РЕЦЕНЗИЯ

за дисертационния труд
Автоматично разпознаване и тагиране на съставни лексикални единици в българския език
на **Ивелина Любенова Стоянова**,
представен за придобиване на образователна и научна степен „**доктор**”

от **проф. д-р Тинко Величков Тинчев**
Софийски университет „Св. Климент Охридски”,
Факултет по математика и информатика

Кратки сведения за Ивелина Любенова Стоянова. Завършва през 2001 год. СУ „Св. Климент Охридски”, факултет по славянски филологии, като бакалавър със специалност Българска филология, а през 2002 год. придобива магистърска степен, спец. Българска филология – Компютърна хуманитаристика. Със заповед 47-2/04.02.2008 год. на директора на ИБЕ „Проф. Любомир Андрейчин”, БАН е зачислена за задочна докторантка по специалността „Общо и сравнително езикознание” (математическа лингвистика), шифър 05.04.11; научен ръководител ст.н.с. д-р Светла Коева; срок на обучение 01.01.2008-01.01.2012 г. Със заповед 357-1/23.11.2011 год. на директора на ИБЕ „Проф. Любомир Андрейчин”, БАН е отчислена с право на защита. В периода на обучение е положила успешно изпит по специалността (отличен 5.50), изпит по английски език (отличен 6), тест WORD (отличен 6) и два изпита с оценка отличен 6 по интердисциплинарен (Основи на логиката) и специализиран курс (Формална семантика) в рамките на проект BG051PO001-3.3-04/27/28.08.2009 г. „Математическа логика и компютърна лингвистика: развитие и взаимно проникване” от ОП „Развитие на човешките ресурси”, признати със сертификати от ЦО – БАН.

Представеният ми за рецензиране дисертационен труд (155 стр.) и автореферат (36 стр.) са изготвени и оформени съгласно изискванията на Правилника за прилагане на ЗРАСРБ, чл. 27, ал. 2. Цитирани са 123 източника, от които 90 са англоезични и 9 елентронни източника за компютърни системи. В автореферата са отбелязани само онези 34 цитирания, които се споменават в него.

Тема на изследването. Обект на изследването са така наречените съставни лексикални единици (СЛЕ) с фокус именните фрази. За успешното изучаване, проведено в дисертационния труд, се върви от по-общото (несвободните фрази) към частното (СЛЕ). Теоретичните изследвания визират въпросите от преглед на наличните теории и терминология до класификации, основани на различни признаци. Тук целта е да се използват, доколкото е възможно, алгоритмично разпознаваеми признаци. Така във втората част на глава 5 намираме описана вече разработената част от компютърна програма, реализираща част от теоретичните резултати. (Работата по програмата е в прогрес.)

Актуалност на темата. Проблемите, свързани с несвободните фрази и в частност със съставните лексикални единици, са както теоретични, така и приложни практически. Основен стимул за активните теоретични изследвания в областта в световен мащаб, освен логиката на вътрешно обусловеното развитие на езикознанието, са нуждите на компютърната лингвистика, които пък са стимулирани от нарасналите (и нарастващи) изчислителни ресурси на компютърната техника и интереса на обществото от информационни продукти, подпомагащи ежедневната му комуникативна дейност, включително и непосредственото му общуване с различни устройства на езиково ниво. Тук имам предвид многообразието от специфични разновидности на задачи за разпознаване на човешка реч, автоматичен превод, автоматично извличане и анализиране на информация и редица други. Проблемът е далеч от окончателно решение и, най-общо казано, се корени в алгоритмичността на пресмятането на нерегулярността на означеното (смисъла) с крайна редица от думи, разглеждано като функция на означените с тези думи и техните лингвистични характеристики. Досегашните изследвания на българския език в тази насока са малко и имат фрагментарен характер. От друга страна, секцията по компютърна лингвистика (СКЛ) към ИБЕ, БАН, има задачи, чието успешно решаване е до голяма степен зависимо от правилното разпознаване на съставните лексикални единици. Така, постигнатите с този дисертационен труд теоретични и практически резултати в областта естествено се вписват в разработваната у нас и в света проблематика. Отбелязаната връзка на изследванията на СЛЕ със задачите на СКЛ имат и обратна страна - за успешното им провеждане е необходима не само подходяща творческа среда, но и богат и разнообразен лексикален ресурс, който от няколко години успешно се изгражда от СКЛ.

