Становище

за дисертацията на Ивелина Стоянова

„Автоматично разпознаване и тагиране на съставни лексикални единици в българския език”
за присъждане на образователната и научна степен „доктор”

от проф. д-р Светла Пенева Коева

Дисертацията на Ивелина Стоянова „Автоматично разпознаване и тагиране на съставни лексикални единици в българския език” изцяло изпълнява поставените за изследване задачи и в някои отношения надминава изискванията за подобен род научни изследвания, като се характеризира с пълнота на анализа, зрелост на направените наблюдения и значимост на получените резултати. Съставните лексикалните единици, състоящи се от две или повече графични думи - обект на разглеждане в дисертацията, - не са изследвани в достатъчна степен не само в българската, но и и световната практика, макар че напоследък предизвикват все по-широк интерес. От друга страна изследването е важно, защото съставните лексикални единици са широко представени и поставят редица проблеми пред компютърната обработка на български език.

Дисертацията е с обем 223 страници, съдържа увод, пет глави, подробна библиография и две приложения - пример за автоматична обработка на текст и списък с предоставените електронни ресурси. Основната цел на дисертацията - създаване на теоретичен модел и приложение му при автоматичното разпознаване на съставните лексикални единици, както и свързаните с нея конкретни задачи - дефиниране на термина *съставна лексикална единица*, класификация на съставните лексикални единици, разработване на методология за разпознаването им и нейното практическо приложение, са добре и ясно формулирани в увода, ограничени по начин, който позволява както фокусирането на изследването към конкретни ясно разграничени области, така и получаването на видими и достатъчно значими резултати.

Във втора глава се предлага фокусиран преглед на литературата, свързана с разглежданата в дисертацията проблематика. Представени са основните теоретични проблеми при разглеждането на несвободните фрази като цяло и съставните лексикални единици в частност, като специално внимание се обръща на тези проблеми, които имат пряко отношение към разработването на методология за анализ и автоматична обработка. Предлага се дефиниция за съставна лексикална единица, която има теоретичен и практически аспект. Анализирани са приликите и разликите на съставните лексикални единици със сложните думи, колокациите, наименованията и термините с цел тяхното еднозначно разграничаване.

В трета глава е представен модел за описание на съставните лексикални единици, които са именни фрази. Моделът се базира на комплексна лингвистична информация - морфологична, синтактична и семантична. Специално внимание е отделено на анализа на идиоматичността като основна отличителна характеристика на съставните лексикални единици. За да се достигне до конкретни обобщения, се изхожда от различни съществуващи класификации за съставните лексикални единици и техните основни характеристики по

семантични и синтактични признаци. Предлаганият модел разграничава успешно съставните лексикални единици от останалите категории несвободни фрази, което има както теоретична, така и практическа стойност. Авторката привежда аргументи против включването на съставните лексикални единици в речниците, които, макар да имат своите основания, не приемаме. Съставните лексикални единици подлежат на системно описание и тяхното включване в речниците не може да бъде препятствано нито от количествени, нито от изчислителни съображения. Напротив - голяма част от лексиката по такъв начин остава извън пълното и непротиворечиво описание на морфологичните, лексикалните, семантичните и синтактичните си характеристики.

В четвърта глава са представени основните методи за разпознаване на съставните лексикални единици - лингвистични, статистически и хибридни. Анализират се факторите, които оказват влияние върху успешното прилагане на методите и качеството на резултатите.

Пета глава описва серия от експерименти, които демонстрират отделните етапи на автоматичното разпознаване на български съставни лексикални единици - именни фрази. Предложена е относително проста методология за разпознаване на последователни стъпки - несвободни фрази, съставни лексикални единици, отделни категории съставни лексикални единици според представените класификации. За различните подзадачи са приложени различни лингвистични и количествени методи, като експериментите показват приложимостта им за разпознаване и разграничаване на лексикалните единици по определени критерии. За целите на дисертацията е разработена компютърна програма, която включва серия от модули за обработка на корпуси, разпознаване на съставни лексикални единици и оценка на резултатите. По този начин е поставена основата за успешно автоматично разпознаване на съставните лексикални единици с необходимата пълнота и непротиворечивост, което разбира се в бъдеще може и трябва да бъде усъвършенствано като част от комплексния анализ на българския език.

Представената разработка като цяло потвърждава факта, че автоматичното разпознаване на съставните лексикални единици езици е сложна и комплексна задача, изискваща комбиниране на лингвистични и статистически методи.

В заключение може да се обобщи, че разглежданата дисертация е с несъмнен приносен характер в областта на теоретичната и компютърната лингвистика. Съставните лексикални единици са представени с оглед на техните основни характеристики в рамките на задълбочен теоретичен анализ и в съпоставка с гранични езикови структури. Направен е комплексен паралел между различни теоретични постановки, който води до ясно и точно представяне на съставните лексикални единици в системата на останалите съпоставими явления, както и до подходящо терминологично описание. Класификацията на съставните лексикални единици във връзка с тяхната идиоматичност дава сведения за начина, по който е формирано значението им, и съответно - за начина, по който да бъдат анализирани при компютърна обработка. Теоретичните наблюдения и обобщения на авторката й позволяват да разработи добра методология за автоматично разпознаване на съставни лексикални единици. Езикът на дисертацията е точен и ясен, представените дефиниции са адекватни, сложната материя е обяснена по един прост и разбираем начин. Концепцията за разработката е добре формулирана и подходящо ограничена, което рефлектира в добрата композиция на труда. В контекста на многообразието на схващания, както и на многозначността, с която с охарактеризират повечето езиковедски изследвания, авторката е съумяла ясно и целенасочено

да представи собствените си виждания, като изгради едно последователно и непротиворечиво изложение.

Ивелина Любенова Стоянова е докторанта по Общо и съпоставително езикознание (компютърна лингвистика) към Секцията по компютърна лингвистика на Института за български език. Докторантката съумя в рамките на периода, предвиден за обучение, не само да завърши научните си изследвания по поставената проблематика и да ги представи за защита като дисертационен труд, но и да покаже, че е вече изграден учен със задълбочени познания в областта на компютърната лингвистика и със собствено мнение по разглежданите въпроси. Ивелина Стоянова се отличава с висока степен на отдаденост към научните изследвания, което се съчетава с умението й да представя наблюдения и анализи си просто и достъпно, независимо от сложния им характер, както и да намира подходящите за защита на тезите си научни аргументи. Не на последно място искам да подчертая съчетанието на теоретични знания в областта на лингвистиката и математиката, както и на практически умения в областта на компютърната обработка на езика, които докторантката демонстрира.

Всичко това ми дава достатъчно основание убедено да предложа на почитаемото Научно жури да присъди образователната и научната степен *доктор* на Ивелина Стоянова за дисертацията й „Автоматично разпознаване и тагиране на съставни лексикални единици в българския език”.

23 юни 2012 г. Светла Коева