

РЕЦЕНЗИЯ

на д.м.н. Георги Атанасов Тотков – професор в ПУ „Паисий Хилендарски”
относно дисертационен труд за присъждане на ОНС „доктор“
в област на висше образование 2.0. Хуманитарни науки,
професионално направление 2.1. Филология,
докторантска програма *Общо и сравнително езикознание*

Докторант: *Георги Емилов Илиев*

Тема: *Езиково мотивирана оптимизация на машинния превод*

Научен ръководител: *проф. д-р Светла Коева*

Със заповед № 344-1/23.09.2014 г. на проф. д-р Елка Мирчева, зам.-директор на Института за български език – БАН (София), съм определен за член на научно жури (НЖ) по повод защитата на ОНС „доктор“ в ПН 2.1. Филология от Георги Илиев.

В електронен формат получих следните документи по процедурата: дисертационен труд, автореферат, копия на публикации, списък на публикации и участия в семинари, заповед за утвърждаване на състава на научното жури (НЖ).

1. Кратки биографични данни

Г. Илиев е завършил магистърска програма по Компютърна лингвистика през 2009 г. след придобиване на ОНС „бакалавър“ по скандинавистика в СУ „Св. Кл. Охридски“. От 1.1.2010 г. до 1.1.2013 г. е редовен докторант към секция „Компютърна лингвистика“ на Института за български език към БАН (София).

От 2014 г. работи като администратор на договори за поддръжка на IBM Technical Support Services за скандинавските страни.

2. Актуалност на тематиката и целесъобразност на поставените цели и задачи

Дисертационното изследване е в актуална за компютърната лингвистика област – машинен превод. Дисертационната тематика е свързана с изследване и решаване на специфични проблеми на машинния превод на български език в контекста на бурно развитие на съвременните технологии за машинен превод, силно доминирани от приложения за „световни“ езици като английски, китайски и арабски. Предмет на работата е машинен превод на български език, осъществяван по метода на фразовия статистически превод, за който са разработени редица софтуерни инструменти с отворен код. Конкретна цел на изследването е машинен превод на текст на шведски език, избран като носител на определени типологически характеристики, които го позиционират между два, в известен смисъл крайни случаи на изходния

език - добре изследван (английски) и не дотам изучен (европейски езици от германското и славянското езиково семейство). Друг аргумент за избора на шведски като изходен език е наличието на богати езикови ресурси, свободно достъпни за академични цели.

В дисертационното изследване е направен аналитичен преглед на състоянието на областта (в исторически и научно-приложен план). В резултат на проведенния анализ не са формулирани основна цел, хипотеза и задачи на изследването.

Резултатите от анализа мотивират избора на конкретната архитектура за машинен превод и са основа за провежданата експериментална работа на следващите етапи.

На **първия етап** от изследването се построява работеща система за фразов статистически машинен превод от шведски на български език. Използват се налични паралелни езикови ресурси за изследваната езикова двойка, както и други референтни системи за машинен превод между различни германски езици и български език.

На **втория етап** от изследването се търсят езиково зависими начини за оптимизиране на построената система за машинен превод от шведски на български език на базата на конкретни типологически характеристики на изследваната езикова двойка. При това на съответните експериментални етапи системата се обогатява с лингвистична информация от свободно достъпни едноезикови ресурси.

3. Степен на познаване на състоянието на проблема

От обзора и анализа, представени в дисертационния труд (на повече от 100 страници в глави 1. и 2). може да се съди, че докторантът обстойно е проучил и изследвал състоянието на изследванията в областта. Списъкът на цитираните литературни източници съдържа 150 заглавия, от които 9 на кирилица и 141 на латиница.

4. Методика на изследването

Основни логически стъпки на дисертационно изследване са: а) проучване в съответната област; б) открояване на проблем (и възможни подходи за неговото решаване); в) формулиране на цели и задачи на изследването за решаване на проблема; г) синтез на решения (ев. съпроводени със създаване/адаптиране на подходящи модели, методи и инструменти); д) планиране и провеждане на експерименти (за установяване на приложимостта на създадените/адаптираните средства); е) анализ (вкл. усъвършенстване и сравнение с резултати на други автори); ж) заключение (решен ли е проблема); з) формулиране на приноси и перспективи, и др. В този смисъл, в методиката, приложена от докторанта отсъстват някои елементи,

между които ясно формулиране на проблем, цели, задачи и перспективи на изследването.

Независимо от посоченото обстоятелство, проведеното изследване е довело до постигане на конкретни значими резултати.

5. Обща характеристика на дисертационния труд

Основният текст на дисертационния труд (общо 208 страници, всяка с около 40 реда по 90 знака) е организиран в Увод, 3 (три) глави, заключение, приложение и библиография (цитирани литературни източници).

