језика (Utvić, 2014) као први корак полуаутоматске методе проширења Српског ворднета новим семантичким релацијама.

Метода проширења паром инверзних релација *specificOf/specifiedBy* коју предлажемо у овом раду, састоји се из следећих корака:

1. Из анотираног корпуса датог природног језика K_l екстраховати лингвистичке конструкције *sims* облика „ПРИДЕВ као ИМЕНИЦА“ и формирати скуп *Sims*, такав да важи следеће:

$$Sims=\{\text{„PRIDEV kao IMENICA“}\},\ sims \in Sims \subset K\_l$$

Из анотираног Корпуса савременог српског језика тако су генерисане 5.952 линије конкорданци облика „ПРИДЕВ као ИМЕНИЦА“ као у следећим примерима:

650810: ri više . - Kakva je ? - <**Bela kao mleko**> . Ona traži isto kao i

54571471: Japanac Takajuki Suzuki, <**brz kao vetar**>, pretrcyao sve domacxe

26045112: i zaturenoj na potiljak , <**crven kao cvekla**> , Platon Rjapčikov i jo

5393576: m leda. Mesec se pojavio <**crven kao krv**>. Pošto su pola časa l

21384848: isjen se vratio srećan i <**lak kao pero**>, sanjao je o slavi. Ne

10660636: slama, žut kao dukat, <**žut kao Mesec**>. Konji ržu, bubne opn

25761433: tresla groznica: bio je <**žut kao pesak**> - kao da ga muči žutic

18206219: dan od zatvorenika; lica <**žuta kao limun**>, radosno polete prema

2. Из скупа Sims ручном методом елиминисати све елементе код којих придеви нису дескриптивни јер су нам само дескриптивни придеви потребни као први део конструкције која чини поређење, тако да је следећи корак:

$$SimsRedycedByADJ=\{sims \in Sims \mid PRIDEV \text{ 'is descriptive'} \}$$

као у следећим примерима где су у питању присвојни придеви:

251511: za taj dan . Jer reč je **<ljudska kao glad>** . Nema uvek istu snagu .

137584415: Drugog? Ljubav **<majčinska kao vernost>** ljubav muško-

После овог корака добили смо $|SimsRedycedByADJ|=2.030$ елемената.

3. Из скупа *SimsRedycedBy*ADJ елиминисати све елементе код којих су именице личне или су замењене скраћеницом, јер такве именице не би биле добри кандидати за формирање реторичке фигуре поређење:

$$SimsRedycedByNOUN=\{sims \in SimsRedycedByADJ \mid IMENICA \text{ 'is a common N'}\}$$

као у следећим примерима:

132719070: Pljevlja bi bila bogata i **<bleštava kao Las>** Vegas - kaže jedan od

40699798: da bude slavna i **<bogata kao Monika>** Seleš . Kako se koja

3992864: je bila tako laka ni tako **<brza kao Elizabet>**, uskoro je izostala, a

87599203: uz reku i da Dunavac bude **<cyist kao Ada>** Ciganlija, a radovi za

68456010: zatvoru u Beogradu , **<opštepoznatom kao CZ>** , naći u poziciji onih

Тако је добијено $|SimsRedycedByNOUN|=1.059$ елемената.

4. Из скупа *SimsRedycedBy*NOUN генерисати подскуп најфреквентнијих елемената:

$$SimsMostFrequent=\{sims \in SimsReducedByNOUN \mid freq(sims) \geq k \} ,$$

где је k минимална фреквенција појављивања структуре „ПРИДЕВ као ИМЕНИЦА“ у посматраном корпусу K_l. У нашем случају, за вредност k=1, укупан број парова ПРИДЕВ-ИМЕНИЦА, кандидата за проширење ворднета је $|SimsMostFrequent|=1.059$.

5. Од датог скупа *SimsMostFrequent* креирати текстуелну датотеку парова ПРИДЕВ-ИМЕНИЦА (Adjective_As_Noun) над којом ће бити примењен алгоритам за полуаутоматско проширење Српског ворднета новим семантичким релацијама (слика 25).

```
Input: Adjective_As_Noun text file
Output: a pair of WordNet mutually inverse
semantic relations (specificOf/specifiedBy)
for each input adjective-noun pair
---------------------------------------------
foreach adjective-noun pair in adjective-noun pairs
if ((adjective exists in Wordnet.adjective.literals)
   and (noun exists in Wordnet.noun.literals)) {
     if ((Wordnet.senses(adjective).Count==1)
        and (Wordnet.senses(noun).Count==1)
        and (Wordnet.sense(adjective).FirstSense)
        and (Wordnet.sense(noun).FirstSense) ) {
         Create_Relation(specificOf,adjective,noun);
         Create_Relation(specifiedBy,noun,adjective);
     }
     else
      foreach (sense in Wordnet.senses(adjective)) {
        add_to_adjective_senses(adjective,sense,synsetId)}
      foreach (sense in Wordnet.senses(noun)) {
        add_to_noun_senses(noun,sense,synsetId)}  }  }
```

Слика 25 – Полуаутоматско проширење Српског ворднета новим семантичким релацијама[136]

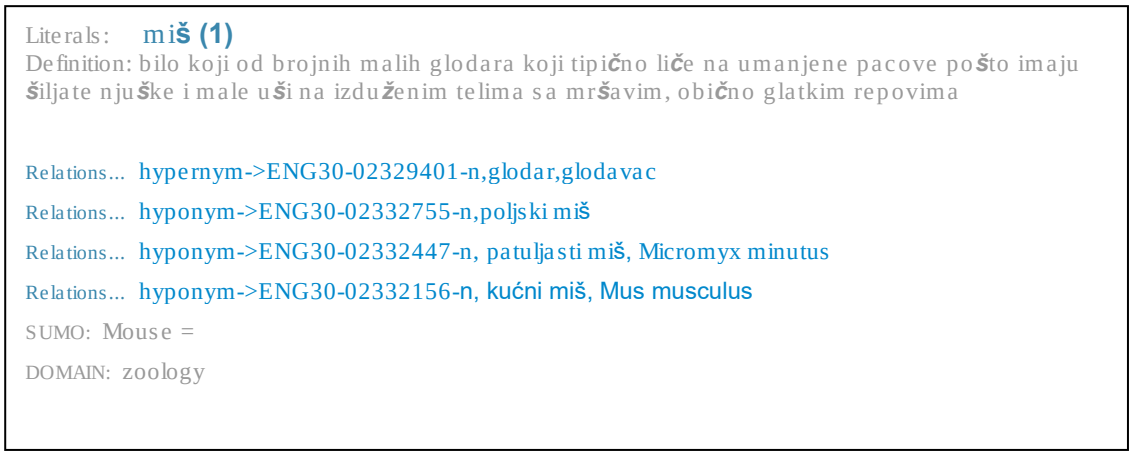
Овај алгоритам за проширење Српског ворднета новим семантичким релацијама, користи се за секвенцијално процесирање унетих кандидата, то јест парова ПРИДЕВ-ИМЕНИЦА. За сваки пар се врши провера на основу које се утврђује да ли у датом ворднету, у нашем случају српском, постоје синсетови тражених придева и именица који су лексикализовани литералима посматраног придева и именице. Затим се примењује процедура аутоматског креирања пара семантичких релација *specificOf/specifiedBy* између синсетова придева и именице, уколико су они лексикализовани само по једним литералом чије је значење прво. Прво значење неке речи је оно семантичко значење дате речи у природном језику које је, корпусом или релевантним речником, дефинисано као најчешће коришћено у датом језику. У случају Српског ворднета то су Копрус савременог српског језика и Речник Матице Српске. Интуиција на којој се заснива ограничење које уводимо односи се на смањење грешке аутоматског упаривања која би се појавила у случају када не постоје жељени синоними у посматраном синсету.

Ако бисмо желели да новом релацијом *specificOf/specifiedBy* упаримо именички синсет {миш} и придевски синсет {плашљив}, тај процес бисмо једнозначно могли да обавимо у случају када у именичком синсету {миш} постоји само један литерал, чије значење је прво. Ако бисмо у ворднету имали два синсета у којима се именица миш појављује као литерал – именица миш у значењу животиња и именица миш као део рачунарске опреме, онда бисмо такве синсетове одвојили у посебну датотеку и спремили

[136] Слика преузета из рада (Mladenović, Mitrović, & Krstev, 2016)

за ручну допуну помоћу посебног алата у оквиру SWNE апликације (о којој говоримо у одељку 2.4.3 овог рада).

У српском ворднету имамо жељени именички синсет представљен на слици 26, док је {računarski miš} литерал засебног синсета, те до горе описаног проблема није дошло.



Слика 26 – Приказ именичког синсета МИШ помоћу SWNE апликације

Вероватноћа за појаву грешке, то јест погрешног упаривања синсетова, постоји уколико бар један од синсетова није исправно допуњен литералима и нема добро додељено значење, или жељено значење није прво или оно не постоји као литерал у ворднету. У том случају, пошто нам је извор могућих грешака познат и ограничен, могуће је извршити проверу структуре ворднета кроз генерисање тестова пре примене алгоритма. Уколико неки од синсетова има више од једног синонима или има један синоним чије значење није прво, такав пар ПРИДЕВ-ИМЕНИЦА се одваја у две независне датотеке: датотеку придева (adjective_senses) и свих његових значења пронађених у ворднету, и у њој аналогну датотеку именица (noun_senses).

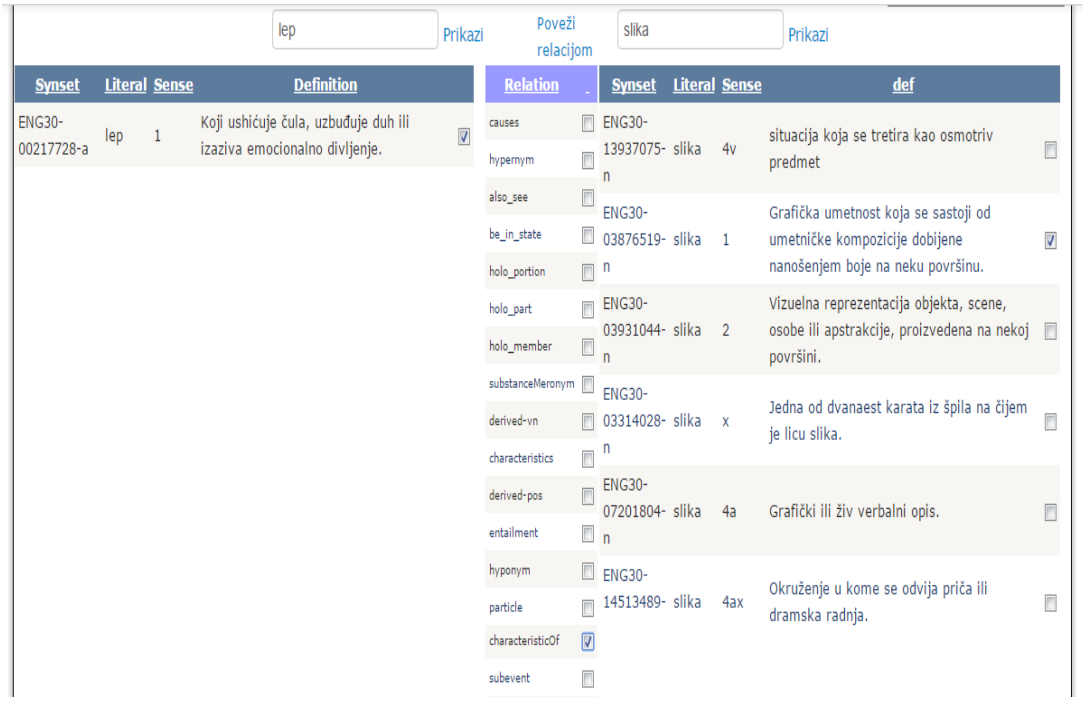
Ако се за неке парове ПРИДЕВ-ИМЕНИЦА на почетку процеса испитивања утврди да уопште не постоје у облику литерала у датом ворднету, они учествују у генерисању кандидата за допуну ворднета на регуларан начин – уносом синсетова.

Пре примене претходно описаног алгоритма за проширење, извршено је испитивање Српског ворднета како бисмо утврдили његову структуру у смислу ограничења која су постављена самим алгоритмом. Утврђено је да наш ворднет садржи преко 22 хиљаде синсетова, од чега је 1.595 синсетова придева са једним литералом, а код 1.452 је значење тог литерала означено као прво[137].

[137] Ови подаци односе се на време спровођења описаних експеримената, јуна 2015. године

Применом предложеног алгоритма, од укупно 1.059 парова ПРИДЕВ-ИМЕНИЦА, са особином „укупно парова чија оба члана имају по једно значење и то значење је прво“ нађено је 69 парова. Парова ПРИДЕВ-ИМЕНИЦА којих има у Српском ворднету, али са више значења или са једним значењем које није прво има укупно 302. Преосталих 688 парова односе се на оне случајеве када бар један члан пара ПРИДЕВ-ИМЕНИЦА не постоји као литерал у SWN. Дакле, предложеном методом Српски ворднет се може одмах, без претходне допуне, проширити са 371 паром релација типа *specificOf/specifiedBy*.

За 302 пара ПРИДЕВ-ИМЕНИЦА којих има у Српском ворднету, али код којих неки од саставних делова има више значења или је значење једно, али не и прво, креирана је страна у веб апликацији коју користимо за изградњу, одржавање и унапређивање Српског ворднета (описане у пододељку 2.4.3) помоћу које смо ручно упарили придеве и именице, бирајући их према одговарајућем значењу и повезујући их паром релација *specificOf/specifiedBy*. Тај процес је описан у пододељку 4.8, док је списак тако додатих веза дат у табели 39 Табела 39која је део прилога 4 овог докторског рада. Овај део апликације корисницима омогућава да уносом речи која означава придев генеришу колону синсетова који су лексикализовани датом речју, а уносом именице генеришу другу колону синсетова који су лексикализовани датом именицом. Одабиром одговарајућих синсетова као и избором жељене релације из треће колоне може се генерисати сама релација. Дакле, за парове ПРИДЕВ-ИМЕНИЦА чији је однос 1:1, то јест и придев и именица имају једно значење и то значење је прво значење, коришћено је мапирање један на један. У случају када је именица вишезначна, или када придев има више значења, софтвер нам омогућава да одаберемо одговарајући синсет, са правим значењем које је потребно за дати пар ПРИДЕВ-ИМЕНИЦА. Тако, на пример, именица *слика*, како је приказано на слици 28, има чак 6 значења, те смо за конструкцију *Леп као слика* одабрали одговарајуће значење – „Графичка уметност која се састоји од уметничке композиције добијене наношењем боје на неку површину“. Слика 27 приказује како се уносом придева „леп“ и именице „слика“ и одабиром жељених опција генерише релација *specificOf* између синсетова ENG30-00217728-a и ENG30-03876519-n, односно добија се семантичка релација s*pecificOf* између придева „леп“ и именице „слика“.



Слика 27 – Аутоматско упаривање литерала „леп“ и „слика“ [138]

4.6 Први пројекат групне расподеле рада и анализа резултата

За потребе овог докторског рада спроведена су два пројекта групне расподеле рада у сврху провере полуаутоматске методе коју смо искористили за проналажење и додавање кандидата за допуну Српског ворднета новим семантичким релацијама. Оба пројекта су у основи спроведена на исти начин, уз измене у другом пројекту које су за циљ имале потенцијално побољшање резултата. Циљ нам је био да оценимо да ли је фреквенција појављивања у Корпусу прихватљив параметар генерисања релација *specificOf/specifiedBy* паровима ПРИДЕВ-ИМЕНИЦА у односу на заступљеност у природном језику. За ово истраживање користили смо онлајн анкету коју смо спровели помоћу сервиса Google Forms[139]. У наставку ћемо објаснити начин спровођења оба пројекта групне расподеле рада и даћемо детаљну анализу добијених резултата.

Поређењем листе (Листа1) која је генерисана аутоматски – помоћу Корпуса савременог српског језика, а потом филтрирана помоћу корака 1-4 у алгоритму за проширење који смо описали у претходном одељку и уређене у опадајућем редоследу

[138] Ова апликација је изграђена у оквиру докторске тезе „Информатички модели у анализи осећања засновани на језичким ресурсима“, ауторке Миљане Младеновић, на Математичком факултету у Београду, 2016. године.

[139] Google Forms. https://www.google.rs/intl/sr/forms/about/

по фреквентности парова, са листом добијеном анонимним анкетирањем (Листа2), желели смо да утврдимо која вредност прага фреквенције *k* обухвата резултате добијене анкетом.

Димензија листе Листа1 одређена је доњим прагом фреквенције појављивања *k*=1 и износила је 1.059 елемената. С обзиром да је из Корпуса савременог српског језика добијено много непотребних података (јер корпус није детаљно анотиран да би дозволио финије упите) који нису у правој мери оцртавали тражене конструкције, а неке конструкције нису носиле жељено значење ван контекста у коме су се налазиле, ручно је одабрано 154 конструкција са Листе 1 за које се сматра да се у неком степену користе у свакодневном језику. Одабир је извршила ауторка овог докторског рада, те је тако настала Листа 2. Неке конструкције са Листе 2 биле су, на пример, следеће: „чист као апотека“, „чист као суза“, „хладан као лед“, „веран као пас“. Конструкције типа „добар као облик“, „добар као писац“, „познат као вођа“ нису узете у обзир јер су представљале случајна појављивања која ван контекста немају жељену семантичку улогу представљену реторичком фигуром поређење. Ручно одабране конструкције искоришћене су за креирање 4 упитника.

С обзиром да нисмо могли да предвидимо колико ће потенцијални учесници бити вољни да помогну, циљ нам је био да бар 30 учесника попуни упитник. Први упитник је имао 30 питања јер смо желели да испитамо ваљаност методе и да утврдимо оптималан број поља у упитнику. У интересу нам је било да за што већи број поређења проверимо да ли се користе у свакодневном језику, али морали смо да ограничимо њихов број у упитницима да се учесници приликом попуњавања не би заморили и изгубили интересовање. Начин формирања упитника и задатака у пројектима групне расподеле рада описан је у (Simperl, 2015).

С обзиром да је ово истраживање било прво истраживање ове врсте на нашим просторима, прво „пуштање“ упитника било је нека врста теста и провера спремности потенцијалних учесника да помогну и допринесу одржавању електронских језичких ресурса за српски језик. Преостали упитници су избалансирани што се тиче броја питања. Број учесника није било могуће одредити унапред – он је зависио од одзива учесника одређеног дана.

