Р Е Ц Е Н З И Я

от проф. дмн Галя Младенова Ангелова,
Институт по информационни и комуникационни технологии (ИИКТ) на БАН
в конкурс за заемане на академичната длъжност „професор” по направление
2.1. Филология (Български език – морфология, синтаксис и корпусна лингвистика), обявен от СУ „Св. Кл. Охридски” и обнародван в Държавен вестник, бр. 9 от 2 февруари 2016 г.

На обявения от СУ „Св. Кл. Охридски” конкурс за професор по Филология (Български език – морфология, синтаксис и корпусна лингвистика) се е явил само един кандидат – доц. д-р Петя Осенова. Както се изисква от Чл. 114 на *Правилника на СУ „Св. Кл. Охридски” за условията и реда за придобиване на научни степени и заемане на академични длъжности*, тя има образователната и научна степен „доктор” от 2000 г. и е заемала академичната длъжност „доцент” в СУ „Св. Климент Охридски” в Катедрата по български език, ФСлФ, повече от пет години (от 2009 г.). Чете основни курсове лекции по морфология и синтаксис, съвременен български език и формални граматики. Ръководител е на двама докторанти, един от които е успешно защитил през 2015 г. Със стипендия „Фулбрайт” е прекарала 5 месеца в Станфордския университет, САЩ. Доц. Петя Осенова е изтъкнат специалист по моделиране на езикови явления в българския език чрез формални граматики и методи на компютърната и корпусна лингвистика, с доказани научни постижения в областта. Представила е за конкурса монографичен труд, издаден през 2016 г., както и 43 оригинални публикации в специализирани научни издания (от тях 41 излезли след 2009 г.), които не повтарят представените трудове за придобиване на образователната и научна степен „доктор” и за заемане на академичната длъжност „доцент”. С това формалните изисквания за заемане на академичната длъжност „професор”, указани в горецитирания *Правилник* на СУ „Св. Кл. Охридски”, са напълно удовлетворени.

Съгласно Чл. 116 на *Правилника*, кандидатът за заемане на академичната длъжност „професор” в СУ „Св. Кл. Охридски” трябва да има активна научноизследователска дейност, която се изразява чрез значимостта и актуалността на изследваните научни теми, създаване на ново направление в науката, качеството на представените публикации, чрез участието в научноизследователски проекти, участия с доклади в международни и национални научни форуми и изнасяне на лекции в чуждестранни университети, приложени в практиката резултати от научни изследвания и т.н. Тъй като не мога да оценя учебната дейност на доц. Осенова като щатен преподават в СУ „Св. Кл. Охридски”, в тази рецензия разглеждам извършваните от нея научни изследвания, свързани с формално моделиране на българския език и разработка на компютърни прототипи (езикови технологии) над получените модели с цел създаване на софтуер за анализ и синтез на български текст.

Представените за конкурса монография и 43 научни публикации отразяват научни резултати на автора, свързани с трите направления в обявения конкурс - морфология, синтаксис и корпусна лингвистика.

Монографията „*Граматическо моделиране на българския език (с оглед на обработката на естествен език)*” точно дефинира научните интереси на автора: моделиране на лингвистично знание за българския език, за целите на компютърната лингвистика и построяването на софтуер за автоматична обработка на езика. Това означава, че фокусът на изследванията на доц. Осенова е насочен към езикови явления, които са по-интересни или нестандарни от гледна точка на създаване на алгоритми за тяхното разпознаване или генерация. Също така би могло да означава, че се разглеждат статистически доминиращи езикови явления, тъй като първата цел при създаване на софтуер е да се осигури коректното му поведение във възможно най-голям процент от извършваните обработки. Авторът определя *ресурсната граматика* като *формална граматика, създадена с оглед автоматична обработка на езика*, и разглежда параметрите на граматическо моделиране на българския език с оглед на конкретни теоретични рамки: опорната фразова граматика и граматика на зависимостите в два варианта (стандартизиран и универсален депендентен модел). Монографията не само представя модели на морфосинтактичните и синтактичните явления в българския език, но и показва изградени ресурси, които са достъпни за ползване. Друг ценен аспект на монографията е, че демонстрира приемствеността в разработките на доц. Осенова, което съответства на задължителния итеративен характер на процеса на натрупване на ресурси. Въвежда се (в известен смисъл неявно) и днешната тенденция в компютърната лингвистика ресурсите да се преформатират (транформират) в термините на други формални теории, за да послужат за други цели или да се вградят в многоезикови системи. Струва ми се, че монографията е много ценен източник на сведения: на българските лингвисти тя показва гледната точка на създателите на лингвистични ресурси с оглед постигане на автоматична обработка на езика, а на информатиците показва дълбочината и важността на лингвистичното знание.

