УНИВЕРЗИТЕТ У НОВОМ САДУ **ОБРАЗАЦ - 11**
ФИЛОЗОФСКИ ФАКУЛТЕТ

ИЗВЕШТАЈ О ОЦЕНИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Кандидат: Страхиња Димитријевић
Тема: Аутоматско одређивање врста ријечи у морфолошки сложеном језику

I ПОДАЦИ О КОМИСИЈИ

1. Датум и орган који је именовао комисију
20.03.2015. год.
Наставно-научно вече Филозофског факултета у Новом Саду

2. Састав комисије са назнаком имена и презимена сваког члана, звања, назива уже научне области за коју је изабран у звање, датума избора у звање и назив факултета, установе у којој је члан комисије запослен:

др Душица Филиповић Ђурђевић, ванредни професор за ужу научну област Психологија, Филозофски факултет у Новом Саду, *председник комисије*

др Петар Милин, ванредни професор за ужу научну област Психологија, Филозофски факултет у Новом Саду, *ментор*

др Александар Костић, редновни професор за ужу научну област Психологија, Филозофски факултет у Београду

II ПОДАЦИ О КАНДИДАТУ

1. Име, име једног родитеља, презиме:
Димитријевић (Бранислав) Страхиња

2. Датум рођења, општина, република:
31.10.1971. год.
Грачаница
Босна и Херецеговина

3. Датум одбране, место и назив магистарске тезе:
27.06.2007. год.
Филозофски факултет Бања Лука
Когнитивне стратегије у обради језика: Примјена контекстуалних језичких информација у задатку аутоматске лематизације

4. Научна област из које је стечено академско звање магистра наука:
Психологија

III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:
Аутоматско одређивање врста ријечи у морфолошки сложеном језику

IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Докторска дисертација се састоји од пет (5) поглавља.

У првом, уводном поглављу, дат је преглед стратегија аутоматске обраде речи, затим је описана улога аналогије и значај фонотактичких информација у продукцији и разумевању језика. На крају овог дела изложена је мотивација и циљ истраживања.

Наредна три поглавља представљају емпиријски део рада, који се састоји од три повезане студије у којима се примењују технике машинског учења, когнитивно засновано рачунарско моделовање и, коначно, психолингивтички експеримент. Прво су представљени резултати истраживања дискриминације врста речи на основу фонотатктичких информација, а уз помоћ машина са векторима подршке. У наредном, трећем поглављу, приказана је студија продукције инфлексионих облика речи с ослонцем на учење засновано на меморији, односно, рачунарски модел који користи аналогије у продукцији нових облика речи. Коначно, у четвртом поглављу представљен је експеримент који је за циљ имао проверу когнитивне веродостојности модела заснованог на меморији (аналогији). Другим речима, резултати друге студије у којој је примењено рачунарско моделовање послужили су за генерисање истраживачке хипотезе која је затим тестирана и експериментално.

У петом поглављу продискутовани су добијени резултати и изведени закључци. Повезани су разултати претходних студија са новим налазима који су произашли из овог рада. Продискутована су отворена питања и дати предлози за будућа истраживања.

Поред наведених пет поглавља, у дисертацији се налазе још и списак коришћене литературе, као и већи број прилога.

Укупан обим дисертације је 130 страна. У дисертацији се налази 13 слика, 19 табела, 17 прилога и 295 рефренци.

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Теоријски део рада кандидат је започео дефинисањем основних појмова, почев од фонеме/графеме и фонотактичких информација, које представљају комбинације ових елементарних јединица које се појављују, односно које су допуштене у неком језику. Постављено је питање да ли овакво ограничење, а посебно чињеница да се допуштене комбинације појављују с неједнаком вероватноћом (неке су веома честе, а друге не), може представљати важну информацију у когнитивној обради језика. Другим речима, да ли се компетентни говорник користи оваквим подацима у свакодневном говору. Посебно детаљно продискутоване су могуће стратегије у различитим задацима обраде речи. Размотрене су предности и ограничења основних типова информација на којима би неки систем за обраду језика могао да се заснива. У раду се полази од подела које су предложили Manning & Schütze (2000) и Милин (2004), те се детаљно разматрају карактеристике система који користе одређени тип језичких информација у задацима обраде природних језика. Посебно је дискутована и когнитивна заснованост поменутих модела.

У посебном пододељку представљен је принцип аналошког учења, са јасним утемељењем у психологији па и лингвистици. Већ је De Saussure (1916), а нешто касније и Bloomfield (1933), говорио о могућностима генерисања нових речи на основу постојећих образаца. У дисертацији је размотрен значај и раширеност поменутог облика учења у области психологије, лингвистике и психолингвистике. Представљени су и најзначајнији модели (Skousen, 1989, 1992; Daelemans & Van den Bosch, 2005), који су упоређени са моделима учења путем дистрибуираног процесирања – конекционистички модели (нпр., Plaut & Gonnerman, 2000; Seidenberg & Gonnerman, 2000), као и другачијим моделима обраде који се темеље на психолошким принципима учења (Baayen, Milin, Filipović Đurđević, Hendrix, & Marelli, 2011).

