Софийски университет „Св. Климент Охридски”
Факултет по славянски филологии

Дипломна работа

на тема:

Машинен превод на глаголни форми от английски на български език

(правила за оптимизация)

Дипломант:
Мариета Красимирова Манева
фак. № 770488
Специалност: Компютърна лингвистика.
Интернет технологии в хуманитаристиката

Научен ръководител:
гл. ас. д-р Атанас Атанасов

Рецензент:
ас. д-р Биляна Радева

София, 2010 г.

Съдържание:

1. Увод

Съвременните тенденции в областта на машинния превод налагат неговото разбиране и реализиране предимно в областта на статистиката. В полето на т. нар. статистически машинен превод са насочени усилията и изследванията на почти всички значими научни институции. Към настоящия момент неслучайно този метод на машинен превод дава по-точни резултати. Статистическият подход работи главно в тясно избрани тематични области и това условие за неговото приложение предопределя високата степен на успешни преводи. Ако се приложи мащабно, идват и проблемите. Пример за това твърдение е машинният превод, създаден от кооперацията Google. Google Translate е широко използвана програма (всъщност това е най-разпространеният продукт на компанията), реализирана изцяло на статистически принципи. Извършените наблюдения върху работата на тази система показват, че тя не разпознава аналитичните глаголни форми с изключение на сегашно прогресивно време (Present Progressive Tense), както и че допуска грешки при съгласуване на подлог и сказуемо дори и при синтетичните времена. Виждат се неправилно преведени глаголни форми и в двете посоки – от английски на български и от български на английски език.

Целта на настоящата работа е изработване на правила и примерен модел за машинен превод на формите на глаголните сказуеми от английски на български език, като по този начин се оптимизира процеса на статистически машинен превод, предлаган от системата Google Translate. Ще се приложи комбиниран модел на машинен превод, който съчетава статистическите и лингвистичните модели за постигане на правилен резултат. Избраният подход е чрез прилагане на правила от английската и българската граматики върху резултата от работата на Google Translate.

За демонстриране на алгоритъма на действие на конструирания модел ще бъде създадена система за машинен превод (със средствата на езика XML), която изпълнява следните задачи:

- Откриване на подлога и сказуемото в изречението от английския текст.
- Определяне на глаголното време и граматичните характеристики на формата.
- Ако глаголната форма е аналитична и предполага място за наречие след спомагателния глагол, това наречие също бива преведено на български език; в

противен случай то не се изследва, защото е извън границите на сказуемото и не създава трудности при статистическия машинен превод.

- Генериране на сказуемото в изречението на българския текст като върху изходния текст от работата на Google Translate се приложат разработените правила относно образуване на подходяща глаголна форма.
- Правилно съгласуване на подлога и сказуемото по род, лице и число и добавяне на обстоятелствено пояснение, ако има такова в английския вариант.

Обект на изследване ще са глаголните форми на английските времена и техните български съответствия. Авторът приема, че английските времена са 16 на брой, като формите им се образуват чрез спомагателните глаголи *to be* и *to have* и форма на пълнозначния глагол. Сложните глаголни сказуеми, както и именните сказуеми, не са обект на изследване на тази работа, чиято цел е да покаже един комбиниран метод за машинен превод, който би работил еднакво добре и за останалите видове сказуемо. Глаголните форми се представени в деятелен залог. Всички времена са в индикатив, защото съответствието на глаголните форми при изразяване на модалност и евиденциалност не е обект на изследване на тази работа. Словоредните промени също няма да бъдат отбелязвани и коментирани. В съответствие с гореспоменатата причина паралелният корпус на български и английски език, който ще бъде създаден за демонстрация на системата, ще съдържа само прости съобщителни изречения без вметнати части или разширения, конструирани по класическата за английския език схема SVO (подлог – сказуемо - допълнение). Ще бъдат направени работни морфологични речници на двата езика, съдържащи думите от корпуса.

Работата е структурирана по следния начин: в първа глава се представят основните подходи към машинния превод; втора глава разглежда в сравнителен план глаголните форми в английски и български език; трета глава описва конструирания модел на превод.

2. Въведение в теорията на машинния превод

2.1. Първи практики в областта на машинния превод

Интересът към машинния превод датира от 40-те години на двадесети век и възниква още с появата на първите компютърни устройства. И до днес това е научна област, привличаща в голяма степен изследователските усилия на лингвисти и компютърни специалисти. В статията на Преслав Наков (Наков ????) се представя историята на машинния превод. Времето на първите опити за генериране на превод чрез машина е периодът непосредствено след края на Втората световна война, по време на която в САЩ, във връзка с необходимостта от декодиране на съдържанието на прихванатите немски съобщения, силно развитие получава теорията на информацията и криптографията. Създаденият в резултат мощен математическо-статистически апарат съвсем естествено се разглежда като средство за постигане на задоволителен автоматичен превод. През 1954 г. в САЩ се създава първата система за превод - руско-английски прототип, който подсилва очакванията за бърз напредък, включително и за други двойки езици. Подходът, който се демонстрира тогава от компанията IBM (International Business Machines), е базиран на лингвистични правила, а методът е превод дума по дума. Изследванията в България също не закъсняват и през 1964 г. се създава специална група за машинен превод между руски и български език под ръководството на проф. Александър Людсканов в Института по математика на БАН. Междувременно във връзка със Студената война и започналото съревнование между СССР и САЩ в овладяването на Космоса, обемът на превежданата научно-техническа документация от руски на английски език нараства значително, а изследванията в областта на автоматичния превод се радват на богато финансиране. През 1966 г. обаче настъпва драматичен обрат - по поръчка на правителството на САЩ Американската академия на науките изготвя доклад за състоянието на изследванията в областта на компютърната лингвистика и на машинния превод в частност, който се оказва силно скептичен. В резултат настъпва дълъг оздравителен период със силно ограничено финансиране първо в САЩ, а после и в световен мащаб. Едва през 1975 - 1985 г. започва постепенно възраждане. Положението се променя през 90-те години на миналия век, когато компютрите и науката за тях се развиват достатъчно, за да бъде проправен пътят на статистическия подход към машинния превод, който е доминиращ и до днес. Благодарение на последвалия значителен теоретичен и практически

напредък, както и поради видимото подобрение в качеството, днес машинният превод отново се радва на значителен изследователски интерес, отчасти мотивиран от икономически очаквания – само годишните разходи за преводи на Европейската комисия са над един милиард евро годишно. Неслучайно най-обемните многоезични корпуси са съставени от архивите с документация на Европейския съюз.

От гледна точка на компютърната лингвистика качественият машинен превод предполага правилен автоматичен анализ и генериране на естествен език на морфологично, синтактично, семантично и прагматично ниво с отчитане на контекста, използване на знание за света и познаване на културата на носителите на езика. Изисква още възпроизвеждане на текст, разрешаване на различни видове многозначност (лексикална и граматическа), разпознаване на собствени имена, транслитерация, анализ на местоимения и изобщо дълбоко разбиране на смисъла на превеждания текст. Заради това методите за статистически машинен превод не са в състояние да обхванат всички езикови явления и съответно резултатите не са съобразени с лингвистичните особености на конкретния изходен език и често са смислово неприемливи или граматически неправилни.

2.2. Основни подходи към машинния превод

През 1949 г. Уорън Уейвър от фондация Рокфелер пише (Уейвър 1955): „Пред мен има текст, написан на руски, но аз ще си мисля, че всъщност е на английски, но е кодиран с някакви странни символи. Всичко, което трябва да направя, за да разчета кодираната информация, е да декодирам текста.” Това разсъждение се оказва изключително полезно и лежи в основата на всички съвременни методи за статистически машинен превод. Процесът на превеждане се разглежда като процес на кодиране и декодиране на информация. Факт е, че днес математическо-статистическите подходи за осъществяване на машинен превод имат по-висока степен на успеваемост в сравнение с чисто лингвистичните. Но процентът на правилно преведени изречения не надхвърля 95% - цифрите са различни за конкретните двойки езици, върху които се работи. Към настоящия момент изследванията в областта на машинния превод в световен план са съсредоточени в усъвършенстване на комбинирани подходи – такива, които да съчетават статистическия и лингвистичния. По-долу са представени характеристиките на основните методи на машинен превод. И статистическия подход, и

подходът, базиран на лингвистични правила, могат да бъдат приложени върху текста на ниво дума, фраза или изречение.