В **Глава 1. Машинен превод - кратък исторически обзор** е анализирана еволюцията на научните изследвания и добри практики в областта от 1940-те години до наши дни. Изложението съдържа ключови разработки и решения - представители на различни поколения системи и научни школи в областта на машинния превод. Представянето обръща внимание и на теоретични разработки (граматика на зависимостите, синтактични теории и др.). На базата на проведен анализ се идентифицират недостатъчно проучени и нерешени въпроси в областта на изследването.

В **Глава 2. Построяване на система за статистически машинен превод на български като целеви език** е направен анализ на математически модели и методи, които са в основата на съвременните системи за статистически машинен превод. Предназначението на анализа е да положи основа за следваща адаптация на разглежданите модели и методи за построяване и оценка на базова система за фразов статистически машинен превод на български език - от събиране и подготовка на паралелни ресурси за обучение на системата до нейното тестване и интерпретиране на резултатите. Съответната система (Moses) се обучава върху версии на паралелния корпус Europarl и 6 (шест) базови системи за превод между езикови двойки както следва: шведско-български, английско-български, шведско-английски, немско-български, датско-български и нидерландско-български. Акцентът е поставен върху превода от шведски на български език. За сравнение са построени и системи за превод от други германски езици, представени в корпуса Europarl, на български, както и система за превод от шведски на английски език. За езиковите двойки шведско-български, немско-български, датско-български и нидерландско-български паралелните данни се подготвят въз основа на съответните корпуси с целеви език английски по метод, разработен от докторанта в съавторство.

В **Глава 3. Езиково зависими компоненти и методи** се търсят езиково зависими методи и компоненти, с които да бъдат обогатени отделните етапи от обучението и работата на системата за статистически машинен превод с цел повишаване на

оценката на резултата при превода на български език. След всеки изведен метод са направени изводи за неговата приложимост. В частност са разгледани методи за **съотнасяне на паралелни корпуси по изречения** (без приложение в експериментите по дисертационния труд) и **формално разширяване на изходната част от тренировъчните данни по езиково зависим метод**. В последния случай е предложено практическо решение при превод от шведски на български език, като за целта се прилагат корпусни техники за анализ на сложни думи и езиково мотивирани техники за „плитък“ синтактичен анализ, морфологичен анализ и пораждане на перифрази с помощта на свободно разпространявани инструменти и ресурси за шведски език. Изведената оценка потвърждава ефективността на разработените методи за анализ и перифразироване на съставни думи на шведски

Експериментите по метод **Съкращаване на пространството за търсене при съотнасяне по думи** водят до последователно извличане на таблици на фразовите преводи с по-малък размер от таблицата на фразовите преводи, извлечена при обучението на базовата система. В съответния раздел се представя възможно решение на задачата за съкращаване на пространството за търсене при съотнасяне на шведско-български паралелни данни по думи по установените модели за статистически машинен превод по думи, посредством лематизация при обучение върху данни с морфологична вариация. В тази връзка е предложено софтуерно решение за лематизация на шведски език, усъвършенствано с метод за лематизация на сложни думи извън морфологичната база данни.

Експериментите с различни методи в глава 3. се съпровождат с резултати от тестове за тяхната статистическа значимост.

Методът за **автоматично синтактично пренаареждане** се основава на строго ограничаване на обхвата на пренаареждане до случаи, за които има основание да се допусне, че автоматичният анализ на зависимостите е с висока точност и максимално точно отразява особеностите на структурата на словоред от тип V2 за шведски като изходен език при превод на език със словоред от тип SVO, български.

Предложените методи за езиково мотивирана оптимизация на машинния превод са реализирани във вид на множество софтуерни инструменти (на Perl и Haskell) – непредставени в разработката, които се прилагат последователно върху тренировъчни и/или тестови данни за постигане на желания резултат. За нормализация на резултата се ползва и модул *BgTagger* за анотация на български текст.

За целите на проведените експерименти по обучение на базовата система Moses са създадени и редица лексикални ресурси (вж. и приложението към дисер-

тационния труд), между които паралелни корпуси: шведско-български, немско-български, датско-български, нидерландско-български. Езиковите ресурси се извличат като сечение на наличния българско-английски и съответните х-английски паралелни корпуси по проекта Europarl.

За целите на експериментите за автоматично съотнасяне на паралелни данни по изречения е построен и проверен „на ръка“ паралелен датско-български корпус-еталон. Освен че с него е възможно оценяване на методи за автоматично съотнасяне по изречения, той позволява и провеждане на бъдещи експерименти за машинния превод от датски на български език.

6. Автореферат

Авторефератът на дисертацията отговаря на изискванията на ЗРАСРБ и Правилника за неговото приложение като отразява в резюмиран вид най-съществените страни на дисертационното изследване. Не всички означения в автореферата са обяснени (въведени), което затруднява неговото самостоятелно четене. Независимо от това, информацията, съдържаща се в автореферата позволява да бъдат обективно оценени основните достойнства и приноси на дисертационното изследване.