Укупно посматрано, уводни део представља обухватни приказ једне интердисциплинарне области, која лежи у пресеку лингвистике, психологије, теорије информација и машинског учења, са пажљивим и јасним тумачењем кључних проблема и отворених питања. Читав овај приказ даје широко разумевање основног проблема који је испитиван у дисертацији.

Други део рада представио је прву емпиријску студију, са задатком дискриминације (разликовања) врста речи помоћу пробабилистичког модела заснованог на векторима подршке (Support Vector Machines – SVM). Модел је доследно користио фонотактичке информације, а систематски је

варирана врста и количина информација (биграми, триграми, специфична позиција у оквиру речи и сл.). Кандидат је показао суверено владање читавим низом статистичких техника и процедура за анализу и припрему података. Осим тога, у овом делу студије примењене су чак и различите варијанте основног алгоритма, варирањем језгрене функције за дискриминацију, да би се упоредиле специфичности и успешност с обзиром на полазне претпоставке. Овај, први емпиријски део, мотивисао је даља испитивања значаја фонотактичких информација у обради језика, а посебно у домену обраде речи.

Одељак "Фонолошка сличност ријечи и проблем продукције инфлексионих облика", представио је и применио још један систем за обраду језика. Овога пута реч је о моделу који има озбиљно психолошко, когнитивно утемељење. За разлику од модела са векторима подршке, за које се не сматра да су когнитивно реални (могући), аналошко учење и модели базирани на начелима аналошког резоновања имају дугу традицију, како у области психологије, тако и у самој лингвистици. Иако бројни налази показују да су модели са векторима подршке вероватно најуспешнији у разноврсним проблемима класификације, у раду је отворена дискусија и дато је поређење успешност овог приступа односу на један когнитивно реални приступ.

Четврти део рада представља експериментално истраживање у којем је примењен задатак лексичке одлуке са визуелном презентацијом стимулуса – речи. Комисија оцењује као посебно вредно и значајно да је изведени експеримент заснован на емпиријским резултатима претходна два истраживачка дела рада. Ово је у методолошком смислу оригинални приступ у којем је кандитат пошао од општијих налаза, ка специфичним: првом студијом, у којој је примењен поступак дискриминационог машинског учења са векторима подршке, дефинисано је најпре шта представља максимум учења, узимајући у обзир прецизно дефинисан задатак и расположиве информације; затим је применом психолошки утемељеног модела учења изведен закључак о томе колико се когнитивни систем приближава овом максимуму и каква је природа и величина одступања; коначно, експерименталним поступком директно је тестирано да ли хумани испитаници, компетентни говорници језика, показују понашање слично ономе које показује и рачунарски модел који користи аналошко учење. Укупно узевши, три емпиријске студије представљају целину која показује висок истраживачки домет и оригиналност кандидата.

Други, трећи и четврти одељак, у којима су представљене три независне, али проблемски повезане студије, представљају целине у којима су поштовани највиши научни стандарди. Сваки одељак даје приказ проблема и мотивацију, а затим детаљно излаже истраживачку методологију (материјале, поступак узорковања, процедуру, опис планираних анализа). Следи приказ резултата и обухватна дискусија која елегантно мотивише потребу за студију која следи у наредном одељку.

Пети део дисертације представља општу, односно обједињену дискусију свих приказаних резултата. У овом делу убедљиво се говори о томе да су фонотактичке информације вероватно *довољне* за успешно обављање задатака из домена обраде речи. Из ове тачке кандидат поново вешто гради дискусију у којој отвара питања која су од централног значаја за читаву област психологије језика. Једно од тих питања тиче се потребе за хипотетичким конструктом *менталног лексикона*. Иако доминантан у области обраде речи, па и шире, овај конструкт је и претходно био доведен у питање, како са лингвистичког (нпр., Bybee, 1985, 2001), тако и са психолошког становишта (Plaut & Gonnerman, 2000; Rumelhart & McClelland, 1986 и други). С тим у вези, од уводних разматрања, преко три емпиријске студије, до завршне дискусије, кандидат исказује подједнаку пажњу различитим приступима (лингвистичком, психолошком и инжењерском), откривајући противуречности и могућа решења. Притом, он показује велики дар за промену угла посматрања и опажање занимљивих истраживачких проблема.

Докторска дисертација завршава закључцима који пружају шири оквир разматрања резултата у којем се указује на теоријске импликације за когнитивне моделе обраде речи.

