Р Е Ц Е Н З И Я

по конкурс за професор в област на висше образование
4. Природни науки, математика и информатика,
Професионално направление: 4.6 Информатика и компютърни науки,
Научна специалност: 01.01.12 – Информатика
(компютърна лингвистика - средства и системи за обработка
на лингвистични знания), обявен в ДВ бр. 58/29.07.2011

Кандидат: доц. д-р Людмила Петрова Димитрова-Рашкова

Рецензент: проф. д.м.н. Петър Любомиров Станчев, Институт по математика и информатика при БАН, e-mail: stanchev@math.bas.bg

Тази рецензия е написана и представена на основание на заповед 254/28.09.2011 г. на Директора на ИМИ, БАН както и на решение на научното жури по процедурата (Протокол от 30.09.2011). Тя е изготвена въз основа на ЗРАСРБ, Правилника за прилагане на ЗРАСРБ, Правилника за развитие на академичния състав на БАН и на ИМИ при БАН.

1. Общо описание на представените материали

Бяха ми представени следните материали на кандидата: професионална автобиография, диплома за завършено висше образование, диплома за придобита образователна и научна степен „доктор“, пълен списък на научните трудове, списък на научните трудове за участие в конкурса, саморъчно подписана авторска справка за научните приноси на трудовете за участие в конкурса, списък цитирания, препис-извлечение от протокола на НС на ИМИ БАН за иницииране на процедурата, копие от Държавен вестник с обявата за конкурса, документи за учебна работа, справки за четени лекции/упражнения, списък на научноизследователски проекти с ръководство или участие на кандидата, копия от трудовете за участие в конкурса, документ, удостоверяващ заемането на академична длъжност “доцент” поне 2 години съгл. чл.29 ал.1 т.2 от ЗРАСРБ

Доц. д-р Людмила Петрова Димитрова-Рашкова е представила за участие в конкурса 29 научни публикации. Тематично те се разделят в следните групи:

А. Математическа лингвистика – логико-граматически формализми и компютърни системи за моделиране на резултати от теория на формалните езици - 3 броя

Б. Компютърна лингвистика – 12 броя

В. Средства и системи за обработка и управление на езикови ресурси - 11 броя

Г. Модели на изследователски Е-инфраструктури -3 броя

От рецензираните общо 29 труда, 10 са в международни списания, 2 в български списания, 15 в сборници на международни конференции, 1 в трудове на български конференции и 1 монография. В 26 от публикациите доц. д-р Людмила Димитрова е първи автор, 2 от публикациите са на български език, а останалите са на английски.

2 Обща характеристика на научната, научно-приложната и педагогическа дейност на кандидата

Доц. д-р Людмила Петрова Димитрова-Рашкова е завършила Софийски университет "Св. Климент Охридски" – магистър по математика, през 1969 г. През 1977 г. защитава дисертация в Московски държавен университет. От 1978 г. е научен сътрудник, и от 1996 г. – старши научен сътрудник, доцент в Института по математика и информатика при БАН.

Ръководила и е участвала в 20 национални и международни проекти. Някои от тях са: **координатор (ръководител на Консорциума) на проект от 7 Рамкова програма на Европейската комисия** *project MONDILEX Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources*; участник и представител на ИМИ в *EU Research Infrastructure CLARIN Common Language Resources and Technology Infrastructure*; участник и представител на ИМИ в *БГ-КЛАРИН Национална интердисциплинарна изследователска Е-инфраструктура за интегриране и развитие на електронните ресурси за български език* като част от Европейския; ръководител на проект „Семантика и съпоставителна лингвистика, ориентирани към разработване на двуезичен електронен речник" с Институт по славистика на Полската академия на науките; ръководител на проект „Електронни корпуси – съпоставително изследване с цел проектиране на българо-словашки електронни езикови ресурси" с Института по лингвистика на Словашката академия на науките.