7. Приноси

Дисертационното изследване определено има **научно-приложен характер**. Базира се на обстоен анализ на известни от литературата методи и техники за провеждане на статистически машинен превод, тяхното модифициране и прилагане за конкретна езикова двойка.

Систематизирани са редица езикови методи, инструменти и ресурси за обработка на шведски и български език, предимно предлагани за свободна употреба за академични цели. На тази база се предлагат конкретни езиково зависими методи с оглед на типологическите особености на съответната езикова двойка и в контекста на известни средства за анализ и фразов статистически машинен превод. Адаптирането на тези езикови инструменти и ресурси и тяхното прилагане за разрешаване на конкретни проблеми на превода от шведски на български език може да се определи като приложен принос на изследването.

Част от описаните експериментални процеси, реализирани във вид на множество външни програмни модули, могат да се разглеждат като хибридна реализация на система за машинен превод, включваща компоненти за предварителна обработка на изходния текст за превод (перифразирание и/или пренареждане на словореда) и/или следваща обработка на резултата (нормализиране).

Друг приложен принос на изследването е нормализацият модул за български език, построен със средствата на граматическия формализъм Grammatical Framework.

8. Преценка на публикациите по дисертационния труд

Резултати на дисертационното изследване са представени в **5 (пет) публикации**, от които 2 (две) статии (в българско списание) и 3 (три) доклада в трудове на конференции (3 на международни конференции в чужбина).

В посочените 5 (пет) публикации Г. Илиев е самостоятелен автор на 3 (три) работи, а в 2 (две) е със съавтор.

В публикации на докторанта са отразени основни резултати, получени в дисертационния труд и може да се счита, че резултатите са апробирани в достатъчна степен пред специализирана научна аудитория.

Не разполагам с данни за цитирания на публикациите от приложения списък.

Нямам основания за съмнения в личния принос на докторанта за получаване на дисертационните резултати.

9. Мнения, препоръки и бележки

В **съдържателен план**: Дисертационният труд съдържа твърде подробен преглед на резултати на други автори в областта на изследването. Това се отнася изцяло за първа глава от дисертационния труд и по-голямата част от втората глава (до раздел 2.7. вкл.) – от 13-та до 106-та страница. На практика отсъстват ясно формулирани проблем и основна цел на дисертационното изследване.

Според мен, трудно може да се приеме, че „научният принос на изследването може да се обобщи в три основни направления, които отговарят на организацията на работата по глави“ и че „първа глава от работата има изцяло теоретичен принос“. На първо място - всеки (дори и блестящ) анализ на състоянието на изследванията в дадена област (глава 1. и до голяма степен глава 2.) е само необходима предпоставка за получаване на нови решения на проблеми в същата област, и не може да бъде определян като принос от какъвто и да е характер – научен, научно-приложен или приложен. Краткият преглед на съдържанието на дисертационния труд, направено по-горе в раздел 5. определено води до извода, че дисертационният труд съдържа основно научно-приложни и приложни приноси, представени във 2-ра (след раздел 2.8) и 3-та глава. От друга страна проблематиката, изследването и конкретните решения са в област, гранична за две научни направления – филология и ин-

форматика, а процедурата е за присъждане на ОНС „доктор“ в първата област. От тази гледна точка формулировките на някои авторски приноси са спорни.

В заключението, както беше посочено и по-горе, се чувства отсъствието на: а) анализи – решени ли са (и до каква степен) поставените задачи; б) конкретни перспективи и възможности за мултиплициране на адаптираните методи и средства; в) систематично представяне и обосновка на характера на приносите – резултат на проведеното изследване. Ще отбележа обаче, че на отделни места в дисертационния труд се коментират някои перспективи на изследването

В редакционен план: От езикова и стилистична гледна точка оформлението на текстовете в дисертационния труд и автореферата се отличава със сравнително добро качество.

Препоръки: Позволявам си да препоръчам на докторанта, в рамките на предстоящата защита да формулира по-прецизно основните цели, задачи и перспективи на дисертационното изследване, като обърне специално внимание на въпроса за систематизиране и определяне на характера на приносите - резултат на неговото провеждане.

10. Заключение

Направените бележки не намаляват качеството и достойнствата на проведеното дисертационно изследване. Определено считам, че представеният за рецензиране дисертационен труд отговаря на изискванията на Закона за РАСРБ и Правилника за неговото прилагане, поради което изразявам своята **положителна оценка** за достойнствата и качеството на проведеното дисертационно изследване. Предлагам на уважаемото научно жури да присъди образователната и научна степен ‘доктор’ на *Георги Емилов Илиев* в област на висше образование *2.0. Хуманитарни науки*, професионално направление *2.1 Филология*, докторантска програма *Общо и сравнително езикознание*.

3. 11. 2014 г.

Пловдив

Подпис:



(проф. д.м.н. Г. Тотков)