Укупно, дисертација садржи 296 референце, релевантне за испитивани проблем истраживања. Референце су одабране тако да приказују кључна становишта и моделе и пружају основу за адекватну компаранцију са резултатима саме докторске дисертације.

Комисија закључује да су сви елементи дисертације написани у складу с научним стандардима.

VI СПИСАК НАУЧНИХ И СТРУЧНИХ РАДОВА КОЈИ СУ ОБЈАВЉЕНИ ИЛИ ПРИХВАЋЕНИ ЗА ОБЈАВЉИВАЊЕ НА ОСНОВУ РЕЗУЛТАТА ИСТРАЖИВАЊА У ОКВИРУ РАДА НА ДОКТОРСКОЈ ДИСЕРТАЦИЈИ

Dimitrijević, S. (2011). Kontekst i vrsta riječi kao faktori tačnosti automatske lematizacije. *Radovi, 4*, 67–87.

Dimitrijević, S., Kostić, A. i Milin, P. (2009). Stability of the syntagmatic probability distributions. *Psihologija*, *41* (1), 53–67.

Dimitrijević, S., Milin, P. i Kostić, A. (2008): Primjena kontekstualnih jezičkih informacija na nivou vrsta riječi u zadatku automatske lematizacije, *XIV Naučni skup Empirijska istraživanja u psihologiji*, Beograd: Filozofski fakultet, 7-8 februar, str. 19.

Dimitrijević, S., Milin, P. i Kostić, A. (2007): Analiza stabilnosti distribucija zavisnih vjerovatnoća na nivou vrsta riječi, *Međunarodni naučno-stručni skup Psihologija i društvo*, Novi Sad: Odsek za psihologiju, Filozofski fakultet, Univerzitet u Novom Sadu, 19-20 oktobar, str. 47- 50.

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА:

Резултати добијени у задатку дискриминације промењљивих врста речи показали су да се на основу учесталости фонотактичких информација – могућих комбинација два односно три фонема/графема, израчунатих на нивоу граматичких типова могу успешно одредити врсте речи којима ти граматички типови припадају. При томе, нису сви биграми односно триграми били подједнако информативни. На основу добијених резултата може се закључити да су биграми минималне фонотактичке јединице на које се можемо ослонити у обради морфолошки сложених речи. Демонстрирана су и детаљно продискутована ограничења у дискриминацији врста речи на основу биграма и триграма који се налазе на крају речи (који су веома често суфикси).

У другој студији демонистрирана је могућност инфлексионе продукције на основу фонотатктичких информација (у овом случају коришћене су информације о последњим слоговима речи), помоћу модела заснованог не аналогији (учење засновано на меморији). Идентификовано је неколико фактора који утичу на успешност овог модела, као што су: врста речи и врста граматичког типа који се обрађује, број инфлексионих наставака за творбу траженог облика, број фонолошких алтернација које се јављају при творби, број примера у оквиру граматичког типа итд.

На основу резултата добијених у другој студији, осмишљен је екперимент у којем је потврђена когнитивна веродостојност учење заснованог на меморији и информација на које се тај модел ослањао у претходном задатку. Показано је да је за речи (именице мушког рода у номинативу множине) за које је модел генерисао погрешан инфлексиони облик потребно дуже време за обраду и да се на њима прави више грешака него на речима за које је продукован исправан инфлексиони облик. Фреквенција облика фацилитира обраду, с тим да само при обради именица из кластера погрешних решења фацилаторну улогу има и фреквенција леме која компензује веће оптерећење које тада стоји пред когнитивним системом. То указује на потребу да когнитивни систем, како би могао да изађе на крај са захтевнијим стимулусима, има на располагању одговарајући ресурс, као што је велики речник, било да је реч о вокабулару појединца или великим језичким корпусима у случају аутоматске обраде речи.

Добијени резултати, с једне стране, указују на потребу да се настави са систематским проверавањем улоге различитих типова језичких информацијама у задацима аутоматске обраде језика. С друге стране, може се закључити да се разумевање и продукција морфолошки сложених ријечи може обавити на основу фонолошких/ортографских и семантичких информације, те да нема потребе увођења посебног домена задуженог за обраду морфологије. Ово има далекосежне импликације, јер указује на бројна ограничења структуралних теорије језика и језичког понашања и доводи у питање претпоставке о менталном лексикону и представама речи са сложеном хијерархијом лингвистичких описа.

VIII ОЦЕНА НАЧИНА ПРИКАЗА И ТУМАЧЕЊА РЕЗУЛТАТА ИСТРАЖИВАЊА:

У методолошком смислу, докторска дисертација кандидата Страхиње Димитријевића следи експериментално-психолошку традицију на најбољи могући начин. На основу постављеног истраживачког проблема изведен је низ од три истраживања, у којима су примењене методе различитих области истраживања. У конкретном случају, примењене су и рачунарски засноване симулације (рачунарски експерименти) и бихејвиорални експеримент на хуманим испитаницима.

