НОВ БЪЛГАРСКИ УНИВЕРСИТЕТ

РЕЦЕНЗИЯ

ОТ ДОЦЕНТ ДОКТОР **ВАЛЕНТИНА ИВАНОВА ГЕОРГИЕВА**

ВОЕННА АКАДЕМИЯ „Г. С. РАКОВСКИ" - ГР. СОФИЯ

доцент в професионално направление „Филология",
научна специалност „Общо и сравнително езикознание"

НА НАУЧЕН ТРУД ЗА ПРИДОБИВАНЕ НА ОБРАЗОВАТЕЛНАТА И НАУЧНА СТЕПЕН „ДОКТОР НА НАУКИТЕ"

по професионално направление: 2.1. Филология (научна специалност: Теория и практика на превода)"

с кандидат ВЕЛИСЛАВА СТОЙКОВА

на тема „Подходи за машинен превод на български език"
("Machine Translation Approaches to Bulgarian Language")

Научният труд на д-р В. Стойкова си поставя като основна изследователска **цел** „да изследва възможностите и начините, които формалните езикови теории (ФЕТ) предоставят за компютърен анализ на българския език, както и границите на приложението им с оглед на машинния превод (МП)" (стр. 2, Автореферат). Тази цел е формулирана в автореферата, който е представен на български език, докато в предговора на дисертацията, написана на английски език, намираме следното твърдение: „Книгата представя контрастивен анализ на начините за машинен крос-езиков превод на български език. Изследването включва анализ на ***няколко*** (курсив мой) софтуерърни приложения за машинен превод на български език, които използват различни подходи и техники за обработка на лингвистични данни. Резултатите от анализа на софтуеъра са представени, за да се открояят ограниченията в обхвата им и възможностите им да подобрят качеството на машинния превод на български език" (стр. vii). Изследването има и **практически задачи**: да представи няколко софтуерърни приложения за машинен

превод на български език чрез представяне на части от техния код източник и сравняване с частите, използвани при превод на целевите езици; анализиране на обхвата на тези приложения и качеството на превода им (стр. 2). Целите и задачите са формулирани по различен начин в дисертацията и в автореферата, така че предложението ми е те да бъдат по-ясно и по-прецизно заявени, както и да бъде изрично посочено в заключението дали са били постигнати в хода на изследването. Трябва да отбележим и факта, че авторефератът, макар и значително по-кратък като обем, е по-прецизен във формулирането на целите, както и в анализа на данните, като напр. в него се казва, че контрастивният анализ изследва ***три*** (курсив мой) подхода за машинен превод на български език (стр. 2 на автореферата).

Значимостта на изследвания проблем също е обоснована само в автореферата: „В научната литература засега няма публикувани подробни изследвания относно границите и ефективността на използването на формални езикови теории за български език, както в теоретичен план (представяне на ново знание), така и в практически аспект (използване за създаване на компютърни програми)." (стр. 2 на автореферата).

Дисертационният труд има **обем** от 95 страници, които включват: теоретична част, библиография от 5 страници и 105 източника; 4 страници с Приложения към главите и 2 страници речник с използваните съкращения. **Съдържанието** на дисертационния труд е структурирано в пет глави: Въведение, DATR език за представяне на лексикална информация и използването му за машинен превод, Универсалният мрежов език (УМЕ) и използването му за машинен превод, Sketch Engine и използването му за машинен превод и Заключение. В резултат, трите основни изследователски глави разглеждат тези три подхода за машинен превод въпреки че, както заявява авторът, „В българската теория и практика по превод има автори, които изследват всички подходи" (стр. 5).

Методологията на изследването е представена във **Въведението,** като основният метод е контрастивният анализ (КА). В. Стойкова заявява: „КА разграничава и описва само сходствата (на базата на анализ на семантичната еквивалентност) от крос-езикова гледна точка" (стр. 4). И добавя: „КА започва там, където свършва общото езикознание", което твърдение е противоречиво, имайки предвид факта, че в следващия параграф авторът заявява: „КА на всички нива на описание (фонетично, фонологично, морфологично и синтактично) е ориентиран към езиковата структура и е фокусиран върху формата" (стр. 4), което е задача и на общото езикознание. Освен метода на контрастивния анализ, авторът въвежда и понятието *tertium comparationis* (ТК) като ключов концепт в контрастивния анализ (стр. 3), тъй като „ТК се основава и оперира с употребата на семантична еквивалентност за да идентифицира или да характеризира като подобни сравняваните езикови феномени" (стр. 4). По този начин, избраната

методология допринася за достоверността и репрезентативността на изследването и е основателна база за анализиране и извеждане на заключения.

След представяне на основните разлики между човешкия и машинен преводи, изследователката стеснява анализа си до два вида машинен превод: Rule-Based Machine Translation (RBMT) и Statistical Machine Translation (SMT). Според нас добре би било анализът да се обогати с актуална информация и данни за Neural MT, Hybrid MT, Example-based MT и/или други видове машинен превод, които да бъдат съпоставени с избраните два и по този начин да се аргументира избора на разгледаните два.

