Рецензия
за дисертационния труд на Ивелина Стоянова
на тема „Автоматично разпознаване и тагиране на съставни лексикални единици в българския език”

Рецензент: доц. д-р Йовка Тишева

Предложеният за защита труд на Ивелина Любенова Стоянова на тема „Автоматично разпознаване и тагиране на съставни лексикални единици в българския език” е подготвен след обучение в задочна форма на докторантура, за срок от 4 години (от 1.01.2008 до 1.01.2012 г.), по специалността „Общо и сравнително езикознание (математическа лингвистика)” за нуждите на Секцията по компютърна лингвистика на ИБЕ „Проф. Любомир Андрейчин”. Научен ръководител е проф. Светла Коева.

При подготовката на докторантката са спазени изискванията по процедурата за придобиване на научна степен „доктор” (чл. 6 (1) и (4) от Закона за развитието на академичния състав в Република България). Ивелина Стоянова е завършила специалността „Българска филология” (бакалавър) в СУ „Св. Климент Охридски” и има магистърска степен „Българска филология – компютърна хуманитаристика” от Факултета по славянски филологии на СУ. По време на докторантурата си, освен задължителните дисциплини и изпити по специалността, е завършила успешно и два курса в Центъра за обучение на БАН – „Формална семантика” и „Основи на логиката”.

Спазени са и нормативните изисквания, свързани със структурата на дисертационния труд. Разработката на Ивелина Стоянова (с обем от 134 стр. основен текст според предоставения ми за рецензия екземпляр с дата 17 април 2012 г.) се състои от увод, четири глави и заключение, в което са представени резултати, изводи и насоки за бъдещи изследвания по темата. Библиографията към разработката – 123 единици, включва заглавия на източници на английски и на български език. Към труда са посочени и две приложения: приложение А, включващо пример за предварителна обработка на файл, съдържащ съставни лексикални единици – тагиране и окончателно XML форматиране, и приложение Б - списък на електронните ресурси: около 27 000 единици от два основни източника – *Български тълковен речник* (1994) и *Уикипедия*.

В уводната част на работата дисертантката много точно и убедително посочва актуалността и значимостта на избраната научна проблематика. Статистически данни показват, че лексикалните единици, състоящи се от две или повече графични думи (наричани тук по-общо *несвободни фрази*), са значителна част от лексикалната система на езика: за WordNet (версия 1.7) те са 41%, за българската семантична мрежа БулНет – 24.49%. Автоматичната обработка на тези единици обаче е свързана с редица проблеми – фактът, че този тип лексеми имат освен семантична и своя синтактична структура, затруднява тяхното разпознаване и анотиране, автоматичното им тагиране, машинния превод, автоматичното резюмиране на текст и т.н. Самостоятелното теоретично изследване на съставните лексикални единици, освен че ще очертае техните специфични

особености, ще бъде и надеждна основа за разработване на адекватни компютърни инструменти за автоматичната им обработка.

Поради факта, че досега в българската компютърна лингвистика не е обръщано такова специално внимание на съставните лексикални единици, изследването на Ивелина Стоянова съвсем основателно може да се определи и като новаторско. Компютърните речници на българския език разпознават простите лексеми (от една графична дума) и ограничен брой устойчиви словосъчетания. Разпознаването на елемент от текста, който има своя синтактична структура – тоест представлява фраза от две или повече графични думи, но чието значение не е „сбор" от значенията на изграждащите го компоненти се свързва с отстраняване на многозначността: избор на едно от значенията и отхвърляне на т.н. „свободни" значения на компонентите. Дисертантката посочва, че разработването на методи за разпознаване и тагиране на съставните лексикални единици ще допринесе за по-точната автоматична обработка на равнището на текста.

Авторката точно и ясно формулира целите и конкретните изследователски задачи, които си поставя за постигането им. Основните цели са изграждането на теоретичен модел за автоматично разпознаване и тагиране на съставни лексикални единици и практическото му приложение върху конкретен материал. Двата аспекта в работата – теоретичен и практически, са с еднаква тежест в изследването: теоретичният модел трябва точно да отразява езиковите факти; приложението му трябва да доведе до качествени резултат. Наблюденията и анализите са направени върху богат и надежден корпус от разнообразни писмени текстове, който включва примери от Българския национален корпус, чрез които се илюстрират определени характеристики на съставните лексикални единици, и от статии от *Уикипедия* (с обем от общо 41 милиона думи).