Када смо први пут објавили позив за учешће у овом пројекту на приватној Фејсбук страници, у првих неколико сати од објављивања није било много заинтересованих учесника. С обзиром да смо користили Google Forms упитнике, могли смо да пратимо свако ново попуњавање упитника и да видимо тачно време сваког

попуњавања. Позив за учешће у истраживању био је означен као јавни (енг. public) што значи да су сви наши пријатељи на Фејсбуку могли да поделе позив на својим Фејсбук „зидовима“ и да тако прошире глас о нашим напорима. Такође, сваки пут када би неко од наших пријатеља означио да му се позив за учешће допада, кликнувши на дугме „свиђа ми се“ (енг. like), сви његови пријатељи су могли да виде о чему се ради и да сами истраже, притисну исто дугме и поделе позив за попуњавање упитника на својим зидовима. Управо због тога се вест о нашем истраживању брзо проширила, јер првог дана је 12 особа поделило позив за учешће на свом профилу и број учесника је после тога почео веома брзо да расте. Како бисмо искористили ту „популарност“ одлучили смо да упитник оставимо на истој адреси и да само променимо и додамо нова питања (у односу на први упитник од 30 питања). Тако је расподела броја питања по упитницима била 30-41-41-42, да би преостала три упитника била колико-толико избалансирана.

Проблем са оваквом врстом учешћа у истраживању јесте у томе што вести постављене на Фејсбук веб страницу врло брзо изгубе своју драж и занимљивост јер се увек појаве новије и занимљивије вести, што је неопходно имати на уму када делимо позив за учешће преко ове друштвене мреже. Природа коришћења Фејсбука је таква да ако је нека вест, слика, објава изузетно популарна данас, сутра ће можда бити потпуно заборављена, ако је замени нешто што је корисницима занимљивије, смешније, шокантније, итд.

У наредна три дана објавили смо још три упитника на истој URL адреси (управо зато што је објава била популарна и дељена) те смо успели да добијемо довољно одговора. Циљали смо на најмање 30 одговора, ради валидности статистичке провере. Другог дана је највећи број учесника попунио упитнике, док је до четвртог дана заинтересованост значајно опала и број учесника је опао, што је само потврдило претпоставку да је најбоље постављати нове упитнике на истој URL адреси.

Анкета је, дакле, рађена у временском периоду од 5 дана, тако што су упитници објављивани сукцесивно на друштвеној мрежи Фејсбук. Анонимни корисници су сваку конструкцију „ПРИДЕВ као ИМЕНИЦА“ постављену у упитницима, оцењивали позитивно или негативно, означавајући их са ДА или НЕ, дајући нам тако до знања да ли дате конструкције користе у свакодневном језику, или их можда користе особе из њиховог окружења.

У Табела 9 је приказана расподела питања према упитницима као и број испитаника који су дали одговоре.

Табела 9 Расподела питања и учесника у првом пројекту

Google упитник	Питања по упитнику	Учесника по упитнику
1	30	46
2	42	138
3	41	150
4	41	100
Укупно	154	434

У процесу анализирања резултата првог пројекта групне расподеле рада прво смо измерили квалитет доприноса учесника и одредили скуп оних учесника чији ће одговори бити узети у обзир као релевантни. У сличним пројектима групне расподеле рада, увек је неопходно извршити адекватну евалуацију доприноса учесника, нарочито ако је у питању волонтерски рад, јер неретко се дешава да неки учесници поступају неодговорно и да намерно дају насумичне или погрешне одговоре. Тако је у првом пројекту групне расподеле рада који смо спровели, неколико учесника давало само позитивне или само негативне одговоре на сва питања, то јест, изјаснили су се или да користе баш све понуђене фразе, или да не користе ниједну, што је, у оба случаја, мало вероватно. Ипак, примећено је да су у пројектима који се заснивају на такозваној групној мудрости, као једној од врста групне расподеле рада, учесници углавном инспирисани алтруизмом, те злонамерних учесника скоро да нема. Неопходно је, дакле, испитати квалитет доприноса учесника, то јест анотатора, како их углавном називамо, те је потребно оценити такозвани степен слагања међу анотаторима.

Наше истраживање је специфично јер није лако дати јасно упутство учесницима пројекта групне расподеле рада, као што је то случај у пројектима анотације текстова на основу врсте речи или у пројектима анотације слика. Одговори у нашем истраживању су у извесној мери лични и зависе и од нивоа образовања, година, итд. Зато је било тешко одредити „златно питање“ као вид ране евалуације, то јест питање на које се очекује позитиван или негативан одговор од свих учесника. У фази припреме упитника за прво истраживање закључили смо тако да израз „*Лукав као лисица*“ није толико широко распрострањен у свакодневном језику као што нам се чинило, те смо одустали од коришћења златног питања на које би одговор морао бити позитиван. Једино питање које смо поставили у свим упитницима било је „Црн као улица“, на њега је очекивани

одговор негативан па смо и све остале одговоре учесника који су на то питање одговорили позитивно изузели из аланизе резултата, то јест њихови одговори нису били релевантни, а учесници су означени као неповерљиви.

Прва линија одабира релевантних учесника заснивала се на утврђивању скупа оних учесника чије одговоре можемо сматрати релевантним. Релевантност је мерена аритметичком средином свих одговора учесника у истраживању, дакле скупова позитивних и негативних одговора (означених са 1 за позитивне и са 0 за негативне одговоре). Други ниво одабира релевантних учесника односио се на анализу одговора свих оних учесника код којих није било значајне разлике између аритметичких средина њихових одговора.

У првој фази одабира користили смо такозвани *Студентов t-тест,* који је најчешће коришћен тест параметријске значајности за тестирање нулте хипотезе и користи се за тестирање значајности разлика између две аритметичке средине. Услови за примену овог теста јесу да обе варијабле морају бити нумеричке и да величина узорка мора бити бар 30 јединица (у супротном њихов распоред треба да буде нормалан или симетричан). Наши подаци испунили су оба ова услова.

Ради прецизније статистичке анализе и мерења појединачног доприноса сваког учесника, четири Гугл упитника поделили смо на 7 подскупова, како је приказано у табели 10.

Сваки подскуп имао је 30 или мање питања и конвертован је у матрицу у којој је сваки ред представљао одговоре једног испитаника, а свака колона је представљала једно питање у облику „ПРИДЕВ као ИМЕНИЦА“. Садржај у свакој ћелији матрице имао је вредност 1 ако је испитаник одговорио на питање са ДА и вредност 0 ако је одговор био НЕ. Редови матрице су онда међусобно упоређени t-тестом за зависне узорке како бисмо утврдили има ли битне разлике између аритметичких средина одговора учесника. Из сваког подскупа изабрали смо по пет испитаника чија су међусобне разлике аритметичких средина свих одговора, на основу резултата t-теста за зависне узорке биле најмање са интервалом поверења 95%.[140]

[140] За ове тестове коришћен је Excel статистички алат t-Test: Paired Two Sample for Means.

Табела 10 Међусобна сагласност на основу Калфа теста

Подскуп упитника	Број учесника	Број питања	Калфа вредност	Број питања означених са ДА
1	5	30	α = 0,7575*	16
2a	5	21	α = 0,713*	17
2b	5	21	α = 0,698*	15
3a	5	21	α = 0,688*	5
3b	5	20	α = 0,484	
4a	5	21	α = 0,434	
4b	5	19	α = 0,375	
Укупно		154		53

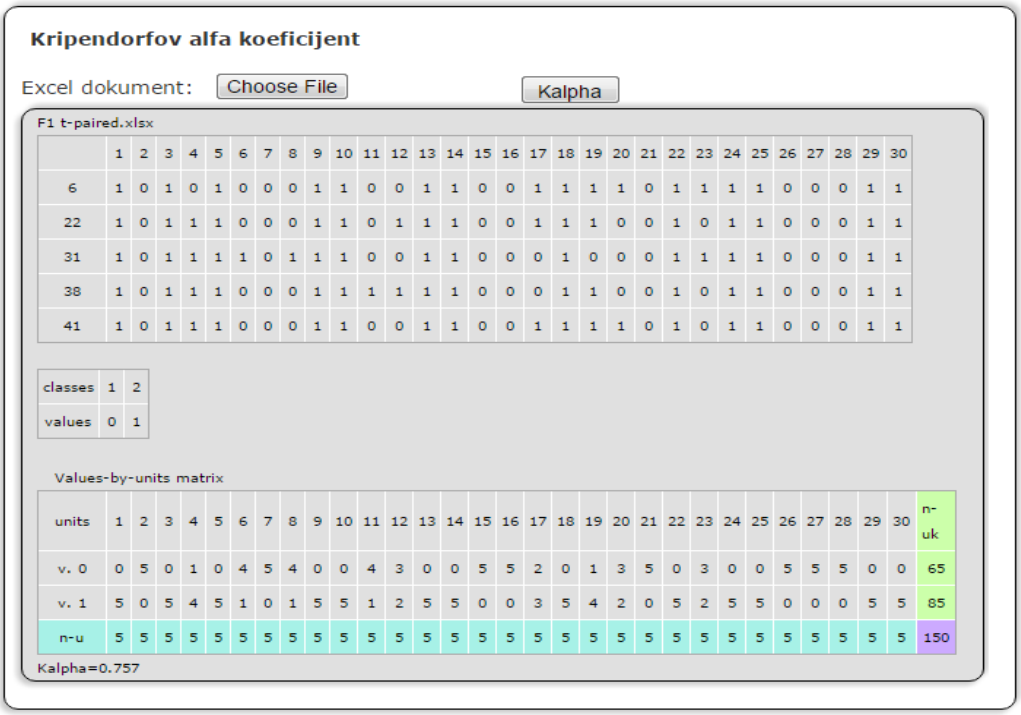
После тог корака, међусобно слагање одговора учесника додатно је измерено рестриктивнијом методом која се назива Крипендорфова алфа, Крипендорфов коефицијент или скраћено Калфа, коју смо описали у пододељку 3.5.2 овог рада.

Резултати које смо добили користећи Калфа тест над скупом добијеним на основу одговора 5 учесника (то су учесници за које се, применом t-теста, показало да су аритметичке средине њихових одговора најприближније) за сваки подскуп упитника приказани су у табели 10. У случају када је вредност Калфа била таква да се међусобно слагање учесника могло сматрати поузданим, што је био случај са прва два упитника и део трећег, за све конструкције „ПРИДЕВ као ИМЕНИЦА“ које су биле понуђене учесницима, које је већина учесника (3 или више од 3, од укупно 5) означила одговором ДА, та конструкција је узимана као елемент Листе 2’. Тако смо добили укупно 53 елемента а њихова расподела у односу на скупове из упитника приказана је у последњој колони табеле 10.

Приметили смо и занимљиву појаву опадања вредности Крипендорфовог коефицијента над истом структуром упитника у односу на време када су упитници попуњавани – што се такође може уочити у табели 10 – наиме, вредност Калфа опада с протоком времена, то јест прогресивно је нижа за сваки наредни скуп који је обрађен. То се може објаснити смањењем интересовања за истраживање, иако не можемо бити сигурни да су упитнике попуњавали исти учесници или су сваки нови скуп питања попуњавали други учесници. Ово запажање не треба занемарити код планирања неког будућег истраживања. Ипак, јасно је да се Крипендорфов алфа коефицијент може користити у другим, сличним, истраживањима, јер смо могли да израчунамо степен слагања међу анотаторима иако се током попуњавања упитника могло догодити да једна

особа попуни сва 4 формулара, или да иста особа попуни један формулар више пута и то различитим одговорима.

Слика 28 даје приказ начина обраде првог подскупа одговора учесника, где је у горњем делу приказан подскуп број 1 као улазни скуп података упитника бр. 1 са 30 питања и одговорима 5 испитаника. У доњем делу исте слике приказани су резултати Калфа теста у облику матрице вредности у односу на јединице мере (енг. values-by-units) где је вредност 0 додељена за одговор НЕ, а вредност 1 за одговор ДА, те се види да је вредност Калфа коефицијента α=0,757. На начин приказан на овој слици обрађено је свих 7 матрица, за све подскупове упитника.



Слика 28 – Обрада првог скупа одговора Калфа тестом

Желели смо и да проценимо колико промена прага фреквенције утиче на релевантност аутоматски одабраних парова ПРИДЕВ-ИМЕНИЦА, мерено на основу резултата које смо добили из упитника. Листа1 је зато редукована тако да садржи упитник 1 и делове упитника 2a, 2b и 3a што укупно чини 93 елемента, то јест све парове ПРИДЕВ-ИМЕНИЦА за које се показало да су релевантни на основу евалуације учесника који су попунили упитнике, коришћењем t-теста и Калфа коефицијента за проверу поузданости. Та листа је названа Листа 1'. С друге стране, Листа 2' садржи само оне парове ПРИДЕВ-ИМЕНИЦА Листе 1' које је лингвистички експерт означио позитивно, то јест као делове конструкција које се користе у свакодневном језику.

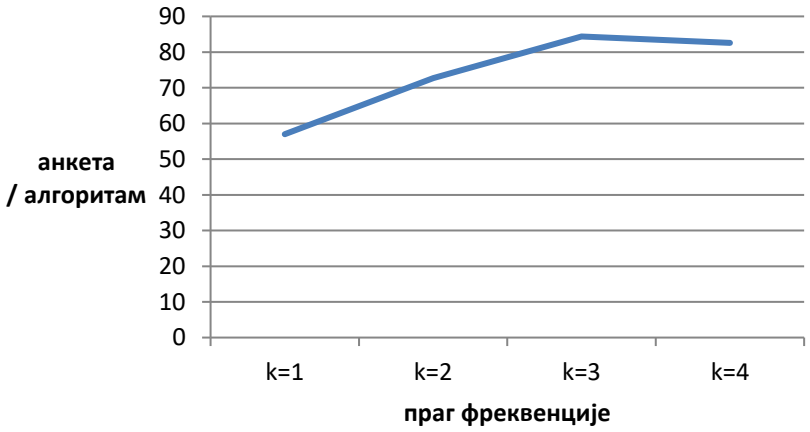
За почетак смо желели да поставимо праг фреквенције на $k = 4$, што је значило да је алгоритам коришћен за процесирање само оних појављивања у Корпусу чија је фреквенција $k \geq 4$. На Листи 1' било је 23 таква пара. Од та 23 пара, 19 парова је било присутно на Листи 2', што значи да учесници у упитницима нису препознали 4 пара која је алгоритам препознао (то су *парови дуг као вечност*, *дубок као бунар*, *јак као град*, *црн као гавран*).

У табели 11 је приказан однос броја парова које смо добили коришћењем алгоритма, у односу на број оних добијених одабиром учесника.

Табела 11 Однос између ручно и аутоматски одабраних парова у односу на промену прага фреквенције

Фреквенција	**Алгоритам**	**Људи**	**Људи/ Алгоритам**
***k*= 1**	93	53	57%
***k* = 2**	44	32	73%
***k* = 3**	32	27	84%
***k* = 4**	23	19	83%

Слика 29 приказује однос између људског одабира у односу на аутоматски одабир када се праг фреквенције мења.



Слика 29 – Однос броја ручно одабраних парова у односу на аутоматски одабране, уз промену фреквенције

На основу података приказаних на слици 29 можемо да закључимо да се на узорку од 93 пара ПРИДЕВ-ИМЕНИЦА са Листе 1', то јест са листе парова који су одабрани на основу Калфа теста, проценат учешћа ручно одабраних парова мења у подскупу који је добијен одабиром само оних парова са исте листе чија је фреквенција једнака или виша у односу на постављени праг, када се праг мења. Остварени резултат од 84% даје нам

ручно измерену прецизност алгоритма за аутоматско проширење ворднета уз праг фреквенције k=3. Надаље закључујемо да ако у методи проширења Српског ворднета заснованој на полуаутоматском моделу проширења ворднета новим семантичким везама, које смо описали у претходном одељку, за праг фреквенције узмемо k=3, можемо очекивати тачност одабира парова у износу од 84%.

Неке од конструкција које су одабране на основу одговора учесника су очекиване, али неке за које смо мислили да ће сасвим сигурно бити одабране, на пример „Брз као мисао“, добиле су веома мало гласова, то јест позитивних оцена. Фреквенција појављивања у Корпусу се показала као добра метода одабира кандидата за допуну ворднета полуаутоматском методом, али наравно да и ту има одступања. Тако, на пример, неке конструкције које су у Корпусу имале фреквенцију појављивања ≥ 4, уопште нису одабране од стране учесника у пројекту групне расподеле рада (у табели 12 означене помоћу *).

Табела 12 Конструкције из анкета

5 од 5 гласова	**≤ 2 гласова**
Tačan kao sat	Brz kao misao*
Hladan kao led	Lak kao ptica*
Hladan kao špricer	Beo kao kreda
Tvrdoglav kao mazga	Debeo kao bure
Lagan kao pero	Blistav kao zvezda

4.7 Други пројекат групне расподеле рада и анализа резултата

Други пројекат спровели смо на исти начин као и први, с тим што су конструкције које смо уврстили у упитнике одабране насумично, то јест нисмо користили помоћ лингвистичког експерта приликом одабира кандидата за упитнике. Дакле, за добијање конструкција „ПРИДЕВ као ИМЕНИЦА“ пратили смо кораке 1-4 алгоритма за проширење ворднета новим семантичким релацијама (пододељак 4.5), те смо тако добили 1.059 конструкција. Надаље смо насумично одабрали 154 конструкције (коришћењем RAND функције у програму Microsoft Word), те смо те конструкције такође понудили потенцијалним учесницима у пројекту групне расподеле рада, и то у односу 30-41-41-42 питања, као што је то учињено и у првом пројекту. Циљ овако

поновљеног истраживања био је да видимо разлику у резултатима када и тај корак обавимо аутоматски, без помоћи људског фактора одабира.

Још једна разлика у односу на прво истраживање је у томе што смо за постављање питања у формуларима, уместо остављања формулара на истој адреси и ручног мењања питања сваког дана, што је био напоран процес, користили такозвани генератор насумичног линка (енг. Random Link Generator), те су корисници посећивањем веб адресе која је са њима подељена преко Фајсбука сваки пут добијали насумично одабрани упитник, један од 4. Тако је овај процес умногоме олакшан.

У другом пројекту групне расподеле рада који смо спровели у склопу овог докторског истраживања, расподела питања, то јест конструкција „ПРИДЕВ као ИМЕНИЦА“ у упитницима била је иста као и у првом истраживању, али је број учесника био значајно другачији. Овог пута је одзив био мањи (219 учесника, у односу на 434 у првом истраживању), што је и приказано у табели 13.

Табела 13 Расподела питања и учесника у другом истраживању

Google упитник	**Питања по упитнику**	**Учесника по упитнику**
1	30	52
2	42	49
3	41	53
4	41	65
Укупно	154	219

Израчунавањем Калфа коефицијента се показало да је скуп учесника у овом, новом истраживању, много поузданији од скупа који је учествовао у претходном истраживању. Због чињенице да су кандидати за упитнике одабрани насумично, а не на основу процене лингвистичког експерта, број веза „ПРИДЕВ као ИМЕНИЦА“ које могу бити кандидати за унос у Српски ворднет знатно је мањи – само 21. У табели 14 представљена је анализа процеса одређивања вредности Калфа коефицијента за све подскупове упитника, као што је то урађено и у претходном истраживању, с тим што је сада, у складу са добијеном вредношћу Калфа коефицијента, закључено да се поузданима могу сматрати одговори учесника из 6 подскупова упитника (док је у првом истраживању било поуздано само 4 подскупа).