2.2.1. Статистически машинен превод (Statistical Machine Translation)

Статистическият подход на машинен превод е базиран изцяло на математически принципи. При него не се използват лексикални или морфологични речници, нито граматики. Преводът се разглежда като кодиране на информация в наличния език и нейното декодиране в изходния. Единственото необходимо условие е наличието на достатъчно обемен съотнесен[1] (най-малко) двуезичен корпус. Съотнасянето, което също се извършва автоматично, може да е по думи, фрази или изречения. В основата на превода лежат закони на статистиката – идеята е да се изчисли вероятността на правилност на разглежданата фраза или на цялото изречение. За целта текстът се разделя на фрагменти от две, три или повече думи (т. н. n-грами) и спрямо първата част се определя вероятността каква да бъде втората, т.е. вероятността някаква последователност от символи от базата данни да отговаря на липсващата част. Системата за разпознаване се изгражда с помощта на невронна мрежа, която има свойството да се тренира и самообучава и постига висока степен на положителен резултат. Широко използван математически метод за този подход е HMM (Hidden Markov Models – Скритите модели на Марков), който успява с голяма точност да определи следващата дума или фраза. На базата на информацията от корпуса автоматично се създава езиков модел, който на статистически принципи изчислява дали генерираната фраза е граматически правилна. На този принцип работи и системата Google Translate. От направените наблюдения може да се направи изводът, че нейният езиков модел, генериращ фрази, работи върху биграми – последователности от две думи. Резултатът на статистическия машинен превод е в пряка зависимост от корпуса – влияние оказват тематичната област на подбраните текстове, тяхното съотнасяне, обемът им. След 2002 година доминиращият начин на оценяване е автоматичен и използва BLEU (Bi-Lingual Evaluation Understudy) – специална оценка, предложена от екип на IBM, която измерва доколко машинният превод е близък на ниво 1, 2, 3 и 4 последователни думи (n-грами) до един или повече еталонни човешки превода.

[1] Английският термин е *alignment* (букв. подреждане) и означава съотнасяне на текстовете от паралелния многоезиков корпус, т.е. всяка фраза от единия език се свързва с конкретния израз, който правилно я превежда. Вярното съотнасяне на корпуса има определяща роля за успешен превод.

Проблемите при статистическия машинен превод възникват при различни граматични особености и словоред на изследваните езици, културни и стилистични разлики при изразяване. При превода по думи проблем е невъзможността за отразяване на контекста - например *interest rate* трябва да се преведе като *лихвен процент*, докато *interest in* – като *интерес към*. Неотчитането на контекста силно затруднява и съгласуването по род, лице, число, падеж, членуване и може да доведе до генериране на фрази като *главен прокурорите*, които невинаги могат да се коригират успешно от езиковия модел.

Преводът с цели фрази дава решение на някои от тези проблеми. Това е генеративен модел, при който първо изречението се разделя на фрази, след това всяка фраза се превежда отделно и после някои от фразите се разместват. С всяка от стъпките е асоциирана съответна вероятност, която се учи от паралелен корпус с преведени и на двата езика изречения. Подходящият превод в конкретно изречение се избира като се взема предвид не само вероятността за превод на двойката фрази, но и вероятността за превод на цялото изречение според езиковия модел.

Интересен модел на статистически машинен превод е преводът, базиран на примери (Example-Based Machine Translation). Работата на такава система се основава на допускането, че ако вече преведено изречение (или фраза – зависи от нивото на съотнасяне на корпуса) се срещне повторно в текста, вероятно същият негов превод би бил отново правилен. Този модел отново не използва речник или друга лингвистична информация, а само фрагменти от вече готови преводи. Алгоритъм за машинен превод, базиран на примери, се прилага успешно в т. нар. преводаческа памет в системите за превод, подпомогнат от компютър[2]. Този подход е широко използван в тесни тематични области – например при превод на климатични прогнози или на техническа литература.

Предимствата на статистическия подход са следните: може да се приложи към всяка двойка езици, за които са събрани достатъчно паралелни изречения; позволява лесно добавяне на нов език; разрешава лексикалната многозначност; открива и разпознава идиомите (ако са включени в корпуса); изисква минимално човешко усилие; не зависи от конкретния език, от който или на който се превежда. Недостатък на статистическия подход е пряката зависимост на качеството на превода от данните в корпуса и начина

[2] Системите за машинен превод, които имат практическо приложение, могат условно да се разделят на две – за превод, подпомогнат от компютър, при който преводач коригира резултата, и за напълно автоматичен висококачествен превод, какъвто още не е постигнат.

на съотнасяне на изреченията, както и неотчитане на лингвистичните характеристики на отделните езици. Като резултат типичните грешки са морфологични и синтактични.

2.2.2. Машинен превод, базиран на правила (Rule-Based Machine Translation)

Този подход на машинен превод се основава на лингвистични правила – относно морфологичните характеристики на думите, съчетаемост във фразата, словоред, семантична съчетаемост. Задължително се използва богат речник с морфологична, синтактична и семантична информация, прилага се и граматика от правила. Тези лингвистични правила обикновено са доста изчерпателни, защото имат за цел описването на всички характеристики на конкретния език, и достигат голям обем. Записват се в различен формат в отделните езици за програмиране, например в PROLOG – логически език интерпретатор. При машинния превод, базиран на правила, началният текст се анализира до ниво на междинно символно представяне, от което се генерира текст на целевия език. В зависимост от това ниво, което се интерпретира, се използват основно три модела – на директния превод, на трансферния превод и на превод с използване на интерлингва.

При директния модел преводът е на ниво лексема. Такъв е първият демонстриран машинен превод. При този модел се използва компютърен двуезичен речник и програма, превеждаща всяка отделна дума с морфологичните й характеристики. После може да се приложат правила относно словореда. Моделът се реализира чрез повърхностен морфологичен анализ. Предимство е сравнително лесното му прилагане и възможността за пряко допълване на нова информация в базата данни. Недостатък е трудното разрешаване на многозначности и проблеми при словореда, което води до ниско качество на превода.

При трансферния модел преводът е на ниво изречение или фраза. При него първо се анализира текста като се групират думите в последователности (т. нар. парсиране – английският термин е *parsing*), определя се тяхната структура (например автоматично се изграждат синтактични дървета) и се прилагат правила за трансформирането й в подходяща структура на целевия език. После чрез други правила се генерира фразата или изречението на изходния език и накрая се превеждат думите. За разлика от директния модел тук работят правила за лексикален и за синтактичен трансфер, извършва се и повърхностен синтактичен анализ. Предимство е разрешаването на

словоредните проблеми, недостатък – необходимостта от конструиране на трансферни правила за всяка двойка езици.

При моделът с използване на интерлингва преводът отново става на равнището на изречението или фразата, но тук междинното ниво представя и значението на думите. Използва се логическо моделиране като цялата граматична и семантична информация на текста се трансформира в независими от конкретния език елементи. Това изцяло абстрактно ниво се нарича интерлингва. При този модел не се прилагат правила за лексикален или за синтактичен трансфер, работи се с понятия и се използват онтологии (йерархии от понятия). По тези причини преводът с помощта на интерлингва намира приложение в конкретни тематични области като хотелски резервации, ресторантски менюта, пътеводители. Предимство на този модел е възможността за сравнително лесното превеждане между повече езици и това, че за добавяне на нов език е необходимо само неговото логическо представяне чрез интерлингва и написването на допълнителни правила единствено за него.

2.2.3. Комбинирани методи на машинен превод

Съгласуването по род, число и падеж, както и някои особености на словореда създават сериозни проблеми на описаните статистически модели, непознаващи понятия като съществително, глагол, подлог и др. Всъщност те не знаят дори какво е дума като лингвистично понятие, за тях няма разлика между лексикална дума и препинателен знак. Затова съвременните усилия в областта на статистическия машинен превод са насочени към прякото моделиране на граматично знание. Това става чрез прилагане на лингвистични правила върху отделни нива на текста. Задачата не е лесна, отчасти защото автоматичният синтактичен анализ е труден сам по себе си - най-добрите синтактични анализатори за английски език работят с 91% точност за вестникарски текст, какъвто са обучени да анализират, но качеството пада значително при други видове текст, например медицински.