Чела е лекции в 3 висши училища. Има богата рецензионна дейност. Участва в редколегия на международно списание и е редактор на 5 научни издания. Нейни приложни разработки са:

1. MTE Language Resources (versions 1 – 4): v. 4 http://nl.ijs.si/ME/V4/doc/index.html#sec-crp
2. CONCEDE Bulgarian LDB: http://www.itri.brighton.ac.uk/projects/concede
3. Slovak-Bulgarian Corpus Linguistics Terminology Data Base: https://data.juls.savba.sk/mltd/
4. Bulgarian-Slovak corpus on the web: http://korpus.sk:8090/
5. Bulgarian-Polish corpus
6. Bulgarian-Polish-Lithuanian corpus
7. Bulgarian-Polish Electronic Dictionary
8. Bulgarian-Polish Online Dictionary

3 Анализ на научните постижения на кандидата

Основните научни, научно-приложни и учебно-методически приноси на Людмила Петрова Димитрова са главно в следните области:

A. Математическа лингвистика – логико-граматически формализми и компютърни системи за моделиране на резултати от теория на формалните езици: 4, 5, 6

Публикациите в тази група са свързани с опитите за формално представяне на естествените езици, които продължават повече от пет десетилетия след работите на Ноам Чомски. В [5] са представени четири вида логически граматики – metamorphosis grammars, definite clause grammars, extraposition grammars, discontinuous grammars и са разгледани

възможностите им за описване на специфични особености на българския език. [6] разглежда в детайли прекъснатите логически граматики. Приносът е в използването на логико-граматически формализми за описание и моделиране на особено тежки за формално описание явления в българския език, които имат пряко приложение към автоматизираната обработка на езика. В [4] са показани резултати, свързващи формални езици с магазинни автомати и е описана диалогова програмна система, предназначена за моделиране на основни резултати от теорията на формалните езици и граматики за нуждите на обучението. Добро впечатление прави умелото боравене с различни формализми.

Б. Компютърна лингвистика: 7, 10, 11, 12, 13, 14, 16, 21, 22, 24, 25, 27

Публикациите в това направление са в представяне и обработка на лингвистични знания и създаване на най-използваните в момента цифрови езикови ресурси – многоезичните корпуси. [10] представя средства за описание и анотиране на лингвистични знания: хармонизирани лексически описания за шест централно- и източно-европейски езици, с акцент на средствата за българския език. Описани са разработените лексикони и специфични езиково-зависими ресурси за български език, необходими за сегментиране на анотирани езикови текстове. Българският лексикон съдържа 55182 речникови статии (50810 словоформи, 4205 – съкращения и имена, 1044 числови данни). Разработеният формализъм дава възможност да се използват програмни средства, създадени в рамките на други европейски проекти. В статията е описано приложението на българските специфични езикови ресурси за автоматично снемане на морфосинтактична многозначност. Прави впечатление огромното количество работа, която е извършена. В [12] са разгледани морфосинтактичните спецификации на някои български глаголни форми (причастията) и са предложени нови, съответстващи на съвременната българска граматика. В [13] са разгледани морфосинтактичните спецификации за български и за словашки език с цел хармонизация и анотиране на паралелни българо-словашки цифрови ресурси и за приложение в съпоставителните изследвания. [7] отразява създаването на всички цифрови езикови ресурси, вкл. корпусите, в рамките на проекта MULTEXT-East на ЕС (http://nl.ijs.si/ME). Ресурсите са разработени в стандартизиран формат, хармонизиращ специфичните характеристики на всеки от шестте езика на проекта, със стандартно маркиране и анотиране съгласно препоръките на TEI (Text Encoding Initiative) Working Group. Паралелният корпус продължава да служи като модел на многоезичен подравнен анотиран ресурс при разработване на нови корпуси за други естествени езици, което е видно от броя на цитиранията - 40. В [11] и [25] са представени в развитие цифрови ресурси с български език, като е описан и разработения за проекта MULTEXT-EAST първи български електронен корпус, включващ паралелен и подравнен корпус MTE-1984.Bg. Корпусът MTE-1984.Bg съдържа около 300 хил. думи, маркирани с SGML в CES-формат. Част от MTE-1984.Bg е анотиран на ниво словоформа корпус с обем 87235 слово-употреби: към всяка словоформа от текста на българския превод на „1984“ е присъединена лингвистична информация, включваща нейната основна форма (лема) и съответното й морфосинтактично описание. Анотираният корпус е получен след автоматично снемане на синтактична многозначност и пост-редакция от авторката (на цифровия ресурс) за отстраняване на грешките. MTE-Fiction.Bg е сравним корпус, с обем 97251 слово-употреби, ръчно анотиран на ниво параграф, като собствените съществителни имена са маркирани за „тип“: лице, място и др. Обемът на извършената работа е огромен.