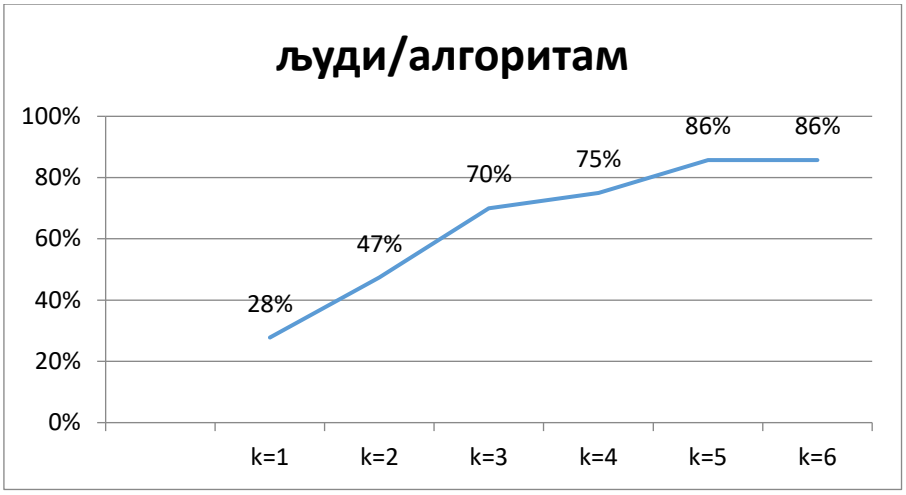
Табела 14 Међусобна сагласност на основу Калфа теста

Подскуп упитника	**Број учесника**	**Број питања**	**Калфа коефицијент**	**Број питања означених са ДА**
1	5	30	α = 0.833*	4
2a	5	21	α = 0.602	
2b	5	21	α = 0.927*	3
3a	5	21	α = 0.692*	3
3b	5	20	α = 0.787*	3
4a	5	21	α = 0.72*	8
4b	5	19	α = 0.525	
Укупно		154		21

И у другом пројекту смо желели да утврдимо колико промена прага фреквенције утиче на релевантност аутоматски одабраних парова ПРИДЕВ-ИМЕНИЦА, мерено на основу резултата које смо добили из упитника. Резултати су приказани у табели 15, док слика 30 даје приказ истог односа.

Табела 15 Однос између ручно и аутоматски одабраних парова у односу на промену прага фреквенције

Фреквенција	**Алгоритам**	**Људи**	**Људи/Алгоритам**
***k*=1**	54	15	28%
***k*=2**	19	9	47%
***k*=3**	10	7	70%
***k*=4**	8	6	75%
***k*=5**	7	6	86%
***k*=6**	7	6	86%



Слика 30 - Однос броја ручно одабраних парова у односу на аутоматски одабране, уз промену фреквенције

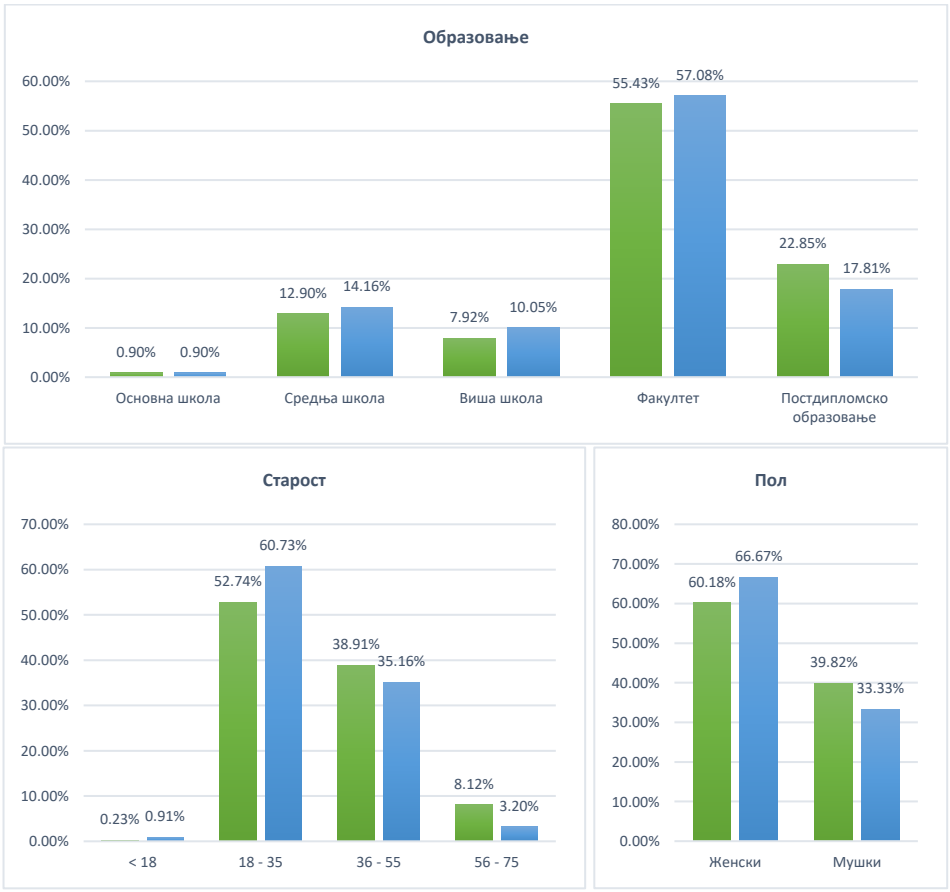
У претходном истраживању се показало да за праг фреквенције $k = 3$, можемо очекивати тачност одабира парова у износу од 84%, док је за исти праг фреквенције у овом другом истраживању очекивана тачност 70%; Праг $k = 5$ је дао добар резултат у другом истраживању, али при оцени овог резултата морамо узети у обзир чињеницу да су у другом истраживању парови ПРИДЕВ-ИМЕНИЦА одабрани насумично. Зато се праг фреквенције $k=3$ може сматрати довољно добрим у смислу резултата добијених у оба ова истраживања.

У првом истраживању, конструкције „Црн као гавран“ фреквенције 5 у Корпусу и „Дубок као бунар“ фреквенције 4, нису биле одабране анкетом, премда је то било очекивано. С друге стране, конструкције „Познат као особа“ и „познат као члан“ су конструкције које су, као део другог истраживања, имале фреквенцију 5 у Корпусу, али, очекивано, нису одабране у анкетама. Такви примери су показали да је ипак неопходно обавити одабир конструкција које ће бити понуђене учесницима пројекта групне расподеле рада, да би се избегло непотребно означавање конструкција које немају смисла.

У другом истраживању је учествовало више женских особа (6,49 % више), било је више учесника из старосне групе 18-35 година (7,99 % више) и више факултетски образованих учесника (само 1,65 % више), али и мање учесника са постдипломским образовањем (5,4 %). Значајна разлика која би могла да буде схваћена као разлог за боље резултате у другом истраживању, у смислу слагања одговора учесника, јесте то што је позив за попуњавање упитника упућен и преко Фејсбук странице Друштва за језичке

ресурсе и технологије[141], те су учесници били више мотивисани да помогну, јер неки од њих имају директан интерес у побољшавању језичких ресурса за обраду природног језика. Ови подаци су приказани у табели 16 док је однос укупног броја учесника који су учествовали у оба описана истраживања приказан у табели 17.

Табела 16 Учесници I истраживања у односу на учеснике II истраживања



[141] JERTEH https:// http://jerteh.rs/

Табела 17 Однос броја учесника у и првом и другом истраживању

Упитник	Бр. питања по упитнику	Учесници I истраживања	Учесници II истраживања	Однос бр. уч. у I и II истраживању
1	30	46	52	0.89
2	42	138	49	0.28
3	41	150	53	0.28
4	41	100	65	0.15
Укупно	154	434	219	0.2

У оба истраживања учествовало је више женских особа у односу на мушке особе. У првом истраживању учествовало је 60,18% особа женског пола у односу на 39,82% особа мушког пола, док је у другом истраживању учествовало 66.67% особа женског пола у односу на 33.33% особа мушког пола. Показало се да у другим значајним истраживањима која за циљ имају прикупљање лингвистичких података за изградњу лексичких ресурса путем модела групне расподеле, као што су игре са сврхом JeuxDeMots и Phrase Detectives (пододељак 3.2.2), већи допринос такође дају учеснице (Gurevych & Kim, 2013).

4.8 Ручно додавање веза

Применом алгоритма за полуаутоматско додавање нових веза у српском ворднету (слика 25) додато је 69 нових веза између синсетова придева и именица. Ти синсетови су приказани у табели 38, у оквиру прилога 5 на крају овог рада.

У процесу ручног повезивања синсетова чији литерали имају више значења, наишли смо на многе занимљиве примере.

Табела 18 приказује повезивање литерала „брз“ и литерала „тигар“. Литерал *тигар* је у српском ворднету представљен у значењу које је у случају овог поређења жељено, то јест као 1)Велика шумска мачка из Азије, али и као 2) река Тигар, због чега ови литерали нису могли бити повезани аутоматски.

У случају повезивања литерала „црвен“ и „трешња“ (табела 19Табела 19), проблем у аутоматском повезивању јавља се зато што је литерал „трешња“ у српском ворднету приказан у два значења: 1) трешњево дрво и 2) плод трешњевог дрвета. Жељено значење које желимо да постигнемо у случају овог поређења – „црвен као трешња“ је друго значење. Вишезначност се јавља у именичном делу поређења и у примеру „црвен као ватра“ (табела 20), јер је литерал „ватра“ у српском ворднету представљен у значењу ватре као једног од основних елемената и ватре као чина отварања ватре на непријатеља.

На примеру додавања везе „дуг као век“ (табела 21) видимо да понекад и именица и придев имају више значења међу којима је потребно одабрати права. Два литерала у српском ворднету лексикализована речју „дуг“ су хомографи, то јест речи истог облика али другачијег значења у зависности од изговора: реч дуг коју изговарамо са дугосилазним акцентом је именица која има значење нечега што се дугује, док исту реч изговорену са краткоузлазним акцентом тумачимо као основно просторно значење, тј. релативно велико простирање.

Сличан је случај додавања везе „хладан као лед“ (табела 22), где имамо два значења придева „хладан“ и два синсета у којима се јавља „лед“, од којих је други синсет{светлоемитујућа диода:1, ЛЕД:1} где је ЛЕД скраћеница за диоду која емитује светло (енг. LED light-emitting diode).

На примеру додавања везе „хладан као шприцер“ видимо да су у Српском ворднету означена и нека пренесена значења речи. Овде у опису другог значења за придев „хладан“ видимо да је то проширено значење и да се односи на психолошку хладноћу, недостатак људске топлине и емоција (табела 23). Ипак, у случају поређења „хладан као шприцер“, жељено значење придева хладан је сличније значењу које се у енглеском језику често користи за придев „cool“, то јест, сталожен, миран, неко ко се не потреса лако, те ћемо и ово значење додати у Српски ворднет. Ради додавања релације „црвен као земља“ бирали смо између чак 9 значења именице „земља“ у Српском ворднету.

Приликом ручног додавања нових релација приметили смо неке непотпуности Српског ворднета, те је створена одлична прилика да унесемо нове синсетове или

значења, као и да побољшамо постојеће компоненте нашег ворднета. Тако, на пример, за придев „сјајан“ немамо значење „онај који сија“ већ само значење „изузетно добар“. Приметили смо и да у Српском ворднету уопште немамо придев „сладак“, који нам је у овом истраживању потребан ради додавања веза „сладак као млеко“ и „сладак као бомбона“. Такође, придев „свеж“ нема значење које би одговарало енглеској речи “fresh”, које нам је потребно за додавање везе „свеж као јабука“.

Занимљиво је да за именицу „жар“ немамо значење које је потребно ради додавања везе „црвен као жар“. Два значења која постоје су 1) осећање интензивне љубави и 2) осећање јаке жудње (обично усмерене на особу или неки повод). Именица „жар“ у Српском ворднету није представљена са значењем „врели комадићи дрвета или угља настали сагоревањем“ које нам је у овом случају потребно.

Српски ворднет је после иницијалне фазе изградње у склопу пројекта BalkaNet проширивaн у складу са потребама које су налагала истраживања у којима је коришћен (Крстев, и др., 2008), што објашњава непотпуности са којима смо се сусрели. Ово истраживање нам помаже да ближе испитамо значења именица и придева који су представљени у нашем ворднету и да ову важну базу знања допунимо значењима која су честа у савременом, говорном српском језику и садржана у често коришћеним поређењима.

Табела 18"Брз као тигар"

Синсет	Литерал	Значење	Дефиниција	Релација		Синсет	Литерал	Значење	Дефиниција
ENG30-00323873-a	брз	1	Наступати или кретати се брзо, или имати способност за брзо наступање или кретање.	specificOf/ specifiedBy		ENG30-02129604-n	тигар	1	Велика шумска мачка из Азије; има смеђежуто крзно са црним пругама; угожена врста.
						ENG30-09458791-n	Тигар	1	Азијска река; улива се у Еуфрат.

Табела 19„Црвен као трешња“

Синсет	Литерал	Значење	Дефиниција	Релација		Синсет	Литерал	Значење	Дефиниција
ENG30-00381097-a	црвен	1	Који има боју која подсећа на боју крви, трешања, парадајза или рубина.	specificOf/ specifiedBy		ENG30-07757132-n	трешња	A1a	Плод дрвета трешње са једном коштицом
						ENG30-12641413-n	трешња	A1b	Неко дрво које рађа мале, меснате, округле плодове са једном тврдом коштицом.

Табела 20„Црвен као ватра“

Синсет	Литерал	Значење	Дефиниција	Релација		Синсет	Литерал	Значење	Дефиниција
ENG30-00381097-a	црвен	1	Који има боју која подсећа на боју крви, трешања, парадајза или рубина	specificOf/ specifiedBy		ENG30-14686186-n	ватра	1	Некада се сматрало да је то један од четири елемента од којих се састоји свемир (по Емпидоклу, архаично).
						ENG30-00986938-n	ватра	2a	Чин отварања ватре на непријатеља.

Табела 21„Дуг као век“

Синсет	Литерал	Значење	Дефиниција	Релација		Синсет	Литерал	Значење	Дефиниција
ENG30-13397174-n	дуг	1	Нешто што се дугује.	specificOf/ specifiedBy		ENG30-15205532-n	век	3	Сто година.
ENG30-01433493-a	дуг	1a	Основно просторно значење; релативно велико простирање или веће него уобичајено.			ENG30-15140744-n	век	1a	Период између рођења и садашњости.

Табела 22 "Хладан као шприцер"

Синсет	Литерал	Значење	Дефиниција	Релација		Синсет	Литерал	Значење	Дефиниција
ENG30-01251128-a	хладан	1	Користи се за физичку хладноћу; који има ниску или неадекватну температуру или има осећај хладноће или који је охлађен (уз помоћ леда или у фрижидеру)..	specificOf/ specifiedBy		ENG30-07914777-n	шприцер	2	Пиће које се прави мешањем вина и минералне воде.
ENG30-01257612-	хладан	2	Проширено значење; посебно се односи на психолошку хладноћу; без људске топлине и емоција.						

Табела 23 „Хладан као лед“

Синсет	Литерал	Значење	Дефиниција	Релација		Синсет	Литерал	Значење	Дефиниција
ENG30-01251128-a	хладан	1	Користи се за физичку хладноћу; који има ниску или неадекватну температуру или има осећај хладноће или који је охлађен (уз помоћ леда или у фрижидеру)	specificOf/ specifiedBy		ENG30-14915184-n	лед	1	Смрзнута вода у чврстом стању.
ENG30-01257612-a	хладан	2	Проширено значење; посебно се односи на психолошку хладноћу; без људске топлине и емоција.			ENG30-03666362-n	светло емитујућа диода, ЛЕД	1	Диода која емитује светло.

5 Примена унапређеног Српског ворднета

Ворднет је ресурс који се примењује у широком спектру задатака у области обраде природног језика. Ворднетове за различите језике је могуће користити једноставно као речнике синонима, претражујући жељене речи и њихове синониме помоћу корисничког сучеља, на пример Ворднет[142], Пољски ворднет[143], Бугарски ворднет[144] или Српски ворднет, који је могуће претраживати преко странице Друштва за језичке ресурсе и технологије[145].

Ворднет се често користи у задацима разрешавања вишезначности (енг. word sense disambiguation) (Fong, 2004), где је улога ворднета да укаже на право значење речи у случају постојања вишезначности. Резултати тог задатка надаље могу представљати улазне податке за аутоматско превођење, проналажење информација, проширивање упита, идентификацију концепата (енг. conceptual identification), одређивање семантичке удаљености (енг. semantic distance) и друге задатке (Lohk, 2015). Ворднет се користи и за семантичко означавање (Cole & Gwizdka, 2008), семантичку класификацију (Gutiérrez, Vázquez, & Montoyo, 2011) и многе друге задатке у којима је важно указати на значење речи или приказати семантичке и лексико-семантичке релације између речи.

Унапређење Српског ворднета које смо приказали у овом докторском раду има за циљ побољшавање резултата система за анализу осећања и ставова и анализу фигуративног језика о чему ћемо говорити у наставку овог рада.

5.1 Аутоматско проналажење ироније и сарказма у корпусу добијеном са друштвене мреже Твитер

Ирониja је фигура речи која утиче на промену значења дела текста у коме се појављује, тако што се то значење тумачи као супротно од очекиваног. У грчком језику реч *eironeia* (гр. ειρωνεία) значи „претварање“, па ирониja, самим тим, претвара реч или појам у њену супротност и доводи је у несклад са контекстом. Ирониja се често јавља у свакодневном говору, где се креће од доброћудне шале до заједљивог сарказма. У говору се ирониja означава нарочитом интонацијом, а у писању понекад знацима навода и другим обележјима, као на пример, коришћењем курзива (енг. italic font). Познавање контекста у оквиру ког се ирониja јавља је уједно и предуслов за разумевање ироније.

[142] http://wordnetweb.princeton.edu/perl/webwn
[143] http://plwordnet.pwr.wroc.pl/wordnet/
[144] http://dcl.bas.bg/bulnet/
[145] http://sm.jerteh.rs/Default.aspx

Унапређени Српски ворднет, то јест онтологија која је генерисана из ове лексичко-семантичке мреже, је недавно искоришћена као део система за детекцију ироничних изјава у процесу машинског учења.

Онтологија SWNonto се генерише аутоматски из семантичке мреже Српски ворднет, помоћу софтверског алата SWNE који користимо за развој и одржавање Српског ворднета (пододељак 2.4.3). С обзиром да овај алат, у зависности од задатака, равноправно користи три формата (OWL, XML и TXT) за репрезентацију Српског ворднета, могуће је серијализовати садржај Српског ворднета у неком од тих формата. Структура и особине ове онтологије детаљно су приказани у (Младеновић, Информатички модели у анализи осећања засновани на језичким ресурсима, 2016).

