Становище

за дисертационния труд на Георги Емилов Илиев

на тема

Езиково мотивирана оптимизация на машинния превод

за присъждане на образователната и научна степен „доктор"

Дисертацията на Георги Илиев *Езиково мотивирана оптимизация на машинния превод* е резултат от целенасочената и ползотворна работа на докторанта по време на неговото обучение в Института за български език „Проф. Любомир Андрейчин". Георги Емилов Илиев е редовен докторант към Секцията по компютърна лингвистика на Института за български език „Проф. Любомир Андрейчин" в периода 1 януари 2010 година – 31 декември 2012 година. Георги Илиев е бакалавър по английска филология (2002 година) и е завършил семестриално скандинавистика (2008 година) в Софийския университет „Св. Климент Охридски". От 2009 година е магистър по специалността *Компютърна лингвистика. Интернет технологии в хуманитаристиката* на Софийския университет „Св. Климент Охридски". Занимава се професионално с превод на текстове от различни стилове и тематични области: наука, публицистика, художествена литература, икономика, право и медицина от и на английски, шведски, датски и норвежки език.

Дисертацията *Езиково мотивирана оптимизация на машинния превод* се състои от 206 страници, разпределени в увод, три глави, заключение, приложение и библиография. Като научен ръководител си позволявам да дам висока оценка на общото оформление на дисертацията: проблематиката е овладяна, структурата е добра, езикът е ясен, терминологията е адаптирана за български. Георги Илиев е съумял да представи сложната материя, предмет на неговата дисертация, в един строго научен текст, в който математическата и лингвистичната част са описани на такъв ясен език, че са разбираеми в целостта си, както от математици, така и от лингвисти. Демонстрира се отлично познаване и анализиране на литературата, свързана с проблематиката. Цитират се 150 заглавия на български, руски, английски, френски и датски език. Експерименталните корпуси, езиковите модели, допълнителните данни, свързани с експериментите, и експерименталните резултати се предоставят заедно с текста на дисертацията.

В увода са посочени актуалността на изследването, неговият предмет, методи и цели, както и основните приноси на дисертационния труд. Обосновава се изборът на шведски език (при машинен превод от шведски на български) с типологическите му характеристики, които го *„отдалечават" от доминиращия изследванията в областта на машинния превод*" английски език и го *„приближават" към европейски езици от германското и славянското езиково семейство, за които не е добре изучен проблемът за машинния превод и в частност за машинния превод на български език*". Формулирани са две основни цели: да се построи работеща система за статистически машинен превод от шведски на български език на базата

на съществуващите ограничени паралелни езикови ресурси и да се потърсят езиково зависими начини за подобряване на построената система.

В първа глава *Машинен превод - кратък исторически обзор* се прави преглед на развитието на машинния превод, разделено на четири етапа, като се засягат теоретични и приложни разработки на различни школи в тази област. Обзорът служи (и може успешно да послужи и при други изследвания) за изграждането на критична позиция към различните подходи в областта: директен превод, трансферен превод, „интерлингвален", по аналогия и статистически превод.

Втора глава *Построяване на система за статистически машинен превод с български като целеви език* представя изграждането на базова система за фразов статистически машинен превод от шведски на български език на базата на наличните паралелни езикови ресурси. Обзорен характер има представянето на някои математически модели, които се използват за статистически машинен превод, направено с цел да се обоснове изборът на базовата система за фразов статистически превод *Moses*. Изложението проследява подробно и ясно етапите на създаването на базовата система: подготовката на паралелни ресурси, обучението на системата и нейното оценяване. Стремежът на докторанта е да покаже възможностите на системата *Moses*, както по отношение на нейните предимства, така и по отношение на нейните ограничения.

Трета глава *Езиково зависими компоненти и методи* представлява сериозно постижение на докторанта - именно в нея той демонстрира отлична лингвистична подготовка, способност за построяване на езиково мотивирани хипотези и умение да се приложи езиковото знание към математическото моделиране. За целите на експериментите се построяват множество паралелни корпуси: шведско-български, немско-български, датско-български, нидерландско-български. Възприет е принципът да се търсят решения за усъвършенстване на статистическия машинен превод без увеличаване на обема на тренировъчните корпуси, а с автоматична обработка на наличните данни и обогатяването им с допълнителна езикова информация. Безспорно е, че всеки един от предложените езиково зависими начини за усъвършенстване на машинния превод може да бъде разширен и подобрен както по отношение на обхвата си, така и спрямо други езици. Само работещите активно в тази област обаче биха могли да оценят огромния труд, който Георги Илиев е вложил, за да предпостави, приложи и анализира предложените езиково мотивирани подобрения на фразовия статистически машинен превод.