Авторката си поставя пет конкретни научни задачи:

- дефиниране на обема и съдържанието на понятието съставна лексикална единица и анализ на гранични случаи, които затрудняват автоматичната обработка (разработена във II глава от дисертацията)
- представяне на морфо-синтактични, семантични, синтактични и прагматични особености на съставните лексикални единици с цел създаване на теоретичен модел за представянето им (в III глава, с уточнението, че се разглеждат само именните фрази)
- класификация на съставните лексикални единици, ориентирана към практическо приложение при описанието и анализа с цел автоматичното им разпознаване (предложена в III глава, но отново с уточнение, че обхваща само групата на именните фрази)
- разработване на методология за автоматично разпознаване на съставните лексикални единици и отделните им категории (разработена в IV глава)
- приложение на разработената методология, анализ на резултатите и оценка на възможностите ѝ (описано в V глава).

Веднага искам да отбележа, че всички изследователски задачи са успешно реализирани. Това твърдение е мотивирано с преглед на основния текст на

дисертацията, в който представям конкретните постижения на Ивелина Стоянова във връзка с всяка от посочените задачи.

Интересът към комбинациите от думи, които се срещат в близост една до друга и имат сравнително висока повторяемост, датира от 50-те години на XX век. Традицията в изследването на този езиков феномен, означаван като колокация, единица/лексема от повече от една дума, фразема, свързана фраза и т.н., в англоезичното езикознание е сравнително дълга. В българската лингвистика те присъстват обикновено при системното описание на лексикалната система, в противопоставянето между свободни и несвободни словосъчетания (фрази). Дисертантката добре се ориентира в терминологичното разнообразие и това личи както при прочита на различни мнения (между които тези на Круз; Синклер; Стъбс; Нунберг, Саг и Уасоу; Мелчук; Муун; Еверт), така и в много точната систематизация на използваната в дисертацията терминология и онагледяване й по най-подходящия начин (таблица 2.2., стр. 16). Прегледът на досегашните изследвания по въпроса води до извода, че при описание на тези гранични елементи на речника са възможни два подхода: теоретичен, при който се търси мястото на несвободните фрази в лексикалната система на езика (напр. в Академичната граматика 1983, Мелчук 1995) или подход, ориентиран към употребата на тези фрази в текстове или корпуси (Синкрел 1998; Мелчук 1998). В тази връзка убедително е мотивиран изборът на ключовия за изследването термин *съставна лексикална единица* пред разложима фраза, фразема или съставна дума.

Тъй като съставните лексикални единици са част от системата на несвободните фрази, авторката тръгва именно от по-общото – описание и класификация на несвободните фрази, за да стигне до подредните елементи, изграждащи тази част от речника. Детайлно са представени класификациите на фразеологичните словосъчетания, изрази и названия в Граматиката на СБКЕ (1998), класификацията на Мелчук (1995) на фраземите, на Болдуин и кол. (2003). В дисертацията се приема последната класификация - според нея несвободните фрази биват неразложими, идеосинкретично разложими и прости разложими - обогатена с още два вида фрази: съставни лексикални единици със служебна функция (съставни съюзи и предлози) и статистически маркирани фрази. Така авторката логично достига до класификацията на лексикалните единици в българския език:

- прости лексикални единици (състоящи се от една дума) – думи, думи с частици и сложни думи (една графична дума, чието значение е формирано от повече от един пълнозначен компонент)
- несвободни фрази.

Имам два въпроса във връзка с тази класификация. Тъй като няма точен паралел между двете групи единици в лексикалната система – прости (лексикални единици) : (несвободни) фрази, дали вече след известно дистанциране от текста не може да се предложат и други названия за тези групи? Сложните думи се оказват в малко по-различно положение спрямо останалите два компонента на първата група заради различните правописни конвенции (слято или разделно писане ва сложни думи). Ако

те се изписват винаги слято, твърдението, че разпознаването им в текста не е проблем (стр. 42), е вярно. Ще бъде ли затруднено автоматичното им разпознаване, ако са изписани разделно, по модела на съставните лексикални единици, но не притежават същата маркираност (например поради „дефективност" по число на първия елемент)? Посочено е, че границата между сложните думи, изписани разделно, и съставните лексикални единици е много тънка (стр. 41); това личи и в примерите за фрази по модела NN – *къща музей* и *бизнес център* (стр. 60); възможно ли да се формулират правила или алгоритми, които да разграничават двете групи?