Машинско учење (енг. machine learning) је дисциплина која омогућава рачунарима да уче без експлицитног програмирања (Samuel 1967). Дефинише се и као генерализација знања на основу претходног искуства (података о појавама и ентитетима који су предмет учења) (Manning & Schütze, 1999) . Добијено знање користи се како би се дали одговори на питања за ентитете или појаве које раније нису виђене. За евалуацију модела машинског учења користе се мере које су познате из области проналажења информација (енг. information retrieval): прецизност (енг. precission), одзив (енг. recall), F-мера (енг. F-measure) и тачност (енг. accuracy).[146] Ове мере се израчунавају на основу могућих исхода – стварно позитивни, СП (енг. true positives, TP), стварно негативни, СН (енг. true negatives, TN), лажно позитивни, ЛП (енг. false positives, FP) и лажно негативни, ЛН (енг. false negatives, FN).

Прецизност је мера тачности (вероватноћа да је класификација случајног документа тачна) и израчунава се по формули:

$$\text{Прецизност} = \frac{\text{СП}}{\text{СП} + \text{ЛП}}$$

Одзив је мера комплетности (вероватноћа да је документ који припада некој категорији тако и класификован). Израчунава се по формули:

$$\text{Одзив} = \frac{\text{СП}}{\text{СП} + \text{ЛН}}$$

Ефикасност једног класификационог модела повећава се повећањем прецизности и одзива.

Мера која представља тежинску хармонијску средину прецизности и одзива назива се *F-мера*. Израчунава се по формули:

[146] https://en.wikipedia.org/wiki/Precision_and_recall

$$\text{Ф} - \text{мера} = \frac{2 * \text{Прецизност} * \text{Одзив}}{\text{Прецизност} + \text{Одзив}}$$

Тачност је однос тачно класификованих докумената и свих класификованих докумената. Израчунава се по формули:

$$\text{Тачност} = \frac{\text{СП} + \text{СН}}{\text{СП} + \text{СН} + \text{ЛП} + \text{ЛН}}$$

Подаци који се користе у процесу машинског учења потребни су за обучавање, валидацију и тестирање модела. Типична подела скупа података је подела на 60% скупа за обучавање, 20% се користи за анализу грешака, а преосталих 20% се користи за тестирање модела машинског учења. Избор узорака је потребно урадити на случајан начин, насумично (енг. random selection).

Модел машинског учења би требало да верно описује појаве или ентитете. Зато се особине и односи у датом домену који истражујемо представљају атрибутима (енг. features), чији правилан одабир представља изазов и обично се врши системом покушавања и исправљања грешака (енг.trial and error) све док се не добије скуп атрибута који на прави начин представља особине и односе. Два најчешћа облика машинског учења су надгледано учење (енг. supervised learning), у коме се користи обележени скуп подаака за тренирање и ненадгледано учење (енг. unsupervised learning), у коме се тренирање алгоритама обавља на основу необележеног скупа података.

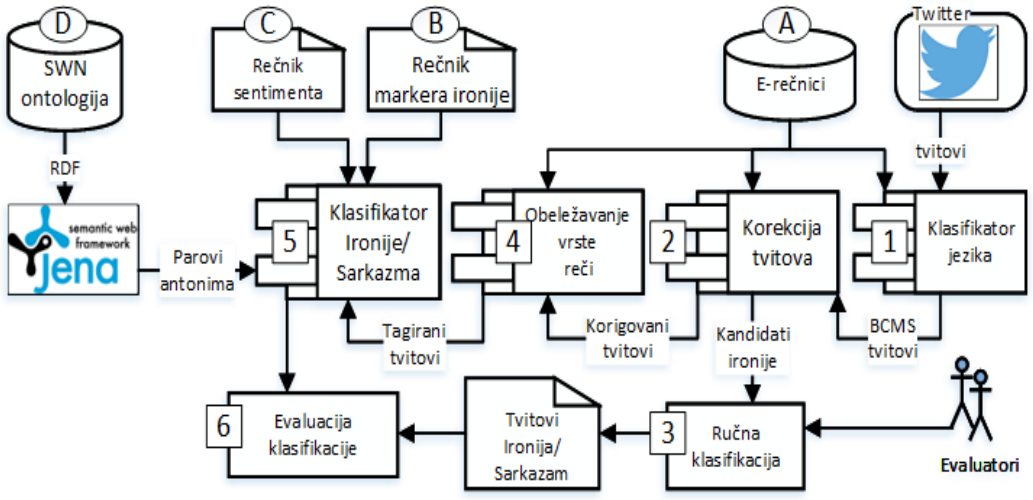
У моделу класификације изјава на друштвеној мрежи Твитер на ироничне и неироничне, коришћен је приступ надгледаног машинског учења и различити језички ресурси: електронски морфолошки речници, речник сентимената, речник маркера ироније и онтологија SWNonto, аутоматски генерисана из Српског ворднета. Атрибути који су коришћени у процесу машинског учења су:

- парови антонима добијени коришћењем правила закључивања над онтологијом добијеном из Српског ворднета (R);
- парови антонима у којима је поларитет осећања једног члана позитиван (енг. positive sentiment polarity) (PPR);
- поларитет речи које исказују позитивна осећања и ставове (PSP);
- уређени низ ознака осећања или ставова (OSA);
- етикете врсте речи (POS) и
- маркери ироније (M).

Корпус кратких порука на коме је вршено истраживање изграђен је на основу онлајн претраге са геолокацијским и временским ограничењем, коришћењем упита:

#ironija near:Belgrade,Serbia within:400km since:2013-01-01 until:2015-10-29

Тако је, после пречишћавања, добијено 2.127 филтрираних твитова, из којих су уклоњени метаподаци, везе према другим твитовима и етикете # (енг. hashtags). Обављена је и унификација алфабета, те су ћирилични твитови преведени на латиницу. Оваква претрага Твитера извршена је да бисмо добили твитове на језицима којима се говори у бившој Југославији, који се понекад називају BCSM језицима (скраћено од Bosnian, Croatian, Montenegrin, Serbian). Маркери иронije или показатељи иронije у српском језику и сродним језицима су, на пример, речце *баш* и *већ*, стилистички маркери иронije, на пример знаци интерпункције, курзивна слоца итд., фразе као што су: „боље не може“, „ултра-мега-гига“ итд. Комплетан скуп маркера иронije део је архитектуре модела за аутоматско препознавање иронije и сарказма, која је приказана на слици 31.

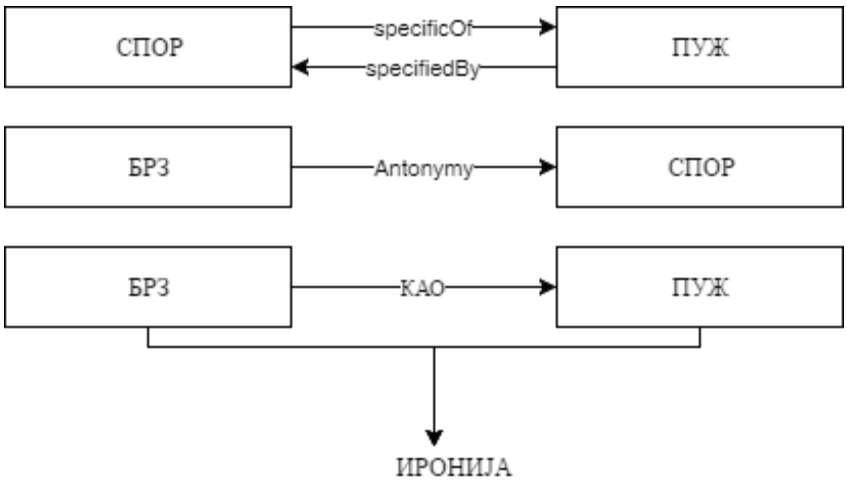


Слика 31 Систем за класификацију иронije[147]

Важан део система на коме се заснива аутоматска детекција иронije и сарказма приказан је, на упрошћен начин, на слици 32, на којој можемо видети да су управо новододате семантичке релације у Српском ворднету, *specificOf/specifiedBy*, суштински важне за препознавање ових реторичких фигура. Ако су придеви *спор* и *пуж* повезани паром ових инверзних релација у ворднету, а релација антонимије свакако постоји

[147] Слика преузета из рада (Mladenović, Krstev, Mitrović, & Stanković, 2017) и прилагођена.

између придева *брз* и *спор*, правилима логичног закључивања (енг. reasoning rules) спроведеним над онтологијом SWNonto, коришћењем свих елемената архитектуре приказане на слици 32, може се аутоматски закључити да су придев *брз* и именица *пуж* део ироничног тврђења.



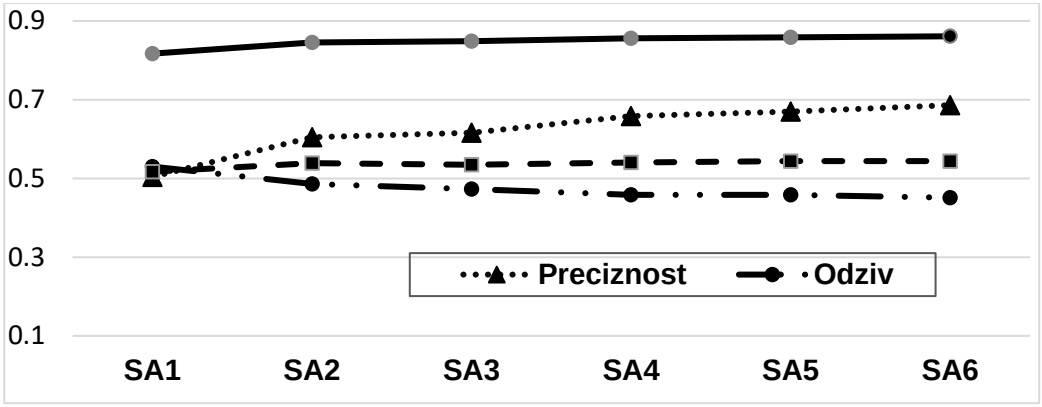
Слика 32 Део система за детекцију иронije заснован на новим везама у СВН

Евалуација је извршена на две колекције твитова, то јест кратких порука са друштвене мреже Твитер, које су ручно означене тако да за сваку поруку у првој колекцији знамо да ли је иронична или не, а у другој колекцији имамо информацију о томе да ли се у порукама користи сарказам или не. За детекцију ироније и сарказма коришћен је исти систем, с тим што је он за иронију показао боље резултате, те би за сарказам у будућим истраживањима требало изменити и прилагодити неке делове тог система.

Резултати евалуације овог истраживања за реторичку фигуру иронија приказани су на слици 33. У табели 24 се види да је најбоља тачност система за детекцију ироније постигнута коришћењем скупа 5 атрибута (PPR, PSP, POS, OSA, M) – acc = 86.1. Систем је тачно класификовао 144 твита, за 66 твитова је погрешно оценио да су иронични, 175 ироничних није класификао као такве и 1347 исправно није означио као ироничне.

Табела 24 Евалуација система за детекцију ироније

Skup atributa	Preciznost	Odziv	F1	Tačnost	Lista atributa
SA1	0.504	0.53	0.517	0.817	POS,OSA,M
SA2	0.605	0.486	0.539	0.845	R,OSA,M
SA3	0.616	0.473	0.535	0.849	PSP,POS,OSA,M
SA4	0.658	0.458	0.54	0.856	R,PSP,POS,OSA,M
SA5	0.67	0.458	0.544	0.858	PPR,POS,OSA,M
SA6	0.686	0.451	0.544	0.861	PPR,PSP,POS,OSA,M

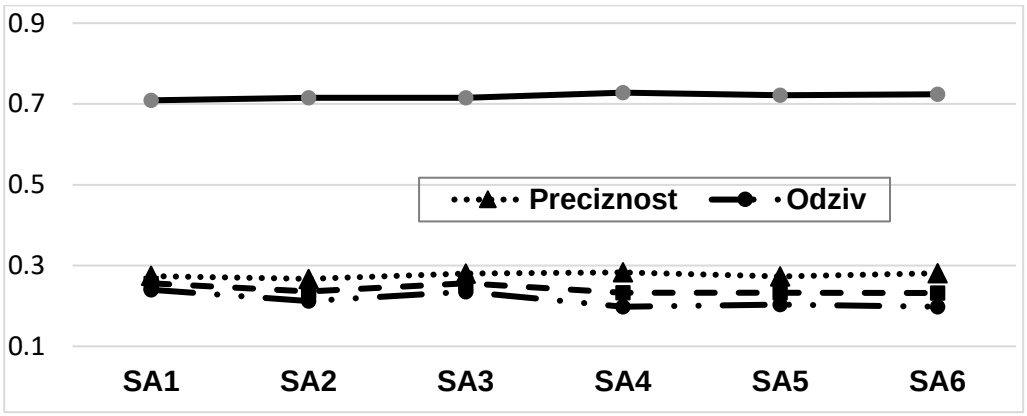


Слика 33 Резултат класификације ироније у твитовима

Резултати евалуације овог истраживања за реторичку фигуру сарказам приказани су на слици 34. У табели 25 се види да је најбоља тачност система за детекцију сарказма постигнута другим, сличним скупом 5 атрибута у односу на скуп којим је постигнут најбољи резултат у детекцији сарказма – (R, PSP, POS, OSA, M) – acc = 72.8.

Табела 25 Евалуација система за детекцију сарказма

Skup atributa	Preciznost	Odziv	F1	Tačnost	
SA1	0.274	0.24	0.256	0.709	POS,OSA,M
SA2	0.267	0.212	0.236	0.715	R,OSA,M
SA3	0.28	0.235	0.256	0.715	PSP,POS,OSA,M
SA4	0.283	0.198	0.233	0.728	R,PSP,POS,OSA,M
SA5	0.273	0.203	0.233	0.722	PPR,POS,OSA,M
SA6	0.281	0.198	0.232	0.724	PPR,PSP,POS,OSA,M



Слика 34 Резултат класификације сарказма у титовима

Детаљан опис овог истраживања дат је у раду (Mladenović, Krstev, Mitrović, & Stanković, 2017).

5.2 Методологија примењена на грчки језик

Истраживање које за циљ има проналажење карактеристичних поређења у српском језику спроведено је и у контексту грчког језика. У сарадњи са Атинским институтом за обраду језика и говора (енг. Institute for Language and Speech Processing – ILSP) спровели смо истраживање на основу методологије примењене у овом докторском раду.

Конструкције облика „именица као придев“ које су у вези са реторичком фигуром поређење добијене су из Хеленског националног корпуса (енг. Hellenic National Corpus (HNC))[148], али и из корпуса добијеног методом прикупљања са веба (енг. web crawling) јер је број поређења у првом ресурсу био недовољан. Око 2.000 поређења добијено је из Хеленског националног корпуса, и из корпуса од око 110 милиона речи добијеног прикупљањем информација са веба (Mastropavlos & Papavasileiou, 2011).

Корпуси су претраживани помоћу обрасца “adjective σα(ν) (Det) noun” (σα(ν) = ‘as’, ‘like’). Фреквенција ових конструкција је ниска што показује податак да је из другог корпуса добијено само 500 погодака, те да је само половина тих погодака заправо представљала устаљено поређење. Већина поређења се у овим корпусима појављивала само једном. Напослетку је број добијених поређења сведен на 154 кандидата како би њихов број био једнак броју кандидата који су учествовали у пројекту групне расподеле рада на српском језику, који је тема овог доктората. Ова одлука је донесена како бисмо за грчке податке могли да употребимо исте мере евалуације и да тако добијене резултате лакше упоредимо.

Конструкције „придев као именица“ на грчком језику понуђене су учесницима у овом пројекту групне расподеле рада на исти начин као и у нашим истраживањима, а расподела учесника по сваком формулару приказана је у табели 26.

Табела 26 Расподела учесника и питања у грчком истраживању

Google упитник	**Питања по упитнику**	**Учесника по упитнику**
1	30	67
2	42	85
3	41	79
4	41	59
Укупно	154	290

Процена међусобне сагласности анотатора је проверена коришћењем Крипендорфовог алфа коефицијента, као и у српском истраживању (пододељак 3.5.2), на начин приказан у табели 27.

[148] Hellenic National Corpus http://hnc.ilsp.gr/

Табела 27 Међусобна сагласност на основу Калфа теста у грчком истраживању

Подскуп упитника	Број учесника	Број питања	Калфа вредност	Број питања означених са ДА
1	5	30	$\alpha = 1$*	20
2a	5	21	$\alpha = 0.736$*	11
2b	5	21	$\alpha = 0.69$*	13
3a	5	21	$\alpha = 0.735$*	10
3b	5	20	$\alpha = 0.696$*	19
4a	5	21	$\alpha = 0.697$*	12
4b	5	19	$\alpha = 0.698$*	9
Укупно		154		94

У табели 27 видимо да је вредност Калфа коефицијента била неочекивано добра. Захваљујући томе, 94 од 154 конструкције из фомрулара су означене као поређења која се користе у свкодневном, говорном грчком језику. Детаљан приказ и опис овог истраживања дати су у радовима (Mitrović, Markantonatou, Mladenović, & Krstev, 2018) и (Mitrović, Markantonatou, Mladenović, & Krstev, 2016).

5.3 Дисеминација резултата истраживања спроведеног у овом докторском раду

У склопу европског програма за сарадњу у домену научних и технолошких информација – COST (енг. European Cooperation in Science and Technology) под називом PARSEME (енг. Parsing and multi-word expressions)[149] који је трајао од 8.03.2013 – 30.04.2017. године, у више наврата смо представили унапређене лексичке ресурсе и алате за обраду српског језика (Savary, et al., 2015).

На редовним састанцима овог значајног пројекта у Атини 2014. године, на Малти 2015. године, као и у Дубровнику 2016. године, наша истраживања су представљена у виду научних постера. Ауторка овог докторског рада је учествовала и на летњој школи чија је главна тема било машинско превођење коришћењем вишечланих језичких јединица (енг. MWEs in Machine Learning), где смо такође представили један научни постер који је међународној научној заједници ближе представио допринос ове

[149] PARSEME http://www.cost.eu/COST_Actions/ict/IC1207

докторске тезе. Учествовање у пројекту PARSEME резултовало је и двонедељним студијским боравком ауторке овог докторског рада у Атини, у сарадњи са Институтом за обраду језика и говора (енг. Institute for Language and Speech Processing – ILSP)[150]. Циљ овог студијског боравка била је размена искустава и настојање да лексичке ресурсе за српски и грчки језик приближимо и искористимо у заједничким пројектима. У складу са тим, методологија коју смо користили за прикупљање релевантних парова придев-именица и додавање нових веза заснованих на реторичкој фигури поређење у српски ворднет, примењена је у случају грчког језика (пододељак 5.2), а примери поређења на грчком језику додати су у онтологију *РетФиг* у циљу проширивања њене базе знања на друге светске језике. Извештај са овог студијског боравка може се пронаћи на веб страници пројекта PARSEME, у делу под називом STSMs (Short Term Scientific Missions)[151]. Из сарадње са институтом ILSP проистекла је и публикација у виду поглавља књиге у издању Атинске академије и њеног Истраживачког центра за неологизме и научне термине (енг. Academy of Athens Research Center for Neologisms and Scientific Terms) (Mitrović, Markantonatou, Mladenović, & Krstev, 2018).