Съотнасянето на паралелни корпуси по изречения по (слабо) езиково зависим метод включва създаването на двуезикови речници от думи - опори и транслитерация на български текстове с цел намирането на когнати в паралелни изречения. Правилен е подходът правилата за транслитерация да се дефинират в зависимост от изходния език, а не да се следва официалната или дадена транслитерация. Трябва да се направи уточнението, че фонетичният принцип затруднява автоматичното съотнасяне с думи от чужд произход от езици, които

използват латинска азбука. *Формалното разширяване на изходната част от тренировъчните данни по езиково зависим метод* включва перифразиране на тренировъчните данни (при превод от шведски на български) с цел да се подобрят резултатите, получени при обучение върху корпуси с ограничен размер. Докторантът прилага морфологичен анализ при анализа на съставни съществителни имена в шведски, като използва най-разпространените словообразователни модели, и „плитък" синтактичен анализ при перифразирането.

Георги Илиев изказва хипотези, които са мотивирани както лингвистично, така и от наблюденията и анализа на резултатите при експериментите. Пример е използването на лематизацията като метод за *автоматично съкращаване на пространството за търсене при съотнасяне по думи* при обучение върху шведско-български паралелни текстове, което да послужи за „сближаване" на пространствата за търсене на двата езика. Резултатите от този експеримент очаквано показват намаляване на покритието и увеличаване на точността. Прави се наблюдението, че *„при оптимизиране на теглата на характеристиките по логаритмично-линейния модел оценката на превода с експерименталните системи не бележи подобрение"*, което всъщност показва, че няма „оптимизиране", а промяна. Следващите експерименти са свързани с *морфологичната информация по модел IBM 4*, по който като параметър се въвежда относително разместване, за да се моделират различни синтагматични отношения между класове думи в различни езици. За целта докторантът обобщава множествата от граматични класове за шведски и български, следвайки определени лингвистични критерии. *Синтактичното пренареждане* преобразува словореда на някои изречения в шведски, за които може да се установи еднозначно отношение между глагол и местоимение - субект, в словоред, характерен за български неемфатични изречения. Правилно е допускането, че при ограничен брой типове в тренировъчните данни, върху които се построява езиковият модел, моделът, съдържащ по-дълги низове, е носител на повече достоверна информация, необходима за оценката на преводния кандидат. Прилага се *езиков модел от морфологични етикети* и поради ограничения брой етикети (анотации) се правят експерименти с два модела, съответно с низове от три и шест етикета. Докторантът изказва и хипотези, които не намират потвърждение в експерименталните данни, с което доказва, че научното търсене се опира както на положителните, така и на отрицателните резултати. Например, експериментите в частта *Факторни модели и техните ограничения* не подкрепят хипотезата, че може да се установи зависимост между последователности от символи и морфологични характеристики. *Нормализирането на резултата с езиково зависим инструмент* само по себе си представлява отделна задача, която е съществен компонент от машинния превод. Георги Илиев ограничава нормализирането до проверката и възможната корекция на съседни думи от определени класове: прилагателно или детерминатор и съществително, детерминатор или прилагателно и прилагателно, като се базира на теоретичните разработки на Й. Пенчев.

Като научен ръководил, който работи с докторант с голям потенциал, признавам, че ми се искаше дисертацията да бъде разширена по посока на допълнително обучение и тестове на сравнителни системи за превод на английски, шведски и български език, както и с разширяване на езиково зависимите начини за подобрение на фразовия статистически машинен превод, независимо от обективните ограничения, основно свързани с достъпността на паралелни езикови ресурси и еквивалентни средства за компютърна обработка на различните езици. Сериозната лингвистична подготовка на докторанта, неговата работоспособност, способността му за анализ и обобщение на резултатите и не на последно място - умението му да представя сложна материя по ясен и достъпен начин, ме правят уверена, че ако Георги Илиев продължи и в бъдеще научните си разработки, той несъмнено ще ги задълбочи и усъвършенства в желаната от мене или друга посока.

Авторефератът отразява основните наблюдения и изводи, направени в дисертацията. Георги Илиев има четири публикации по темата на дисертацията, две от които самостоятелни.

В заключение на изложеното дотук убедено предлагам почитаемото жури да вземе решение за присъждането на образователната и научна степен „доктор" на Георги Емилов Илиев.

София, 3 ноември 2014 г.

(проф. д-р Светла Пенева Коева)