Сваки корак у поступку истраживања приказан је брижљиво и детаљно, као што стандарди и захтевају, па је на основу описа могуће поновити студије. Примењен је читав низ различитих статистичких поступака, чиме је кандидат показао сувереност у овом аспекту. Резултати су представљени прегледно и на систематски начин, како табелама, тако и графички. Осим тога, ове сумарне приказе пратио је и адекватни опис у самом тексту. На основу добијених резултата изведени су закључци који дају одговоре на постављене циљеве истраживања.

IX КОНАЧНА ОЦЕНА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

1. Дисертација је написана у складу са образложењем наведеним у пријави теме

2. Дисертација садржи све битне елементе

3. По чему је дисертација оригиналан допринос науци

   Постоји већи број разлога због којих се докторска дисертација Страхиње Димитријевића мора сматрати оригиналним научним доприносом. Као прво, кандидат је у центар свог истраживања поставио веома актуелан истраживачки проблем. Овај проблем је додатно занимљив за истраживања на српском и сличним језицима, са богатом инфлексијом, а тиме и специфичном тежином фонотактичких информација. Друго, дисертација је на убедљив начин показала да се потенцијале интердисциплинарног приступа у истраживањима језика. Теоријски увод се суверено кретао у материји лингвистике и психологије, а делимично и у области машинског учења. У истраживачком делу, такође, примењена је разноврсна методологија, која је укључивала психолошки експеримент, технике статистичког учења и рачунарско моделовање когнитивних процеса. Треће, сами резултати представљају значајан допринос у разумевању начина на који когнитивни систем обрађује природни језик. Важан је и онај део налаза који се односи на резултате учења заснованог на аналогији. С једне стране, резултати су упоређени са понашањем испитаника, чиме је потврђена когнитивна релевантност оваквог и сличнох модела. С друге стране, ти исти резултати су доведени у везу са успехом приступа који се сматра најуспешнијим за задатке овог и сличног типа. Према томе, налази су вредни како за когнитивну психологију, тако и за област машинског учења, па и за развој интелигентних система заснованих на језику. Четврто, као што је већ и претходно истакнуто, рад представља јединствен допринос и у методолошком смислу, с обзиром да је у раду изведен бихејвиорални експеримент у функцији тестирања предикција које су произашле из рачунарске симулације помоћу модела заснованог на аналогији. Како је рачунарски модел потпуно спецификован, тако и саме предикције постају врло конкретне и, што је од суштинског значаја за науку, оповргљиве (тестабилне). Другим речима, у раду се није пошло од широких и неспецифичних хипотеза и конструката које је најчешће немогуће емпиријски оповргнути, већ је постављен и изведен један функционални истраживачки приступ. Коначно, осим што су резултати рачунарског модела тестирани и експериментално, задатак лексичке одлуке омогућио је и генерализију и повезано промишљање о везама између продукције и рецепције (обраде) језика: рачунарски модел је симулирао задатак продукције нових облика речи, а затим је у експерименту испитано да ли испитаници греше на сличан начин у рецепцији, тј. обради речи.

4. Недостаци дисертације и њихов утицај на резултат истраживања

   Докторска дисертација кандидата Страхиње Димитријевића представља једну заокружену целину, којој је тешко наћи недостатке. Можда има смисла говорити о потреби да се овде приказани резултати доведу у везу са другим важним рачунарским моделима у области обраде језика. Пре свега, ту мислимо на моделе паралелне обраде (конкционистички модели). Такође, занимљиво би било упоредити и тардиционалне моделе, а посебно оне који су значајни

представници модела с менталним лексиконом, као основном когнитивном структуром. Сва ова питања, међутим, више се тичу очекивања да кандидат настави даља истраживања у истом правцу. У том смислу, има смисла говорити о предлозима за нова испитивања овог и сродних феномена, док сам рад нема недостатака који би ограничавали веродостојности основних резултата.

X ПРЕДЛОГ:

На основу укупне оцене дисертације, комисија предлаже да се докторска дисертација Страхиње Димитријевића под називом "Аутоматско одређивање врста ријечи у морфолошки сложеном језику" прихвати, а кандидату одобри одбрана.

ПОТПИСИ ЧЛАНОВА КОМИСИЈЕ:

др Душица Филиповић Ђурђевић, ванредни професор за ужу научну област Психологија, председник комисије

др Петар Милин, ванредни професор за ужу научну област Психологија, ментор

др Александар Костић, редовни професор за ужу научну област Психологија, члан