Резултати овог докторског рада представљени су и на Међународној радионици из области Рачунарске реторике (енг. Computational Rhetoric)[152], релативно нове области обраде природног језика у којој се највише пажње посвећује обради фигуративног језика и реторичких фигура. Значај онтологије *РетФиг* и модела додавања нових веза у ворднет је у томе што се тако могу представити и детектовати реторичке фигуре на многим светским језицима, уз неопходна прилагођавања која налажу синтаксичка правила тих језика – велико интересовање за сарадњу по овом питању за сада су испољиле колеге из Словачке, Кине и Канаде, те се надамо да ће пројекти изградње ресурса за препознавање фигуративног језика, на основу резултата наших истраживања, бити спроведени и за друге светске језика.

С обзиром да је главни ресурс који је коришћен и унапређен током израде овог докторског рада Српски ворднет, резултате овде представљених истраживања смо представили на најважнијој међународној конференцији која је посвећена Ворднету и свим другим лексичко-семантичким мрежама које су у свету настале по узору на Ворднет. Ова конференција се одржава сваке две године, те смо радове објавили и

[150] ILSP http://www.ilsp.gr/en
[151] STSMs https://typo.uni-konstanz.de/parseme/index.php/stsm-grants/finished-stsms
[152] http://computationalrhetoricworkshop.uwaterloo.ca/#

представили на овој конференцији 2014. године (Mladenović, Mitrović, & Krstev, 2014) и 2016. године (Mladenović, Mitrović, & Krstev, 2016).

6 Закључак и будући рад

У овом докторском раду представили смо методу језички независног полуаутоматског додавања пара инверзних семантичких релација у Српском ворднету, на основу семантичког знања садржаног у реторичкој фигури поређење. Те релације, назване *specificOf/specifiedBy* формирају се између именичког и придевског синсета. Предложени метод користи анотирани корпус за истраживање семантичког знања садржаног у лингвистичким конструкцијама које имају улогу реторичке фигуре поређење. На основу фреквенције појављивања поређења у корпусу, предложили смо додавање пара нових релација, које повезују синсет именице и синсет придева који представља карактеристични атрибут те именице. Овај приступ је проверен методом провере оцењивача, говорника српског језика, путем модела групне расподеле рада како би валидност аутоматске методе била процењена.

Евалуација је показала да је 84% аутоматски одабраних и најфреквентнијих језичких конструкција облика „именица као придев“, чији је праг фреквенције појављивања у Корпусу савременог српског језика 3, такође одабрано од стране оцењивача, учесника у овде описаним истраживањима спроведеним путем модела групне расподеле рада.

Предложена нова метода за унапређивање језичких ресурса, нарочито ворднета, заснована на провери путем модела групне расподеле рада је врло погодна јер за њену реализацију нису потребна значајна финансијска средства пошто је заснована на доприносу волонтера. Значај ове методе је у томе што се може применити и за доградњу језичких ресурса на другим светским језицима – очигледан следећи корак је коришћење ове методе за доградњу Ворднета, чиме би била омогућена детекција реторичких фигура на енглеском језику, помоћу правила описаних у пододељку 5.1, с тим што бисмо, уместо онтологије SWNonto, користили онтологију генерисану над Ворднетом.

Планиран је и рад на проширењу лексичко-семантичке мреже типа ворднет за немачки језик – GermaNet (Kunze & Lemnitzer, 2002), уз неопходна прилагођавања методологије додавања нових семантичких веза, узимајући у обзир посебну структуру лексичко-семантичке мреже GermaNet, у којој су, на пример, синсетови придева организовани другачије него у Ворднету, уз коришћење хијерархијског приступа, сличнијег начину структурирања мрежа именица и глагола у Ворднету. Семантичке релације које се користе за организовање придева на немачком језику у семантичку мрежу су антонимија, хипонимија, релација асоцијације, пертонимија и партицип.

Модел кооперативног рада на доградњи Српског ворднета који је спроведен у првим годинама након завршетка пројекта Балканет, коришћењем модела групне расподеле рада би могао би бити убрзан коришћењем модела групне расподеле рада, а контрола квалитета унетих синсетова би свакако била олакшана, захваљујући новом скупу алата за одржавање и доградњу Српског ворднета. Један овакав пројекат планирамо у блиској будућности.

Структура Корпуса савременог српског језика, као референтне збирке текстова на савременом српском језику показала се као довољна за истраживање представљено у овом докторском истраживању. Ипак, изградњом специјализованог корпуса књижевних дела свакако бисмо добили боље резултате, као и могућност примене исте методологије за неке друге реторичке фигуре. С обзиром на непотпуну морфолошку анотацију корпуса и немогућност постављања финијих морфолошких упита, приликом претраге овог корпуса добили смо много непотребних података, што је захтевало ручно филтрирање. Ипак, чињеница да је корпус анотиран на нивоу врсте речи и леме је била веома корисна, јер смо у почетној фази истраживања добили потенцијалне комбинације придева и именица у српском језику које повезане везником „као“ формирају реторичку фигуру поређење, такве да се именица налази на крају посматране лингвистичке структуре, што је значајно прецизније и ефикасније у односу на слична истраживања која су била спроведена слањем упита на Гугл претраживач (Veale & Hao, 2008). План за будуће истраживање је коришћење алата Unitex и електронских морфолошких речника српског језика за претраживање Корпуса савременог српског језика ради могућности постављања финијих упита.

Резултати овог докторског рада искоришћени су за инстанцирање атрибута у онтологији *РетФиг* (пододељак 2.2.2), то јест за додавање примера поређења која су прикупљена пројектима групне расподеле рада описаним у овом раду. Тако је овај важан ресурс обогаћен са скоро 400 примера поређења на српском језику за која можемо рећи да се користе у живом, српском језику, те су значајна за многе примене у рачунарској обради српског језика. Додавање нових семантичких релација у Српски ворднет испунило је очекивања о побољшању система за аутоматску детекцију реторичких фигура у српском језику, приказаног у пододељку 5.1.

Алгоритам за полуаутоматско проширење ворднета паром нових семантичких релација (слика 25) може бити побољшан у кораку 5 коришћењем неке од метода за разрешавање значења речи (енг. word-sense disambiguation) чиме бисмо и литерале са више од једног значења могли да повежемо аутоматски, без потребе за ручним

повезивањем као што је то описано у пододељку 4.8, а учињено помоћу веб алата за одржавање Српског ворднета (пододељак 2.4.3).

Планирано је и слично истраживање за конструкције „глагол као именица“, као нпр. „Поцрвенети као булка“. Прелиминарна истраживања Корпуса савременог српског језика показала су да је број оваквих конструкција које би биле корисне за додавање у језичке ресурсе веома мали, па зато нисмо покушали да пратимо исту методологију која је коришћена за конструкције „ПРИДЕВ као ИМЕНИЦА“. С обзиром да се конструкције „ГЛАГОЛ као ИМЕНИЦА“ у свакодневном говору користе скоро исто колико и конструкције „ПРИДЕВ као ИМЕНИЦА“, у наредним истраживањима ћемо изменити начин прикупљања података, то јест, од учесника ћемо тражити да нам кажу које изразе тог облика користе у свакодневном говору. Поређења типа „чисто је као у апотеци“ су такође занимљива и биће део будућих истраживања. Оваква поређења су лингвистички комплексније, па у овом тренутку није јасно у којој мери ће методологија примењена у истраживањима представљеним у овом докторском раду бити измењена и прилагођена.

Инспирацију за могућу следећу фазу доградње Српског ворднета семантичким релацијама на основу реторичких фигура добили смо од неких учесника у првој фази пројекта; наиме, учесници су радо остављали коментаре испод Фејсбук објаве која је садржавала везу према нашим упитницима, и давали су конкретне примере које они сами користе у говору, а који нису били садржани у упитницима. Тако би у неком наредном истраживању, један ред упитника могао да изгледа овако: „Паметан као ...“, те би учесници сами попуњавали други део конструкције, чиме бисмо, у теорији, могли да добијемо заиста актуелно, „живо“, семантичко знање српског језика. Ипак, овакав начин прикупљања података био би још захтевнији за евалуацију и групна расподела рада отвореног типа због тога можда не би била добар избор. Највеће богатство нашег језика крије се управо у народној мудрости и неке изразе које можемо чути од наших бака и дека, свакако не можемо пронаћи у литератури нити у речницима. Зато ово остаје отворено питање, чије смо решење истраживањем представљеним у овом докторском раду само начели, а има изузетан значај за очување наше културне баштине.

7 Литература

Angioni, M., Demontis, R., Deriu, M., & Tuveri, F. (2008). Semanticnet: a WordNet based Tool for the Navigation of Semantic Information. *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, (стр. 21-34).

Antonić, S., & Krstev, C. (2008). Serbian WordNet for Biomedical Sciences . *INFORUM 2008: 14th Conference on Professional Information Resources.* Prague.

Arazy, O., Morgan, W., & Patterson, R. (2006). Wisdom of the Crowds: Decentralized Knowledge Construction in Wikipedia. *16th Annual Workshop on Information Technologies & Systems.* Milwaukee, USA.

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of International Conference on Computational Linguistics COLING* , (стр. 101-108). Geneva, Swizerland.

Biemann, C., & Nygaard, V. (2010). Crowdsourcing Wordnet. *Proceedings of the 5th Global WordNet conference.* Mumbai.

Black, M. (1962). *Models and Metaphors.* Cambridge University Press.

Bobrow, D. G. (1964). *A Question-Answering System for High School Algebra Word Problems.* New York: Proceedings of AFIPS Conference, 26. FJCC, Part I.

Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*, 75-90.

Brabham, D. C. (2008). Moving the crowd at iStockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application. *First Monday*.

Brabham, D. C. (2013). *Crowdsourcing.* MIT Press.

Braslavski, P., Mukhin, M., Ustalov, D., & Kiselev, Y. (2016). YARN: Spinning-in-Progress. *Proceedings of the 8th Global WordNet Conference*, (стр. 58-65). Bucharest, Romania.

Buecheler, T., Sieg, J. H., Füchslin Rudolf M, R. M., & Pfeifer, R. (2010). Crowdsourcing. Open Innovation and Collective Intelligence in the Scientific Method. *Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems*, (стр. 679-686). Odense, Denmark.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, *22*(2).

Causer, T., Tonra, J., & Wallace, V. (2012). Transcription maximized; expense minimized: Crowdsourcing and editing The Collected Works of Jeremy Bentham. *Literary and Linguistic Computing, 27*(2).

Chamberlain, J., Kruschwitz, U., & Poesio, M. (2012). Motivations for Participation in Socially Networked Collective Intelligence Systems. *Collective Intelligence*.

Charles, W. G. (1988). The categorization of sentential contexts. *Journal of Pszcholinguistics Research, 17*(5), 403-411.

Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break It Down: A Comparison of Macro and Microtasks. *Human-Computer Interaction.* Seoul, Republic of Korea.

Chklovski, T. (2005). Collecting paraphrase corpora from volunteer contributors. *Timothy Chklovski. 2005. Collecting paraphrase corpora from volunteer contributors. In Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05)* (стр. 115-120). New York, NY, USA: ACM.

Chklovski, T., & Mihalcea, R. (2002). Building a sense tagged corpus with open mind word expert. *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions - Volume 8* (стр. 116-122). Association for Computational Linguistics.

Christ, O. (1994). A modular and exible architecture for an integrated corpus query system. *Proceedings of COMPLEX '94:3rd Conference on Computational Lexicography and Text Research*, (стр. 23-32). Budapest, Hungary.

Christodoulakis, D. N. (2004). *BalkaNet Final Report IST-2000-29388 1*. University of Patras, Greece. Преузето са http://www.dblab.upatras.gr/balkanet/deliverables/finalreport_sub.pdf

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cole, M. J., & Gwizdka, J. (2008). Tagging Semantics: Investigations with Wordnet. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL).* Pittsburgh, PA.

da Costa, L. M., & Bond, F. (2016). Wow! what a useful extension to wordnet! *Language Resources and Evaluation Conference (LREC)*, (стр. 4323-4328). Portorož.

Dawid, A., & Skene, A. V. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics, 28*(1), 20-28.

Devedžić, V. (2010). *Semantic Web and Education. Integrated Series in Information Systems.* New York: Springer-Verlag.

Dinucci, D. (1999). *Fragmented Future*. Преузето 2013 са http://darcyd.com/fragmented_future.pdf

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM 54(4): 86.*, *54*, стр. 86.

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2002). Towards an Integrated Crowdsourcing Definition. *Journal of Information Science*, *38*(2), 189-200.

Esuli, A. E., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *In Proceedings of the LREC-06, 5th conference on language resources and evaluation,*, (стр. 417–422). Genova, ITs.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge: MIT Press.

Fellbaum, C., Gross, D., & Miller, K. (1993). Nouns in wordnet. Five papers on WordNet. (C. F. G. A. Miller, Ур.)

Fellbaum, C., Gross, D., & Miller, K. J. (1993). Five Papers on WotdNet. MIT Press.

Filatova, E. (2012). Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. *Language Resources and Evaluation (LREC).* Istanbul.

Fillenbaum , S., & Jones, L. V. (1965). Grammatical Contingencies in Word Association. *Jouirnal of Verbal Learning and Verbal Behavior, 4*, 248-255.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fong, S. (2004). Semantic Opposition and Wordnet. *Journal of Logic, Language and Information, 13*, 159-171.

Fort, K., Adda, G., Sagot, B., & Mar, J. (2011). Criticisms About Amazon Mechanical Turk Overpowering Use. У *HLT Challenges for Computer Science and Linguistics* (стр. 303-314). Springer.

Godfrey, J. J., & Zampolli, A. (1996). Language Resources. У A. R. Cole, *Survey of the State of the Art in Human Language Technology.* Преузето са http://www.lt-world.org/hlt-survey/master.pdf

Grier, D. A. (2011). Not for All Markets. *Computer, 44*(5), 6-8.

Grier, D.A. (1998). The Math Tables Project of the work projects administration: the reluctant start of the computing era, *IEEE Annals of the History of Computing*, 20(3), 33-50. IEEE.

Gross, D., & Miller, K. J. (1990). Adjectives in WordNet. *International Journal of Lexicography, 3*(4), 265-277.

Gurevych, I., & Kim, J. (Уредници). (2013). *The People's Web Meets NLP*. Springer Berlin Heildeberg.

Gurevych, I., & Kim, J. (2013). *The People's web meets NLP: Collaboratively constructed language resources.* Heilderberg: Springer.

Gutiérrez, Y., Vázquez, S., & Montoyo, A. (2011). Sentiment Classification Using Semantic Features Extracted from Wordnet-Based Resources. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA '11*.

Hanks, P. (2005). Similes and Sets: the English Preposition like. У C. a. and, & R. B. (eds.) (Ур.), *Jazyky a jazykověda (Languages and Linguistics: Festschrift for Professor Fr. Čermák.* Prague: Philosophy Faculty of the Charles University.

Hao , Y., & Veale, T. (2010). An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes. *Minds and Machines, 20*(4), 635–650.

Hardie, A., Koller, V., Rayson, P., & Semino, E. (2007). Exploiting a Semantic Annotation Tool for Metaphor Analysis. У P. R. M. Davies (Ур.), *Proceedings of the Corpus Linguistics 2007 Conference.*

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures, 1*(1), 77-89.

Hladká, B., Mírovský, J., & Schlesing, P. (2009). Designing a language game for collecting coreference annotation. *In Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09)* (стр. 52-55). Stroudsburg, PA, USA: Association for Computational Linguistics.

Horák , A., & Smrž, P. (2003). VisDic - Wordnet Browsing and Editing Tool. *Proceedings of the Second International WordNet Conference (GWC)* (стр. 136-141). Brno, Czech Republic: Masaryk University.

Horák, A., & Smrž, P. (2004). New Features of Wordnet Editor VisDic. *Romanian Journal of Information Science and Technology, 7*(No. 1-2), Horak, A., Smrz, P., New Features of Wordnet Editor VisDic, Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2), 2004.

Howe, J. (2006). The Rise of Crowdsourcing. Преузето January 10, 2014 са http://www.wired.com/wired/archive/14.06/crowds.html

Howe, J. (2008). *Crowdsourcing: How the Power of the Crowd is Driving the Future of Business.* London: Random House Business Books.

Ipeirotis, P., Provost, F., & Wang, J. (2010). Quality management on Amazon Mechanical Turk. *Proceedings of the ACM SIGKDD Workshop on Human Computation*, (стр. 64-67).

Jurafsky, D., & Martin, H. J. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd изд.). NJ, USA: Pearson Education International.

Kazai, G. (2011). In Search of Quality in Crowdsourcing for Search Engine Evaluation. *Proceedings of the 33rd Proceedings of the European conference on Advances in Information retrieval*, (стр. 165-176.).

Kittur, A., Chi, E., & Suh, B. (2008). Crowdsourcing User Studies Using Mechanical Turk. *Human Computer Interaction* (стр. 453-456). ACM.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., . . . Horton, J. (.). The future of crowd work. *Proceedings of the 2013 conference on computer supported cooperative work* (стр. 1301-1318). San Antonio, Texas: ACM.

Koeva, S., Krstev, C., & Vitas, D. (2008). Morpho-semantic Relations in WordNet. A Case Study for two Slavic Languages. *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, (стр. 239–253).

Kolte, S. G., & Bhirud, G. (2009). Exploiting Links in Wordnet Hierarchy for Word Sense Disambiguation of Nouns. *Proceedings of the International Conference on Advances in Computing, Communication and Control, ICAC3 '09.*

Krauwer, S. (1998). *The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap.* Utrecht: Utrecht Institute of Linguistics.

Kripendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability.* University of Pensilvania.

Krstev, C. (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije.* Beograd: Univerzitet u Beogradu, Matematički fakultet.

Krstev, C. (2006). Specifični koncepti Balkana u semantičkoj mreži Wordnet. У L. S. al (Ур.), *Zbornik radova "Susreti kultura"* (стр. 275-285). Novi Sad: Univerzitet u Novom Sadu, Filozofski fakultet.

Krstev, C. (2008). *Processing of Serbian - Automata, Texts and Electronic Dictionaries.* Belgrade: Faculty of Philology.

Krstev, C., & Vitas, D. (2005). Corpus and Lexicon - Mutual Incompletness. *Proceedings of the Corpus Linguistics Conference.* Birmingham.

Krstev, C., & Vitas, D. (2005). Corpus and Lexicon - Mutual Incompletness. У e. P. Wagenmakers (Ур.), *Proceedings of the Corpus Linguistics Conference.* Birmingham.

Krstev, C., & Vitas, D. (2009). An Aligned English-Serbian Corpus. У e. N. Vujić (Ур.), *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality). Volume I*, стр. 495-508. Belgrade: Faculty of Philology.

Krstev, C., Obradović, I., & Vitas, D. (2006). Developing Balkan specific concepts within BalkaNet - a multilingual database of semantic networks. У S. K. Dimitrova-Vulchanova (Ур.), *Proceedings of the 5th International Conference Formal Approaches to South Slavic and Balkan Languages, FASSBL* (стр. 94-98). Sofia, Bulgaria: Institute of Bulgarian Language.

