НОВ БЪЛГАРСКИ УНИВЕРСИТЕТ
ДЕПАРТАМЕНТ ЧУЖДИ ЕЗИЦИ И КУЛТУРИ

Велислава Стойкова

Подходи за машинен превод на български език

Автореферат
на
дисертация за присъждане на научната степен „доктор на науките“

2.1. Филология. Теория и практика на превода

София 2022

1. Увод

В съвременните изследвания на българския език *езиковите технологии* са една интердисциплинарна област, която се развива все по-бързо. В последните години се появиха нови публикации, научни резултати и компютърни програми за българския език, свързани с приложението на различни формални езикови теории (ФЕТ) за автоматична обработка и анализ на български език. Една голяма част от тези ФЕТ са унифицирани универсални формални теории, които имат програмни платформи и приложения за много езици и са подходящи за машинен превод (МП). По-известните от тях са: 1) DATR език за представяне на лексикална информация, 2) Опорни фразово-структурни унификационни граматики, 3) Универсален мрежов език и др., като всички от изброените са били използвани за автоматична обработка и анализ на български език.

2. Цел, задачи и подходи на изследването

В научната литература засега няма публикувани подробни изследвания относно границите и ефективността на използването на ФЕТ за български език, както в теоретичен план (представяне на ново знание), така и в практически аспект (използване за създаване на компютърни програми). Представеният труд „Подходи за машинен превод на български език“ (Machine Translation Approaches to the Bulgarian Language) си поставя точно такава цел, а именно да изследва възможностите и начините, които ФЕТ предоставят за компютърен анализ на българския език, както и границите на приложението им с оглед на МП [1]. Настоящото изследване използва като метод контрастивния анализ (КА) [6], широко прилаган в сравнителното езикознание [28], за да изследва три подхода за машинен превод на български език: 1) DATR език за представяне на лексикална информация (DATR), 2) Универсалния мрежов език (УМЕ) и 3) Sketch Engine (SE) (Скетч Енджин) – статистическа търсачка за обработване на

двуезични електронни текстови корпуси. Използва се понятието *терциум компарационис* (ТК) [29], въз основа на което се правят сравненията. Компютърни програмни приложения за български език и на трите подхода се анализират с оглед на начина, по който формално са представени езиковите данни, с оглед на спецификата и възможностите те да бъдат използвани за МП (да генерират подходящ преводен еквивалент), с оглед на това, как са представени различни езикови фрагменти в различните ФЕТ, с оглед на измерването на качеството на превода и възможностите за подобряване на резултата му, както и с оглед на времето и ресурсите, необходими за прилагането и използването им.

3. Машинен превод – принципи и приложение

Съвременният превод, като човешка дейност, е изследван както с оглед на теорията и практиката му [2, 31, 37], така и в културологичен и антропологичен аспект [27]. Разделянето на превода на художествен и технически [9, 36] отразява разликите и спецификата му, както и разглеждането на особеностите на превода между родствени езици като славянските [30] отразява други особености.

В същото време развитието и усъвършенстването в областта на компютърната техника и информационните технологии (ИТ) откриват нови възможности за използването им в сферата на превода. Възникналото направление на машинния превод (МП) използва различни подходи в опит да автоматизира превода като компютърно моделира тази човешка дейност, главно знанията за изходния и целевия език. По принцип МП представлява употреба на различни езикови програмни продукти, с които преводът на практика се извършва автоматично. Основната идея е да се представи изходния език символно – като езиково-независимо формално представяне, главно с помощта на ФЕТ или набор от стандартни статистически измерители, а след това това представяне трябва да се трансформира като езиково към целевия език.

Така при МП [8, 11] основната задача е компютърното представяне както на изходния, така и на целевия език чрез използването на общи и за двата езика ФЕТ или статистически техники. Първоначално МП е създаден за компютърен технически превод [9], но впоследствие се ползва и за превод на семантично по-сложни текстове, въпреки че употребата му е ограничена, тъй като все още не може да се моделира успешно когнитивната страна на превода като човешка дейност. В традицията на българския език МП се изследва и анализира в [32, 33].

В следващото изложение ще представим контрастивен анализ на две ФЕТ и набор от статистически техники, използвани за МП на български език към руски, словашки, сръбски и английски език.

4. DATR език за представяне на лексикална информация и използването му за машинен превод

DATR език за представяне на лексикална информация е създаден през 1989 г. в Университета на Съсекс, Великобритания [3] като опит да бъде подобрена трансформационната граматика на Чомски. Той е замислен като универсална ФЕТ с помощта на която могат да се представят различни езикови явления или техни характеристики (фонетика, морфология, синтаксис и др.) като се задават компютърно чрез граматичните им правила (с определена степен на обобщеност) за действие в рамките на общ компютърен езиков модел.

За тази цел се използва структурата на *семантична мрежа*, с която се представя съответното граматично знание за конкретния език (чрез граматичните му правила) като работеща компютърна програма, чийто принцип на работа е изравняването на зададените езикови стойности (фонеми, морфеми, думи и др.) с дефинираните междинни стойности от конструираните правила, като крайният резултат са нови фонемни промени, словоформи или синтактични цялости. С този работещ механизъм DATR подобрява значително трансформационната граматика, защото теоретически включва възможност за

представянето и на изключенията от правилата (което е характерно за естествения език) чрез задаването на нови правила за тях и в този смисъл се разглежда като универсална ФЕТ, с която могат да се представят езикови характеристики както на родствени, така и на неродствени езици.

DATR позволява граматичните правила на избрания естествен език да бъдат представени чрез структурата на семантична мрежа, в която всеки възел включва определени правила, които могат да бъдат наследявани или променяни чрез свързването му с друг, съседен възел. Така граматичното описание се представя по възможно най-кратък начин.

Отделно от това за DATR е разработена платформа за директна работа (в диалогов режим), а за много езици като английски, немски, испански и др. има направени програмни приложения, което го прави много удобен за многоезичен МП. Особено подходящи за това са славянските езици, повечето от които като български, руски, полски, чешки, словенски и др. също са представени в DATR чрез тази обща платформа. По същество тези приложения представят предимно словоизменение, което е характерно за славянските езици.

По-нататък ще сравним и анализираме представянето на словоизменителната морфология на български и руски език за целите на МП като се използва DATR като обща ФЕТ и терциум компарационис (основа за сравнението).

4.1. Използване на DATR за представяне на словоизменението на съществителните имена в български и руски език за машинен превод

Както вече бе отбелязано трансформационната граматика, която също е универсална ФЕТ, има ограничени възможности за представянето на словоизменителна морфология, защото е създадена главно за синтактична обработка на естествения език. В същото време славянските езици, повечето от които са падежни езици, са характерни със своето словоизменение. В този

смисъл DATR има възможности за представяне на славянска словоизменителна морфология, която се използва и за МП между тези езици [12, 13].

4.1.1. Представяне на компютърна морфология с DATR

По начало словоизменителната морфология е сложно да бъде представена компютърно, защото в процеса на словоизменението не само се прибавят флективни морфеми, но също се извършват и различни фонетични промени и затова процесът е труден за компютърно моделиране.

Съществуват два основни подхода за компютърно представяне на словоизменителна морфология: а) първият приема словоизменителната основа за постоянна и към нея прибавя словоизменителните морфеми, а б) вторият приема словоизменителната основа за вариращa и я прибавя към постоянна флективна морфема (най-често е използван, когато има звукови промени в основата на думата). За представянето на словоизменението на съществителните имена в българския език с DATR е използван още по-сложен подход, който съчетава вариращa основа и вариращи словоизменителни морфеми, за да се интерпретират сложните фонетични промени, които съпътстват процеса, които са исторически обусловени.

4.1.2. Използване на DATR за представяне на словоизменението на съществителните имена в български език

Подробен анализ на словоизменението на съществителните имена в българския език, представено с DATR е вече публикуван в [34], където са описани граматичните основания за формалното представяне и е даден програмният код на софтуера.

Тук накратко можем да отбележим, че в основата на това формално представяне е възприет типологичният подход, който представя флексията на българските съществителни имена като словоизменителни типове, всеки един от които е отделен възел в семантичната мрежа и съдържа своите граматични

правила за словоизменение на включените към него думи, които също са отделни възли от мрежата. Този подход не е традиционен, доколкото в българския език съществителните имена нямат падежи и съответно склонения, но имат словоизменение по отношение на род, число и определеност / неопределеност. Затова тези граматични категории на съществителните са възприети в DATR представянето точно в тази йерархия и служат в програмата за тригери, които променят стойността на получените словоформи.

Във връзка с това основният възел на семантичната мрежа, който съдържа правилата за словоизменението на съществителните имена съдържа променливата <gender>, която може да приема стойност на мъжки, женски или среден род и в зависимост от това се получават съответните словоформи. Останалите възли от семантичната мрежа съдържат граматични правила съответно за смяна на стойността на рода, числото или членната морфема. По този начин са дефинирани всички словоизменителни типове съществителни в българския език. Пример за словоизменението на българската дума „закон“, която използва словоизменителните правила на основния възел е както следва:

```
Zakon:

<> = Noun.
<root> = zakon.

Zakon: <gender> == masc_1.
Zakon: <sing undef> == zakon.
Zakon: <plur undef> == zakon_i.
Zakon: <sing def_1> == zakon_ut.
Zakon: <sing def_2> == zakon_a.
Zakon: <plur def_1> == zakon_i_te.
```

4.1.3. Използване на DATR за представяне на словоизменението на съществителните имена в руски език

Словоизменението на съществителните имена в руския език в DATR е създадено от известния британски езиковед и славист Г. Корбет. В основата му е идеята за компютърно представяне на склонението на съществителните имена в

руския език, главно според речника на Зализняк. Формалната структура на семантичната мрежа включва възли със словоизменителните правила на съществителните имена в руски език за всички склонения. Като основни са приети граматичните им категории род, падеж, число и одушевеност / неодушевеност и в зависимост от смяната на техните стойности, компютърната програмата дава съответните словоформи.