За по-добри резултати се използват методи, съчетаващи статистически подходи и подходи, базирани на лингвистични правила. Пример за такъв вид метод е превод с използване на синтактична информация – това е статистически машинен превод, който използва като основа на изчислението на вероятността синтактичен анализ на ниво изречение. Най-широко приложение намират т.нар. синтактични дependентни дървета (dependency trees), изградени спрямо граматиката на зависимостите. В Google Research

Center към настоящия момент се работи върху подобрение на системата за превод чрез използване на такива синтактични структури за представяне на информацията.

Перспективна алтернатива на гореизложените синтактични подходи са факторните модели, които позволяват просто моделиране на морфологични и лексикални характеристики на ниво отделна дума. Например процесът на превод на английската дума *dogs* като *кучета* включва анализ, транслиране и генериране. Първо формата *dogs* се анализира като съществително *dog* в множествено число и съответно окончание *-s*. След това лемата, синтактичната и морфологичната информация се транслират поотделно в английските си еквиваленти. Накрая те заедно генерират правилната българска форма *кучета*. Основно предимство на факторните модели е, че позволяват отделен лингвистичен модел на ниво морфология (за правилно съгласуване по род и число), на ниво част на речта (за граматично правилна последователност), на ниво лема (за семантично правилна последователност) и на ниво дума (за допълнително изглаждане). Тези лингвистични модели могат да се използват като правилата се приложат върху корпуса преди статистическия подход или върху резултата от превода за постигане на граматична правилност. Това обаче е свързано със значително общо забавяне на системата за машинен превод, което прави факторните модели приложими само при малки обеми тренировъчни данни.

Друг комбиниран метод на машинен превод е разработен от изследователската група на проф. Вернер Винивартер (Винивартер 2007) във Факултета по научни изчисления на Виенския университет (Department of Scientific Computing, University of Vienna). Учените представят система за превод от японски на английски език, базирана на лингвистични правила (Винивартер 2007). Подобрението на класическия машинен превод е в автоматично генериране на трансферни правила от паралелния корпус вместо ръчното им предварително конструиране отделно от корпуса. Има възможност за преглед и корекция на получените правила от базата данни. Друга група изследователи от Microsoft Research Center в САЩ подобряват статистическия метод за машинен превод, базиран на примери. Те изработват инструмент, който автоматично трансформира структурата на вече преведените примери в семантично представяне, наречено “логическа форма” (Брокет 2002). Той се прилага след парсирането и така се разрешават неточностите при превода, причинени от различните синтактични особености на разглежданите езици. За генериране на синтактично дърво и на изречение на изходния език също се прилага набор от лингвистични правила. Така

изработеният комбиниран метод разрешава проблемите на статистическия подход и повишава степента на правилност на машинния превод.

Целта на настоящата работа е изграждане на модел за превод на глаголните форми от английски на български език чрез лингвистични правила. Направеният модел ще е независим от текста, към който се прилага, тъй като правилата са универсални и не се влияят от конкретиката на базата данни. Написаните правила биха могли да се включат като компонент на математическите формули, генериращи изходния текст, и по този начин да се отстранят голям брой неточности, които статистическият подход не отбелязва като грешки.

2.3. Избор на XML език за конструиране на лингвистични правила

Избраната среда за изграждане и демонстриране на правилата за машинен превод е езикът XML. XML (Extensible Markup Language) е метаезик или буквално преведено "маркиращ език" за структуриране на документи, разработен от Уеб консорциума (World Wide Web Consortium) като спецификация на SGML (Standard Generalized Markup Language). XML е създаден да подрежда, пренася и запазва информация. Всеки документ, написан с този стандарт, има еднаква структура – състои се от елементи и техни атрибути, подредени спрямо изискванията на консорциума. Характерна особеност на метаезика е способността му да се самоописва чрез разширението DTD (Document Type Definition), което позволява свобода при избора на имена на елементите и атрибутите, прави използването на XML лесно и универсално въпреки налагането на стандарта. В допълнение към езика XML са разработени модулите DTD (дефиниция с изисквания към съдържанието на XML документа), XSL (Extensible Stylesheet Language – език, даващ стилова информация за всеки таг, елемент и атрибут), XSLT (Extensible Stylesheet Language Transformations – език за трансформиране на информацията в друг вид документ), XPath (синтаксис за навигация в структурата на XML документа и за търсене на определени елементи и атрибути и тяхното съдържание) и XQuery (Extensible Query Language - език за изпращане на заявки към данните в документа).

XML е универсална среда за съхраняване и транспортиране на данни, която позволява различни операции с тях. XSLT трансформира отделни части от документа по зададени параметри независимо от контекста. Това е добра предпоставка за избора на XML, защото предлага възможността изследването да бъде съсредоточено само

върху темпоралните характеристики на сказуемото вместо да се създава система за превод на цялото изречение, каквато вече е разработена от Google. Правилата за превод ще бъдат написани чрез XSLT, резултатите – визуализирани. Езикът XML е по-гъвкав и предлага по-голяма свобода при обработка на данните в сравнение с езиците за обектно програмиране и едновременно с това осигурява среда за демонстрация на положителния резултат от прилагането на правилата за машинен превод на глаголните форми, каквато е и една от задачите на работата.

3. Сравнителен анализ на глаголните форми в английски и български език

3.1. Особености на английската граматика, които оказват влияние при осъществяване на машинен превод на глаголните форми

При разглеждането на структурата на глаголното сказуемо в английски и български език би следвало да се вземат под внимание някои основни характеристики на английските морфология и синтаксис.

3.1.1. Фиксиран словоред на изречението

Английското съобщително изречение обикновено има фиксиран словоред, който може да се представи през формулата SVO(C)A. В нея S (Subject) отбелязва мястото на подлога, V (Verb) – на сказуемото, O (Object) – на допълнението (пряко или непряко), C (Complement) – на сказуемното определение и A (Adverbial) – на обстоятелственото пояснение. Пример за изреченска конструкция, при която всички позиции са заети, е *John painted the door green yesterday*. Невъзможността за промяна на позициите на подлога и сказуемото играе определяща роля при откриването им и съответно ситуирането на тези позиции в българския превод. Единственият случай, при който английското изречение не започва с подлог, е когато началната позиция е запълнена от обстоятелство, най-често за време или начин. Формулата в този случай е ASVO(C), като винаги след обстоятелството се намира подлогът и последователността SVO не се нарушава[3]. Всъщност обстоятелственото пояснение за време има относителна свобода при позиционирането си в изречението, но само в определени ситуации и на определени места като това важи само за отделен клас наречия. Интерес представлява наблюдението, че различното място на обстоятелственото пояснение в изречението променя цялостния му смисъл. Случаите на начална позиция на обстоятелственото пояснение нямат да бъдат разглеждани в тази работа, тъй като то остава извън рамките на сказуемото, същото се отнася и за случаите на позициониране между подлога и сказуемото. Интерес представляват аналитичните глаголни форми, които позволяват вмъкване на обстоятелство за време или начин или изразяващо модалност между

[3] Изключение от това правило има само при отрицание и полуотрицание - *Never in his life was he so grossly insulted*; *It is neither pleasant, nor is it healthy to bathe in this cold water*.

формата на спомагателния глагол *to be* или *to have* и формата на пълнозначния глагол. Така определеното място има статут на задължително за обстоятелството дори когато сказуемото се състои от два спомагателни глагола - няма значение дали е след първия или втория спомагателен глагол: *John is slowly parking the car; I have never played golf; The building had been severely damaged*. Този тип словоред се представя чрез формулата SVauxAVO(C).

Тази особеност на английската морфология е от първостепенно значение за постигане на успешен машинен превод. Ако една дума е пред глаголна форма и не е наречие, със сигурност тя изпълнява функцията на подлог или е част от неговата структура. А определянето на граматическите характеристики на тази дума води до правилно съгласуване с глаголната форма.