Krstev, C., Pavlović-Lažetić, G., & Vitas, D. (2004). Using Textual and Lexical Resources in Developing Serbian Wordnet. *Romanian Journal of Information Science and Technology, 7*(1-2), 147-161.

Krstev, C., Pavlović-Lažetić, G., Obradović, I., & Vitas, D. (2003). Corpora Issues in Validation of Serbian Wordnet. У P. M. eds. Vaclav Matoušek (Ур.), *Proceedings of the 6th International Conference TSD 2003 : Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence*, стр. 132-137. Springer, Berlin.

Krstev, C., Stanković, R., Obradović, I., Vitas, D., & Utvić, M. (2010). Automatic Construction of a Morphological Dictionary of Multi-Word Units. *7th International Conference on NLP, IceTAL*. Reykjavik.

Krstev, C., Stanković, R., Vitas, D., & Obradović, I. (2006). WS4LR: A Workstation for Lexical Resources. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, (стр. 1692-1697). Genoa, Italy.

Krstev, C., Vitas , D., & Erjavec, T. (2004). MULTEXT-East Resources for Serbian. У J. Z. Tomaž Erjavec (Ур.), *Zbornik 7. mednarodne multikonference "Informacijska druzba IS 2004", Jezikovne tehnologije.* Ljubljana, Slovenija.

Krstev, C., Vitas , D., & Pavlović-Lažetić, G. (2003). Resources and methods in the morphosyntactic processing of Serbo-CroatianI. У G. Z. (eds.) (Ур.), *Formal Description of Slavic Languages: The Fifth Conference*, (стр. 3-11). Leipzig.

Krstev, C., Vitas, D., & Trtovac, A. (2011). Orwell's 1984 – the Case of Serbian Revisited. *Proceedings of 5th Language & Technology Conference*, (стр. Cvetana Krstev, Duško Vitas, Aleksandra Trtovac, "Orwell's 1984 – the Case of Serbian Revisited", in Proceedings of 5t570-574). Poznań.

Krstev, C., Vitas, D., Obradović, I., & Utvić, M. (2011). E-Dictionaries and Finite-State Automata for the Recognition of Named Entities. *9th International Workshop on Finite State Methods and Natural Language Processing.* Blois, France.

Krstev, C., Zečević, A., Vitas, D., & Kyriakopoulou, T. (2013). NeRosetta -- an Insight into Named Entity Tagging. *6th Language and Technology Conference*, (стр. 168-172). Poznań.

Kunze, C., & Lemnitzer, L. (2002). GermaNet – representation, visualization, application. *Proceedings of LREC 2002, main conference.*

Kuti, J., Varasdi , K., Gyarmati, A., & Vajda, P. (2008). Language Independent and Language Dependent Innovations in the Hungarian WordNet. *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, (стр. 254–269).

Lakoff, G. (1990). Some Empirical Results About The Nature Of Concepts. *Mind and Language*.

Lakoff, G. (2012, June 25). *Metaphor and Health Care: On The Power to Make Metaphor Into Law*. Преузето са Huffington Post: http://www.huffingtonpost.com/george-lakoff/health-care-ruling_b_1623753.html

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago University Press.

Laporte, E. (2003). The RELEX Network. Преузето са http://infolingu.univ-mlv.fr/Relex/Relex.htm

Lohk, A. (2015). *A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries.* Tallinn: Tallinn University of Technology.

Lombard, M., Snyder-Duchand, J., & Campanella Bracken, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 587–604.

Maggetti, M. (2013). Regulation in Practice: The defacto Independence of Regulatory. *Swiss Political Science Review, 19*(1), 111-113.

Magnini, B., & Cavaglià. , G. (2000). Integrating Subject Field Codes into WordNet. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC).* Athens, Greece: European Language Resources Association (ELRA).

Maji, S. (2011). *Large Scale Image Annotations on Amazon Mechanical Turk.* University of California at Berkley.

Manning, D. C., & Schütze, H. (1999). *Foundations of statistical language processing.* MIT Press.

Marcus, G. K. (1994). The Penn Treebank: annotating predicate argument structure. *Workshop on Human Language Technologies*, (стр. 114-119).

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., & Mendes, S. (2006). WordNet.PT new directions. *Proceedings of the 3rd International Global Wordnet Conference (GWC2006)*, (стр. 319–321).

Maslow , A. H. (1943). A Theory of Human Motivation. *Psychological review* (50).

Mason, Z. J. (2004). CorMet: A Computational, Corpus-based Conventional Metaphor Extraction System. *Computational Linguistics, 30*(1), 23-44.

Mastropavlos, N., & Papavasileiou, V. (2011). Automatic Acquisition of Bilingual Language Resources. *Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece*. Komotini, Greece.

Maurel, D. (2008). Prolexbase: A Multilingual Relational Lexical Database of Proper. *Language Resources and Evaluation LREC* , (стр. 334-338). Marrakech, Morocco.

Maziarz, M., Szpakowicz, S., & Pia, M. (2012). Semantic Relations among Adjectives in Polish WordNet 2.0: A New Relation Set, Discussion and Evaluation. *Cognitive Studies / ´Etudes Cognitives, 12*, 149–179.

Mendes, S. (2006). Adjectives in WordNet. *Proceedings of the 3rd International Global Wordnet Conference (GWC2006)*, (стр. 225–230).

Mihalcea, R. (1998). *SEMCOR semantically tagged corpus. Unpublished manuscript.*

Miller, A. G., Becwith, R., Fellbaum, C., Gross, D., & Miller, J. K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235-312.

Miller, G. (1998). *Nouns in WordNet in WordNet: An Electronic Lexical Database.* MIT Press.

Miller, G. A. (1979). Images and models, similes and metaphors. У A. Ortony (Ур.), *Metaphor and Thought.* Cambridge University Press.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of ACM, 38*(11), 39-41.

Miller, G. A. (1995). WordNet: A Lexical Database for English., *38*, стр. 39-41.

Miller, G. A. (1998). *Nouns in WordNet: An Electronic Lexical Database.* MIT Press.

Miller, G. R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Five Papers on WordNet. *Special Issue in International Journal of Lexicography, 3*(4).

Mitrović, J. (2014, March 10-11). Mitrović, J. (2014). Electronic Tools and Resources for Multi-Word Unit Detection and Research in serbian. *The 2nd PARSEME General Meeting*. Athens, Greece.

Mitrović, J., Markantonatou, S., Mladenović, M., & Krstev, C. (2016). Greek and Serbian Similes – Enrichment of Lexical Resources via Crowdsourcing, poster presentation. *PARSEME 6th general meeting.* Dubrovnik, Croatia.

Mitrović, J., Markantonatou, S., Mladenović, M., & Krstev, C. (2018). A Cross-linguistic study on Greek and Serbian MWEs and Enrichment of lexical resources using the crowdsourcing model. У *Volume on MWEs in Greek and other languages: from theory to implementation.* Academy of Athens, Research Center for Neologisms and Scientific Terms.

Mitrović, J., Mladenović, M., & Krstev, C. (2015, Septemner 23-24). Adding MWEs to Serbian Lexical Resources Using Crowdsourcing, Poster presentation. *the 5th PARSEME General Meeting*. Iasi, Romania.

Mitrović, J., O'Reilly, C., Mladenović, M., & Handschuh, S. (2017). Ontological Representations of Rhetorical Figures for Argument Mining. (R. A. Marco, Ур.) *Argument and Computation, 8*(3), Ontological Representations of Rhetorical Figures for Argument Mining.

Mladenović, M. (2016). Ontološko prepoznavanje retoričkih figura. *Časopis za digitalnu humanistiku , 16(1-2).*

Mladenović, M., & Mitrović, J. (2013). Ontology of Rhetorical Figures for Serbian. *Lecture Notes in Computer Science and Artificial Intelligence. 8082*, стр. 383-393. Springer-Verlag Berlin Heilderberg.

Mladenović, M., & Mitrović, J. (2014). Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool. У С. К. Gordana Pavlović-Lažetić (Ур.). Belgrade: University of Belgrade, Faculty of Mathematics.

Mladenović, M., Krstev, C., Mitrović, J., & Stanković, R. (2017). Using Lexical Resources for Irony and Sarcasm Classification. *Miljana Mladenović, Cvetana Krstev, Jelena Mitrović, Ranka Stanković, "Using Lexical Resources for Irony and Sarcasm Classification". In Proceedings of the Proceedings of the 8th Balkan Conference in Informatics (BCI '17). Article 13, 8 pages. 2017.* New York, NY, USA: ACM.

Mladenović, M., Mitrović, J., & Krstev, C. (2014). Developing and Maintaining a WordNet: Procedures and Tools. *Proceedings of the Global WordNet Conference*, (стр. 55-62). Tartu, Estonia.

Mladenović, M., Mitrović, J., & Krstev, C. (2016). A Language-Independent Model for Introducing a New Semantic Realation between Adjectives and Nouns in a WordNet. *Proceedings of the 8th Global WordNet Conference* (стр. 218-225). Bucharest: Romanian Academy Research Institute for Artificial Intelligence.

Mladenović, M., Mitrović, J., Krstev, C., & Vitas, D. (2015). Hybrid Sentiment Analysis Framework for a Morphologically Rich Language. *Journal of Intelligent Information Systems, 46*(3), 599-620.

Mladenović, M., Stanković, R., & Krstev, C. (2017). A WordNet Ontology in Improving Searches of Digital Dialect Dictionary . *New Trends in Databases and Information Systems: ADBIS 2017 Short Papers and Workshops - SW4CH (Semantic Web for Cultural Heritage. 767*, стр. 373-383. Nicosia, Cyprus: Springer International Publishing.

Munro, R., Bethard, S., Kuperman, V., Tzuyin, L., Melnick, R., Potts, C., . . . Tily, H. (2010). Crowdsourcing and Language Studies: The New Generation of Linguistic Data. *Human Language Technologies: Conference of the North American Chapter of the Association for Computational Linguisticsc, Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.*

Murugesan, S. (2007). *Understanding web 2.0.* Преузето 2014 са http://hcotuk.etu.edu.tr/bil554/Understanding_web_2.pdf

O'Donnell, J. (2002). *John Harrison and the Longitude Problem*. Преузето са National Maritime Museum website: http://www.nmm.ac.uk/harrison

Obradović, I., & Stanković, R. (2008). Obradović, I. Softverski alati za korišćenje jezičkih resursa za srpski jezik. *Infoteka, IX*(1-2).

Obradović, I., Krstev, C., Pavlović-Lažetić, G., & Vitas, D. (2004). Corpus Based Validation of WordNet Using Frequency Parameters", in Proceedings of the GWC : Second International WordNet Conference, Brno, Czech Republic, January 20-23, 2004, eds. P. У K. P. Peter Sojka (Ур.), *Proceedings of the GWC: Second International WordNet Conference* (стр. 181-186). Brno, Czech Republic: Masaryk University.

Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., & Biewald, L. (2011). Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Proceedings of the Third Association for the Advancement of Artificial Intelligence.*

Paumier, S. (2011). *Unitex - Manuel d'utilisation.*

Paumier, S., Nakamura, T., & Voyatzi, S. (2009). UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources. *In eLexicography in the 21st century: new challenges, new applications (eLEX'09)*, (стр. 173–175).

Pazienza, M. T., Stellato , A., & Tudorache, A. (2008). A Bottom-up Comparative Study of EuroWordNet and WordNet 3.0 Lexical and Semantic Relations. *Proceedings of the*

Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA).

Pease, A. (2011). *Ontology: A practical Guide.* Angwin, CA: Articulate software Press.

Pianta, E., Bentivogli, L., & Girar, C. (2002). MultiWordNet: Developing an aligned multilingual database. *Proceedings of the 1st International WordNet Conference*, (стр. 293-302). Mysore, India.

Piasecki, M., Szpakowicz, S., & Broda, B. (2009). *A Wordnet from the Ground Up.* Oficyna Wydawnicza Politechniki Wrocławskiej.

Popović, L., & Vitas, D. (2003). Konspekt za izgradnju referentnog korpusa standardnog srpskog jezika. *Zbornik Naučni sastanak slavista u Vukove dane*, (стр. 221-227).

Proctor, N. (2013). Crowdsourcing – an Introduction: From Public Goods to Public Good. *The Museum Journal* , 105–106.

Quinn, A. J., & Bederson, B. B. (2009). *A taxonomy of distributed human computation.* Technical report, University of Maryland.

Richards, I. A. (1965). *The Philosophy of Rhetoric.* New York: OUP.

Ruan, S., Di Marco , C., & Harris, R. A. (2016). Rhetorical figure annotation with XML. *REF RUan S. Ruan, C. Di Marco and R.A. Harris, Rhetorical figure annotation with XML, in: Computational Models of NaturalArgumentation (CMNA) 16, a Workshop at the 2016 International Joint Conference on Artificial Intelligence (IJCAI).* New York.

Ruan, S., Di Marco, C., & Harris, R. A. (2016). Rhetorical Figure Annotation with XML. *Computational Models of Natural Argumentation (CMNA)16, A Workshop at the 2016 International Joint Conference on Artificial Intelligence (IJCAI).*

Samuel, A. (1967). Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress. *IBM Journal of Research and Development, 11*(6), 601- 617.

Sarasua, C., Simperl, E., & Noy, N. F. (2012). Crowdmap: Crowdsourcing Ontology Alignment with Microtasks. *11th International Semantic Web Conference.* Boston.

Savage, N. (2012). Gaining Wisdom from Crowds. *Communications of ACM, 55*(3), 13-15.

Savage, N. (2012). Gaining Wisdom from Crowds. *Communications of the ACM, 55*(3), 13-15.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., . . . Sangati, F. (2015). PARSEME – PARSing and Multiword Expressions within a European multilingual network. *In the Proceedings of the 7th Language & Technology Conference (LTC 2015).* Poznań, Poland.

Saxton, G. D., Oh, O., & Kishore, R. (2013). Rules of Crowdsourcing: Models, Issues, and Systems of Control. *Information Systems Management, 30*, 2-20.

Schaer, P. (2012). Better than their reputation? on the reliability of relevance assessments with students. *Third international conference on Information Access Evaluation: multilinguality, multimodality, and visual analytics (CLEF'12)* (стр. 124-135). Berlin-Heidelberg: Springer-Verlag.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.* Manchester, UK.

Schreibman, S., Siemens, R., & Unsworth, John, J. (Уредници). (2004). *A Companion to Digital Humanities*. Преузето 2015 са http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1&toc.depth=1&toc.id=ss1-2-1&brand=default

Ševa , N., & Kostić , A. (2003). Annotated corpus and the empirical evaluation of probability estimates of grammatical forms. *Psihologija, 36*(3), 255-270.

Shutova, E., Teufel , S., & Korhonen, A. (2013). Statistical Metaphor Processing. *Computational Linguistics, 39*(2), 301-353.

Silberztein , M. D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris.

Simperl, E. (2015). How to Use Crowdsourcing Effectively: Guidelines and Examples. *LIBER Quarterly*, 18–39.

Siorpaes , K., & Martin , H. (2008). Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems, 23*(3), 50-60.

Škaljić, A. (1989). *Turcizmi u srpskohrvatskom jeziku.* Sarajevo: Svijetlost.

Solar, M. (2005). *Teorija književnosti* (T. XX izdanje). Zagreb: Školska knjiga.

Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Cristea, D., Tufis, D., . . . Grigoriadou, M. (2002). Balkanet: A Multilingual Semantic Network for Balkan Languages. *1st International Global WordNet Conference.* . "Balkanet: A Multilingual Semantic Network for Balkan Languages" Paper presented at the 1st International Global WordNet Conference, Mysore, India, January 21-25 2002.

Stanković, R. (2009). *Modeli ekspanzije upita nad tekstuelnim resursima, doktorska disertacija.* Boegrad: Matamatički fakultet.

Stanković, R., Krstev, C., Obradović, I., & Utvić, M. (2012). A tool for enhanced search of multilingual digital libraries of e-journals. *8th International Conference on Language Resources and Evaluation LREC.* Istanbul, Turkey.

Stanković, R., Krstev, C., Vitas, D., Vulović, N., & Kitanović, O. (2017). Keyword-Based Search on Bilingual Digital Libraries. У D. G. Andrea Calì (Ур.), *Semantic Keyword-Based Search on Structured Data Sources - Second COST Action IC1302 International KEYSTONE Conference. Revised selected Papers* (стр. 112-123). Cluj-Napoca: Springer, LNCS 10151. doi:10.1007/978-3-319-53640-8_10

Stanković, R., Obradović, I., & Trtovac , A. (2012). An Approach to Development of Bilingual Language Resources. *Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages.* Novi Sad.

Stanković, R., Obradović, I., Krstev, C., & Vitas, D. (2011). Production of Morphological Dictionaries of Multi-Word Units Using a Multipurpose Tool. У P. W. K. Jassem (Ур.), *Proceedings of the Computational Linguistics-Applications Conference* (стр. 77-84). Jachranka, Poland: Polish Information Processing Society.

Stanković, R., Vitas, D., & Krstev, C. (2007). The Nooj System as Module within an Integrated Language Processing Enironment. У X. Blanco, & M. Silberztein (Ур.), *Proceedings of the 2007 International Nooj Conference* (стр. 228-248). Cambridge Scholars Publishing.

Stevens, S. S. (1946). On the Theory of Scales of Measures. *Science, New Series, 103*(2684), 677-680.

Stuart, T. (2011). *9 Examples of Crowdsourcing, Before 'Crowdsourcing' Existed*. Преузето 4.3.2016. са MemeBurn: http://memeburn.com/2011/09/9-examples-of-crowdsourcing-before-%E2%80%98crowdsourcing%E2%80%99-existed/

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics – Simulation and Computation, 5*(1), 55-64.

Tufis, D., Cristea, D., & Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology, 7*(1-2), 9-43.

Utvić, M. (2011). Annotating the Corpus of Contemporary Serbian. *INFOtheca, 12*(2), 36a-47a.

Utvić, M. (2013). *Izgradnja referentnog korpusa savremenog srpskog jezika, Doktorska disertacija.* Beograd: Filiološki fakultet.

Utvić, M. (2014). Liste učestanosti Korpusa savremenog srpskog jezika. *Naučni sastanak slavista u Vukove dane* (стр. 241-262). Filološki fakultet, Univerzitet u Beogradu.

Utvić, M., Stanković, R., & Obradović, I. (2008). Integrisano okruženje za pripremu paralelizovanog korpusa. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, 563-578.

Veale , T., & Hao, Y. (2008). Enriching WordNet with folk knowledge and stereotypes. *The 4th International Global Wordnet Conference (GWC2008)*, (стр. 453–461).

Veale, T., & Hao, Y. (2012). In the Mood for Affective Search. *Proceedings of WWW'2012, the 21st World-Wide-Web conference.* Lyon, France.

Veale, T., & Hao, Y. (2013). Talking about Similes and Stereotypes: Subjective Talking Points. У T. Veale, K. Feyaerts, & C. Forceville (Уредници), *Creativity and the Agile Mind: A Multi-Disciplinary Study of a Multi-Faceted Phenomenon in Applications of Cognitive Linguistics* (T. 21). De Gruyter Mouton.