Пример за словоизменението на руската дума „закон“ според зададените граматични правила е както следва:

Zakon:

<> == N_I
<infl_root> == zakon
<sem animacy> == inanimate.

Zakon: <gloss> = law.
Zakon: <mor nom sg> = zakon.
Zakon: <mor acc sg> = zakon.
Zakon: <mor gen sg> = zakon _a.
Zakon: <mor dat sg> = zakon _u.
Zakon: <mor inst sg> = zakon _om.
Zakon: <mor loc sg> = zakon _e.
Zakon: <mor nom pl> = zakon _i.
Zakon: <mor acc pl> = zakon _i.
Zakon: <mor gen pl> = zakon _ov.
Zakon: <mor dat pl> = zakon _a _m.
Zakon: <mor inst pl> = zakon _a _m’i.
Zakon: <mor loc pl> = zakon _a _x.
Zakon: <syn gender> = masc.
Zakon: <syn animaey> = inanimate.

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 (независимо от това, че някои от словоформите в двата езика са омонимни), 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.

4.2. Използване на DATR за сравняване на представянето на притежателните и възвратното притежателно местоимение в български език

Представянето на притежателните и възвратното притежателно местоимение в българския език с DATR е част от първия модел на словоизменението [34] и следва общата идея то да бъде моделирано с оглед на граматичните категории род, число и определеност / неопределеност. Подробният преглед и кодът на софтуера включват също анализ и изясняване на сложната природа на граматичните категории, които тези местоимения имат.

Особеността в граматиката на притежателните местоимения е свързана със семантиката им да изразяват семантичната универсалия *притежателност*, която включва *притежател* и *притежаемо*. Затова те имат два типа категории род и число – словоизменителни и несловоизменителни, и в зависимост от това, коя от тези категориални йерархии е приета за основна, е възможно те да бъдат представени в DATR по два различни начина.

При първото представяне на притежателните местоимения в DATR са взети за основа техните словоизменителни категории и е следван модел, в който всички правила за словоизменението им ползват тези, създадени за прилагателните, а самите местоимения са зададени като отделни възли на семантичната мрежа чрез своите словообразувателни основи и несловоизменителните си категории род, число и лице.

За основа на второто представяне на тези местоимения в DATR [19] е взета йерархията на несловоизменителните им категории род, число и лице и е добавен един нов възел, който съотнася словоизменителните им категории към несловоизменителните такива, за да се получат съответните им словоформи.

Отделните местоимения са зададени към семантичната мрежа само със словообразувателните си основи.

Така двете представяния на притежателните местоимения в DATR се различават освен по начина на използването на вътрешната йерархия на граматичните им категории, а и по начина на представянето на самите местоимения – в първия случай несловоизменителните им категории са част от базата данни, а във втория те са част от граматичните правила.

Същата идея е използвана и за DATR представянето на възвратното притежателно местоимение. В случая DATR се използва като основа за сравнение (терциум компарационис) на двете представяния, което дава ново знание за начините и подходите, които могат да се прилагат за формалното представяне на тези части от речта в български език за целите на МП.

5. Универсалният мрежов език (УМЕ) и използването му за машинен превод

Универсалният мрежов език (УМЕ) е ФЕТ, която също като DATR представлява *семантична мрежа*, но с различна структура и повече възможности за използване в областта на МП. УМЕ е замислен като универсална формална езикова теория, с помощта на която може да бъде представяно по езиково независим начин граматично знание за съответния език (за всички нива на анализ – фонетика, морфология и синтаксис) и в този смисъл той се използва за МП. От създаването си досега [26] той бива усъвършенстван както по отношение на механизмите за представяне на граматично знание, така и по отношение на платформата му, с помощта на която е възможно да се създава езиков софтуер. Структурата на УМЕ включва три раздела: *универсални думи*, *атрибути* и *отношения*, които са свързани помежду си за работа в обща програмна среда.

Универсалните думи са отделни възли в семантичната мрежа и са лексикалната база данни или речника на УМЕ. Те съответстват на частите на

речта в естествения език и включват класове като съществителни, прилагателни, числителни и др., които са йерархично свързани помежду си и изграждат базата от знание (онтология) на УМЕ. *Универсалните думи* се попълват за всеки един естествен език поотделно като отразяват лексикалната му система, но не включват граматичните му правила.

Атрибутите са връзките на съответния възел в семантичната мрежа, които задават определена специфична за него информация, съответстваща на граматичните категории род, число, време и др. Те се задават чрез определени анотации и с тях този възел участва в компютърната обработка.

Отношенията са връзките, които свързват един възел с друг в семантичната мрежа, включват възможности за изразяване на семантична информация (отношения) между двата възела и са част от структурата на семантичната мрежа. Отношенията, с които възлите могат да бъдат свързани са предикативни, атрибутивни, пространствени и др. и те са краен брой. В различните подобрения на УМЕ има тенденция този брой да се намалява.

Комбинирането на *универсалните думи*, *атрибутите* и *отношенията* са основните формални средства, с които един естествен език може да бъде представен. УМЕ има разработени формални средства за компютърна обработка на много езици като английски, руски, български [35], словашки и др. като повечето от тях са направени в рамките на международния проект „The Little Prince Project (LPP)" [7] за целите на МП.

5.1. Използване на УМЕ за представяне на словоизменението на съществителните имена в българския език

Формалното представяне на словоизменителната морфология на съществителните имена в българския език с DATR, който също е семантична мрежа, е анализирано в 4.1.2., където е отбелязано, че в основата на приетата йерархия са граматичните категории род, число и определеност / неопределеност. Представянето на словоизменителната морфология на

съществителните имена в българския език с УМЕ е анализирано в [14], където са дадени и общите характеристики на модела, а сравнителен анализ на DATR и УМЕ представянията е публикуван в [18].

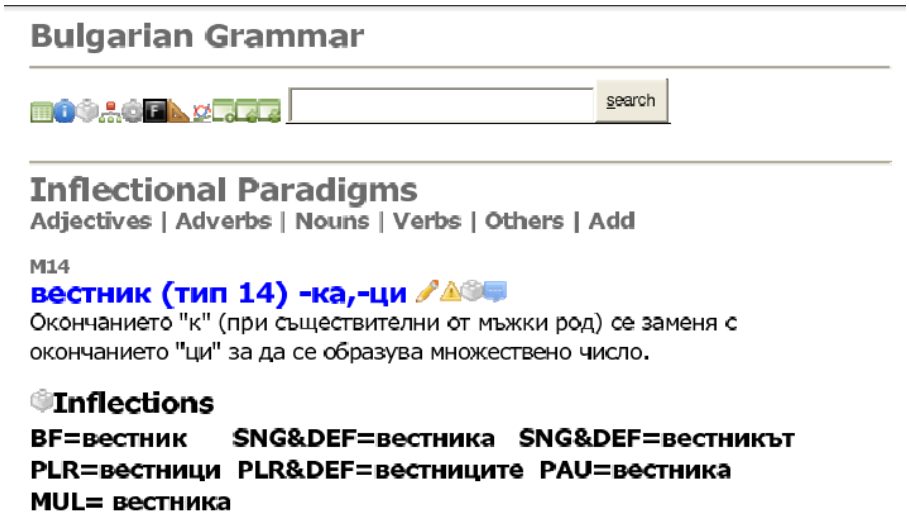
Важна особеност на представянето на съществителните имена в УМЕ е използването на стандартните правила за словоизменение, които тази формална теория предлага и които са четири типа: а) за прибавяне на представки, б) за добавяне на наставки, в) за вмъкване на морфеми и г) за заместване. Тези правила ограничават представянето на българското словоизменение до чисто техническото им използване. Словоизменението на съществителните имена е най-обемната част от компютърното представяне на българския език в УМЕ и е както следва:



Statistics

Class	UNL-NL Dictionary			NL Dictionary
	UWs	Lemmas	Polysemy	
Adjectives	367	603	1.64	397
Adverbs	179	237	1.32	203
Nouns	3,671	4,257	1.16	754
Verbs	587	915	1.56	692
Other	27	69	2.56	100
All	4,831	6,081	1.26	2,146

То съдържа 74 словоизменителни типа (значително повече от тези за DATR), защото те са формулирани чрез използването на стандартните словоизменителни правила на УМЕ. Примерът за получените словоформи на думата „вестник“ е както следва:



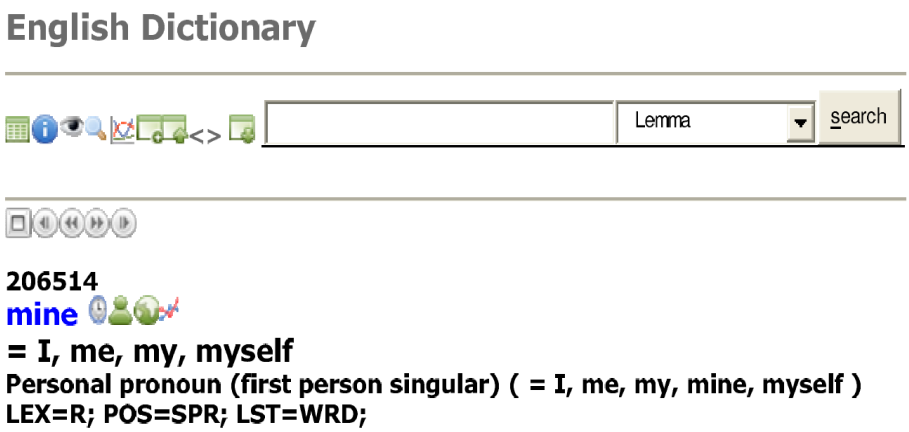
В същото време УМЕ предлага по-големи възможности за представяне на езикова информация, защото включва представянето както на граматични, така и на семантични (отношения) езикови характеристики. Двете формални представяния на словоизменението на българските съществителни - DATR и УМЕ си приличат по това, че и двете включват в структурата си словоизменителни типове и лексикална база данни, състояща се от съществителните имена, зададени като отделни възли в семантичната мрежа, но УМЕ предлага възможност и за задаване на семантична информация за тях.