3.1.2. Задължителна експлицитност на подлога

В английското изречение подложната позиция винаги се запълва експлицитно – от съществително име, местоимение или фраза на съществителното име. Тази особеност на синтаксиса помага в голяма степен при откриването на думата или фразата, изпълняваща функция на подлог, и вярното й превеждане на друг език.

3.1.3. Особености на категорията вид

Въпреки че глаголната форма в английски език не носи експлицитно видова информация с изключение на прогресивните форми, той не е „безвидов". Приема се, че категорията вид не е граматическа за този език за разлика например от славянските езици, в които формате на глагола показва и вида му (Ницолова 2008). В английските граматики видът се разглежда като синтактична категория, независимо дали рефлектира с перфектността или перфектността и прогресивността (Грийнбаум 1995). В тази работа се следва теорията на Кабакчиев (Кабакчиев 1992), в чийто труд видът се определя като хареактеристика, присъща по-скоро на изказването или на изречението, отколкото само на глагола. Пак там авторът обобщава факторите, които оказват влияние върху видовата характеристика на изречението. Отношение към проблема за вида имат формалната структура и граматическата оформеност на подлога и допълнението, лексикалното значение на подлога, допълнението и сказуемото, противопоставянето на лични и нелични местоимения, различаването на пределни и непределни глаголи, членуването. Обстоятелственото пояснение е май-мощния

показател за вида. Например в израза *I often sang opera before* видовото значение е несвършено, докато в *I sang opera yesterday* – свършено. Може да се направи извода, че английският глагол има маркирани откъм вид форми само за изразяване на несвършени действия и това са прогресивните форми, образувани със сегашното причастие на глагола. А видовото значение в изречение с немаркирани откъм вид глаголни форми зависи от редица фактори. По тези причини в тази работа засега видът няма да се разглежда, защото правилното трансформиране на видовото значение на английското изречение във форма на български глагол изисква по-мащабно изследване и е обект на бъдеща работа в тази област.

3.1.4. Граматическа омография на глаголните форми

Понеже подлогът винаги е експлицитен, в английския език глаголната форма не носи информация за лицето, числото или рода на субекта (с изключение на формите за трето лице, единствено число, сегашно време, които имат наставка -*s*). Допълнителен проблем е фактът, че инфинитивно глаголите не изразяват категорията вид. Така че глаголните форми за различните времена често съвпадат и една форма на английския глагол съответства на няколко на българския.

Посочените характеристики на английската морфология отличават начина на изразяване на подлог и сказуемо в двата разглеждани езика. Заради тези разлики най-честите грешки при машинния превод са именно в областта на съгласуването. Отстраняването на този тип грешки е една от конкретните задачи на настоящата работа.

3.2. Общ преглед на българската темпорална система

Ще бъде направен сравнителен преглед на английската и българската темпорални системи с цел да се съпостави всяко от глаголните времена от английската граматика със съответното в българската. Характеристиките на времената ще се разгледат в контекст, който осигурява необходимата информация за формално сравнение и верен превод. От значение ще са начините на образуване на глаголната форма – дали е синтетична или аналитична, и темпоралното значение на времето – отношение към момента на говорене и към някакъв друг минал или бъдещ момент. Това знание ще помогне за конструиране на правила за откриване, образуване и съгласуване на

глаголните форми в двата езика като по този начин значително се повиши степента на точност при превода.

При разглеждането на темпоралните системи на двата езика се следва моделът на проф. Руселина Ницолова (Ницолова 2008). Основният момент в него е начинът на подреждане на събитията – хронологично по темпоралната ос. Локализацията на събитията, означени с глаголните форми на различните времена, се извършва с помощта на три момента (интервала):

- момент на говорене – ще го означаваме с T (Talking moment), спрямо него събитията се определят като настоящи, минали или бъдещи
- интервал или момент на действието – означен с E (Event), показва времето на случване или протичане на действието
- интервал на референтност – означен с R (Reference moment), това е моментът или интервалът, за който се говори или мисли; R се посочва допълнително в изказването и служи за по-точно ситуиране на E по темпоралната ос, както и за отбелязване на резултативността при част от времената

Спрямо тези три интервала бихме могли да характеризираме действията по темпоралната ос, независимо дали те са изразени от български или английски глаголни форми. Избраният модел на представяне на значението на времената улеснява формалното им сравнение като основа за правилен машинен превод. Има две общи характеристики на българската и английската темпорални системи, които оказват влияние на процеса на превода – продължителност (наричана още континуативност) и резултативност на действието. Признакът продължителност е диференциален за българските времена само при употребата на минало несвършено време, докато в английския език той удвоява броя на времената. Наличието на някакъв резултат отделя т. нар. резултативни времена. Те функционират и в двата разглеждани езика по сходен начин – резултатът от действието се локализира по темпоралната ос спрямо допълнителен момент, дори при формалното изразяване има прилика – образуват се от спомагателен глагол в съответното хронологично време и минало причастие. Може да се отбележи наблюдението, че единствено при превода на тези форми винаги има точно съответствие в простото изречение, независимо от речниковото значение на глагола и контекста. Затова при превода на тези английски форми със съответните им български

няма изключения или повече от един приемлив вариант, което води и до значително по-малък брой забелязани грешки при статистическия машинен превод, осъществяван от системата Google Translate. Ето четирите английски времена, които имат точни български съответствия: *I have played → аз съм играл, I had played → аз бях играл, I will have played → аз ще съм играл, I would have played → щях да съм играл*.

3.3. Общ преглед на английската темпорална система

Английската темпорална система притежава характеристики, които я сближават до българската, както и особености на идейно конструиране и на формално граматично представяне, отличаващи се от българските.

Свършеността и несвършеността на действието не се показват експлицитно чрез глаголната форма (с изключение на прогресивните форми), а се изразяват систематично по различни начини – лексикални, морфологични и синтактични в зависимост от контекста (Кабакчиев 1998). Правилният машинен превод на вида на глагола на български език изисква друг вид разглеждане на английската граматика, което да обхване и семантичната страна на изречението. Това е следващата стъпка към усъвършенстването на машинния превод и поле за бъдещи изследвания. На този етап ще се спазят ограниченията в употребите на времената и ще се отбележат най-честите случаи на начина на протичане на действието, за да се избере подходящия български глагол от видовата двойка. Например фактът, че английските прогресивни времена изразяват само несвършени действия, води до превода им с български глаголи от несвършен вид независимо от конкретното време по хронологичната ос. Всъщност наименованията на английските времена показват и начина на протичане на действието. Прогресивността се асоциира с категорията вид, а перфектността – с допълнителна референтност и/или резултативност. По тези причини повече от едно българско време превежда едно английско и няколко английски времена получават еднакъв български превод. Настоящата глава е опит за съпоставка на закономерностите на явленията, които да послужат за основа на по-точен машинен превод.

При конструиране на модела на английската темпорална система, поради гореизложените явления, се вземат предвид някои допълнителни фактори. Освен момента на говорене T, интервала на референтност R и интервала на действието E някои от времената се определят с помощта на допълнителен четвърти ориентационен момент. Глаголните форми на английските времена съдържат информация за времето

на извършване на събитието и за начина на протичане на действието (което може да се подразбира и от контекста). Сегашността, миналостта, бъдещността и бъдещата миналост са характеристики на времето, докато неопределеността, прогресивността, перфектността и перфектната прогресивност – на действието.

Друга основна характеристика на английската граматика е правилото за съгласуване на времената. То действа главно при структуриране на сложни изречения и последователни изказвания и е по-скоро особеност на изграждане на текст. Принципът, който задължително се следва, е базиран на локализацията на действията върху темпоралната ос. Според него времето на сказуемото се определя спрямо времето на глагола в главното или в предишното изречение. Ако първият глагол е в минало време например, другият се ориентира спрямо него, а не спрямо момента на говорене. По тази причина се появяват разлики при превод на изречения като *They proved he **has stood** behind the murder – Те доказаха, че той **стои** зад убийството* вместо пряк превод на глаголната форма, при който тя би имала вида *е стоял*. Други примери за това явление са:

*They said on Thursday their countries **were planning** to strengthen…. – Казаха в четвъртък, че техните страни **планират** да засилят... (*вместо *са планирали*)

*He said it **would be better** now – Той каза, че сега **ще е по-добре*** (вместо *щеше да е по-добре*)

*How your job **has been going** – Как **върви** твоята работа* (вместо *е вървяла*)

*He told **he had taken** part in the organization – Той каза, че **е участвал** в организацията* (вместо *бил участвал*)

3.4. Съпоставка на българските и английските глаголни времена и техните формални показатели

В тази част ще бъдат схематично представени формите на глаголните времена в двата езика. Граматичните принципи, които имат значение за настоящето изследване, са относно начина на образуване на формите и основната употреба на конкретното време, която се определя от позициите на Т, Р и Е интервалите върху темпоралната ос. Всички сравнения между българските и английските глаголни времена ще са в тези формални граници, които имат за цел правилен машинен превод. Употребата на

времената при възможни транспозиции, както и модалните употреби, не се разглеждат, тей като единствено формата е обект на изследване, а тя се запазва в такива случаи. Няма да бъде обект на настоящата работа и употребата на времената в подчинени изречения, при която, както вече се показа, се проявяват разлики.