Vesselinov, R., & Grego, J. (2012). *Duolingo Effectivness Report.* Преузето 2014 са http://tcrn.ch/VshbIC

Vitas , D., & Krstev, C. (2005). Derivational Morphology in an E-Dictionary of Serbian. (Z. Vetulani, Ур.) *Proceedings of 2nd Language & Technology Conference*, 139-143.

Vitas , D., & Krstev, C. (2012). Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne, LXIII*, 279-292.

Vitas, D. (1979). Prikaz jednog sistema za automatsku obradu teksta. *Informatica* , 7-11.

Vitas, D. (1981). Generisanje imenickih oblika u srpskohrvatskom jeziku. *Informatica*(3), 49-55.

Vitas, D., & Krstev, C. (1998). *Electronic Edition of Serbian Translation of Orwell's 1984 Aligned with 7 languages.* (L. a. Erjavec, Ур.)

Vitas, D., & Krstev, C. (2006). Literature and Aligned Texts, in Readings in Multilinguality. У G. A. eds. Milena Slavcheva (Ур.), *Duško Vitas, Cvetana Krstev, "Literature and Aligned Texts", in Readings in Multilinguality, eds. Milena Slavcheva, Galia Angelova and Kiril Simov, pp. 148-155,* (стр. 148-155). Sofia, Bulgaria: Institute for Parallel Processing, Bulgarian Academy of Sciences.

Vitas, D., Koeva, S., Krstev, C., & Obradović, I. (2008). Tour du monde through the dictionaries. *Actes du 27eme Colloque International sur le Lexique et la Gammaire.* L'Aquila.

Vitas, D., Krstev, C., Obradović, I., Popović, L., & Pavlović-Lažetić, G. (2003). An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts. У V. K. S.

Piperidis (Ур.), *Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki, Gr*, (стр. 97- 104). Thessaloniki, Greece.

Vitas, D., Krstev, C., Pavlović-Lažetić, G., & Nenadić, G. (1998). Recent Results in Serbian Computational Lexicography. У N. Bokan (Ур.), *Proceedings of the Symposium "Contemporary Mathematics"*, (стр. 113-130). Belgrade: Faculty of Mathematics.

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *SIGCHI Conference on Human Factors in Computing Systems*, (стр. 319-326).

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*, стр. 58-67.

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: a game for collecting common-sense facts. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)* (стр. 75-78). New York, USA: ACL.

Vossen , P., Bloksma, L., Rodriguez , H., Climent, S., Calzolari, N., Roventini, A., . . . Peters, W. (1997). *The EuroWordNet Base Concepts and Top Ontology, LE-4003, Deliverable D017, D034, D036.* Amsterdam: University of Amsterdam.

Vossen, P. (1998). *Eurowordnet: A Multilingual Database with Lexical Semantic Networks.* Norwell, MA, USA: Kluwer Academic Publishers.

Vossen, P. (1999). *EuroWordNet General Document. EuroWordNet,Final Document Deliverable D032D033/2D014.* Amsterdam: University of Amsterdam, The Netherlands.

Vossen, P., Bloksma , L., Rodriguez H., H., Climent, S. C., Calzolari, N., Roventini, A., . . . Peters, W. (1997). *The EuroWordNet Base Concepts and Top Ontology, LE-4003, Deliverable D017, D034, D036.* University of Amsterdam.

Vuković, M., & Bartolini, C. (2010). Towards a Research Agenda for Enterprise Crowdsourcing. Leveraging Applications of Formal Methods, Verification, and Validation. *4th International Symposium on Leveraging Applications.* Heraklion, Crete.

Vukovic, M., Kumara, S., & Greenshpan, O. (2010). Ubiquitous Crowdsourcing.

Wang, A., Hoang, C. D., & Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources $ Evaluation, 47*, стр. 9-31.

Yuen, M., Chen, L. J., & King, I. (2009). A survey of human computation systems. *IEEE International Conference on Computational Science and Engineering*, (стр. 723-728).

Zarrouk, M., Lafourcade, M., & Joubert, A. (2013). Inductive and deductive inferences in a Crowdsourced Lexical-Semantic Network. *Proceedings of the International*

Conference Recent Advances in Natural Language Processing RANLP 2013, (стр. 740–746). Hissar.

Аристотел. (1987). *Аристотел: Реторика 1/2/3 (превео Марко Вишић)* (Т. 40). Београд: Независна издања .

Белић, А. (1998). *О језичкој природи и језичком развитку* (Изабрана дела изд., Т. први том). Београд: Завод за уџбенике и наставна средства.

Витас, Д., Поповић, Љ., Крстев, Ц., Обрадовић, И., Павловић-Лажетић, Г., & Станојевић, М. (2012). *Српски језик у дигиталном добу, Серија белих књига.* (H. U. Georg Rehm, Ур.) Српски језик у дигиталном добу, Душко Витас, Љубомир Поповић, Цветана Крстев, Иван Обрадовић, Гордана Павловић-Лажетић, Младен Станојевић, Серија белих књига, 2012, Georg Rehm, Hans Uszkoreit (уредници, editors) Springer: Springer.

Вуловић, Н. (2015). *Српска фразеологија и религија* (Монографија 23 изд.). Београд: Лингвокултуролошка истраживања, Институт за српски језик САНУ.

Голяк, С. (2009). Устаљена поређења у српском и белоруском језику са аспекта идиоматичности. *Научни састанак слависта у Вукове дане 38/1.* Београд: Филолошки факултет.

Грицкат, И. (1967). Стилске фигуре у светлу језичких анализа. *Наш језик, 51*(4), 117-135.

Драгићевић, Р. (2010). *Вербалне асоцијације кроз српски језик и културу.* Београд: Друштво за српски језик и књижевност Србије.

Ињац-Малбаша, В. (2012/2013). Crowdsourcing – одређење појма, типологија и сродни термини. *Гласник Народне библиотеке Србије*, 239-255.

Квинтилијан, М. Ф. (1985). *Образовање говорника.* Сарајево: Веселин Маслеша.

Ковачевић , М. (1998). *Стилске фигуре и књижевни текст.* Београд: Требник.

Крстев, Ц., Ђорђевић, Б., Антонић, С., Ивковић-Берчек, Н., Зорица, З., Црногорац, В., & Мацура, Љ. (2008). Кооперативан рад на доградњи Српског Wordneta. *Infotheca, IX*(1-2), 57-75.

Митровић, Ј. (2015). Модел групне расподеле рада у дигиталној хуманистици. *Дигитална хуманистика.* Београд: Филолошки факултет.

Младеновић, М. (2016). *Информатички модели у анализи осећања засновани на језичким ресурсима.* Београд.

Младеновић, М. (2016, август). Онтолошко препознавање реторичких фигура. *Инфотека, 16*(1-2).

Мршевић-Радовић, Д. (1982). Фразеолошка је диница и њен синоним. *Зборник радова, НССУВД , 12*(1), 123–130.

Солар , М. (1981). *Теорија књижевности.* Загреб: Школска књига.

Телија, В. Н. (1996). *Русская фразеология, Семантический, прагматический и лингвокультурологический аспекты.* Москва: Языки русской кул

Тртовац, С. А. (2016). *Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама, докторска дисертација.* Београд: филолошки факултет, Универзитет у Београду.

Утвић, М. (2011). Анотација Корпуса савременог српског језика. *ИНФОтека, 12*(2), 39-51.

8 Прилози

Прилог 1

Списак 25 јединствених почетних синсетова (unique beginners) за именице у Ворднету

1. {act, action, activity}
2. {animal, fauna}
3. {artifact}
4. {attribute, property}
5. {body, corpus}
6. {cognition, knowledge}
7. {communication}
8. {event, happening}
9. {feeling, emotion}
10. {food}
11. {group, collection}
12. {location, place}
13. {motive}
14. {natural object}
15. {natural phenomenon}
16. {person, human being}
17. {plant, flora}
18. {possession}
19. {process}
20. {quantity, amount}
21. {relation}
22. {shape}
23. {state, condition}
24. {substance}
25. {time}

Прилог 2

Електронски морфолошки речници српског језика

Табела 28Електронски морфолошки речник српског језика – просте речи[153]

Просте речи			
	леме	**облици**	**однос**
Српски	93.817	3,968.57	42,30
Властита	10.862	270.984	24,95
Лична	20.886	283.663	13,58
Страна	7.796	58.440	7,50
Укупно	133.361	4.581.657	34,36
Префикси	114	114	1,00
Мере у кувању	95	1572	16,55
Стандардне мере	213	1936	9,09
Светске валуте	277	1946	7,03

Табела 29Електронски морфолошки речник српског језика – вишечлане речи[154]

Вишечлане речи			
	леме	**облици**	**однос**
Српски	10.081	222.735	22,09
Властита	1.653	15.627	9,45
Непром	608	608	1,00
Abb-new	82	711	8,67
Укупно	12.424	239.681	19,29
Мере у кувању	9	150	16,67
BI-специфично	1.284	22.855	17,80
Укупно	13.717	262.686	19,15

[153] Величина речника дана 10.12.2014. године
[154] Величина речника дана 10.12.2014. године.

```
bridža,bridž.N:ms2q
brifingu,brifing.N:ms3q
briga,.N:fp2q
briga,.N:fs1q
briga,brigati.V+Imperf+It+Ref:Ays
briga,brigati.V+Imperf+It+Ref:Pzs
brigade,brigada.N:fw2q
brigade,brigada.N:fp5q
brigade,brigada.N:fs2q
brige,briga.N:fw2q
brige,briga.N:fw4q
brige,briga.N:fp1q
briljantan,.A:akms1g
briljantan,.A:akms4q
briljira,briljirati.V+Imperf+Perf+Tr+Iref:Ays
briljira,briljirati.V+Imperf+Perf+Tr+Iref:Azs
briljira,briljirati.V+Imperf+Perf+Tr+Iref:Pzs
briljirao,briljirati.V+Imperf+Perf+Tr+Iref:Gsm
brine,brinuti.V+Imperf+It+Ref+Iref:Pzs
brinemo,brinuti.V+Imperf+It+Ref+Iref:Pxp
brinu,brinuti.V+Imperf+It+Ref+Iref:Ays
brinu,brinuti.V+Imperf+It+Ref+Iref:Azs
brinu,brinuti.V+Imperf+It+Ref+Iref:Pzp
brinula,brinuti.V+Imperf+It+Ref+Iref:Gwm
brinula,brinuti.V+Imperf+It+Ref+Iref:Gwn
brinula,brinuti.V+Imperf+It+Ref+Iref:Gpn
brinula,brinuti.V+Imperf+It+Ref+Iref:Gsf
brinuti,.V+Imperf+It+Ref+Iref:W
brisa,bris.N:mw2q
brisa,bris.N:mw4q
brisa,bris.N:ms2q
brisa,brisati.V+Imperf+Tr+It+Iref:Ays
brisa,brisati.V+Imperf+Tr+It+Iref:Azs
Brisel,.N+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brussel:ms1q
Brisel,.N+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brussel:ms4q
Brisela,Brisel.N+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brussel:ms2q
briselske,briselski.A+PosQ+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brusse
l:aefw2g
briselske,briselski.A+PosQ+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brusse
l:aefw4g
briselske,briselski.A+PosQ+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brusse
l:aefp1g
Briselu,Brisel.N+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brussel:ms3q
Briselu,Brisel.N+NProp+Top+Gr+CC2=BE+Val=Bruxelles+Val=Brussel:ms7q
Britanac,.N+Hum+NProp+Top+Inh+CC2=UK:ms1v
Britanca,Britanac.N+Hum+NProp+Top+Inh+CC2=UK:mw2v
Britanca,Britanac.N+Hum+NProp+Top+Inh+CC2=UK:mw4v
Britanca,Britanac.N+Hum+NProp+Top+Inh+CC2=UK:ms2v
Britanca,Britanac.N+Hum+NProp+Top+Inh+CC2=UK:ms4v
Britanija,.N+NProp+Top+Reg:fs1q
Britanija,Britani.N+NProp+Hum+Last+EN+Val=Brittany:ms2v
```

Слика 35 Извод из речника DELAF простих речи (речник облика са граматичким категоријама)

```
a da,.CONJ+Comp
a kamoli,.CONJ+Comp
a ono,.CONJ+Comp
a to,.CONJ+Comp
Adis Abebi,Adis Abeba.N+Comp+NProp+Top+Gr+CC2=ET:fs3q
Adis Abebi,Adis Abeba.N+Comp+NProp+Top+Gr+CC2=ET:fs7q
administrativne postupke,administrativni postupak.N+Comp:mp4q
ADSL-u,ADSL.ABB+Acr+Noun+D:m3sq:m7sq
advokatska kancelarija,.N+Loc+Comp:fs1q
agencije za nacionalnu bezbednost,Agencija za nacionalnu
bezbednost.N+NProp+Org+CC2=RS+Comp:fs2q
agencije za privredne registre,Agencija za privredne
registre.N+NProp+Org+CC2=RS+Comp:fs2q
akcionarska društva,akcionarsko društvo.N+Org+Comp:nw2q
akcionarske skupštine,akcionarska skupština.N+Comp:fs2q
akcionarsko društvo,.N+Org+Comp:ns1q
akcionarskoj skupštini,akcionarska skupština.N+Comp:fs3q
ako i,.CONJ+Comp
Al Kaide,Al Kaida.N+Comp+NProp+Org:fs2q
Aleksa Šantic,.N+Comp+NProp+Top+Gr+CC2=RS:ms1v
alkoholna pica,alkoholno
pice.N+Comp+Conc+Drink+Food+Prod+DOM=Culinary:nw2q
ambulantnim kolima,ambulantna kola.N+Conc+Comp:np3q
americki kongres,.N+NProp+Org+Comp:ms1q
americkog dolara,americki
dolar.N+Cur+CC2=US+CC2=EC+CC2=SV+CC2=HT+CC2=MH+CC2=FM+CC2=PW+CC2=PR+CC2
=TR+ISO=USD+Comp:ms2q
americkoj saveznoj državi,americka savezna država.N+Top+Dr+US+Comp:fs3q
americkoj saveznoj državi,americka savezna država.N+Top+Dr+US+Comp:fs7q
anarho-sindikalisti,anarho-sindikalist.N+Hum+Cr+Comp:mp1v
aneks ugovora,.N+Comp:ms1q
arhijerejski sabor,.N+Org+Comp:ms4q
arhijerejski sinod,.N+Org+Comp:ms1q
arhijerejskog sabora,arhijerejski sabor.N+Org+Comp:ms2q
arhijerejskog sinoda,arhijerejski sinod.N+Org+Comp:ms2q
atletski savez Srbije,Atletski savez
Srbije.N+NProp+Org+CC2=RS+DOM=Sport+Comp:ms1q
atomsku bombu,atomska bomba.N+Comp+Conc:fs4q
audio-vizuelne,audio-vizuelni.A+Comp+PosQ:aefw2g
audio-vizuelni,.A+Comp+PosQ:aems5g
auto-delova,auto-delovi.N+Comp+Conc+Coll:mp2q
auto-delove,auto-delovi.N+Comp+Conc+Coll:mp4q
auto-industrijama,auto-industrija.N+Comp:fp3q
auto-moto saveza,auto-moto savez.N+Comp:mw2q
autobuska linija,.N+Comp:fs1q
autobuskoj stanici,autobuska stanica.N+Comp:fs3q
automatske puške,automatska puška.N+Conc+Comp:fp5q
automatske puške,automatska puška.N+Conc+Comp:fs2q
automatskog oružja,automatsko oružje.N+Conc+DOM=Mil+Comp:ns2q
automobilski saobracaj,.N+Comp:ms1q
autonomne pokrajine,autonomna pokrajina.N+Top+Reg+Comp:fw2q
avio kompanije,avio-kompanija.N+Comp+Org:fw2q
```

Слика 36 Извод из речника DELACF вишечланих речи (речник облика са граматичким категоријама)

Табела 30Семантичке ознаке у SrpMed

Семантичка ознака	Пример
+Hum људско биће	девојка
+Being биће	утвара
+Zool животиња	крокодил
+Bot биљка	лешник
+Micro микроорганизам	бактерија
+Food храна	хлеб
+Org организација	министарство
+Char карактер	астериск
+Lng језик	српски
+Loc локација	раскршће
+Prof занимање	учитељ
+Zgrada грађевина	штала
+FoS област науке	тригонометрија
+Diagnosis дијагноза	гастритис

Прилог 3

Расподела синсетова у Српском ворднету

Табела 31 Расподела синсетова у Српском ворднету према врсти речи[155]

Укупно синсетова	**21877**
Именица (n)	17922
Глагола (v)	2209
Придева (a)	1622
Прилога (b)	129

Табела 32 Расподела синсетова у Српском ворднету према врсти речи[156]

Укупно синсетова	**22530**
Именица (n)	18248
Глагола (v)	2249
Придева (a)	1907
Прилога (b)	126

Табела 33 Раподела литерала у синсетовима Српског ворднета[157]

Број литерала	**Број синсетова**
1	11982
2	2042
3	6950
4	585
5	201
6	64
7	38
8	10
9	2
10	2
13	1

[155] Стање забележено 30.03.2107. године.
[156] Стање забележено 10.12. 2017. године
[157] Стање забележено 30.03.2017. године.

Табела 34 Расподела значења према врсти речи у Српском ворднету[158]

Именице	
Број значења	**Број синсетова**
1	1 9759
2	2 5769
3	3 1653
4	4 478
5	5 165
6	6 55
7	7 32
8	8
9	2
10	1
Глаголи	
1	945
2	821
3	315
4	89
5	27
6	6
7	3
8	1
10	1
13	1
Придеви	
1	1192
2	335
3	64
4	16
5	9

[158] Стање забележено 30.03.2017. године.

