РЕЦЕНЗИЯ

от проф. дмн Галя Младенова Ангелова, ИИКТ-БАН

за дисертацията на Каменка Атанасова Стайкова

"*Лингвистични и семантични ресурси при компютърно генериране и анотиране на български текстове*"

представена за присъждане на образователната и научна степен "доктор"

Дисертацията е свързана с някои от най-актуалните тенденции в компютърната лингвистика:

- поставяне във фокус на ресурсите, които позволяват реализация на системи за автоматична обработка на текста с по-високо качество;
- създаване на ресурси за естествени езици, говорени от сравнително по-малко хора, за които има недостатъчно лингвистично осигуряване (т.нар. less resourced languages – какъвто се явява и българският език); и
- експерименти с платформи за многоезиков анализ или генерация, с цел установяване на стандарти за обработка на много естествени езици едновременно.

Дисертацията съдържа резултати, получени при съвместна работа с водещи европейски експерти в областта, и допринася за вграждането на компоненти за обработка на български език в известни платформи за компютърна генерация на текст.

Съгласно Правилника за специфичните условия за придобиване на научни степени и за заемане на академични длъжности в Института по информационни и комуникационни технологии (ИИКТ) при БАН, кандидатът за получаване на образователната и научна степен "доктор" трябва да има "*поне 3 научни публикации, поне една от които да е в списание с импакт фактор или в специализирано международно издание*". Резултатите от дисертацията са представени в 12 публикации, като:

- една от тях е в серията Lecture Notes на Шпрингер;
- една е в Сборника трудове на Световната конференция по компютърна лингвистика COLING;
- три са в списанието Cyberbetics and Information Technology;
- две са в Сборници трудове на конференцията RANLP и свързани с нея мероприятия;
- две са в Сборници трудове на конференцията CompSysTech;
- една от тях е в списание "Български език" (съвместно с проф. Й. Пенчев).

Две от статиите в списание Cyberbetics and Information Technology и един технически отчет са самостоятелни публикации на автора. При повърхностно търсене с Google Scholar се забелязват поне 5 цитирания на статията, публикувана в Сборника трудове на Световната конференция по компютърна лингвистика COLING. Така изискванията на Правилника на ИИКТ-БАН към кандидатите за получаване на образователната и научна степен "доктор" са изпълнени.

Трудът съдържа 129 страници текст и 50 страници приложения. Текстът е организиран в увод, 4 глави и заключение, списък на използваните термини, съкращения и означения, както и библиография с 81 заглавия. Петте приложения съдържат модел на предметната област, в която работи системата за автоматична генерация; формално представяне на процедурен текст; онтологии в областта на иконографията; анотационни граматики и резултати от семантичното анотиране.

Уводът представя актуалността на темата и мотивира целите и задачите на дисертацията. Главната цел е реализация на компютърно генериране на български текст чрез създаване на лингвистични ресурси за основни обекти на българския език и на семантични ресурси за конкретни предметни области. В Глава 1 е направен подробен обзор на подходите за компютърно генериране на текст на естествен език, като са въведени и лингвистичните ресурси, необходими в процеса на генерация. Тази глава съдържа и кратко описание на семантичните технологии, използвани при обработка на естествен език.

В Глава 2 "Подготовка на лексико-граматически ресурс за компютърно генериране на български текстове" (с обем 46 страници) се представят резултати, свързани с подготовка на лингвистичен ресурс, който се зарежда в платформата KPML за многоезикова генерация – след което тя може да генерира български текст паралелно на руски и чешки текстове. Предметната област, в която се извършват експериментите, е генериране на CAD/CAM документация. Езиковите явления се формализират и описват чрез системично-функционалната граматика. Работи се със сравнително малък корпус (достатъчен за целите на проекта) от 194 изречения с 1219 думи. Определени са граматическите конструкции, използвани в изреченията (които са подробно анализирани в текста на труда). Предложен е начин за описване на явленията в термините на системично-функционалната граматика. Разработени са формалните декларативни спецификации на основни езикови явления, необходими за реализиране на генерация на документация (т. нар. instructional texts). В секция 2.2 подробно се показва начинът за създаване на лингвистичния ресурс за български език при интерактивна работа със средата KPML. Чрез примери е илюстрирано генерирането на изречения, включващи разнообразни езикови явления, в това число: изречения в деятелен и страдателен залог, с глаголи от свършен или несвършен вид, в заповедно или разказно наклонение, номинализация на глаголната група, съгласуване по род и число в номинални фрази, членуване, както и номинализация на глаголната група. Въведена е "да-конструкция" за генериране на модалните глаголни групи "мога да ...", "трябва да ...". Текстовете се генерират в сегашно време. Показана е генерация на кохерентен дискурс за постъпков процес с вметнати изрази, напр. *първо - след това – накрая*. Генерират се и сложни изречения главно-подчинено, например "*Натиснете Return, за да затворите полилинията*". В тази глава е показано, че създаденият формален модел на български граматични явления е достатъчно голям, за да се генерират сравнително разнообразни изречения и по този начин да се подготви основа за генерация на процедурни текстове. Както е написано в дисертацията, граматиката е „прагматично ограничена“ да покрива конкретния стил на текстовете от корпуса, но от друга страна, ресурсът е отворен за повторно използване и по-нататъшно

развитие. Много добро впечатление прави съвместната работа с проф. Йордан Пенчев, водещ специалист по формално моделиране на българския синтаксис с конституентни граматики, и съвместната публикация на автора с него в списание "Български език".