Характеристиките на българските глаголни времена са представени главно по Ницолова (Ницолова 2008), английските – по Кабакчиев (Кабакчиев 1998).

3.4.1. Глаголните времена в български език

При формалното описание на българските глаголни време не се изследва видът на глагола и неговото влияние върху възприемането на цялостното значение на изказването. Понеже фокусът е върху превода на английските глаголни форми на български език, се разглеждат само основните употреби на българските времена с цел максимално изчерпателна съпоставка с английските форми.

- **Сегашно време – *играя***

Образуване – от сегашната основа и окончанието за съответните лице и число

Примерни форми – *аз играя, ти играеш, той / тя / то играе, ние играем, вие играете, те играят*

Употреба – изразява в основната си употреба действия или състояние, които се извършват в момента на говорене, и периодът на референтност включва този момент – T принадлежи на R и E включва R.

В момента играя тенис. Обикновено чета само сутрин.

- **Минало свършено време – *играх***

Образуване – от аористната основа и окончанието за съответните лице и число

Примерни форми – *аз играх, ти игра, той / тя / то игра, ние играхме, вие играхте, те играха*

Употреба - изразява действия или състояния, минали спрямо момента на говорене, може да означава еднократни и последователни действия; тук R е преди (е минал спрямо) T и E се включва в или съвпада с R.

Снощи вечерях и гледах телевизия. Вчера той донесе доклада.

- **Минало несвършено време – *играех***

Образуване – от имперфектната основа и окончанието за съответните лице и число

Примерни форми – *аз играех, ти играеше, той / тя / то играеше, ние играехме, вие играехте, те играеха*

Употреба – изразява действия или състояния, минали спрямо момента на говорене, може да са еднократни и повторителни, свършени и несвършени, но винаги има характеристиката продължителност при основната си употреба; често се използва за фон на други свършени действия; тук R е преди T и R принадлежи на E.

Четири месеца той ставаше всяка сутрин точно в пет. Когато бях дете, играех всеки ден навън.

- **Минало неопределено време – *съм играл***

Образуване – от формите на спомагателния глагол *съм* в сегашно време и минало свършено деятелно причастие на пълнозначния глагол.

Примерни форми – *аз съм играл, ти си играл / играла, той е играл, тя е играла, то е играло, ние сме играли, вие сте играли, те са играли*

Употреба – изразява действия или състояния, минали спрямо момента на говорене и с неконкретизирано време на случване, като резултатът е наличен в този момент; може да означава еднократни и повторителни действия; тук T принадлежи на R, E е преди R и резултатът от E е наличен в R.

Анна си е дошла и го чака. Никога не съм ходила в Италия.

- **Минало предварително време – *бях играл***

Образуване – от формите на спомагателния глагол *съм* в минало несвършено време и минало свършено деятелно причастие на пълнозначния глагол

Примерни форми – *аз бях играл, ти беше играл / играла, той беше играл, тя беше играла, то беше играло, ние бяхме играли, вие бяхте играли, те бяха играли*

Употреба – изразява действия или състояния, извършвани или извършени преди минал момент, и резултатът е наличен в този минал момент; R е преди T, E е преди R и резултатът от E е наличен в R.

Когато влязох, детето беше заспало. Те бяха говорили много и съжаляваха.

- **Бъдеще време – *ще играя***

Образуване – от частицата *ще* и формите на сегашно време на пълнозначния глагол, отрицателните форми се образуват с помощта на глагола *няма*

Примерни форми – *аз ще играя, ти ще играеш, той / тя/ то ще играе, ние ще играем, вие ще играете, те ще играят, аз няма да играя, ти няма да играеш, той / тя / то няма да играе, ние няма да играем, вие няма да играете, те няма да играят*

Употреба – изразява бъдещи спрямо момента на говорене действия или състояния; тук R включва E и R е след T.

Утре по това време ще е в града. Утре ще пиша цял ден. Ще препиша текста утре.

- **Бъдеще време в миналото – *щях да играя***

Образуване – от формите на глагола *ща* в минало несвършено време, частицата *да* и сегашно време на спрегаемия глагол; отрицателните форми се образуват от минало несвършено време на глагола *няма*, частицата *да* и сегашно време на пълнозначния глагол.

Примерни форми – *аз щях да играя, ти щеше да играеш, той / тя / то щеше да играе, ние щяхме да играем, вие щяхте да играете, те щяха да играят, аз нямаше да играя, ти нямаше да играеш, той / тя / то нямаше да играе, ние нямаше да играем, вие нямаше да играете, те нямаше да играят*

Употреба – изразява действия или състояния, бъдещи спрямо минал момент; тук R е преди T и E е след R.

Ако не ме беше прекъснал, щях да напиша писмото по-бързо.

- **Бъдеще предварително време – *ще съм играл***

Образуване – от формите на спомагателния глагол *съм* в бъдеще време и минало свършено деятелно причастие на пълнозначния глагол, отрицателните форми се образуват чрез глагола *няма*.

Примерни форми – *аз ще съм играл / играла, ти ще си играл / играла, той ще е играл, тя ще е играла, то ще е играло, ние ще сме играли, вие ще сте играли, те ще са играли, аз няма да съм играл / играла, ти няма да си играл / играла, той няма да е играл, тя няма да е играла, то няма да е играло, ние няма да сме играли, вие няма да сте играли, те няма да са играли*

Употреба – изразява действия или състояния, бъдещи спрямо момента на говорене, които ще са се извършили преди даден момент или период, който от своя страна също е бъдещ спрямо Т и в него ще е наличен резултатът от действието; тук R е след Т, Е е преди R и след Т и R включва резултата на Е.

До шест часа ще съм прочел цялата книга. До утре той ще е говорил с всички свои приятели.

- **Бъдеще предварително време в миналото – *щях да съм играл***

Образуване – от формите на спомагателния глагол *съм* в бъдеще време в миналото и минало свършено деятелно причастие на пълнозначния глагол, отрицателните форми са образуват от глагола *няма* в минало несвършено време

Примерни форми – *аз щях да съм играла, ти щеше да си играл / играла, той щеше да е играл, тя щеше да е играла, то щеше да е играло, ние щяхме да сме играли, вие щяхте да сте играли, те щяха да са играли, аз нямаше да съм играл / играла, ти нямаше да си играл / играла, той нямаше да е играл, тя нямаше да е играла, то нямаше да е играло, ние нямаше да сме играли, вие нямаше да сте играли, те нямаше да са играли*; възможно е глагола *съм* да се замени с глагола *бъда*, но тези форми засега няма да се разглеждат като алтернативен резултат от прилагането на правилата за машинен превод.

Употреба – изразява действия или състояния, бъдещи спрямо момент или период R, който от своя страна е минал спрямо момента на говорене, като резултатът от действието е наличен в R1; тук R и R1 са минали спрямо Т, Е е между R и R1.

Те щяха да са нарисували картините само след два дни.

3.4.2. Глаголните времена в английския език

- **Сегашно неопределено (просто) време - *I play tennis every day***

Образуване - от инфинитива на глагола като в трето лице единствено число се добавя – *s*. Във въпросителните и отрицателните форми се използва спомагателният глагол *do, does* за 3 л, ед. ч. и съкратените форми *don't* и *doesn't.* Изключение са модалните глаголи, но те са обект на бъдещи изследвания.