Очертаването на мястото на съставните лексикални единици в общата лексикална система води авторката до стесняване и конкретизация на обема на понятието чрез няколко предложения за дефинирането му. В основата им са типовете маркираност, посочени от Болдуин при съставяне на модел за описание на лексикалните единици: лексикална, синтактична, семантична, прагматична и статистическа маркираност. За изясняване на същността на съставните лексикални единици е важно да се посочат процесът на конструиране на значението чрез съчетаване на значенията на конституентите и възможностите за възстановяване на значенията на отделните конституенти от значението на фразата. Композиционалността, разбирана като степента, до която характеристиките на компонентите на фразата комбинирано обясняват характеристиките на цялата фраза, е определена от авторката като основна характеристика на съставните лексикални единици. При това композиционалността покрива всички аспекти на маркираността без статистическата. Другият аспект при характеристиката се свързва с разложимостта, тоест степента, до която особеностите на фразата могат да се обяснят чрез характеристиките на съставящите я компоненти. Ивелина Стоянова извежда четири дефиниции за съставните лексикални единици. В дисертацията се работи основно с дефиниция 3., определена като теоретична (стр. 38), защото обобщава наблюденията върху езиковите характеристики на съставните лексикални единици, и дефиниция 4 – с практически характер (стр. 39), представяща характеристиките на същите единици при реализацията им в речта, поради което за тях могат да бъдат намерени методи за количествена и качествена оценка (представени подробно в IV глава).

Лингвистичното описание и класификацията на съставните лексикални единици са представени в трета глава от дисертацията. Тук авторката е направила уточнението, че по-нататъшната ѝ работа ще се свързва само с именните фрази в групата на съставните лексикални единици, като мотивира това, от една страна, с факта, че този тип фрази имат най-значителен дял в корпуса, и от друга – с желанието си за по-голяма целенасоченост при съставянето на теоретична рамка, която да позволи разработването на методи за автоматично разпознаване и тагиране. Тези уточнения биха могли да бъдат заявени по-ясно още в началото на работата, при представяне например на нейната структура или на корпуса, върху който се базират анализите.

Лингвистичното описание на съставните лексикални единици следва представяния вече във втора глава модел на Болдуин (2004), който се основава на видовете

маркираност на фразите. При изясняване на семантичните особености отново е дискутиран и въпросът за композиционалността и възможностите за вариране при изграждане на значението. При синтактичното описание авторката прави още едно стесняване на обема на разглежданите обекти, този път свързано с дължината на именните фрази – до пет думи. Подробно са разгледани трите начина за образуването на тези съставни лексикални единици: чрез присъединяване на предпоставено съгласувано определение прилагателно име към опората, чрез задпоставно определение предложна фраза и чрез приложение. Интересни са коментарите за семантичните и граматичните ограничения върху подчинените части и прилагането на тези знания при разработването на методите за автоматична обработка. Прагматичната маркираност на разглежданите фрази се проявява в тяхната обвързаност с определена комуникативна ситуация, отношения между изказванията в дискурса или между участниците в него. Този тип маркираност се разглежда и във връзка със степента, до която значението на съставната лексикална единица се формира с участие на референции към обекти от извънезиковата действителност. Като най.съществено за формообразувателните особености на представяните фрази се извеждат парадигматичните ограничения, най-често във връзка с категорията число. Включването в параграфа за прагматичните особености на класификацията на наименованията, спомената и в предходни части, като че ли остава малко встрани от основната проблематика на раздела. Дискусията за това дали съставните лексикални единици трябва да се включват в речника, с която приключва прегледът на лингвистичните особености, би могла да се ориентира към по-общите части от работата, още при изясняване на мястото на тези елементи в групата на несвободните фрази, тъй като посочените аргументи се отнасят не само до фразите от типа NP. Иначе напълно споделям мнението на авторката по въпроса и приемам посочените от нея аргументи (стр. 67).

Преди да пристъпи към класификацията на съставните лексикални единици, Ивелина Стоянова прави преглед на три различни подхода: семантичен – според отношенията между опората и останалите конституенти, синтактичен - според синтактичната структура на фразата, и идиоматичен – според степента на идиоматичност. Най-голяма внимание е отделено на семантичната класификация. Тя е представена в различни варианти: чрез пренасяне на детайлната класификация на наименованията (стр. 45-46); чрез седемте семантични релации, представени в WordNet и илюстрирани с примери в таблица 3.2 (стр. 69); чрез видовете номинацизации (стр. 70-71). Нито една от тези класификационни рамки обаче не обхваща всички особености на примерите от корпуса, затова авторката допълва семантичните релации и предлага свой модел, основаващ се на значението наподчинената част и отношението с главната част в съставната лексикална единица. При останалите две класификации се подчертава, че те имат приложна (практическа) насоченост във връзка с оптимизиране на автоматичното разпознаване на разглежданите фрази в текст.