В глава 3 "Реализация на процеса за генериране на български текстове" се разглежда процесът на автоматично създаване на текстове-инструкции на български език. Генерацията се извършва от системата AGILE, която е разработена от международен научен колектив в рамките на проект, финансиран от Европейската комисия в програмата Коперник. Именно тази система използва граматичния ресурс за български език, представен в Глава 2, и генерира текстове на български, руски и чешки език. Генерацията работи над база знания от декларативно представени понятия, свързани с Обобщения модел (понятийна йерархия от високо ниво). Понятията от предметната област се описват като *процеси*, *неща* и *елементи, структуриращи текста* (процедури, методи, списъци от методи и списъци от процедури). В процеса на работа по проекта е направен анализ на понятията от предметната област на CAD/CAM приложенията, като изграденият понятиен модел е свързан директно с Обобщения модел, който съответства пряко на Системично-функционалната граматика. В секция 3.2 е описан подходът за избор на съдържанието (content selection, what to say), който изцяло се управлява от потребителя, задаващ във входния интерфейс на системата AGILE набор от слементи, структуриращи текста, и набор от фокусирани понятия. Този процес се нарича "затворено планиране". Повърхнинната реализация (how to say it) се осъществява с помощта на текстови шаблони, които дефинират стила на текста в термините на лексико-граматичния модел. След анализ на текстовия корпус в предметната област са забелязани редица особености на стила: явно указване на агента, извършващ действието; различни начини за изразяване на връзката с читателя (явно адресиране или не); изрази за реализиране на инструкции и различни по сложност лингвистични изрази. Въведени са два стила на шаблоните: персонален императивен и безличен изявителен. Получените резултати са демонстрирани при генерирането на кохерентен текст. Подкрепям твърдението на автора, че те са уникален опит за автоматично генериране на български текст чрез използване на една от най-развитите (и вече класически) платформи за многоезикова генерация. Тук бих добавила, че личните ми резерви към сложността на интерфейса (в който потребителят задава почти ръчно комуникативната цел и фокуса на бъдещия дискурс) не касаят постиженията и качеството на настоящата дисертация.

В Глава 4 "Работа с български текстове, семантични технологии и ресурси" се разглеждат приложения на семантичните технологии за индексиране на специализирани текстове на български език, които биха подпомогнали не само процеса на компютърно генериране, но също така и процеси на семантично търсене и управление на метаданни в дигитални архиви. В секция 4.1 е представено формализирано описание на знания в областта на иконографията, структурирано като "Онтология на български иконографски обекти" (ОБИО). ОБИО е съвместима в CIDOC-CRM, модел стандартизиран от Международния съвет на музеите, и се използва при създаване на семантични модели на мултимедийни учебни обекти в платформа, разработена по проекта СИНУС, финансиран от Националния Фонд за научни изследвания. В секция 4.2 е предложен подход "от онтология към текст" за семантично анотиране на

специализирани български текстове с платформата CLaRK. Създадени са две граматики от регулярни изрази и маркери: за разпознаване на индивидите и на онтологичните класове в предметната област на проекта СИНУС. След проведени тестове с корпус от предметната област е показано, че при обучението на граматиките са постигнати точност 0,98% и покриваемост 0,82%. Реализираната процедура за автоматично индексиране разпознава най-често срещаните онтологични термини в текста и създава адекватни семантични анотации.

Заключението резюмира коректно научните и научно-приложни приноси на дисертацията, като изброява резултатите, представени в дисертацията. Авторефератът отразява коректно съдържанието на труда.

Дисертационният труд е прегледно подреден и ясно написан. След няколко редакции изложението в Глави 1-3 прави много добро впечатление: многобройните примери и илюстрации в текста помагат на читателя да разбере същността на разработката, което не е лесно в специфичния (и сравнително затворен) свят на KPML. Резултатите са свързани с работата на автора по два проекта: AGILE и СИНУС. По темата на дисертацията са изнесени презентации на седем научни събития, едно от тях в чужбина.

Бих направила и следните коментари и бележки:

Жалко е, че след завършването на проекта AGILE авторът не е успял да продължи разработките си в областта на компютърната генерация на български език, което според мен се дължи отчасти на трудностите да се работи самостоятелно с толкова голяма платформа, базирана на декларативни модели, а също така и на прехвърляне на интереса – в световен мащаб - към усъвършенстване на методите за извличане на информация от текста чрез частичен анализ и статистически методи. В този контекст новите разработки върху компютърната генерация, която остава една базирана върху правила технология, се броят на пръсти през последните 15 години. Може би този факт обяснява и липсата на ясно изразени солидни намерения за бъдеща работа на автора в областта на генерацията, а са изказани само общи намерения в "безличен стил" (ако използваме терминологията за стила на шаблоните от Глава 3).

Освен това, при колективна работа в проекти с много изпълнители (каквито са AGILE и СИНУС) би следвали по-ясно да се разграничат оригиналните приноси на автора. Но от материалите по дисертацията е видно, че г-жа Стайкова несъмнено е компютърният лингвист, който е "превел" лингвистичните закономерности на българския език в термините на системично-функционалната граматика и е направил възможно функционирането на българския компонент на системата AGILE. Приемам това като неин личен принос в проекта и настоящата дисертация, а също и като доказателство за компетентността й в областта на компютърната генерация на текст, а също и като заслуга за успешното завършване на проекта AGILE.

Заключение

Считам, че получените резултати и публикуваните статии доказват наличието на експертиза и качества за извършване на самостоятелна научна и научно-приложна работа, които се изискват от ЗРАСРБ за присъждане на образователната и научна степен "доктор". **Подкрепям с положително заключение присъждането на образователната и научна степен "доктор" на г-жа Каменка Стайкова и предлагам на членовете на Научното жури единодушно да гласуват в подкрепа на такова решение.**

22 юни 2015 г.
София

Член на Научното жури