Примерни форми – *I play, you play, he / she / it plays, we play, you play, they play*

Употреба – изразява предимно несвършени действия или състояния независимо от значението на глагола или останалите думи и изрази; може да означи краткотрайни или дълготрайни действия, еднократни действия, обичайни и общовалидни действия, повтарящи се действия, повтарящи се ограничен или неограничен брой пъти. Превежда се на български език с глаголи в сегашно време от свършен и несвършен вид.

- **Сегашно прогресивно време – *I am playing tennis at the moment***

Образуване – от форма на спомагателния глагол *to be* в сегашно време за съответните лице и число и сегашно деятелно причастие на пълнозначния глагол.

Примерни форми – *I am playing, you are playing, he / she / it is playing, we /you / they are playing*

Употреба – изразява само несвършени действия или състояния, конкретно извършващи се в настоящ момент или период; обикновено означава фонови действия или състояния за други действия или състояния, които се предават в глаголи в сегашно неопределено време. Превежда се на български език със сегашно време на глаголи само от несвършен вид.

- **Сегашно перфектно време – *I have played tennis since 2001 year***

Образуване – от форма на спомагателния глагол *to have* в сегашно просто време и минало причастие на пълнозначния глагол.

Примерни форми – *I / you / we / you / they have played; he / she / it has played*

Употреба – изразява действия или състояния, вече извършвани или извършени, които имат значение за настоящия момент и внасят допълнителен елемент в цялостното

тълкуване на действието в изречението и на резултата от него. Съответства на българското минало неопределено време по това тълкуване. При представянето на това време на темпоралната ос се въвежда четвърти момент или период, в който се реализира резултатът. Тази особеност важи също и за българския перфект. Интерес представлява фактът, че не винаги е удачно формите на английското перфектно време да се преведат на български език с формите на минало неопределено време, въпреки че означават еднакви по същността си действия и състояния. Причината за това може да се открие в правилото за съгласуване на английските времена, в разликата в изразяването на вида при двата езика и в лингвистичния факт, че на български език са приемливи употреби на различни времена за едно и също събитие (което води до промяна на неговото възприемане). Разглежданото време се превежда на български с минало неопределено време от свършени и несвършени глаголи, със сегашно време и с минало свършено време в зависимост от контекста. Все пак формите на българския перфект отговарят в най-висока степен по значение и употреба на разглежданите английски и това ще е основното при конструирането на подходящи правила. Ето няколко примера за възможни преводи:

*I **have worked** for this institute for three years. – **Работя** в този институт от три години / **Работил съм** в този институт за три години.*

*I **have played** tennis since 2001 year. – Аз **играя** тенис от 2001 година.*

*John **has** already **washed** the dishes. – Джон вече **изми (е измил)** чиниите.*

*I **have known** him for two weeks. – **Познавам** го от две седмици.*

*It **has** recently **opened** all positions for citizens – Тя **отвори (е отворила)** врати отскоро за граждани.*

*I **haven't passed** my exam yet – Не **съм** си **взел** изпита още.*

*The doctor **has** never **treated** the similar case – Докторът никога **не е лекувал** подобен случай.*

- **Сегашно перфектно-прогресивно време – *I have been playing tennis***

Образуване – от формите на спомагателния глагол *to be* в сегашно перфектно време и сегашно причастие на пълнозначния глагол

Примерни форми – *I / you / we / you / they have been playing, he / she / it has been playing*

Употреба – изразява действия или състояния, които са започнали в неопределен момент или период в миналото и продължават в момента на говорене като резултатът от тях е наличен в настоящия момент или период. Означава само несвършени действия и се превежда на български език със сегашно време на глаголи от несвършен вид.

*I **have been working** for this institute for three years – **Работя** в този институт от три години.*

*John **has been washing** the dishes for ten minutes – Джон **мие** чиниите десет минути.*

- **Минало неопределено (просто) време – *I played tennis***

Образуване – от инфинитива на пълнозначния глагол, към който се добавя наставката – *ed* има група глаголи, които образуват тези форми по друг индивидуален начин – група на неправилните глаголи в английския език.

Примерни форми – *I / you / he / she / it / we / you / they played*

Употреба – изразява действия или състояния, извършвани или извършени в някакъв минал спрямо момента на говорене момент или период и вече приключили; може да означава свършени и несвършени действия, краткотрайни и дълготрайни действия, еднократни действия, обичайни и общовалидни действия, повтарящи се действия, повтарящи се ограничен и неограничен брой пъти. Значение се получава от комбинацията между речниковото значение на глагола и изреченския контекст. Английското минало неопределено време няма българско съответствие, превежда се с форми на минало свършено и минало несвършено време на глаголи от свършен и несвършен вид. Ето няколко примера:

*John often **wrote** letters when he was younger – Джон често **пишеше** писма, когато беше по-млад.*

*John **wrote** letters for an half of hour – Джон **писа** писма половин час.*

За правилният български превод на вида на глагола влияние оказват определеността на допълнението и контекста на изречението:

*The artist **painted** landscapes most of his life – Художникът **рисуваше** пейзажи през по-голямата част от своя живот.*

*The artist **painted** a rose - Художникът **нарисува** роза.*

*The waiter **cleared** tables in the corner – Сервитьорът **почистваше** масите в ъгъла.*

*The waiter **cleared** the table in the corner – Сервитьорът **почисти / разчисти** масата в ъгъла.*

- **Минало прогресивно време – *I was playing tennis***

Образуване – от формите на спомагателния глагол *to be* в минало просто време и сегашно причастие на спомагателния глагол.

Примерни форми – *I was / you were / he (she / it) was / we were / you were / they were playing*

Употреба – изразява несвършени действия или състояния, извършващи се в определен минал момент или период, който е посочен или се подразбира от контекста; използва се като фон на други свършени действия; понеже означава единствено несвършени действия, на български език се превежда с форми на минало несвършено време на глаголи от несвършен вид.

*The child **was eating** an apple when his mother came in – Детето **ядеше** ябълка, когато майка му влезе.* (сравнено с *The child **ate** an apple – Детето **изяде** една ябълка*)

*Yesterday John **was working** in the garden – Вчера Джон **работеше** в градината.*

- **Минало перфектно време – *I had played tennis***

Образуване – от формите на спомагателния глагол *to have* в минало просто време и минало причастие на пълнозначния глагол

Примерни форми – *I / you / he / she / it / we / you / they had played*

Употреба – изразява действия или състояния, извършени или извършвани преди ориентационния момент или период, който от своя страна е минал спрямо момента на говорене; съответства на българското минало предварително време по значение, често се превежда с неговите форми на глаголи от свършен и несвършен вид; понякога по-

удачен е преводът на български език с минало свършено или минало несвършено време.

*When I met him, he **had** already **found** a new job – Когато го срещнах, той вече си **беше намерил** нова работа.*

*John **had worked** at the factory for 5 years when he got a promotion – Джон **беше работил** във фабриката 5 години преди да получи повишение.*

*By six the housewife **had** already **prepared** the sandwiches - До шест часа домакинята вече **беше приготвила** сандвичите.*

- **Минало перфектно-прогресивно време – *I had been playing tennis***

Образуване - от формите на спомагателния глагол *to be* в минало перфектно време и сегашно причастие на пълнозначния глагол

Примерни форми – *I / you / he / she / it / we / you / they had been playing*

Употреба - изразява действия или състояния, които са започнали в неопределен момент или период и са минали спрямо други действие или период, които от своя страна също са минали спрямо момента на говорене; означените действия или състояния продължават в момента, спрямо който се определят като минали и резултатът от тях е наличен в този момент или период. Това английско време няма български аналог и понеже изразява само несвършени действия, се превежда с минало предварително време на глаголи от несвършен вид. Всъщност по значение и употреба тези времена съвпадат. Но понякога, в зависимост от речниковото значение на глагола и от изреченския контекст, е подходящо значението на английското време да се предаде на български с форма на глагола в минало несвършено време.