[16] представя поредна версия на първия българо-полски/полско-български цифров корпус, разработен в рамките на междуакадемичното сътрудничество между ИМИ-БАН и Института по славистика на ПАН. В работата са обсъдени проблемите на разработка на паралелни корпуси. Двуезичният корпус съдържа художествена литература (белетристика, публицистика) и специална лексика (текстове на документи на Европейската Комисия и на Европейския Съюз). В [14] е направено сравнение между две системи за анотиране на текстове (на ниво словоформа) на два славянски езика: български език и полски език. В [27] са представени два двуезични корпуси, съставени от паралелни текстове на три славянски езика – български, полски и словашки. Това са първите българо-полски/полско-български и българо-словашки/словашко-български корпуси. Българо-словашкият корпус, разработен в рамките на междуакадемичното сътрудничество между ИМИ-БАН и Института по лингвистика на САН, с обем 1 млн. словоупотреби, съдържа два вида текстове: оригинални текстове на български език или на словашки език с преводите им, съответно, на словашки или на български, и текстове, преводи на български и словашки от трети език. В статията са описани подробно приложенията на двуезичните корпуси в съпоставителните изследвания – за съпоставяване на славянски езици: един аналитичен (български) с два синтетични (полски и словашки). [21], [22] и [24], са посветени на първия българо-полски-литовски корпус (с обем около 3 млн. словоупотреби). Той съдържа две части: паралелен (около 1 млн. словоупотреби) и сравним корпус. Част от паралелния корпус съдържа текстове (белетристика), подравнени на ниво „изречение“ и текстове (специална лексика), подравнени на ниво „параграф“, които образуват подравнен триезичен корпус. Сравнимият корпус е съставен от текстове, публикувани от електронни медии в Интернет, които са сравними не по обем, а по съдържание: те описват едно и също събитие. Като опорен език към трите езика е присъединен английски език. Подравненият триезичен корпус дава възможност за съпоставяне на езици от две различни езикови групи: славянската (български и полски) и балтийската (литовски). Триезичният корпус има широко приложение в съпоставителните изследвания. [21] е достъпна на Интернет-страницата на международната организация Association for Computational Linguistics (ACL). Обемът на създадените средства и приложимостта на резултатите са значителни.

В. Средства и системи за обработка и управление на езикови ресурси: 1, 2, 3, 8, 9, 15, 17, 18, 19, 20, 29

В тази група са събрани разработки за автоматична обработка на български език с доказан научен и научно-приложен принос в областта. В [1] е представена първата система за автоматично сегментиране на българските словоформи без използване на лексемен речник и без каквито и да е ограничения, наложени върху входния текст. Системата реализира описания в [2] конструиран от авторите първи алгоритъм за автоматичното сегментиране на словоформите в текста. Алгоритмът е комбинация от евристични съображения и точно определени множества от езикови елементи (множествата на българските окончания), той е разработен въз основа на анализа на 700 000 български словоформи, които като обем представят съдържанието на няколко тълковни речника и могат да се образуват от 70-те хиляди единици на Обратния речник на българския език. [3] представя две версии на програмна система за автоматичен анализ на български текст. За първи път за български език е реализиран алгоритъм за автоматично снемане омонимията на окончанията (автоматично причисляване на дадена словоформа към определен граматичен клас). Системата, работеща в режим на диалог на български език, дава възможност за изследване вида на лингвистичната информация, която може да се извлече