5.2. Използване на УМЕ за представяне на притежателните и възвратното притежателно местоимение в български, английски и руски език

Както беше вече посочено в раздел 4.2. при представянето на притежателните и възвратното притежателно местоимение в DATR принципите, които са използвани за него имат само граматични основания (йерархията на категориите род, число и определеност / неопределеност). Също бе отбелязано, че УМЕ дава възможност и за семантични основания за представяне на езиковата информация, което съществено допълва формалното описание. Представянето на притежателното и възвратното притежателно местоимение с УМЕ за български, английски и руски език е дадено и анализирано в [15, 17].

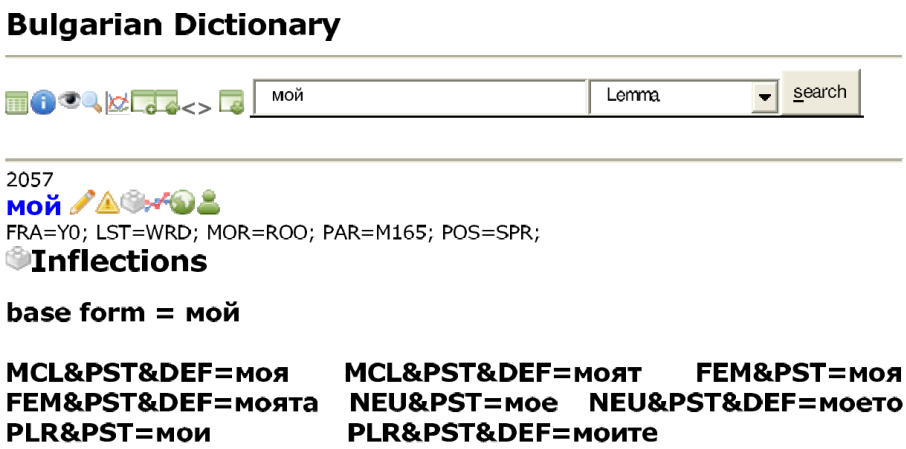
Трябва да отбележим, че притежателните местоимения в английски и руски език също изразяват универсалната семантика на *притежание* (в която различаваме *притежател* и *притежаемо*) чрез своите граматични категории. За английски език тази разлика е синтактична и е представена чрез различните части на речта като притежателни прилагателни и притежателни местоимения, които имат различни синтактични функции, а за руски език това са граматичните категориите падеж, род и число.

УМЕ предоставя възможност всичките тези граматични и семантични особености на притежателните местоимения и за трите езика да бъдат представени в обща компютърна програма за МП. Така английското притежателно местоимение *mine* в УМЕ е зададено по следния начин:



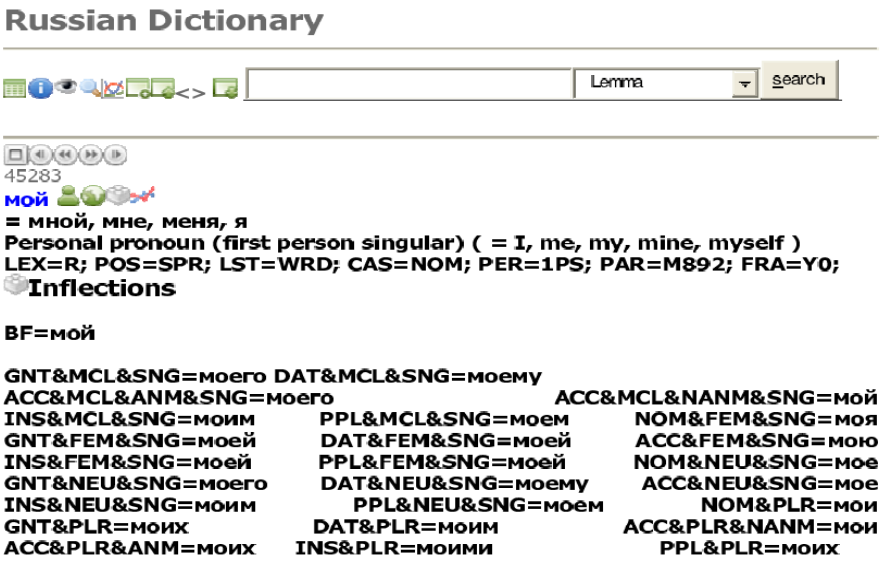
То е представено като лексема (универсална дума), която е съотнесена с други местоимения от 1 л., ед. ч., (I, me, my, myself) както и чрез спецификациите: LEX=R, означаваща, че принадлежи към лексикалната категория на местоименията, POS=SPR, означаваща, че изпълнява синтактичната функция на спецификатор и LST=WRD, означаваща, че принадлежи към списъка на универсалните думи.

А българското притежателно местоимение *мой* е зададено в УМЕ така:



То също е представено като лексема (в основна форма) и чрез спецификациите: LST=WRD, означаваща, че принадлежи на списъка от универсални думи, MOR=ROO, означаваща, че за словоизменението му се използва основата на думата, PAR=M165, означаваща, че за словоизменението се използват правилата на флективен тип M165, POS=SPR означаваща, че изпълнява синтактичната функция на спецификатор.

Задаването на руското притежателно местоимение *мой* в УМЕ е:



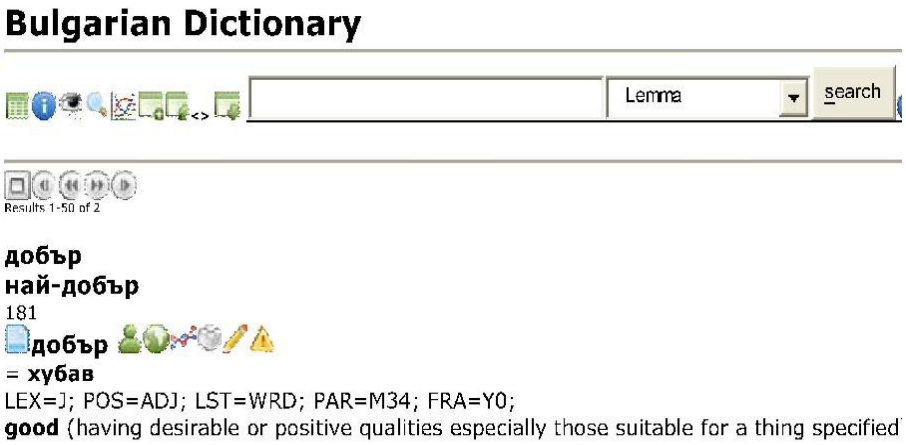
То е представено като лексема (в основна форма) и включва спецификациите: CAS=NOM, означаваща, че формата е в именителен падеж, PER=1PS, означаваща, че формата е от 1 л., ед. ч. Останалите спецификации са същите, както за английското и българското притежателно местоимение – LEX=R, POS=SPR и LST=WRD, и служат за свързването им при изпълнението

на програмата (логическа еквивалентност) за машинен превод, което осигурява семантична и съответно преводна еквивалентност.

5.3. Използване на УМЕ за машинен превод между български и словашки език

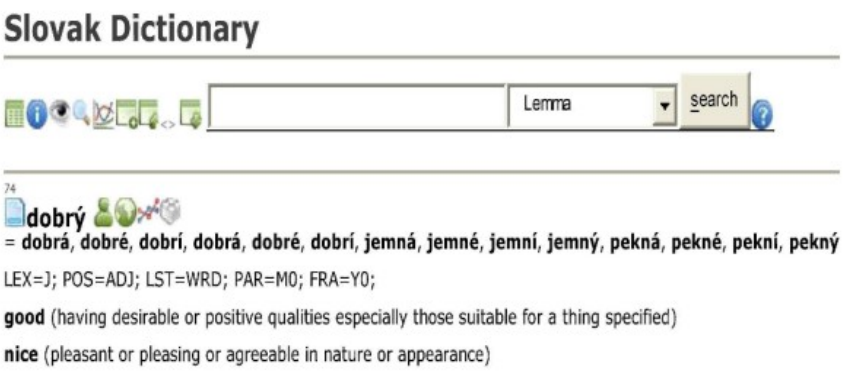
В рамките на подхода, който в УМЕ се използва е възможно да бъде осъществен МП между български език и други родствени славянски езици като словашкия език. Условието за това е и двата езика да са представени формално в рамките на УМЕ и анализ на такова представяне е публикуван в [16, 25], който обхваща както граматичните, така и семантичните подходи, използвани за двата езика в рамките на проекта LPP [7].

И за двата езика са разработени формални граматични правила за всички части на речта и лексикална база данни, включваща и семантични отношения. Така например българското прилагателно „добър“ е част от лексикалната база данни и е зададено както следва:



Освен граматични характеристики като основна форма и спецификациите LEX=J, POS=ADJ, LST=WRD, PAR=M34 и др. за това прилагателно е зададена и семантична информация чрез дефиницията му на английски език.

По сходен начин е представено и словашкото прилагателно за „добър“:



За него е представена граматична информация за основната и всичките му флективни форми и спецификациите LEX=J, POS=ADJ, LST=WRD и др. и семантична информация чрез дефиницията му на английски език и чрез задаване на синоними (и на английски и на словашки език).

Структурата на УМЕ включва представяне на семантична информация не само чрез *универсалните думи* (онтологията), но и чрез *отношенията*. За българското прилагателно *добър* тези отношения са представени така:



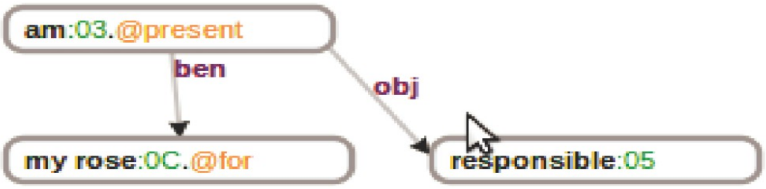
Те включват лексикално-семантичната дефиниция на прилагателното, както и неговите синоними и антоними. Този механизъм, допълнен с общите спецификации, зададени за прилагателното в двата езика, осигуряват съответствието на семантичен и преводен еквивалент при МП.