Приносите в областта на морфологията са представени в девет от приложените статии. Въвежда се представяне на многокомпонентни думи (multiwords) в българския език чрез описание с катена и се предлага кодиране на тези думи в речника, както и начини за свързването им със синтактичното поведение и семантиката, чрез представяне на взаимодействието между поведението на тези думи в речника и текста. Заключението е, че морфологичните катени, каквито са сложните думи, са стабилни откъм състав, докато при синтактичните може да се вмъква допълнителен материал.

Кандидатът е основен дизайнер на лингвистичното знание при разработка на автоматично средство за морфосинтактичен анализ (т.нар. морфологичен анализатор) и автоматично средство за разрешаване на морфосинтактичната многозначност при морфологичния анализ (т.нар. Part-of-speech tagger, или POS-tagger). Под лингвистично знание тук се разбира речник, правила за последователно срещане на частите на речта в текста, описание на типовете многозначности, система от анотационни маркери за означаване на граматически категории и разработка на корпуси с учебни и тестови данни. Също така доц. Осенова анализира грешките при морфосинтактичния анализ с оглед на лематизацията (привеждане на словоформата в основната форма), и указва начини за редуцирането им чрез добавяне на допълнителни правила или промени в алгоритмите. Описани са основните типове морфологични многозначности от перспективата на отношението между лексема и словоформа на основата на морфологично анотиран корпус. Създаденият през 2012 г. POS-tagger за българския език, описан в статия [26], е един от най-точните в света, което е голямо

постижение особено като се има предвид, че за английски такива програми се създават от около 30 години, а морфосинтактичните характеристики за българския език са 680 на брой като комбинации от част на речта и нейните граматически характеристики. (Ще си позволя отклонение от темата и ще кажа, че по принцип е трудно да се демонстрира иновация в информационна технология, която се развива с десетилетия; но авторите на статия [26] са го постигнали).

Разгледани са и частите на речта в Опорната фразова граматика с акцент върху семантиката, като се показва, че от семантична гледна точка представителите на една и съща част на речта се делят на различни групи, а граматическите характеристики се моделират на две равнища – морфосинтактично и семантично.

Резултатите на автора в областта на синтаксиса са представени в 16 от приложените статии. Дадена е формализация на явлението *катена* като поддърво в синтактично представяне. Представено е моделирането на многокомпонентните думи в синтактичен ресурс (Бултрибанк, като авторът е един от основните разработчици на банката) и в речник. Авторът изучава взаимодействието между представянето на думите в речника и синтактичното им поведение в текста.

В няколко статии се разглеждат въпроси, свързани със синтактичния ресурс Бултрибанк и негови прекодирания в други формализми или влагането му в многоезичен сценарий чрез създаване на паралелни синтактични ресурси. Конвертирането на Бултрибанк от оригиналния конституентен корпус до дependентно представяне позволява създаването на семантични структури на основата на дependентен синтактичен анализ за българския език. Автоматичният модул, който извършва голяма част от прекодирането, също използва правила, създадени с водещо участие на доц. Осенова. Въвеждането на дependентните граматики като формализъм за моделиране на българския синтаксис е сравнително нова теоретична рамка, тъй като дependентният синтаксис не е систематично прилаган за моделиране на български изречения (макар че има някои по-ранни работи, свързани с отделни езикови явления).

Друг принос на автора е разработка на оригинален подход за анотиране на синтактичния корпус Бултрибанк със семантични роли върху аргументите на предиката – подлог, пряк и непряк обект, като семантичните роли са взети от ресурса VerbNet.

Развитието на семантичното описание позволява извличане на валентен речник от синтактично анализирания корпус. Този речник е свързан с онтология. Показани са най-честите валентни типове в ресурса и най-честите онтологични ограничения (*лице, обект, събитие, място и пр.*) върху аргументите на предиката.

С цел създаване на български модул в система за многоезичен машинен превод е разработен българско-английски паралелен синтактичен корпус на основата на Бултрибанк. Създаването на корпуса включва избор на анотационна схема и дизайн на правила за „подравняване” (построяване на преводно съответствие) между думи и изрази на двата езика.

Създаден е и синтактичен ресурс в определена тематична област (текстил и вътрешен дизайн) за целите на конструиране на подходяща онтология в същата област чрез извличане на подходящи релации.

Научните резултати на кандидата в областта на корпусната лингвистика са представени в 20 от приложените статии. Тематиката им покрива резултати, свързани със създаване на корпуси или резултати от корпусно базирани изследвания вкл. машинен превод; анотиране на корпуси с различни цели вкл. семантична анотация на Бултрибанк; подготовка на данните и оценката на резултатите при различни модули за автоматична обработка на българския език.