от текста в резултат на този вид автоматичен анализ. В [8] са отразени резултатите от разработка на лексикографските спецификации за българския език съгласно международните стандарти за кодиране на електронни речници, подбора на заглавните думи за едноезичната лексическа база данни (LDB) за български език, комплектуването и кодирането на речниковите статии за тези думи. CONCEDE LDB за българския език съдържа 2 700 речникови единици, избрани по определена методика, според изискванията на проекта CONCEDE, базирана е на Българския тълковен речник и използва лексикографските спецификации, разработени за проекта MULTEXT-EAST. Разработката на лексическата база данни за българския език е определена като **едно от най-важните научно-приложни постижения на ИМИ-БАН през 2000 г**. Приносът на [17] се състои в разработка на модел на двуезична лексическа база данни. Работата отразява проектирането и разработването на първата двуезична лексическа база данни за поддържане на българо-полски/полско-български онлайн речници. В [20] са представени резултатите от работата върху многоезична терминологична база данни на термините от областта на корпусната лингвистика. Софтуерната среда за представяне на терминологична база данни е базирана на PHP и Python, което позволява разширяване на обема на данните, както по отношение на броя на езиците, така и по отношение на съдържанието. Три работи са посветени на стандартизираното представяне на класификаторите в речниковата статия на цифровите многоезични речници: в [9] са обсъдени проблемите, свързани с хармонизацията на класификаторите в речниковата статия. В [15] е описано представянето на българския глагол в цифровите българо-полски/полско-български речници. Предложени са нови класификатори (разграничаващи форми и значения) за описване на специфични характеристики и езикови явления в българския език. Дискусията за стандартизирано представяне на класификаторите продължава и в [18]. Обсъдени са примери от текущата експериментална версия на българо-полския онлайн речник. Приносът на [19] и [20], представящи първия българо-полски онлайн речник, е научно-приложен. [19] описва разработването на двуезичния онлайн речник. Уеб приложението за представяне на речника е базирано на лексическа база данни, чието проектиране и разработване е описано детайлно в [17]. Текущото състояние на разработката, с разширена и обновена лексическа база данни, е представено в [29]. И тази група изследвания показват огромен обем работа и значителна приложимост на резултатите.

Г. Модели на изследователски Е-инфраструктури: 23, 26, 28

[23] описва целите и задачите на инициатива за създаване на модел на изследователска Е-инфраструктура за езикови ресурси, разработени за шест славянски езика: български, полски, руски, словашки, словенски и украински. В работата са представени и постиженията на участващите в инициативата институции в областта на езиковите ресурси и технологии. В [28] са представени най-важните резултати от изпълнението на проекта MONDILEX, вклитело неговия научен и социален принос. Специално място е отделено на възможностите, които предлагат Grid-базираните технологични платформи за обработка на стандартизирани многоезични ресурси. Приносът на монографичното изследване [26] е в предложения концептуален модел на Grid-базирана Европейска изследователска Е-инфраструктура, поддържаща цифрови езикови ресурси за славянска лексикография. Моделът обхваща всички съвременни езикови технологии и видове цифрови езикови ресурси, които трябва да бъдат включени в такава изследователска Е-инфраструктура: корпуси (едно- и многоезични, паралелни и съпоставими, анотирани и подравнени), речници (едно- и двуезични, електронни и

онлайн), бази данни, лексикони, тезауруси, онтологии. Разгледани са стандартите за представяне на цифрови езикови ресурси, създадени с цел тяхното многократно използване в съвременни многоезични системи. Като технологична платформа за поддръжка на изследователската Е-инфраструктура е предложена една модерна среда – Knowledge Grid, приложение на Grid-технологията за обмен и обработка на многоезични ресурси. Предназначението на Концептуалния модел е да интензифицира съвместните научни изследвания в областта на цифровата лексикография, осигурявайки възможност за разпространение на иновациите независимо от географското положение. Изследванията, представени в тази група трудове, са на много високо научно ниво. Разработката на Концептуалния модел е определена като **едно от най-важните научно-приложни постижения на ИМИ-БАН през 2010 г**.