*I **had been playing** tennis for two hours when Ann came – Аз **бях играл** тенис два часа, когато Ан дойде.*

*John **had been washing** the dishes – Джон **беше измил** чиниите.*

*John **had been washing** the dishes for ten minutes – Джон **миеше** чиниите десет минути.*

- **Бъдеще неопределено (просто) време – *I will / shall play tennis tomorrow***

Образуване – от спомагателния глагол *will / shall* (за 1 л., ед. ч.) и инфинитива на пълнозначния глагол.

Примерни форми – *I / you / he / she / it / we / you / they will play*

Употреба – изразява действия или състояния, бъдещи спрямо момента на говорене; може да означава свършени и несвършени действия, краткотрайни и дълготрайни действия, еднократни действия, обичайни и общовалидни действия, повтарящи се действия, повтарящи се ограничен и неограничен брой пъти. Превежда се на български език с формите на бъдеще време на свършени и несвършени глаголи.

*He **will knock** three times on the door - Той **ще почука** три пъти на вратата.*

*He **will** always **knock** on the wall if you make noise - Той **ще чука** винаги на стената, ако вдигаш шум.*

- **Бъдеще прогресивно време – *I will / shall be playing***

Образуване – от формите на спомагателния глагол *to be* в бъдеще неопределено време и сегашното причастие на спрегаемия глагол

Примерни форми - *I / you / he / she / it / we / you / they will be playing*

Употреба – изразява действия или състояния, които ще се извършат в конкретен бъдещ спрямо момента на говорене момент или период, които са посочени или се подразбират от контекста; често е фон за други свършени действия. Превежда се на български език с бъдеще време на глаголи в несвършен вид.

*If you call me after eight I will not pick up the phone because I **will be watching** TV - Ако ми се обадиш след осем, няма да вдигна телефона, защото **ще гледам** телевизия.*

- **Бъдеще перфектно време – *I will have played tennis***

Образуване – от формите на спомагателния глагол *to have* в бъдеще неопределено време и минало причастие на пълнозначния глагол.

Примерни форми - *I / you / he / she / it / we / you / they will have played*

Употреба – изразява действия или състояния, бъдещи спрямо ориентационния момент или период, който от своя страна също е бъдещ спрямо момента на говорене; превежда се на български език с бъдеще предварително време, главно с глаголи от свършен вид.

*When John gets a promotion, he **will have worked** there for five years - Когато Джон получи повишение, **ще е работил** там пет години.*

*By the time the guests have arrived, the housewife **will have prepared the sandwiches** - Докато гостите пристигнат, домакинята **ще е приготвила** сандвичите.*

*John **will have washed the dishe**, when you see him – Джон **ще е измил съдовете**, когато го видиш.*

- **Бъдеще перфектно-прогресивно време – *I will have been playing tennis***

Образуване - от формите на спомагателния глагол *to be* в бъдеще перфектно време и сегашно причастие на пълнозначния глагол.

Примерни форми - *I / you / he / she / it / we / you / they will have been playing*

Употреба - изразява действия или състояния, бъдещи спрямо ориентационния момент или период, който от своя страна също е бъдещ спрямо момента на говорене; характерното е, че действието е само несвършено и се включва в ориентационния момент или период; превежда се на български език с бъдеще предварително време на глаголи от несвършен вид.

*By the time the guests have arrived, the housewife **will have been preparing** the sandwiches - Докато гостите пристигнат, домакинята **ще е приготвяла** сандвичите.*

*John **will have been washing** the dishes when you see him – Джон **ще е мил** съдовете, когато го видиш.*

- **Бъдеще в миналото неопределено (просто) време – *I would play tennis***

Образуване – от формата на спомагателния глагол *will* в минало просто време *would* и инфинитив на пълнозначния глагол.

Примерни форми - *I / you / he / she / it / we / you / they would play*

Употреба – изразява действия или състояния от различен тип - свършени и несвършени действия, краткотрайни и дълготрайни действия, еднократни действия, обичайни и общовалидни действия, повтарящи се действия, повтарящи се ограничен и неограничен брой пъти, които са бъдещи спрямо ориентационния момент, който от своя страна е минал спрямо момента на говорене; на български език се превежда с бъдеще време в миналото от свършени и несвършени глаголи. Понякога в сложни изречения е по-удачен друг избор на подходящо време на глаголните форми на български език – значението се запазва, но различните морфологични принципи откриват и друга възможност за превод. Тази особеност няма да бъде обект на изследване сега и само се отбелязва като съществуваща.

*They **would soon** ask her to leave – Те **щяха да я помолят** скоро да напусне.*

*Three weeks later I **wouldn't stand** my colleagues' manners any more – След три седмици **нямаше да понасям** обноските на колегите си.*

*I knew he **would call** me later – Знаех, че той **ще се обади / щеше да се обади** по-късно.*

*John **would call** his friends when he needed money – Джон **се обаждаше / щеше да се обади** на приятелите си, когато се нуждаеше от пари.*

- **Бъдеще в миналото прогресивно време – *I would be playing***

Образуване – от формите на спомагателния глагол *to be* в минало в бъдещето неопределено време и сегашно причастие на пълнозначния глагол.

Примерни форми - *I / you / he / she / it / we / you / they would be playing*

Употреба - изразява действия или състояния, които ще се извършват в конкретен бъдещ спрямо друго действие момент или период като ориентационния момент или период е минал спрямо момента на говорене, често е фон за други свършени действия; превежда са на български език с бъдеще в миналото от несвършени глаголи.

*I knew he **would be watching** television, so I did not phone him – Знаех, че той **ще гледа** телевизия, затова не му се обадих.*

- **Бъдеще в миналото перфектно време – *I would have played***

Образуване – от формите на спомагателния глагол *to have* в минало в бъдещето неопределено време и миналото причастие на пълнозначния глагол.

Примерни форми - *I / you / he / she / it / we / you / they would have played*

Употреба - изразява действия или състояния, бъдещи спрямо ориентационния момент или период, който от своя страна е минал спрямо момента на говорене; резултатът от действието е наличен в или оказва влияние на ориентационния момент, като в повечето случаи то вече е приключило; на български език се превежда с бъдеще в миналото предварително време, главно от свършени глаголи.

*He **would have worked** in the factory for five years before getting a promotion – Той **щеше да е работил** във фабриката пет години, преди да получи повишение.*

*By six o'clock the housewife **would have prepared** the sandwiches - До шест часа домакинята **щеше да е приготвила** сандвичите.*

*I knew he **would have fallen** asleep, so I did not call him – Знаех, че той **ще е заспал**, затова не му се обадих.*

*By six o'clock John **would have washed** the dishes - До шест часа Джон **щеше да е измил** съдовете.*

- **Бъдеще в миналото перфектно-прогресивно време – *I would have been playing***

Образуване – от формите на глагола *to be* в бъдеще перфектно време в миналото и сегашното причастие на пълнозначния глагол.

Примерни форми - *I / you / he / she / it / we / you / they would have been playing*

Употреба – изразява само несвършени действия или състояния, бъдещи спрямо ориентационния момент или период, който от своя страна е минал спрямо момента на говорене; действието протича и в ориентационния момент; превежда се на български език с бъдеще в миналото предварително време от глаголи в несвършен вид.

*I knew that by six o'clock the housewife **would have been preparing** the sandwiches - Знаех, че до шест часа домакинята **щеше да е приготвяла** сандвичите.*

By six o'clock John ***would have been washing*** *the dishes for ten minutes - До шест часа Джон* ***щеше да е мил съдовете*** *десет минути.*

3.5. Превод на английските глаголни форми със съпоставими български

В тази част ще съотнесем всяко от английските времена с правилната форма на български език. Ще се следва употребата на времето в английското просто изречение като база за правилен превод.