Резултати от изследването. От категорията на несвободните фрази са отделени СЛЕ. Направен е теоретичен анализ и са показани техните основни характеристики в семантичен, синтактичен и прагматичен аспект. Терминологичната система е синхронизирана с използваните в областта от различни изследователи, включително и българските. Дадено е лингвистично описание на СЛЕ, които са именни фрази, и са разгледани техните семантични, синтактични, прагматични и формообразувателни особености. Направена са класификации по семантични и синтактични признаци. Особено място заема класификацията според идиоматичността, защото тя е основа за предложената методология за разпознаването на различни СЛЕ, които са именни фрази.
Проектирана е и е разработена на JAVA система от компютърни модули за автоматично извличане на ресурси и разпознаване на СЛЕ.

Кратък преглед по глави. Глава 1 е уводна. Съдържа описание на целта, задачите и принципите на изследването. Описани са използваните ресурси и структурата на дисертационния труд.
В глава 2 се въвеждат основните възгледи и терминология в литературата, включително и на българските изследователи по въпросите за несвободните фрази и СЛЕ. Установява се съответствие между използваните термини. Дефинира се

обекта на изследването – СЛЕ. Описват се отношенията им със сложните думи, колокациите и наименованията.
В трета глава изследването се фокусира върху именните фрази, които са СЛЕ. Тук са и най-съществените оригинални теоретични приноси в дисертационния труд. Описват се семантичните, синтактични, прагматични и формообразувателни особености на СЛЕ. Дават се класификации по семантични и синтактични признаци. Централен резултат, определящ успеха на приложната част, е получената класификация по признака идиоматичност, която е основата на предлаганата от дисертантката методология за разпознаване на именните фрази, които са СЛЕ.
В глава 4 се дава характеристика на основните практически значими методи. Описани са използваните математически понятия. Направен е анализ на съществените фактори и ефективността на методите за разпознаване.
В глава 5 са описани известните програми, с които може да се решават някои от поставените проблеми. Същественият научноприложен принос е разработената на JAVA система от модули, с която автоматично се извличат и обработват езикови ресурси.

Критични бележки. Съществените ми критични бележки се отнасят към глава 2.
1. Неточно описание на понятието N-грам – това не е „комбинация от тоукъни", а „последователност от N тоукъна". Защо всъщност се въвежда тази чуждица (не ни ли стига „тоукън"!)? – просто можем да казваме „N-редица". (стр. 12)
2. Да се наричат „свободните фрази" просто „фрази или словосъчетания" създава сигурни предпоставки за грешки, освен в твърде ограничен контекст. (стр. 13)
3. „Тук се приема емпиричното разбиране за колокациите като последователност от думи, които директно могат да бъдат наблюдавани в текста, ..." вероятно трябва да се разбира като дефиниция на понятието „колокация" в рамките на дисертационния труд. От последващото изложение си мисля, че с корекцията „която се наблюдава достатъчно често", това е наистина външна (не лингвистична), статистически базирана, дефиниция, която дисертантката ползва. (стр. 13)
4. В първия параграф вероятно „устойчивите съставни наименования" и „съставните термини" се отъждествяват, което е неправилно. (стр. 14)
5. На стр. 14 редове 10-12 отдолу нагоре, двете срещания на „единици" трябва да се заменят с „фрази".
6. На стр. 14, ред 3 отдолу нагоре, изразът „като разложими несвободни" трябва да се замени с „като прости разложими несвободни". Това недоглеждане се наблюдава и в таблицата на стр. 16.
7. На стр. 15 се обяснява защо е предпочетен терминът „СЛЕ" пред „разложима фраза" (неточно: трябва да е „проста разложима несвободна фраза"). Аргументите са несъстоятелни. За целта е достатъчно да погледнем Класификация 1 на стр. 33 и да видим, че на простите лексикални единици се противопоставят несвободните фрази, като части от тях са СЛЕ и свободните колокации!
8. Пак на стр. 15 един параграф е посветен на дефиницията на „новоконструирания термин свободна колокация". Този параграф е просто абсурден! Започва с това, че свободните колокации са несвободни фрази и завършва с удачността на термина – насочвал, според дисертантката, към това, че са свободни фрази. Всъщност, целият параграф е странна смес от лингвистични и статистически съображения и учудване, че те не са взаимно сводими.