Глава 2 детайлно представя работния механизъм на **езика DATR** за представяне на лексикална информация и използването му за машинен превод. Един от приносите на дисертацията е свързан с анализ на лингвистичния и компютърен модели на флективната морфология (2.3) и предложението да се прилага „смесен" подход за МП, тъй като и основите, и суфиксите се изменят. В 2.4. авторът представя граматичните правила и йерархията за словоизменение на съществителните имена в българския език като семантична мрежа в DATR, докато в 2.5. по подобен начин са представени особеностите на словоизменение на съществителните имена в руски език в DATR. Важното заключение, формулирано в резултат на семантичния анализ на думата 'закон', е представено отново само в автореферата: „Семантичният анализ за целите на МП показва, че макар и омонимна и в двата езика, думата „закон" в български език не може да бъде „объркана" от компютърната програма с думата „закон" в руски език [...], защото и двете думи и съответните им словоформи са представени чрез специфичните им в съответните езици граматични категории, промяната на които променя и словоформите. В този смисъл може да се направи заключението, че DATR е надеждна ФЕТ за създаване на софтуер за многоезичен МП между славянските езици" (стр. 9 на автореферата). Заключение, близко до формулираното в автореферата, е направено след представянето на особеностите на словоизменение на съществителните имена в полски език в DATR, което е по-схематично от това за български и руски език: „... сродните езици могат да бъдат формално представени чрез използване на сходни идеи и техники за кодиране, защото споделят сходни граматически характеристики" (стр. 20).

Друг безспорен приносен момент от тази глава е кодът на софтуеъра на DATR за изясняване на граматичните категории на притежателните и възвратното притежателно местоимение в българския език.

Глава 3 е посветена на анализ на възможностите и особеностите на **Универсалния мрежов език (УМЕ)** за машинен превод. Авторът подчертава факта, че УМЕ „също като DATR представлява семантична мрежа, но с различна структура и повече възможности за използване в областта на МП" (стр. 10 на автореферата) и

доказва това твърдение с анализ на характеристиките му в 3.1., като и с обобщението в 3.2.2. че „УМЕ е онлайн уеб-базирана интелигентна система за мениджмънт на знанието, която позволява да се използват различни видове семантични връзки за организиране на лингвистични данни" (стр. 32). В главата са представени разработените формални средства за компютърна обработка на български, английски, руски и словашки с УМЕ.

Приносният елемент в тази глава е съпоставителният анализ на двете формални представяния на словоизменението на българските съществителни - DATR и УМЕ - и в откроването на ограниченията, както и на предимството на УМЕ да задава семантична информация за словоизменителни типове и лексикалната база данни на съществителните.

В 3.3. В. Стойкова анализира представянето на притежателното и възвратното притежателно местоимение с УМЕ за български, английски и руски език. След теоретичните описания на типовете трансформационни флективни правила на УМЕ, процесът на кодиране/задаване в компютърна програма е представен нагледно с извадки от речника на УМЕ за английските местоимения 'mine' и 'myself', на българското притежателно местоимение 'мой' и на руското притежателно местоимение 'мой'. Изводът от анализа е, че УМЕ позволява адекватно кодиране (и съответно преводна еквивалентност) не само на близкородствени езици (като български и руски), но и на неродствени езици (като английски и български) [...] и може успешно да се използва за МП (стр. 37).

В 3.4. се разглежда използването на УМЕ за машинен превод между български и словашки език. Приносният момент тук е представянето на формалните граматични правила на български и словашки езици чрез прилагането на УМЕ кодиране, което е представено интерлингвистично на базата на преводите на книгата „Малкият принц". Формалното представяне в рамките на УМЕ е онагледено с представянето на българското прилагателно 'добър' и на словашкото 'dobrý'. Направеният извод е, че механизмът на УМЕ, допълнен с общите спецификации, зададени за прилагателното в двата езика, осигуряват семантичен и преводен еквивалент при МП (стр. 42).

Последната секция 3.5. представя използване на УМЕ за измерване и подобряване на качеството на машинния превод на притежателните местоимения в български и английски език. Авторът заявява, че „Платформата на УМЕ е създадена така, че да позволява статистическо измерване на точността на преводите". Нейният принос е в разработването на синтактични правила за трансформиране с цел разграничаване на притежателните местоимения според синтактичната им функция, което значително подобрява точността на превода (стр. 48).