5.4. Използване на УМЕ за измерване и подобряване на качеството на машинния превод

Платформата на УМЕ е създадена така, че да включва двата основни процеса при МП, а именно представяне на изходния естествен език с ФЕТ (което става с намесата на човека) и трансформирането на това формално представяне в преводен еквивалент на целевия език (което става автоматично).

За двата процеса се използват различни програми и затова е възможно компютърно те да бъдат сравнени чрез сравняване на изходния и получения превод с формалното представяне и по този начин да бъде измерено качеството на превода. Също чрез подобряването на това съответствие е възможно да се подобри семантичният, съответно преводният еквивалент.

Пример за това е представянето на притежателните местоимения в английски и български език и по-специално граматичните правила (словоизменителни и синтактични) за МП на изречения, които конвертират езиковата информация от английски език в запис на УМЕ. Пример за такова конвертиране в УМЕ на изречението *I am responsible for my rose* е както следва:



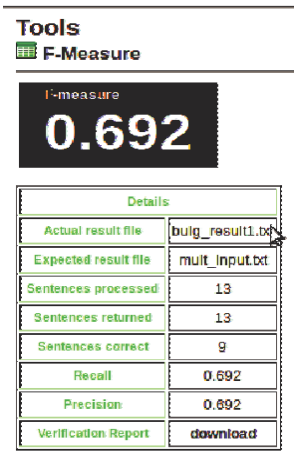
Това изречение е представено формално чрез граматиката на включените в него думи, свързани от отношенията *ben* и *obj*. Този формален запис може да бъде автоматично деконвертиран в превод на български език както следва:

```
[S:1]
{org}

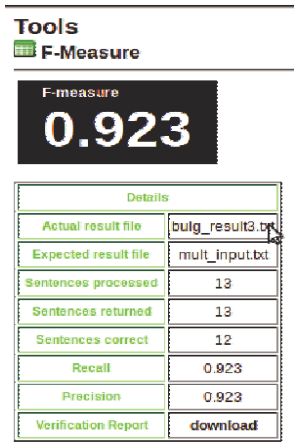
{/org}
{bul}
Аз съм отговорен за моята роза
{/bul}
{eng}

{/eng}
{unl}
ben("am":03.@present, "my rose":0C.@for)
 obj("am":03.@present, "responsible":05)
{/unl}
[/S]
Dictionary Lookup Time 0 seconds, 4 milliseconds.
Disambiguation Time 0 seconds, 0 milliseconds.
Transformation Time 0 seconds, 243 milliseconds.
Total Time 0 seconds, 247 milliseconds.
```

Измерването на съответствието на формалното представяне на изходния език и това на целевия език става с коефициента F-measure. Резултати от такова измерване на изречения за МП между английски и български език са:



След подобряване на граматичните правила и за двата езика коефициентът F-measure се повишава, както и качеството на превода чрез подобряване на семантичната еквивалентност, което подобрява преводната еквивалентност.



6. Sketch Engine и използването му за машинен превод

Съвременните подходи за търсене в електронни езикови бази данни използват статистически подходи, чиято цел е да се получи резултат, който е семантично възможно най-близък до търсения. В това отношение има разработени различни статистически техники, основани главно на определяне на статистическа близост и вероятности между зададения и получения при търсенето резултат, който обикновено е дума или израз от естествения език.

В същото време такива статистически подходи се използват и за откриване на смислови връзки и отношения (откриване на ново знание) между

думите от обработваните електронни документи (текстове). Това се основава на приемането на *статистическата близост* за *семантична близост*.

Затова тези подходи могат да бъдат използвани и за МП като търсенето се извършва в двуезични паралелни или сравнителни електронни текстови корпуси и за двата езика, а *семантичната близост* между зададения на изходния език и получения на целевия език резултат се приема за *преводен еквивалент*. Също така използваните статистически подходи се приемат за езиково независимо универсално представяне.

За резултатите от МП, които се представят и анализират по-нататък се ползва машината за търсене Sketch Engine (SE). Тя е създадена [4] за подпомагане на лексикографската работа, свързана с определянето на семантичните характеристики на думите въз основа на статистическо търсене в големи електронни текстови масиви (електронни корпуси). За това е използван набор от статистически техники (T-score, MI-score и MI3-score), с които SE позволява да се определят главно статистическа близост между думи и вероятности, с които техники се търсят и получават *конкордансии*, *колокации*, *семантични отношения* и *тезаурус* на търсената ключова дума.

6.1. Използване на SE за машинен превод на екологични термини между български и словашки език

Както бе отбелязано вече едно от първите успешни приложения на МП е било в областта на техническия превод [10], който е свързан с превода на специализирана лексика (термини). Оттогава МП активно се използва за такъв превод, което е свързано с превод на термини от различни административни текстове и от съответните им документи.

Затова изследването ни [22] използва специализирания електронен многоезичен корпус EUROPARL 7, който съдържа преводи от 21 европейски езика на текстовете от речите на депутатите от Европейския парламент, които са структурирани като отделни паралелни едноезични корпуси и които не са

анотирани и не ползват формални граматики. По-специално използваме 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.

Екологичната терминология има широка сфера на употреба и е интересно да бъде изследвана във връзка с административната и политическата лексика. Тук представяме резултати от статистическо търсене на конкорданси със SE за български екологични термини и откриването на техните словашки преводни еквиваленти като проследяваме и симетричността в превода.

Трябва да отбележим, че търсенето на по-сложни екологични термини като например *биоценоза* не дава резултат. При паралелното търсене в българската част на EUROPARL 7 за термина *околна среда* се получава следния резултат:



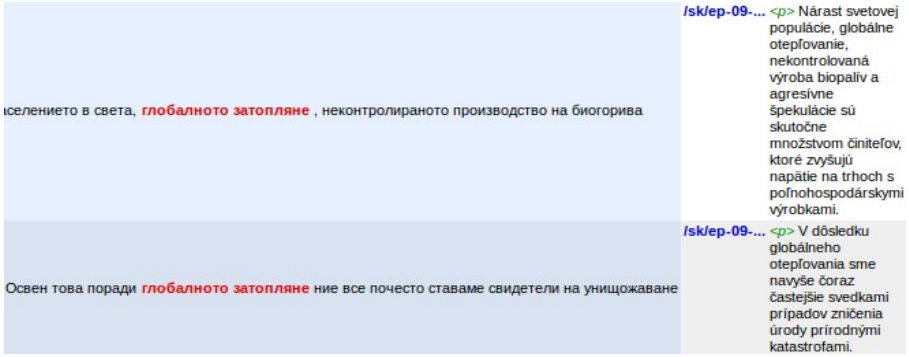
Той включва част от наименованието *Европейска агенция за околната среда* и съответния ѝ словашки превод, както и словашкия превод на самия термин *životne prostredie*.

Резултатът от търсенето за термина *биологично разнообразие* е:



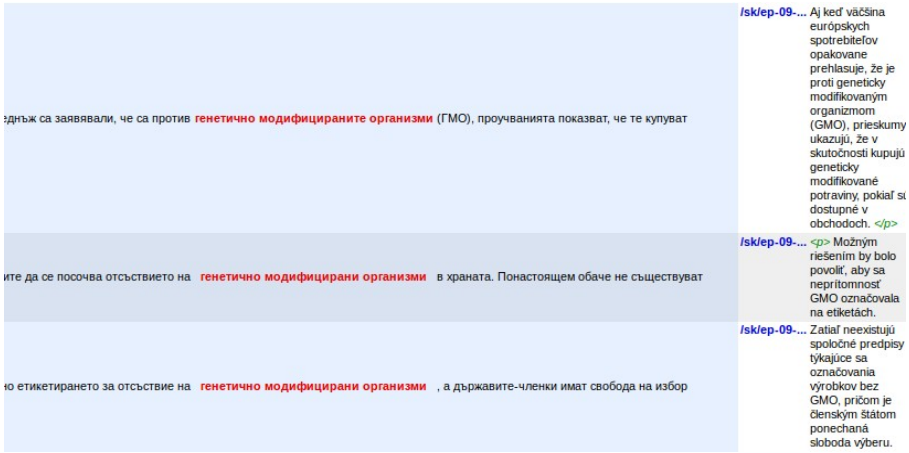
Той дава като словашки преводен еквивалент възприетия от английски език термин *biodiversity*.

Резултатът от търсенето за термина *глобално затопляне* е:



Той дава словашкия превод на термина, който е *globalne otepľovanie*.

Резултатът от търсенето за термина *генетично модифициран организъм (ГМО)* е:



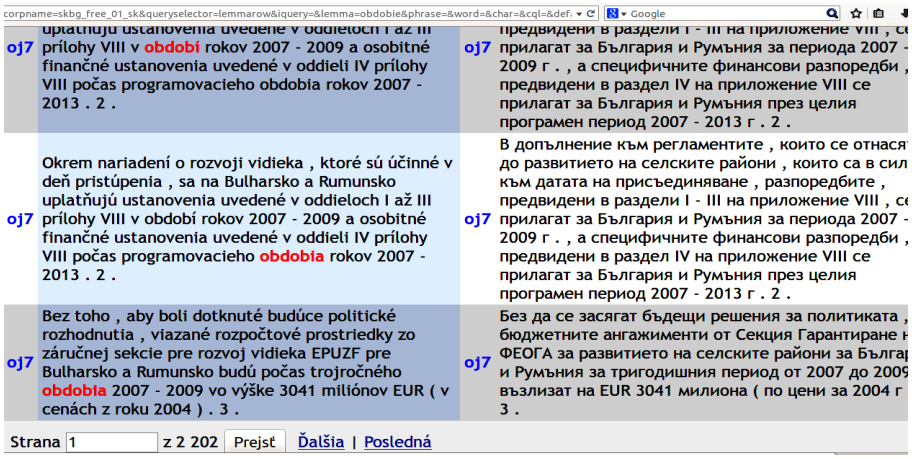
Той определя за словашкия превод на термина абревиатурата *GMO*.