6	3
7	2
8	1
Прилози	
1	86
2	25
3	10
4	2
7	1

Табела 35 Расподела класа сентимената у Српском ворднету[159]

Позитивних	**Негативних**
1982	2070

Табела 36 Расподела класа сентимената Српском ворднету[160]

Позитивних	**Негативних**
2146	2247

[159] Стање забележено 30.03.2017. године.
[160] Стање забележено 10.12. 2017. године

Прилог 4

Везе у Српском ворднету

Табела 37 Врста и број лексичко-семантичких релација у Српском ворднету[161]

Врста релације	**Број**	**Врста релације**	**Број**	**Врста релације**	**Број**
also_see	19546 217	holo_portion	224	specificOf	224
attribute	19694 2	hypernym	20338	specifiedBy	221
be_in_state	288	hyponym	328	subevent	79
category_domain	1069	mero-portion	1967	substanceHolonym	12
causes	66	mero_member	9	substanceMeronym	7
derived-gender	38	near_antonym	1147	topicDomain	1
derived-pos	45	particle	10	topicDomainMember	107
derived-vn	3	partMeronym	36	Usage-domain	17
eng_derivative	2987	pertainym	11	Verb-group	182
entailment	1	region_domain	164	УКУПНО синсетова	
holo_member	3925	region_Domain	2	21877	
holo_part	1843	similar_to	223		

[161] Стање забележено 30.03.2017. године

Табела 38 Аутоматски додате везе[162]

Синсет	Литерал	Релација	Синсет	Литерал
ENG30-14635722-n	bakar	specifiedBy	ENG30-00381097-a	crven
ENG30-09827683-n	beba	specifiedBy	ENG30-00479330-a	ugodan
ENG30-02403325-n	bik	specifiedBy	ENG30-01170243-a	zdrav
ENG30-02795169-n	bure	specifiedBy	ENG30-02410393-a	debeo
ENG30-02897820-n	cigla	specifiedBy	ENG30-00381097-a	crven
ENG30-03234306-n	crtež	specifiedBy	ENG30-00217728-a	lep
ENG30-05084201-n	daljina	specifiedBy	ENG30-01433493-a	dugačak
ENG30-12164363-n	dinja	specifiedBy	ENG30-01170243-a	zdrav
ENG30-00831191-n	disanje	specifiedBy	ENG30-01580050-a	neophodan
ENG30-12946849-n	dren	specifiedBy	ENG30-01170243-a	zdrav
ENG30-03485997-n	drška	specifiedBy	ENG30-02410393-a	debeo
ENG30-00034213-n	fenomen	specifiedBy	ENG30-01343918-a	zanimljiv
ENG30-06685456-n	garancija	specifiedBy	ENG30-03151302-a	ostavljen
ENG30-00757080-n	greh	specifiedBy	ENG30-00217728-a	lep
ENG30-02879517-n	gudalo	specifiedBy	ENG30-02311544-a	kriv
ENG30-07679356-n	hleb	specifiedBy	ENG30-00828779-a	jestiv
ENG30-07679356-n	hleb	specifiedBy	ENG30-01170243-a	zdrav
ENG30-03027001-n	hronometar	specifiedBy	ENG30-01749320-a	savršen
ENG30-05923696-n	ideal	specifiedBy	ENG30-01114658-a	velik
ENG30-05833840-n	ideja	specifiedBy	ENG30-00323873-a	brz
ENG30-12731401-n	jablan	specifiedBy	ENG30-02310895-a	prav
ENG30-11501381-n	kiša	specifiedBy	ENG30-00323873-a	brz
ENG30-01747885-n	kobra	specifiedBy	ENG30-00323873-a	brz
ENG30-02200198-n	komarac	specifiedBy	ENG30-01391351-a	mali
ENG30-05399847-n	krv	specifiedBy	ENG30-00381097-a	crven
ENG30-02883344-n	kutija	specifiedBy	ENG30-01391351-a	mali

[162] Стање забележено 30.03..2017. године.

ENG30-02441942-n	lasica	specifiedBy	ENG30-00323873-a	brz
ENG30-04118776-n	lenjir	specifiedBy	ENG30-02310895-a	prav
ENG30-02274259-n	leptir	specifiedBy	ENG30-00854255-a	nežan
ENG30-07376257-n	lupa	specifiedBy	ENG30-01170243-a	zdrav
ENG30-01974773-n	ljuskar	specifiedBy	ENG30-00381097-a	crven
ENG30-12695572-n	mahagon	specifiedBy	ENG30-00381097-a	crven
ENG30-02390258-n	mazga	specifiedBy	ENG30-00225564-a	pakostan
ENG30-02131653-n	medved	specifiedBy	ENG30-01170243-a	zdrav
ENG30-07106800-n	metafora	specifiedBy	ENG30-00043765-a	stvaran
ENG30-10320863-n	ministar	specifiedBy	ENG30-01863066-a	privilegovan
ENG30-09426788-n	more	specifiedBy	ENG30-01135914-a	žučan
ENG30-02219486-n	mrav	specifiedBy	ENG30-02551380-a	suv
ENG30-02137549-n	mungos	specifiedBy	ENG30-00323873-a	brz
ENG30-03800563-n	muzej	specifiedBy	ENG30-01343918-a	zanimljiv
ENG30-09548632-n	nimfa	specifiedBy	ENG30-00217728-a	lep
ENG30-04359589-n	oslonac	specifiedBy	ENG30-00995775-a	povoljan
ENG30-02411705-n	ovca	specifiedBy	ENG30-01892953-a	krotak
ENG30-05567217-n	palac	specifiedBy	ENG30-02410393-a	debeo
ENG30-01846331-n	patka	specifiedBy	ENG30-01405047-a	jasan
ENG30-11928549-n	pelin	specifiedBy	ENG30-02396098-a	gorak
ENG30-10444194-n	pesnik	specifiedBy	ENG30-01343918-a	zanimljiv
ENG30-13129165-n	peteljka	specifiedBy	ENG30-02551380-a	suv
ENG30-11410625-n	posledica	specifiedBy	ENG30-01405047-a	jasan
ENG30-05566504-n	prst	specifiedBy	ENG30-02100987-a	osnovni
ENG30-07199583-n	pršuta	specifiedBy	ENG30-02551380-a	suv
ENG30-01503061-n	ptica	specifiedBy	ENG30-01061489-a	slobodan
ENG30-12620031-n	Rosa	specifiedBy	ENG30-01674464-a	mlad
ENG30-12620196-n	ruža	specifiedBy	ENG30-00217728-a	lep
ENG30-12620196-n	ruža	specifiedBy	ENG30-01170243-a	zdrav

ENG30-02987492-n	sablja	specifiedBy	ENG30-02311544-a	kriv
ENG30-02974697-n	sanduk	specifiedBy	ENG30-01391351-a	mali
ENG30-13831441-n	sever	specifiedBy	ENG30-02021905-a	bogat
ENG30-05585383-n	skelet	specifiedBy	ENG30-02551380-a	suv
ENG30-11444117-n	smrt	specifiedBy	ENG30-01126291-a	rđav
ENG30-03743902-n	spomenik	specifiedBy	ENG30-01235859-a	uspravan
ENG30-10375214-n	starac	specifiedBy	ENG30-00033574-a	neaktivan
ENG30-05665146-n	tehnika	specifiedBy	ENG30-01749320-a	savršen
ENG30-12136720-n	trska	specifiedBy	ENG30-02551380-a	suv
ENG30-10284064-n	tvorac	specifiedBy	ENG30-01996377-a	odgovoran
ENG30-09503282-n	veštica	specifiedBy	ENG30-00225564-a	pakostan
ENG30-11525955-n	vetar	specifiedBy	ENG30-00323873-a	brz
ENG30-11525955-n	vetar	specifiedBy	ENG30-01061489-a	slobodan
ENG30-12685831-n	zdravac	specifiedBy	ENG30-01170243-a	zdrav

Табела 39 Кандидати за ручно повезивање

Придев	**Именица**
бесан	рис
благ	девојка
блистав	живот
близак	контекст
богат	посластица
бојажљив	кокош
брз	мисао
брз	ракета
брз	тигар
брз	ватра
брз	бол
брз	зец
црвен	цвекла
црвен	корал
црвен	трешња
црвен	ватра
црвен	шкрге
црвен	уста
црвен	земља
црвен	каранфил
црвен	жар
дебео	пух
дебео	дете
добродушан	дете
дебео	бундева
дебео	пух
добродушан	дете
дуг	штап
дуг	година

дуг	век
глуп	делфин
глуп	сом
глуп	во
глуп	гуска
горак	лек
горак	тоник
хладан	додир
хладан	камен
хладан	кип
хладан	лед
хладан	смрт
хладан	стена
хладан	шприцер
хладан	туна
хладнокрван	риба
храбар	лав
храбар	војник
јак	бик
јак	биво
јак	брат
јак	држава
јак	град
јак	колектив
јак	коњ
јак	ратник
јак	стабло
јак	уговор
јак	земља
јак	банка
јак	дрво
јак	град

јак	смрт
јак	ждребица
јак	сребро
јак	стена
јак	во
јак	земља
јак	ракија
јак	камила
јасан	дан
јасан	књига
јасан	месец
јасан	звоно
лак	перо
лак	птица
лак	ваздух
лак	веверица
лак	ластавица
леп	слика
леп	младост
леп	сан
љут	рис
мален	веверица
мален	зрно
млад	сунце
напрасит	рак
натуштен	облак
нем	дух
нем	кип
нем	риба
нем	сфинга
нем	гроб
неопходан	ваздух

неопходан	живот
непријатан	сведок
несигуран	извор
неутралан	место
нежан	мирис
одложен	отпад
одвратан	појава
осетљив	девојка
остављен	плен
остављен	заштита
освојен	супарник
паметан	народ
плашљив	срна
плашљив	зец
поносан	паун
популаран	адвокат
популаран	писац
популаран	дело
популаран	певач
популаран	средство
популаран	коцкарница
постојан	рука
потпун	промена
висок	дрво
повољан	третман
познат	особа
познат	писац
познат	поборник
познат	противник
познат	упориште
познат	вођа
познат	концепт

познат	књига
познат	газда
познат	аутор
познат	филозоф
познат	краљ
познат	легенда
познат	песник
познат	певач
познат	сликар
познат	уредба
познат	уредник
познат	закон
познат	поступак
познат	правац
познат	уметност
познат	место
познат	модел
познат	стециште
познат	операција
познат	теорија
прав	јаблан
прав	јела
прав	стрела
прав	основ
прихватљив	основа
прихватљив	партнер
прихватљив	пријатељ
прихватљив	страна
прихватљив	заговорник
прихватљив	доказ
прихватљив	представник
прихватљив	савезник

прихватљив	део
прихватљив	експеримент
припремљен	салата
припремљен	пуњење
присутан	аутор
присутан	део
присутан	сведок
признат	инструмент
прљав	прасе
прљав	свиња
продоран	лет
променљив	мода
променљив	време
прост	пасуљ
расположив	датотека
раван	површина
раван	стакло
раздражљив	пас
раздражљив	врхунац
разумљив	резултат
сигуран	град
сигуран	смрт
сјајан	ватра
слаб	дете
слаб	девојка
слаб	вода
слаб	апарат
сладак	бомбона
сладак	млеко
славан	глумац
снажан	бик
снажан	глас

снажан	лав
снажан	медвед
снажан	утицај
снажан	ветар
спокојан	дете
спор	сукоб
сталан	планина
стар	ципела
стар	лоза
стар	време
стидљив	девојка
страствен	танго
стваран	небо
сув	грана
супротан	бумеранг
сув	штап
сув	лист
сув	папир
свеж	јабука
свеж	уметник
свеж	зора
шарен	тигар
широк	стих
широк	језеро
широк	град
широк	небо
широк	шума
широк	крило
широк	шака
широк	врата
тајанствен	пут
тајанствен	време

танак	филм
танак	хартија
танак	хрт
танак	конац
танак	сламка
танак	девојка
танак	стрела
танак	длака
тежак	камен
тежак	буре
тежак	олово
тежак	рад
тежак	во
тужан	пас
тужан	смрт
тужан	туга
управан	бор
управан	дело
управан	стуб
управан	бедем
узбудљив	истраживање
ужасан	понашање
важан	део
важан	намирница
важан	савезник
важан	сегмент
важан	симбол
важан	тренд
велик	брдо
велик	црква
велик	галаксија
велик	поклопац

велик	шума
велик	живот
велик	аутомобил
велик	кутија
велик	наклоност
велик	одговорност
велик	планина
велик	шака
велик	трофеј
велик	дрво
велик	дворац
велик	планина
велик	победа
велик	врата
велик	кула
велик	црква
велик	кит
велик	море
велик	брдо
велик	небо
велик	простор
велик	глава
велик	језеро
велик	капија
велик	котао
велик	шума
велик	кабао
висок	буква
висок	кула
висок	планина
висок	торањ
занимљив	сведок

занимљив	тема
занимљив	вид
занимљив	илустрација
занимљив	извор
занимљив	ритам
здрав	вук
здрав	змај
згужван	кора
жив	река

Прилог 5

Уводни текст упитника у оба пројекта групне расподеле рада

„Поштовани, попуњавањем овог упитника помажете да сачувамо српски језик у дигиталном окружењу! За то ће Вам бити потребно само неколико минута. Резултат ће бити изрази који се најчешће користе у свакодневном говору, које ћемо додати у језичке ресурсе за рачунарску обраду српског језика. Молимо Вас да одговорите да ли наведене изразе користите (тада ће одговор бити Да), или не користите (одговор ће бити Не) у свакодневном говору, или их можда користе особе у Вашем окружењу. Линк према овом упитнику слободно можете проследити својим колегама и познаницима. Хвала и уживајте!“

Прилог 6

Речник термина

енглески језик	српски језик
API – Application Program Inteface	апликационо програмско сучеље за веб сервисе
ASCII – American Standard Code for Information Interchange	амерички стандардни код за размену информација
BILI (Balkan Interlingual Index)	нтерлингвални индекс језика Балкана
CQP – Corpus Query Processor	алат за обраду упита над корпусом
crowdsourcing	групна расподела рада
DTD – document type definition	дефиниција типа документа
EWN (EuroWordNet)	Еуроворднет
ILI (Interlingual Index)	интерлингвални индекс језика
natural language processing (NLP)	обрада природног језика
NLTK – Natural Language Tool Kit	скуп алата за обраду природног језика
OWL (Web Ontology Language)	веб језик за онтологије
PWN (Princeton WordNet)	Принстонски ворднет (у овом раду Ворднет)
RDF – Resource Description Framework	оквир за описивање ресурса у Семантичком вебу
RetFig	онтологија реторичких фигура за српски језик
SGML – Standard Generalized Markup Language	стандардни генерализовани језик за означавање
Simile	поређење
SPARQL – рекурзивна скраћеница од SPARQL Protocol and RDF Query Language	протокол и језик за RDF упите
SWNonto	онтологија Српски ворднет
WSLR – Work Station for Lexical Resources	радна станица за језичке ресурсе
XML – Extensible Markup Language	прошириви језик за означавање
XPath – XML Path Language	XML језик за навигацију кроз елементе и атрибуте XML документа
XSD – XML scheme definition	дефиниција XML схеме

9 Биографија аутора

Јелена Митровић је рођена 25.11.1980. године у Ужицу, где је завршила основну школу и гимназију друштвено-језичког усмерења. Основне студије на Филолошком факултету, на одсеку за Неохеленске студије, где је студирала модерни грчки језик и књижевност, старогрчки језик и енглески језик, завршила је 2004. године. Дипломске академске студије - мастер на Филолошком факултету завршила је на одсеку за Библиотекарство и информатику. Мастер рад са темом „Пројекат Гугл књиге“ одбранила је код професорке Цветане Крстев 2010. године. Докторске студије на модулу Култура уписала је 2011. године на истом факултету, а 2013. године је, положивши све испите прописане програмом докторских студија са највишом оценом, стекла право да пријави тему докторске дисертације.

Јелена Митровић је као студент радила у Канцеларији председника Општине Ужице, као преводилац за енглески и грчки језик, а касније и у школи страних језика Калифорнија, такође у Ужицу. По завршетку основних студија, радила је у Интернационалном центру за стране језике и уметност на Бермудским острвима (Bermuda International Languages and Art Institute), где је две године била наставник грчког језика, старогрчког језика и енглеског као страног језика. Касније је, по повратку у Србију, радила као преводилац и стручни сарадник у издавачкој кући Лагуна у Београду. Поред учешћа на неколико групних пројеката, самостално је превела и адаптирала књигу „Седам доба жене“, ауторке Розмари Ленард. Радила је и ради као преводилац и лектор за многе каталоге изложби и књиге које представљају активности Трећег Београда, уметничког удружења које предводи др. Селман Тртовац, директор библиотеке Гете института у Београду.

У Универзитетској библиотеци „Светозар Марковић“ Јелена Митровић радила је као информациони стручњак што је подразумевало разна задужења – лектор за научне радове и приказе на енглеском, грчком и српском језику, уредница званичне веб презентације Библиотеке, менаџер културних дешавања – изложби, поставки, стручних и научних предавања, преводилац, стручњак за односе с јавношћу.

Као сарадник у оквиру Темпус пројекта „New Library Services at Western Balkan Universities“, Јелена Митровић је била ангажована као званични преводилац и лектор, током 2011. и 2012. године.

Јелена Митровић радила је и у Градској библиотеци Панчево, на пројекту ретроактивног уноса библиотечког фонда и његовом означавању пругастим кодом, који

је финансирало Министарство рада и запошљавања Републике Србије после чега је положила стручни испит у библиотечко-информационој делатности. У истој библиотеци била је ангажована као сарадник уредништва научног часописа „Читалиште“ (некада „Панчевачко читалиште“).

Члан је Савета редакције научног часописа Инфотека, као лектор за научне радове писане на енглеском језику. Бави се самосталном преводилачком и лекторском делатношћу и радила је као Сарадник у настави на катедри за Библиотекарство и информатику, Филолошког факултета у Београду, на предметима Информатички практикум 1, Мултимедијални документ и Дигитални текст 2, зимског семестра школске 2014/2015. године.

Активно је учествовала на многим значајним домаћим и међународним конференцијама и скуповима: Inforum, Prag, 2011; SEEDI, 2011, Zagreb; „Text, Speech and Dialogue“, 2013, Plzen; „Kulture u dijalogu“, 2013, Filološki fakultet, Beograd; „35 godina računarske lingvistike u Srbiji“, 2013, Beograd; „Global WordNet Conference“, 2014, Tartu, Estonia, 2014; PARSEME Meeting, Athens, 2014.; EUROLAN Summer school on Linguistic Linked Open Data, Sibiu, Romania, 2015; „Global WordNet Conference“, 2016, Bucharest, Romania; PARSEME Summer School in Machine Translation, La Rochelle, France, 2016; „Computational rhetoric workshop – Computing figures, figuring computers“, University of Waterloo, Canada – предавач по позиву; „Current trends in figurative language“ Tuebingen, Germany, 2016.

Од 1.10. 2016. године Јелена Митровић је запослена као истраживач и предавач на Факултету за рачунарство и математику, Универзитета Пасау у Немачкој, на катедри за информатику, дигиталне библиотеке и веб информационе системе. Активан је члан Групе за обраду природног језика и семантичко рачунарство Универзитета Пасау. Одржава активну научну сарадњу са члановима Групе за језичке ресурсе и технологије, Универзитета у Београду.

Говори енглески, грчки, мађарски и немачки језик. Познаје и старогрчки и француски језик.

10 Изјаве о докторској дисертацији

Прилог 1.

Изјава о ауторству

Име и презиме аутора **Јелена Д. Митровић**

Број индекса **11087 Д.**

Изјављујем

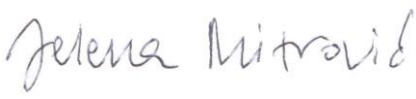
да је докторска дисертација под насловом

Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 17.01.2017



Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора **Јелена Д. Митровић**

Број индекса 11087 Д.

Студијски програм Култура – Библиотекарство и информатика

Наслов рада **Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада**

Ментор **проф. др Цветана Крстев**

Потписана Јелена Д. Митровић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива 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 аутора

У Београду, 17.01.201

Jelena Mitrović

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)
2. **Ауторство – некомерцијално (CC BY-NC)**
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.

Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 17.01.2018.

Jelena Mitrović

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 прерада**. 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 прерада**. 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.