4 Принос на кандидата

Някои от приносите на Людмила Димитрова в зависимост от чл.2 т.6 от Правилника на ИМИ за приложение на ЗРАСРБ са:

- трудове на други учени, публикувани в авторитетни издания, които съществено използват резултати на кандидата: резултати на кандидатката, представени в 7 от трудовете, са използвани съществено при разработка на програмни продукти или езикови ресурси от други учени; 16 работи на други учени, публикувани в авторитетни издания, цитират 2 статии на кандидатката (съгласно представения списък на цитиранията).

Някои от приносите на Людмила Димитрова в зависимост от чл.3 от Правилника на ИМИ за приложение на ЗРАСРБ са:

- ръководство и участие в международни и национални научноизследователски проекти: Координатор (ръководител на Консорциума) на проект от 7 Рамкова програма на Европейската комисия: MONDILEX Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources, както и в още 19 проекта
- участие в програмни и организационни комитети на научни мероприятия (от последните 5 години): 7
- членство в авторитетни творчески и/или професионални организации в съответната научна област: 2
- участия с доклади в международни и национални научни форуми (от последните 6 години): 26
- участия в редколегии на научни издания и редактор на научни издания: 6
- аудиторни занятия във висши училища – лекции и семинари: в 3 ВУ

  четени курсове в Софийски университет, ФМИ, специалност информатика: дискретна математика 1978 – 1990; математически основи на информатиката 1988 – 1989; математическа и компютърна лингвистика 1991 – 1994;

  четени курсове в Софийски университет, ФКНФ Магистърска програма: математически и логически основи на компютърната лингвистика 1995 – 1997;

  четени курсове в НБУ, бакалавърски факултет, специалност информатика: автомати, езици, изчисление, 1996 – 2000;

  четени курсове в ИМИ – ВТУ Маг. програма “Езикови и мултимедийни технологии” 2002 – 2007: логика, език, информация, бази данни, компютърна

лексикография, програмиране в Интернет и професионални комуникации, дискретна математика и елементи на математическата логика, компютърни системи и архитектури;

- разработване на лекционни курсове (създадени и прочетени за първи път в България курсове във ВУ): 3
- дейности, свързани с научното развитие на докторанти, дипломанти и студенти:
  - Докторанти – ръководство: 1
  - Ръководител на дипломанти по магистърски програми: 10
- Приложни разработки – създаване на информационните продукти: 7
- Участие в експертна дейност, рецензиране: 8

Доц. д-р Людмила Петрова Димитрова има цитирани 15 работи в 102 научни труда.

Представени са и отзив-препоръка от ИЛ–САН и 2 писма за участие в Редколегията на международното списание Cognitive Studies/Études Cognitives.

5 Заключение

Познавам кандидатката и работите й. Впечатленията ми от работата на кандидатката са отлични. Искам да изтъкна умението на кандидатката да работи и ръководи международни проекти в областта на информатиката. Давам **положителна оценка** на всички представени материали от доц. д-р Людмила Петрова Димитрова-Рашкова. Считам, че тя напълно удовлетворява всички изисквания на ЗРАСРБ, правилника на МС за прилагането му, правилника на БАН и правилника на ИМИ за условията и реда за заемане на академичната длъжност професор в ИМИ на БАН в област на висше образование 4. Природни науки, математика и информатика, професионално направление 4.6. Информатика и компютърни науки, научна специалност 01.01.12 Информатика (компютърна лингвистика – средства и системи за обработка на лингвистични знания).

София, 1.11.2011

Рецензент:
/проф. дмн Петър Станчев/