Във II и III глава от дисертацията личат задълбочените теоретични познания на Ивелина Стоянова в областта на лингвистичното описание на изследнатите обекти. В IV глава тя показва, че е много добре запозната и с използваните методи за разпознаването на несвободните фрази. Изборът на собствена методика за автоматична обработка логично е предхождан от преглед на трите групи най-широко прилагани методи и анализ на факторите, които влияят върху ефективността им. Лингвистичните методи се основават на разработени вече езикови ресурси и модели (списъци от думи, речници, лексикално-семантична мрежа, системи от правила, описващи явлението). Тъй като са свързани най-вече с лексикална информация, те имат ограничено приложение в конкретната разработка, наапример за описание на отделни групи от фрази. Количествените методи – представени като алтернатива на лингвистичните, не изискват предварително разработен езиков ресурс, но пък и не дават особено добри резултати, затова авторката прави извода, че се налага по-широко използване на хибридни методи. Прилагането на синтактичен филтър ограничава възможните кандидати, което освен че намалява времето за автоматично извличане на съставните лексикални единици, повишава точността на резултатите. За разграничаване на съставните лексикални единици от другите два вида несвободни фрази особено подходящо се оказва векторното представяне на значението.

В последната част от IV глава авторката обективно преценява възможностите за автоматичното разпознаване на изследваните фрази с прилагането на формулираните в предходните две глави дефиниции, направените сравнения, напр. със свободните колокации, и представените хибридни методи. Така последната изследователска задача се конкретизира до разработване на методи за автоматично разпознаване на несвободни фрази, които в съставения корпус са предимно съставни лексикални единици, а разграничаването на отделните категории в групата на несвободните фрази е една от перспективите за продължаване на изследването. Малкият брой на неразложимите и на идеосинкретично разложимите фрази не позволява качествено приложение и оценка на методи за тяхното разпознаване.

Приложната част от работата на дисертантката обхваща поредица от експерименти, демонстриращи възможностите на разработената комплексна методика върху конкретен корпус – Уики1000+. Експериментите са проведени със специално разработена за целите на дисертацията компютърна програма bgMWE, която включва инструменти за обработка на текстови корпуси, разпознаване и тагиране на съставни лексикални единици и отчитане на резултатите. Тази програма може да се прилага за цялостна обработка и анализ на езикови корпуси – преобразуване на формата на текстовете, тагиране, прилагане на лексикални ресурси, анализ и оценка на резултатите и т.н. Всеки модул от нея може да се използва като отделна програма за изпълнение на конкретна задача. Освен да се прилагат последователно, в някои случаи могат и да се интегрират. За разпознаването на отделните групи несвободни фрази са предложени конкретни правила и е разработен опростен алгоритъм за автоматично определяне.

В последвата част от дисертацията Ивелина Стоянова обобщава постигнатите резултати и достига до извода, че задачата за разпознаването и тагирането на съставните лексикални единици изисква комбиниране на лингвистични и статистически методи. Посочени са пет приноса на разработката, които отговарят на теоретичните анализи и проведените експерименти.

Авторефератът към дисертацията обективно и изчерпателно запознава със структурата и съдържанието на представения за защита труд.

Авторката има необходимия брой самостоятелни публикации по темата на дисертацията. Препоръчвам към публикациите в съавторство да се приложат документи, посочващи дяловото участие в тях на всеки от авторите (по раздели, страници и под.).

Заключение: Дисертационният труд на Ивелина Стоянова „Автоматично разпознаване и тагиране на съставни лексикални единици в българския език" съдържа значими резултати в теоретичен аспект, свързани с дефинирането, характеристиката и класификацията на несвободните фрази в българския език и на съставните лексикални единици в частност. Разработената комплексна методика за автоматично разпознаване и тагиране на тези фрази, както и на система от компютърни модули за комплексна обработка на корпуси са важни научноприложни резултати от работата на дисертантката. Дисертационният труд показва, че Ивелина Стоянова притежава задълбочени теоретични знания в областта на лингвистиката и на компютърната лингвистика, които умее да прилага за постигане на конкретни научноизследователски задачи. Всичко това мотивира моята положителната оценка по процедурата.

Рецензент:

доц. д-р Йовка Тишева
Катедра по български език, ФСФ
СУ „Св. Климент Охридски"