Така можем да обобщим, че статистическите подходи на SE за търсене в двуезичния паралелен Българско-словашки EUROPARL 7 корпус за МП на екологични термини дават добри резултати по отношение на преводната еквивалентност, макар терминът от изходния език и полученият преводен еквивалент да не са симетрични. Допълнително предимство на този подход е, че той е много бърз и не изисква разработване на формални граматики за преводните езици, а само достатъчно голям и специализиран двуезичен електронен текстов корпус.

6.2. Използване на SE за машинен превод на изрази за време между български и словашки език

Статистическите подходи на SE могат да бъдат използвани и с други двуезични електронни корпуси за превод на други видове лексика. Така например изразите за време са специфични в повечето езици и правилният им превод често е от голямо значение за информативността на текста.

В следващото изследване [21] се прави анализ на възможностите за МП на изразите за време в словашки и български език като се представят резултати от статистическо търсене със SE в съществуващите Българско-словашки електронни корпуси на *конкорданси* и *колокации*. Използват се два двуезични корпуса, чиито резултати се сравняват - паралелен Българско-словашки EUROPARL 7 корпус и първият създаден паралелен Българско-словашки корпус. Той включва текстове от административни периодични издания като „Official Journal of European Union” и др., преведени на словашки и български език като в словашката си част съдържа 77 764 528 словоформи, а в българската си част съдържа 84 947 528 словоформи и използва също SE за статистическо търсене. Резултатът от двуезичното паралелно търсене в този корпус на словашката дума *obdobie* е както следва:

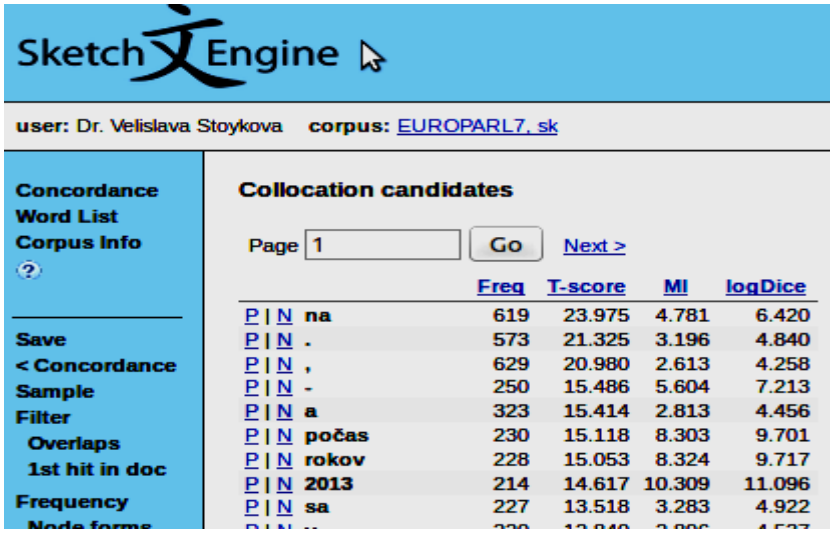


Той показва, че думата (и словоформите ѝ) има 44 039 срещания в словашката част от корпуса, а в българската част нейният преводен еквивалент е думата *период*. За да могат обаче да бъдат изследвани изразите, в които тази дума участва е необходимо да бъде извършено статистическо търсене за тази дума чрез определяне на нейните колокации. Резултатите от това търсене са:

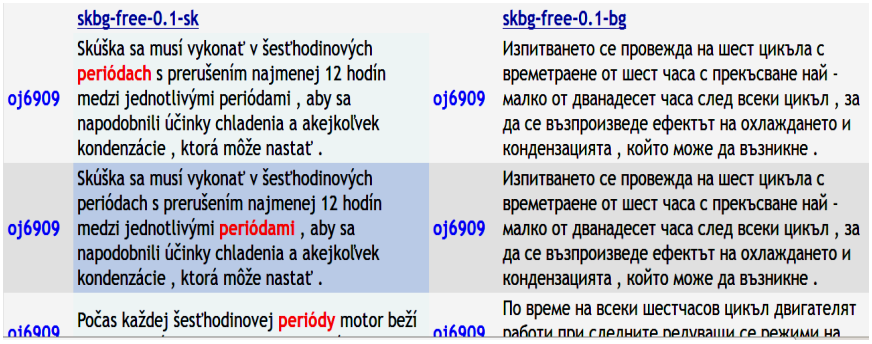


Те показват, че думите, които най-често се срещат с думата *obdobie* (които лексикално-семантично се съчетават с нея) са *rokov*, *počas* и др. като образуват словосъчетанията *obdobie rokov* и *počas obdobie*, които на български се превеждат като *годишен период*, *през* (*време на*, *целия*) *период*(*а*).

Същите резултати се получават и при статистическо търсене за колокациите на думата *obdobie* в паралелния Българско-словашки EUROPARL 7 корпус, които са както следва:



Важно е да се отбележи, че в словашки език съществува дума *peroida*, която също се използва в изрази за време. Резултатът от статистическото търсене за тази дума в Българско-словашкия корпус е:



Този резултат задава българският превод на тази дума да е *цикъл*. Примерът показва, че статистическите подходи на SE могат да помогнат както за търсене на преводни еквиваленти на отделна дума или изрази с нея, така и за различаването при случаи на междуезикова омонимия („лъжливите приятели на преводача“) чрез повторно търсене за същата дума, особено при превод между родствени езици.

6.3. Използване на SE за машинен превод между български и словашки език

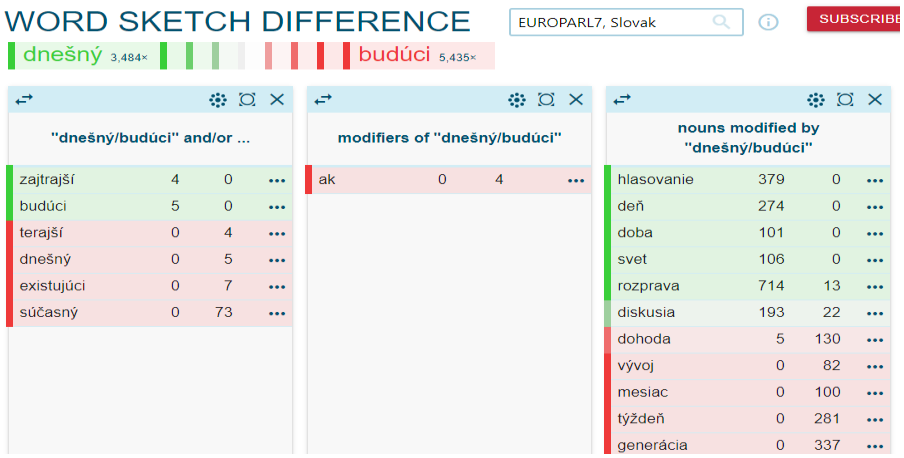
Статистическите подходи на SE предлагат и по-сложни техники за семантично търсене в електронни текстове за получаване на граматични и семантични характеристики на търсената дума, главно чрез статистическо

измерване на нейните контексти. В следващото изследване [24] се представя анализ на резултати от използването на такива техники за търсене в двуезичния паралелен Българско-словашки EUROPARL 7 корпус чрез генериране на конкорданси, колокации, семантични отношения и тезаурус на ключовата дума. За изследването е избрана словашката дума *dnešný* (*днешен*) и за нея последователно се правят статистически търсения, които на всеки следващ етап разширяват търсенето. Резултатът за колокациите на думата *dnešný* е:

	Word	Cooccurrences	Candidates	T-score	MI	MI3
1	rozprava	160	3,199	12.61	8.51	23.16
2	rozpravy	132	2,418	11.46	8.64	22.73
3	rozprave	99	1,526	9.93	8.89	22.15
4	dňa	78	1,130	8.81	8.98	21.55
5	hlasovanie	119	3,091	10.87	8.14	21.93
6	dobe	56	830	7.47	8.95	20.56
7	hlasovaním	48	713	6.91	8.94	20.11

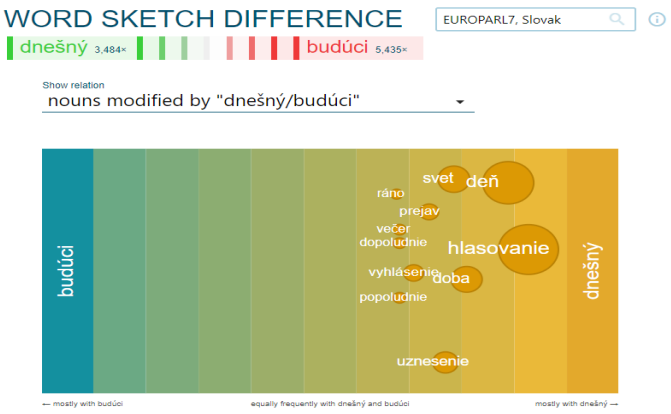
Този резултат включва думите, които образуват словосъчетания със зададената дума като *dnešná rozprava* (днешно разискване), *dnešný deň* (днешен ден), *dnešnô hlasovanie* (днешно гласуване), *dnešná doba* (днешно време). Следователно търсенето за колокации дава лексикалната съчетаемост на думата, което показва, че с използването на определени статистически подходи е възможно да се покажат семантични характеристики на думата.