9. В края на стр. 25 на два пъти се говори за „прагматичните лексеми”. Вероятно се имат предвид „прагматичните фраземи”.
10. На стр. 34 се казва, че „свободните колокации от лингвистична гледна точка са свободни съчетания”. Дали наистина е така?
11. На стр. 38, в таблица 2.6 се казва, че СЛЕ са „рядко синтактично маркирани”, а малко преди това се твърди, че „често са синтактично маркирани”.
12. На стр. 38 условието (2) от дефиниция 3 е необосновано сложно изказано. По-разбираемо е да се формулира така: „Маркирана е статистически и поне по още един от признаците: лексикално-граматически, семантично, синтактично, прагматично”. Така също условието (4) е логически неправилно; да се замени с „тя е семантично разложима фраза ...”
13. Фигура 2.2 на стр. 39 вярно представя съотношенията между понятията, но е подвеждаща. С кръговете се създава погрешното очакване, че те представят обема на понятията, а това е така само за СЛЕ.
14. Логически неблагополучия с дефиниция 4 на стр. 39. Ако наистина искаме да изкажем дефиниция, то тя трябва да започне примерно така: „СЛЕ се наричат онези редици от думи, които показват следните особености...”. Тогава трябва да аргументираме, че едни и същи редици от думи удовлетворяват деф. 3 и деф. 4. Бихме могли да кажем, че това е твърдение: „Една редица от думи е СЛЕ точно тогава, когато ...”
Освен това, има припокриваме на условията (3) и (5).
15. На стр. 41 е казано, че „водопад = /явление, при което/ вода пада”. Съмнявам се. Струва ми се, че „водопад =/място, където/ вода пада”.
16. На стр. 44 отношението между колокации и СЛЕ е изразено по безсмислен графичен начин.
17. На стр. 50 на 5-ти ред след заглавието: изразът „има несвободни фрази, както и СЛЕ” да се замени с „има несвободни фрази, измежду които и СЛЕ”.

Заключение. Въпреки отправените критични бележки смятам, че те не влияят на окончателното ми становище: целите на планираното изследване с този дисертационен труд са постигнати и той съдържа оригинален принос в науката. Предвид демонстрираното отлично и задълбочено познаване на теоретичните и практически изследвания в света по темата на дисертацията и уменията за творческото им развитие и прилагане към спецификите на българския език, наред с огромната по обем практическа работа намирам, че дисертантката притежава способности за самостоятелни научни изследвания. Така смятам, че **изискванията** на ЗРАСРБ и Правилника за прилагане на ЗРАСРБ **към дисертационните трудове и авторите им са напълно удовлетворени**. Ето защо **убедено препоръчвам на Ивелина Любенова Стоянова да бъде присъдена образователната и научна степен „доктор”** по научна специалност „Общо и сравнително езикознание” (математическа лингвистика), шифър 05.04.11.

25 юни 2012 год. Рецензент:
София (проф. Т. Тинчев)