Глава 4 представя машината за търсене **Sketch Engine** (СЕ) и използването й за машинен превод. Началото на главата е посветено на разкриване на принципа на действие на СЕ, който се основава на приемането на статистическата близост на думите за семантична близост, с което наподобява принципа на Google Translator, а в 4.1.2. са разкрити характеристиките на СЕ за статистически цели, което позволява използването й като *tertium comparationis.*Благодарение на възможностите на СЕ е осъществен машинен превод и контрастивен анализ на екологични термини между български и словашки език. Чрез използване на паралелния Българско-словашки EUROPARL 7 корпус, който в българската си част съдържа 9 215 000 словоформи, а в словашката си част – 13 000 000 словоформи, са представени резултати от статистическо търсене на конкордънси със СЕ за български екологични термини и откриването на техните словашки преводни еквиваленти, като е проследена и симетричността в превода. Изводът на изследвателката е, че резултатите показват добра преводна еквивалентност, въпреки че преводните еквиваленти на екологични термини не са винаги симетрични (напр. ‘биологично разнообразие’ (бълг.) – ‘biodiversity’ (словашки).

Следва анализ на използването на СЕ за машинен превод на изрази за време между български и словашки език чрез статистическо търсене със СЕ в съществуващите Българско-словашки електронни корпуси на конкордънси и колокации. Разгледаните примери на превод на словашката дума ‘obdobie’ показват, че статистическите подходи на СЕ могат да помогнат както търсене на преводни еквиваленти на отделна дума или изрази с нея, така и различаването при случаи на междуезикова омонимия чрез повторно търсене за същата дума, особено при превод между родствени езици.

В 4.4. се анализира използването на СЕ за машинен превод между български и словашки език и се доказва възможността за откриването на по-сложни йерархични семантични връзки на изследваната дума, които могат да бъдат визуализирани с различни цветове. Възможно е и двуезично паралелно търсене за тезаурус на дадена дума. Изводът е, че статистическите подходи на СЕ за двуезично паралелно търсене в електронни корпуси могат успешно да служат за МП. Подчертано е предимството на този подход, а именно, че е много бърз и не изисква разработване на формални граматики за преводните езици, а само достатъчно голям и специализиран двуезичен електронен текстов корпус.

Секцията 4.5. също съдържа приносен момент с разработването на двуезични сравнителни електронни корпуси Bulgarian Mathematical Wikipedia Corpus (MathWikiBG) и Serbian Mathematical Wikipedia Corpus (MathWikiSR) и представянето на статистическите подходи на СЕ за МП на математически термини между български

и сръбски език. Разработени са алгоритмични стъпки за търсене по ключова дума. Авторът демонстрира как със статистическите подходи на СЕ за определяне на семантични отношения и тезаурус на ключовата дума е възможно да се определят преводните ѝ еквиваленти при МП между български и сръбски език.

Научният принос на дисертацията за придобиване на образователната и научна степен „доктор на науките" беше подчертаван в експозето на рецензията, но е важно да се подчертае, че изследването не само представя и обобщава свойствата на трите техники за машинен превод, приложени към българския език (т.е. DATR, UNL и CE), но също така сравнява и анализира техните плюсове и ограничения на базата на достатъчно примери. Не можем да не се съгласим със самооценката на кандидата за основните приноси, формулирани в резюмето на стр. 34. Важно е да се подчертае, че В. Стойкова създава корпус от специализирани текстове по математика на български език MathWikiBul, който се използва за статистически машинен превод между български и сръбски език.

Библиографията с теоретични изследвания показва дълбоко разбиране на концептуалната рамка по въпроси, които са необходими за постигане на целта на дисертацията, но някои от тях са доста остарели (напр. от 1989 г.), докато технологичните решения на различни изследователски проблеми се модернизират изключително бързо.

Кандидатът представя 12 публикации, които допълват и разширяват изследването, представено в дисертационния труд.

Препоръки и предложения:

- Бих насърчила д-р Стойкова да подчертае по-ясно приносните моменти на академичното изследване, вместо да използва безподложни изречения.

- Специализираният академичен текст ще бъде по-лесен за възприемане, ако няма такова прекомерно използване на съкращения, които затрудняват вникването в аргументите.

- На места има препратка към източници, което изисква допълнително търсене на статия, която не е достъпна онлайн, напр. „Ние използваме ТК, както е дефинирано в [37]", „Нашият подход към флективната морфология на притежателните местоимения се основава на кодирането, което вече е публикувано в [63]" (стр. 21), „Подробното изследване на DATR на българския номинална флективна морфология вече е публикувана в [59]" (стр. 30).

Авторефератът отразява точно съдържанието на дисертационния труд, но някои от формулировките и постановките в него са по-прецизни и ясни.

Въз основа на посочените приноси и изтъкнатите по-горе положителни качества може да се направи изводът, че рецензираният труд удовлетворява изискванията за дисертационен труд за получаване на образователната и научна степен „доктор на науките". Това ми дава основание да препоръчам на Научното жури да присъди на д-р Велислава Стойкова образователната и научна степен „доктор на науките" в професионално направление: 2.1. Филология, научна специалност: Теория и практика на превода

Дата: 12.09.2022 г. Подпис: ……..

/Доц. д-р Валентина Георгиева/