Резултат от търсенето на семантични отношения за думата *dnešný* е:

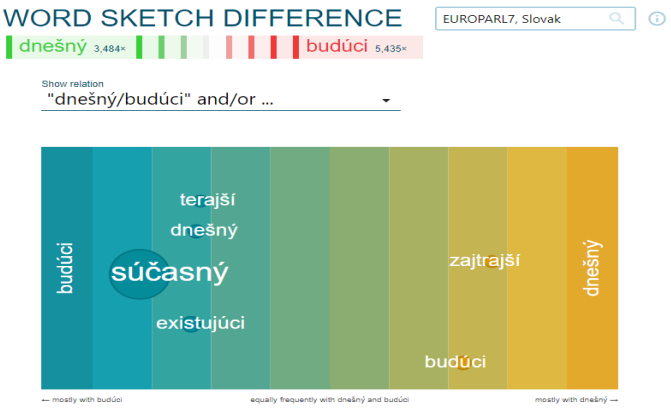


Той показва, че изследваната дума участва в два главни типа семантични отношения „*and/or*“ (синонимни) и „*nouns modified by*“ (атрибутивни) и дава

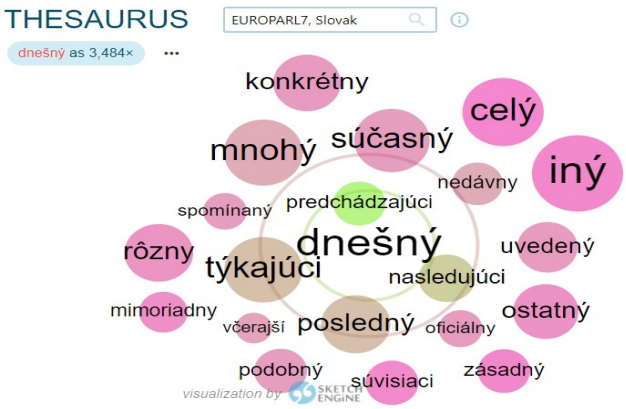
списъци от думи и за двете отношения. Например списъкът от думи за второто отношение съдържа по-разширен списък от думите, които вече бяха получени при първото търсене и с които изследваната дума образува словосъчетания. Допълнително резултатът е визуализиран така:



Списъкът от думи за първото отношение (синонимия) съдържа думи, които са синоними на изследваната дума като *terajši* (сегашен), *existujúci* (съществуващ), *sučasný* (съвременен) и др., както и техни антоними (изразяващи същото отношение, но с по-ниска честота) и е визуализиран така:



В SE е възможно и откриването на още по-сложни семантични връзки на изследваната дума като се използва търсене за получаване на тезаурус. То е основано на факта, че ако две думи имат общи колокации, то те са семантично свързани. Резултат от търсенето на тезаурус на думата *dnešný* е визуализиран:



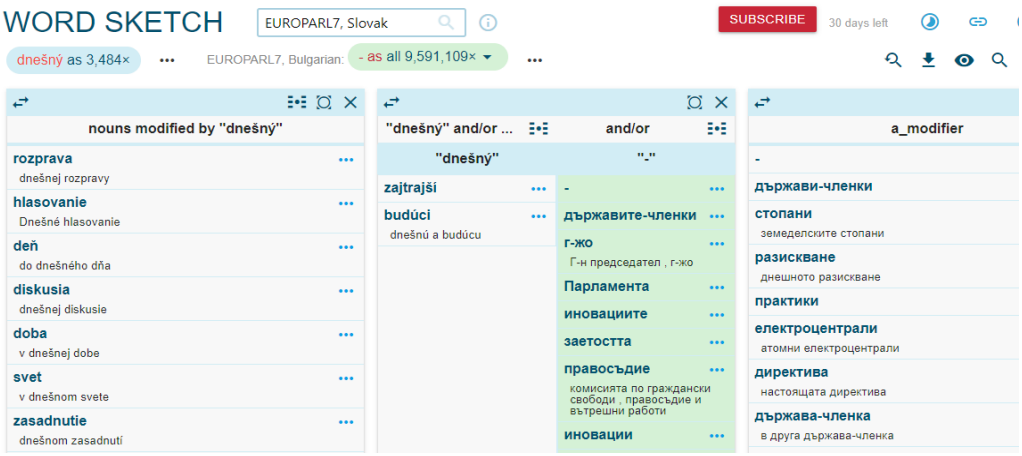
Той дава още по-сложни семантични отношения като йерархичните връзки, които тази дума има и които са визуализирани с различни цветове.

Разгледаните статистически техники за търсене могат да бъдат използвани и за търсене в българската част на корпуса. Резултати от търсенето на колокации на думата *днешен* са както следва:

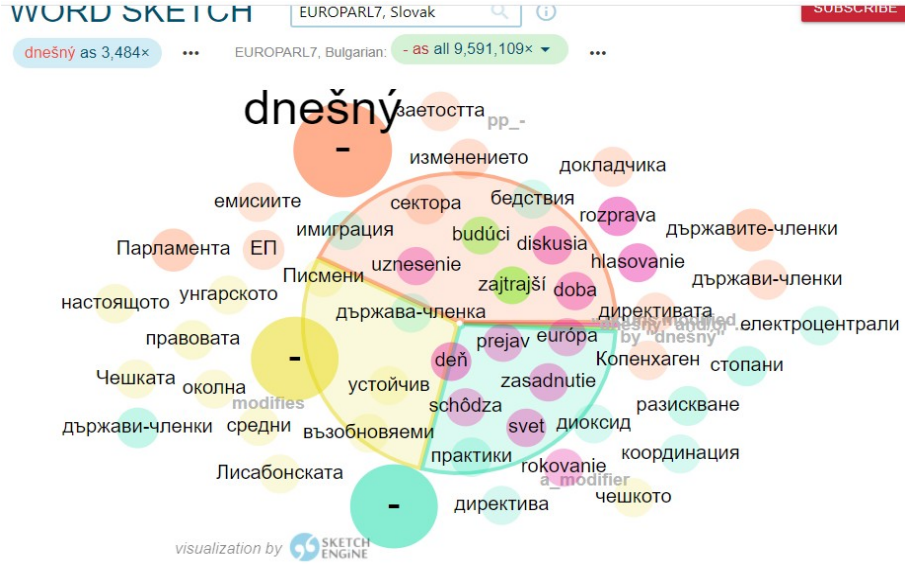
	Word	Cooccurrences	Candidates	T-score	MI	MI3
1	разискване	337	4,017	18.32	9.07	25.86
2	гласуване	197	2,303	14.01	9.09	24.34
3	дата	99	348	9.94	10.83	24.09
4	ден	88	1,600	9.35	8.46	21.38
5	вот	54	1,304	7.32	8.05	19.56
6	заседание	57	1,573	7.52	7.85	19.52
7	свят	54	2,132	7.30	7.34	18.85

Те представляват точен превод на тези, получени за словашката дума *dnešný* като *разискване, гласуване, време* и *др*., които образуват и същите словосъчетания – *днешно разискване*, *днешно гласуване*, *днешно време* и др.

По същия начин е възможно да се направи двуезично паралелно търсене за семантичните отношения на думата *dnešný* в целия корпус, което дава всичките преводни еквиваленти във връзка с отделните семантични отношения:



Възможно е и двуезично паралелно търсене за тезаурус на думата *dnešný*, чиято визуализация е както следва:



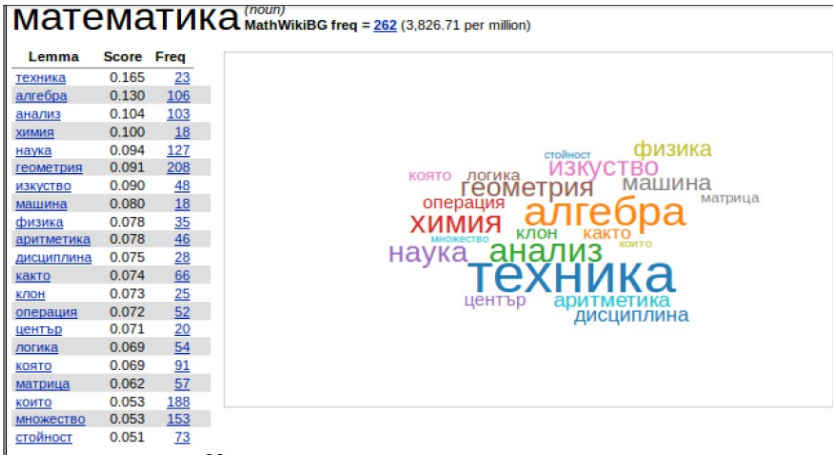
В заключение трябва да се отбележи, че статистическите подходи на SE за двуезично паралелно търсене в електронни корпуси могат да бъдат използвани за откриване на сложни семантични характеристики на думите, свързани с тяхната лексикална съчетаемост и независимо от това, че не използват формални граматични или семантични теории могат успешно да служат за МП.

6.4. Използване на SE за машинен превод на математически термини между български и сръбски език

Както бе посочено статистическите подходи на SE могат да определят и много сложни семантични характеристики на изследваната дума като нейните синоними (антоними), лексикална съчетаемост и тезаурус (йерархични връзки). Също бе отбелязано, че със SE могат успешно да се получават преводни еквиваленти на екологични термини при МП.

В следващото изследване [23] се представят и анализират резултати от използването на двуезични сравнителни електронни корпуси и статистическите подходи на SE за МП на математически термини между български и сръбски език. Електронните корпуси за това изследване са специално създадени – това са Български математически уикипедия корпус (MathWikiBG) – 90 000 словоформи и Сръбски математически уикипедия корпус (MathWikiSR) – 103 000 словоформи. Двата корпуса не са анотирани, но включват кратки формални граматики за синтактичен анализ съответно за български и сръбски език. Статистическото търсене включва определянето на семантични отношения и тезаурус за термина *математика* и от двата корпуса.