Специално бих желала да отбележа създаването на „българския pipeline"-последователност от софтуерни компоненти за обработка (наречен „цялостен модул за автоматична обработка на български текстове"), включваща разпознаване на словоформите в текста, разделяне на изречения, морфологичен анализ, лематизация и снемане на морфосинтактичната многогозначност и автоматичен синтактичен анализ. В набора компоненти има автоматичен модул за разпознаване и категоризиране на собствени имена на български език чрез използване на богата лингвистична информация. Правени са експерименти с кореферентно свързване между съседни изречения след извършване на автоматичен синтактичен анализ. Създадени са корпуси и речници в отделни предметни области (политически жаргон). Няколко разработки, представени в приложените статии, са свързани с многоезикови задачи по специфични проекти – например машинен превод.

В плановете за бъдеща работа на автора, изброените разнобразни дейности се свързват в обща кохерентна задача: разработка на лингвистичното осигуряване на базисния пакет от езикови ресурси и технологии за българския език, в рамките на европейската инфраструктура CLARIN. Необходимите модули за изпълнение на ангажиментите на България в тази инфраструктура са вече налице, като прототипи или завършени компоненти, и основната задача сега е свързана с тяхната интеграция, създаване на уеб-достъп към тях, адаптиране към различни групи потребители и т.н. Във всички споменати дотук разработки **доц. Осенова е основен дизайнер на лингвистичното знание и осъществява или ръководи процеса на натрупване на необходимите лингвистични ресурси**. Тъй като дейностите се развиват след 2009 г., приемам, че те са извършени след хабилитацията й за доцент и са информационен (електронен) продукт, наличен като резултат от нейните научни изследвания, представени в приложените за конкурса публикации. Същевременно трябва да отбележим, че представените за конкурса монография и 43 публикации отразяват научни постижения на автора през последните 6 години, които не са изчерпателни от гледна точка на създаваните езикови модели. Настоящите резултати надграждат постигнатото преди 2009 г. и са основа за бъдещи постижения.

След запознаване с трудовете не мога да не спомена референциите към многобройни актуални източници с автори световно-признати чуждестранни учени. Списъкът от 69 цитирания за периода 2009-2016 г. показва, че публикациите на кандидата са интересни и полезни за множество експерти от целия свят. Изявите на доц. Осенова при лекции и презентации на национални и международни научни форуми показват, че тя е постигнала над-езиково ниво на абстракция при моделиране на морфологични и синтактични явления и постиженията й като разработчик на лингвистични ресурси се оценяват по достойнство от чуждестранни експерти. Вече сме забравили годините, в които разработките (демонстратори, proof-of-concept prototypes) се създаваха на единствен персонален компютър с речници-играчки от няколкостотин езикови единици. В европейската инфраструктура CLARIN се

отчита, че българският не е сред езиците, които са „бедни откъм ресурси”. В съществена степен това се дължи на целенасочените усилия на доц. Петя Осенова (и доц. Кирил Симов, с когото си сътрудничат от десетилетия). Със задоволство отбелязвам, че по този начин следващото поколение български компютърни лингвисти успешно се вписва в европейския контекст и се надявам да остане там десетилетия, уважавано за непрестанния си труд, упоритост и постижения независимо от оскъдното финансиране. Също така вярвам, че до няколко години ще има изграден парсер (автоматичен синтактичен анализатор) за български текст, публично достъпен както са парсерите за широко разпространените езици – английски, немски, испански и др., който ще позволи да се създават алгоритми и софтуер за семантичен анализ на български текст.

Заключение

Съдържанието на представените за конкурса материали отговаря на изискванията, определени в *Правилника на СУ „Св. Кл. Охридски” за условията и реда за придобиване на научни степени и заемане на академични длъжности*. Вижда се, че доц. Осенова не се занимава с дребни или периферни научни задачи, а се насочва към последователно създаване на формална граматика на българския език, която да бъде стане основа за разработка на автоматичен анализатор за български текст. Считам, че научните резултати на кандидатката, публикуваните статии, цитатите, участието в проекти с лидерска роля, изнесените у нас и в чужбина презентации и лекции доказват наличието на задълбочени знания, качества за ръководене на групи в проекти, способност за формиране на творчески колективи, водеща роля при формулиране на амбициозни изследователски цели, както и постоянство, прецизност и стремеж към достигане на световното ниво, които се предполагат като присъщи на "професор" в най-добрия български университет. **Подкрепям убедено избора на доц. Петя Осенова за "професор" по направление 2.1. Филология (Български език – морфология, синтаксис и корпусна лингвистика) и предлагам на членовете на Научното жури единодушно да гласуват в подкрепа на такова решение.**

5 юни 2016 г.
София

Член на Научното жури
за процедурата:
проф. дмн Галя Ангелова