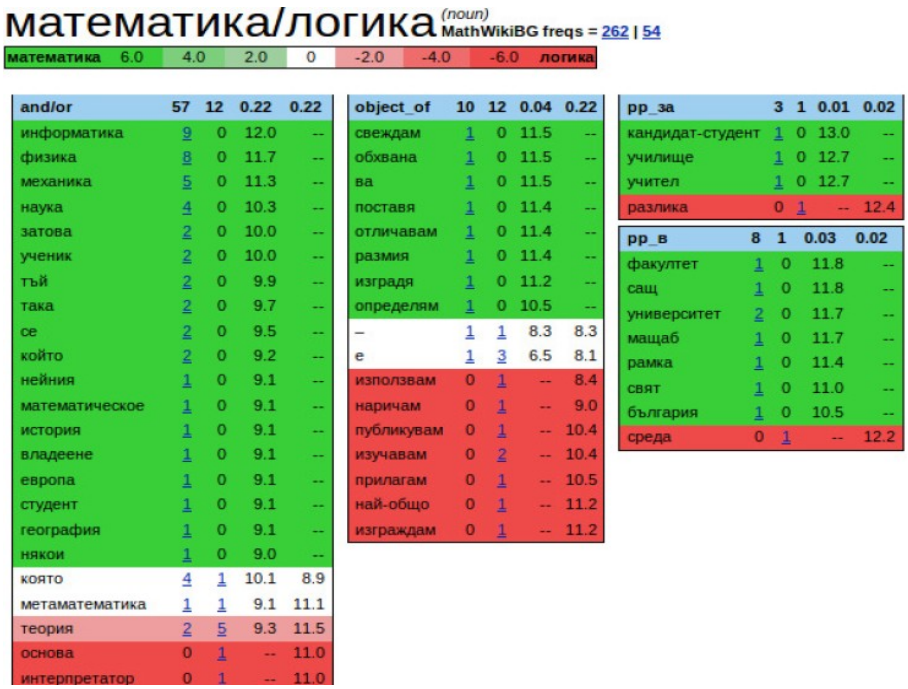
Терминът *математика* има 262 срещания в корпуса MathWikiBG и 230 срещания в корпуса MathWikiSR. По-нататък изследваме връзките на този термин (понятие) с други термини (понятия) в двата корпуса. Резултатите са както следва:





Те показват, че терминът *математика* е семантично свързан с *дисциплина* (SR – oblast), *наука*, *техника*, *алгебра*, *геометрия*, *логика* и др. Тези резултати изразяват ясна йерархия помежду си и дават тезаурусни отношения.

В същото време е възможно за изследваната дума *математика* да бъдат проследени и други видове семантични отношения като „and/or” и “modified by”. Резултатите от търсенето за семантични отношения са както следва:



математика/логика *(noun)* MathWikiBG freqs = 262 | 54

математика 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 логика

and/or	57	12	0.22	0.22
информатика	9	0	12.0	--
физика	8	0	11.7	--
механика	5	0	11.3	--
наука	4	0	10.3	--
затова	2	0	10.0	--
ученик	2	0	10.0	--
тъй	2	0	9.9	--
така	2	0	9.7	--
се	2	0	9.5	--
който	2	0	9.2	--
нейния	1	0	9.1	--
математическое	1	0	9.1	--
история	1	0	9.1	--
владеене	1	0	9.1	--
европа	1	0	9.1	--
студент	1	0	9.1	--
география	1	0	9.1	--
някои	1	0	9.0	--
която	4	1	10.1	8.9
метаматематика	1	1	9.1	11.1
теория	2	5	9.3	11.5
основа	0	1	--	11.0
интерпретатор	0	1	--	11.0

object_of	10	12	0.04	0.22
свеждам	1	0	11.5	--
обхвана	1	0	11.5	--
ва	1	0	11.5	--
поставя	1	0	11.4	--
отличавам	1	0	11.4	--
размия	1	0	11.4	--
изградя	1	0	11.2	--
определям	1	0	10.5	--
–	1	1	8.3	8.3
е	1	3	6.5	8.1
използвам	0	1	--	8.4
наричам	0	1	--	9.0
публикувам	0	1	--	10.4
изучавам	0	2	--	10.4
прилагам	0	1	--	10.5
най-общо	0	1	--	11.2
изграждам	0	1	--	11.2

pp_за	3	1	0.01	0.02
кандидат-студент	1	0	13.0	--
училище	1	0	12.7	--
учител	1	0	12.7	--
разлика	0	1	--	12.4

pp_в	8	1	0.03	0.02
факултет	1	0	11.8	--
сащ	1	0	11.8	--
университет	2	0	11.7	--
мащаб	1	0	11.7	--
рамка	1	0	11.4	--
свят	1	0	11.0	--
България	1	0	10.5	--
среда	0	1	--	12.2



matematika/logike MathWikiSR freqs = 52 | 26

matematika 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 logike

words before "%w"	33	10	0.63	0.38
Diskretna	6	0	12.3	--
Finansijska	4	0	11.8	--
Primenjena	4	0	11.8	--
primenjena	4	0	11.8	--
i	4	0	4.6	--
da	3	0	3.7	--
se	4	3	4.9	4.4
je	4	4	4.4	4.4
od	0	3	--	7.0

words after "%w"	47	17	0.90	0.65
matematike	4	0	9.3	--
iz	3	0	7.8	--
može	3	0	7.6	--
od	4	0	7.1	--
za	3	0	5.9	--
je	9	0	5.8	--
na	4	0	5.6	--

Те показват, че за свързващи термини по отношението „and/or” се получават тезаурусните отношения между термините: *информатика*, *физика*, *механика*, *наука* и др. За отношението “modified by” се получават колокациите на изследваната дума, които са подвидовите термини като *приложна математика* (SR - primenjena matematika), *чиста математика*, *дискретна математика* (SR – diskretna matematika), *изчислителна математика* и др.

В заключение трябва да се отбележи, че със статистическите подходи на SE 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.

7. Заключение

В представените примери на подходи за МП на български език се анализират използването на семантичните мрежи и съответните им ФЕТ, както и статистическите подходи като универсално езиково независимо междинно представяне на езиковите данни както на изходния, така и на целевия за превода език, с помощта на които чрез компютърна обработка (автоматично) се получават преводните еквиваленти. Качеството и точността на тази *преводна еквивалентност* е свързано с *логическата еквивалентност* и *семантичната еквивалентност*, зададени предварително чрез компютърната програма.

В тази връзка представянето на резултатите върви успоредно с контрастивния анализ на възможностите на трите подхода да бъде получен най-точния преводен еквивалент. Докато DATR и УМЕ изискват предварително формално представяне на граматиките и на двата, участващи в превода езика (чрез граматични правила), то SE използва статистически техники за изчисляване на статистическа близост (приета като семантична близост) и вероятности, с помощта на които могат да бъдат определяни много по-сложни лексикално-семантични и понятийни отношения на думите. В този смисъл DATR и УМЕ дават граматично по-верен преводен еквивалент, докато SE дава семантично по-адекватен преводен еквивалент.

Съществена страна и на трите разгледани подхода за МП е и фактът, че те не включват онези страни от превода като човешка дейност, при които се използват по-сложни когнитивни механизми като допълнителни знания, интерпретиране на културологичен и антропологичен контекст и др. Това е така, защото и при трите подхода действието на компютърната програма е ограничено от предварително зададения модел на езиковите данни.

Основни научни резултати и приноси в дисертацията:

За първи път в изследователската практика и в граматичната литература за българския език се представят:

1. Подробен и систематичен анализ на подходите за машинен превод на български език във взаимовръзката им със съвременните формални езикови теории, съответните им програмни приложения и с използваните статистически подходи, както и зависимостта на машинния превод от начина на представянето на езиковите данни, използвани за това.
2. Сравнителен анализ на програмни приложения (софтуер) за машинен превод на български език, създадени въз основа на използването на формални езикови теории, който е в рамките на традицията за контрастивните езиковедски изследвания като основата за сравнение (*терциум компарационис*) са както формалните теории и подходи, така и различните формални представяния на едни и същи езикови фрагменти (напр. словоизменението на съществителните имена или на притежателните местоимения) в рамките на различни формални езикови теории.
3. Установява се взаимовръзката и единството между *семантична еквивалентност*, *логическа еквивалентност* и *преводна еквивалентност* като главно и задължително условие за коректността на машинния превод на български език.
4. Сравнителният анализ на резултатите от работата на компютърните програми за машинен превод на български език е направен с оглед на изходните и получените езикови данни, които са разгледани, анализирани и дадени като отделни примери.

5. Определя се начин за компютърно измерване на качеството на резултата от машинния превод на български език, както и възможностите за подобряването му чрез подобряване на формалното представяне (езиковите правила) с оглед на по-точния и адекватен избор на преводния еквивалент.

6. Създаден е електронен корпус от специализирани математически текстове MathWikiBG, който е използван за статистически машинен превод и са представени резултати от такъв превод между български и сръбски език.

7. Установяват се границите на използването на машинния превод на български език като допълнителен и подпомагащ превода, извършван от човека.

Публикации, които са представени към дисертацията:

1. Stoykova, V. 2011. Common Formal Framework for Multilingual Representation of Inflectional Morphology for two Related Slavonic Languages. In: R. Garabik (ed.) Proceedings of 6th International Conference of NLP Multilinguality, Tribun EU, Bratislava, 146-155. ISBN: 978-80-263-0049-6

2. Stoykova, V. 2011. Representing Nominal Inflectional Morphology for Slavonic Languages in DAlTR In: Ganchev, H., Lowe, B., Norman, D., Sosokov, I. and Soskova, A. (eds.), Models of Computation in Context: 7th Conference on Computability in Europe, CiE 2011, 234-245, St. Kliment Ohridcki University Press, Sofia ISBN: 978-954-07-3232-9

3. Stoykova, V. 2013. Representation of Lexical Information for Related Languages with Universal Networking Language. Proceedings from the 6th Language Technology Conference, 514-517. ISBN: 978-83-07-932640-3-2 Google Scholar

4. Stoykova, V. 2013. Multilingual Representation of Possessive Pronouns in Universal Networking Language. Recent Advances in Computer Engineering Series, vol. 14, 130-133. ISSN: 1790-5109 Google Scholar

5. Stoykova, V. 2013. Approaches to Semantic Networks Interpretation of Bulgarian Nominal Inflectional Morphology. Recent Advances in Computer Engineering Series, vol. 14, 124-129. ISSN: 1790-5109 Google Scholar

6. Stoykova, V. 2014. Interpretation of Possessive and Reflexive-possessive Pronouns of Bulgarian Language in DATR. IFIP Advances in Information and Communication Technology, vol. 436, 528-536, Springer-Verlag Berlin Heidelberg, ISSN 1868-4238 SCOPUS SJR 0.209 Q3

7. Stoykova, V. 2015. Semantic Model of Possessive Pronouns Machine Translation for English to Bulgarian Language. Lecture Notes in Computer Science (LNCS), vol. 9416, Springer, 349–356. ISSN: 0302-9743 WoS SCOPUS SJR: 0.427 Q2

8. Stoykova, V., M. Simkova, D. Majchrakova and K. Gajdosova. 2015. Detecting Time Expressions for Bulgarian and Slovak Language from Electronic Text Corpora. Procedia Social and Behavioral Sciences, 186, 257–260, Elsevier, ISSN:1877-0428 WoS IF 0.46

9. Stoykova, V., Simkova, M. and Majchrakova, D. 2016. Extracting Ecological Terms for Bulgarian and Slovak Language Using Parallel Text Corpora, In: 13th National Conference with International Participation, Elektronika 2016, Proceedings, 82–85. ISSN: 1313-3985

10. Stoykova, V., Stankovic, R. 2019. Using Query Expansion for Cross-Lingual Mathematical Terminology Extraction. Artificial Intelligence and Algorithms in Intelligent Systems, Advances in Intelligent Systems and Computing, vol. 764, 154-164, Springer, ISSN:2194-5357, WoS, SCOPUS (SJR 0.174) Q3

11. Stoykova, V., Majchrakova, D. 2021. Query Expansion for Slovak to Bulgarian Language Machine Translation Using Parallel Seach. WSEAS Transactions, on Systems and Control, vol. 16, 351-357, ISSN 1991-8763 SCOPUS SJR: 0.174 Q3

12. Stoykova, V. 2022. Universal Networking Language Approach to Machine Translation between Bulgarian and Slovak Language. Proceedings of ICTCS 2021, Lecture Notes in Networks and Systems, LNNS. Springer (in press) ISSN: 2367-3370 SCOPUS SJR: 0.17 Q4

Библиография

А. Библиография на латиница

1. Arnold, D., Balkan, L., Meijer, S., Lee Humphreys, R., Sadler, L.: Machine Translation: an Introductory Guide, Blackwells-NCC, London (1993)

2. Alexieva, B.: Readings in General Translation Theory, SU Sveti Kliment Ohridski, Sofia (1993)

3. Evans, R., and Gazdar, G.: DATR: A Language for Lexical Knowledge Representation. *Computational Linguistics* 22.2, 167–216. (1996)

4. Kilgarriff, A. et al.: The Sketch Engine: Ten Years On. Lexicography, 1, 7-36. (2014)

5. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. Proceedings from MT Summit 2005, 79–86. (2005)

6. Krzeszowsky, T.: Contrasting Languages. The Scope of Contrastive Linguistics. Mouton de Gruyer, Berlin – New York (1990)

7. Martins, R.: Le Petit Prince in UNL. Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2012, pp. 3201–3204. ACL Anthology (2012)

8. Newmark, P.: Approaches to Translation, Pergamon Press, Oxford (1981)

9. Nirenburg, S., Somers, H., Wilks, Y. Readings in MT. MIT Press, (2003)

10. Noncheva, V., Stancheva, Y., Stoykova, V.: The Little Prince Project - Encoding of Bulgarian Grammar. Tech. rep., UNDL Foundation www.undl.org (2011)

11. Sommers, H.: Computers and Translation. A Translator's Guide, John Benjamins, London (2003)

12. Stoykova, V.: Common Formal Framework for Multilingual Representation of Inflectional Morphology for two Related Slavonic Languages. In: R. Garabik (ed.) Proceedings of 6th International Conference of NLP Multilinguality, Tribun EU, 2011, Bratislava, 146-155.

13. Stoykova, V.: Representing Nominal Inflectional Morphology for Slavonic Languages in DATR In: Ganchev, H., Lowe, B., Norman, D., Sosokov, I. and Soskova, A. Eds., Models of

Computation in Context: 7th Conference on Computability in Europe, CiE 2011, 234-245, St. Kliment Ohridcki University Press, Sofia
14. Stoykova, V.: Bulgarian Inflectional Morphology in Universal Networking Language. In: M. Kay, C. Boitet (Eds.) Proceedings of 24th International Conference on Computational Linguistics (COLING 2012): Demonstration Papers, pp. 423—430. ACL Anthology (2012)
15. Stoykova, V: Representation of Possessive Pronouns in Universal Networking Language. In: L. Iliadis, H. Papadopoulos, C. Jayne (Eds.) Engineering Applications of Neural Networks (EANN 2013), Communications in Computer and Information Science, vol. 384, pp. 129-137, Springer (2013)
16. Stoykova, V.: Representation of Lexical Information for Related Languages with Universal Networking Language. Proceedings from the 6th Language Technology Conference, pp. 514-517. (2013)
17. Stoykova, V.: Multilingual Representation of Possessive Pronouns in Universal Networking Language. Recent Advances in Computer Engineering Series, vol. 14, 130-133. (2013)
18. Stoykova, V.: Approaches to Semantic Networks Interpretation of Bulgarian Nominal Inflectional Morphology. Recent Advances in Computer Engineering Series, vol. 14, 124-129. (2013)
19. Stoykova, V.: Interpretation of Possessive and Reflexive-possessive Pronouns of Bulgarian Language in DATR. IFIP Advances in Information and Communication Technology, vol. 436, Springer-Verlag, Berlin Heidelberg, 528-536. (2014)
20. Stoykova, V.: Semantic Model of Possessive Pronouns Machine Translation for English to Bulgarian Language. In: I. Ciuciu I. et al. (Eds.) On the Move to Meaningful Internet Systems: OTM 2015 Workshops, OTM 2015, Lecture Notes in Computer Science, vol. 9416, 349-356. Springer (2015)
21. Stoykova V., Šimkova, M., Majchrakova, D., Gajdošova, K.: Detecting Time Expressions for Bulgarian and Slovak Language from Electronic Text Corpora. Procedia Social and Behavioral Sciences, 186, 257-260, Elsevier (2015)
22. Stoykova, V., Simkova, M., Majchrakova, D. 2016. Extracting Ecological Terms for Bulgarian and Slovak Language Using Parallel Text Corpora, In: 13th National Conference with International Participation, Elektronika 2016, Proceedings, 82–85. (2016)

23. Stoykova, V., Stankovic. R.: Using Query Expansion for Cross-Lingual Mathematical Terminology Extraction. Artificial Intelligence and Algorithms in Intelligent Systems, Advances in Intelligent Systems and Computing, 764, 154-164, Springer (2019)
24. Stoykova, V., Majchrakova, D.: Query Expansion for Slovak to Bulgarian Language Machine Translation Using Parallel Search. WSEAS Transactions, on Systems and Control, vol. 16, 351-357, WSEAS Press (2021)
25. Stoykova, V.: Universal Networking Language Approach to Machine Translation between Bulgarian and Slovak Language. Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Lecture Notes in Networks and Systems (LNNS), vol. 4, Springer. (2022) (in print)
26. Uchida, H., Zhu, M., Della Senta, T.: Universal Networking Language. UNDL Foundation (2005)

Б. Библиография на кирилица

27. Кръстева, И.: Вавилонски отклонения, Преводът между лингвистика и антропология. Изток-Запад, София (2016)
28. Легурска, П.: Съпоставителни анализи и национална специфика на лексиката. Издателство „Ето“, София (2011)
29. Легурска, П.: Терциум компарационис и съпоставителни лексикални изследвания. Съпоставително езикознание, кн. 1, 46 – 65. (2011)
30. Ликоманова, И.: Славяно-славянският превод. Лингвистичен подход към художествения текст. УИ „Св. Климент Охридски“, София (2006)
31. Ликоманова, И.: Преводът между теорията и практиката. УИ „Св. Климент Охридски“, София, (2002)
32. Людсканов, А.: Превеждат човекът и машината, Избрани трудове. Народна култура, София (1980)
33. Паскалева, Е.: Компютър и превод. Наука и изкуство, София (1987)
34. Стойкова, В.: Компютърна морфология на определителния член в българския език. УИ „Св. Климент Охридски“, София, (2012)
35. Стойкова, В.: Представяне на българските местоимения с Универсалния мрежов език. УИ „Св. Климент Охридски“, София, (2019)

36. Рецкер Я. И.: Методика технического перевода. Издательство НТП, Москва (1934)
37. Рецкер Я. И.: Теория перевода и переводческая практика, Очерки лингвистической теории перевода. Международные отношения, Москва (1974)

Съдържание

1. Увод 2
2. Цел, задачи и подходи на изследването 2
3. Машинен превод – принципи и приложение 3
4. DATR език за представяне на лексикална информация и използването му за машинен превод 4
4.1. Използване на DATR за представяне на словоизменението на съществителните имена в български и руски език за машинен превод 5
4.1.1. Представяне на компютърна морфология с DATR 6
4.1.2. Използване на DATR за представяне на словоизменението на съществителните имена в български език 6
4.1.3. Използване на DATR за представяне на словоизменението на съществителните имена в руски език 7
4.2. Използване на DATR за сравняване на представянето на притежателните и възвратното притежателно местоимение в български език 9
5. Универсалният мрежов (УМЕ) език и използването му за машинен превод 10
5.1. Използване на УМЕ за представяне на словоизменението на съществителните имена в българския език 11
5.2. Използване на УМЕ за представяне на притежателните и възвратното притежателно местоимение в български, английски и руски език 13
5.3. Използване на УМЕ за МП между български и словашки език 16
5.4. Използване на УМЕ за измерване и подобряване на качеството на машинния превод 17
6. Sketch Engine и използването му за машинен превод 19
6.1. Използване на SE за машинен превод на екологични термини между български и словашки език 20

6.2. Използване на SE за машинен превод на изрази за време между български и словашки език 23
6.3. Използване на SE за машинен превод между български и словашки език 25
6.4. Използване на SE за машинен превод на математически термини между български и сръбски език 30
7. Заключение 33
Основни научни резултати и приноси в дисертацията 34
Публикации, които са представени към дисертацията 36
А. Библиография на латиница 38
Б. Библиография на кирилица 40
Съдържание 42