Английска глаголна форма	Българска глаголна форма
Сегашно неопределно време – *I sing*	Сегашно време – *пея*
Сегашно прогресивно време – *I am singing*	Сегашно време на глаголи само в несвършен вид – *пея*
Сегашно перфектно време – *I have sung*	Минало неопределено време – *съм пял*
Сегашно перфектно-прогресивно време – *I have been singing*	Сегашно време на глаголи само в несвършен вид – *пея*
Минало неопределено време – *I sang*	Минало свършено време – *пях* Минало несвършено време – *пеех*
Минало прогресивно време – *I was singing*	Минало несвършено време на глаголи само в несвършен вид – *пеех*
Минало перфектно време – *I had sung*	Минало предварително време – *бях пял*
Минало перфектно-прогресивно време – *I had been singing*	Минало предварително време на глаголи само в несвършен вид – *бях пял*
Бъдеще неопределено време – *I will sing*	Бъдеще време – *ще пея*
Бъдеще прогресивно време – *I will be singing*	Бъдеще време на глаголи само в несвършен вид – *ще пея*
Бъдеще перфектно време – *I will have sung*	Бъдеще предварително време – *ще съм пял*

Бъдеще перфектно-прогресивно време – *I will have been singing*	Бъдеще предварително време на глаголи само от несвършен вид – *ще съм пял*
Бъдеще в миналото неопределено време – *I would sing*	Бъдеще време в миналото – *щях да пея*
Бъдеще в миналото прогресивно време – *I would be singing*	Бъдеще време в миналото на глаголи само от несвършен вид – *щях да пея*
Бъдеще в миналото перфектно време – *I would have sung*	Бъдеще в миналото предварително време – *щях да съм пял*
Бъдеще в миналото перфектно-прогресивно време – *I would have been singing*	Бъдеще в миналото предварително време на глаголи само от несвършен вид – *щях да съм пял*

Информацията в горната таблица илюстрира правилните български форми за английските времена. Тъй като засега категорията вид остава извън предмета на изследване и полето на превода, тя няма да бъде заложена като фактор в изградените правила. Генерираните български сказуеми ще съдържат глаголни форми само от несвършен вид, определян като базов (ГСБКЕ 1983). По същата причина английското минало просто време ще бъде превеждано като минало свършено (изборът на аорист или имперфект при българския превод зависи от вида като синтактична категория в английски и е обект на бъдещи изследвания). Подлогът няма да се показва при превода на формите в 1 и 2 л., ед. ч. Друго необходимо уточнение е, че се избира само формата на глагола за ед.ч. при превод на английското местоимение *you*. Изграденият модел за демонстрация на машинен превод няма претенция да обхване всички езикови явления, свързани с употребата на сказуемото и неговите форми, а има за цел да създаде правила за оптимизиране на механизните за откриване на правилното английско време от една страна, и вярно генериране на форма на съответното българско от друга.

4. Конструиране на правила за превод и изграждане на модел за прилагането им

4.1. Основни типове грешки при статистическия машинен превод на системата Google Translate

Инструментът за машинен превод, разработен от компанията Google, до който има свободен достъп в интернет пространството на адрес http://translate.google.com (българският домейн се намира на страница *http://translate.google.bg*), осъществява двупосочен превод между петдесет и седем езика. Подходът е изцяло статистически, нивото на съотнасяне в многоезичния корпус е фразово, като за фраза се възприема произволна последователност от два знака (например две думи, точка и дума или дума и запетая). Друг извод от направените наблюдения е, че системата реално превежда единствено от и на английски и друг език, т.е. преводът от румънски на чешки език например всъщност е превод от румънски на английски и резултатът бива преведен на чешки. Понеже се работи с биграми, в повечето случаи контекстът не оказва влияние върху резултата, няма разлика между превод на сказуемо в просто и сложно изречение. Забелязват се еднакви типове грешки:

- Повечето аналитичните глаголни форми не се откриват, само сегашно перфектно и сегашно прогресивно време биват разпознати понякога, но дори и в тези случаи липсва съгласуване с граматичните характеристики на подлога - *he has read the book → той е прочел книгата, she has read the book → тя е прочел книгата, I have sung → имам пеят.*

- Често сегашното причастие се превежда като съществително име - *I am singing →аз съм пеене, she is reading a book → тя е четене на книга, she is slowly parking the car → тя е бавно паркинг на автомобила.*

- Несъгласуване на формите на сказуемото с подлога при наличие или липса на обстоятелствено пояснение между тях – *you sing → пееш*, но *you often sing → често пее, we are always singing → ние сме винаги пее.*

Действащият алгоритъм на статистическия подход, базиран на структури от биграми, дава обяснение за появата на тези неточности. Прави впечатление фактът, че различните лексеми в еднакъв контекст (т.е. при едни и същи темпорални

характеристики на произволно избрани глаголи) получават различна конструкция при българския превод. Понякога дори поставянето на главна буква или на запетая променя резултата. Основната причина за допусканите грешки е в принципа на действие на използвания алгоритъм – той намира липсващата дума според предишния знак в биграмата без значение на това какво има около разглежданата фраза. Друг фактор е ограничеността на функциониране на статистическо-математическия принцип – той също действа само в границите на конкретната биграма. Затова начинът на преодоляване на допусканите неточности е чрез прилагане на правила върху резултата, които да откриват формите на глагола, независимо какво има между тях, и да отчитат граматическата информация, носена от подлога, върху сказуемото. Употребите на английските времена ще бъдат отбелязани в създаден за целта корпус – XML документ, направен със средствата на редактора Oxygen XML Editor. Останалите приложения също ще бъдат написани с тази програма. Ще бъдат изградени два морфологични работни речника и работен двуезичен английско-български речник.

4.2. Изграждане на паралелен корпус

За изпълнение на задачите на работата е необходим паралелен двуезиков корпус, който да съдържа формите на английския глагол в шестнадесетте времена и съответния им превод, извършен от системата Google Translate. Идеята е да се регистрират грешните варианти, които подлежат на обработка, както и да се използва резултата от статистическия превод. Корпусът представлява XML документ от 598 преведени израза (50% фрази и същият процент изречения) се намира на приложения на диска файл **corpus.xml**. Състои се от форми на английските глаголи във всички лица и числа за разглежданите времена. Също така има и примери с включено наречие, те са избрани да илюстрират възможните му употреби, когато то е позиционирано между спомагателния и пълнозначния глагол, защото това негово положение създава трудности при откриването на вярната глаголна форма при статистическия метод на превод. Въпреки че липсват примери, илюстриращи словоред от вида SAVO(C) – когато обстоятелството е веднага след подлога и преди всички форми на сказуемото, изградените правила за вярно генериране на българската глаголна форма работят и тук. Формите в корпуса са представени в отделни елементи самостоятелно като фрази и като сказуемо в изречение. За постигането на целите в настоящата работа избраните изречения са съобщителни, без вметнати части или разширения, а сказуемите – в положителна форма. Формално XML документът е изграден от елемента <ph>, който

има дъщерните елементи <en> и <bg> за английския израз и съответния му български превод, направен от Google. Началото на файла изглежда по следния начин:

```
 <?xml version="1.0" encoding="UTF-8"?>
<corpus>
 <ph>  <en>I sing</en><bg>Пея</bg></ph>
<ph>
<en>I sing every day.</en><bg>Аз пея всеки ден.</bg></ph>
<ph>
<en>I often sing</en><bg>аз често пее</bg></ph>
<ph>
<en>I barely sing.</en><bg>Едва пее.</bg></ph>
 <ph>
  <en>you sing</en><bg>пееш</bg></ph>
<ph>
 <en>You sing every day.</en><bg>Можете пеят всеки ден.</bg></ph>
<ph>
 <en>you often sing</en><bg>често пее</bg></ph>
<ph>
<en>You barely sing.</en> <bg>Едва се пее.</bg></ph>
```

Ето и част от корпуса за някои от аналитичните глаголни форми:

```
<ph>
  <en>he will be singing</en><bg>той ще се пее</bg></ph>
 <ph>
  <en>He will be singing tonight.</en><bg>Той ще се пее тази вечер.</bg></ph>
 <ph>
  <en>he will be loudly singing</en><bg>той ще бъде шумно пеене</bg></ph>
<ph>
  <en>He will be loudly singing tonight.</en><bg>Той ще бъде силно пее тази вечер.</bg></ph>
<ph>
  <en>they would have been singing</en><bg>те биха били пеене</bg> </ph>
 <ph>
  <en>They would have been singing with all their friends last year.</en><bg>Те биха били пеене с
всички техни приятели миналата година.</bg></ph>
 <ph>
  <en>they would have been loudly singing</en><bg>те биха били силно пеене</bg></ph>
 <ph>
  <en>They would have been loudly singing last night.</en><bg>Те биха били силно пеене
снощи.</bg> </ph>
 <ph>
  <en>they would actually have been singing</en><bg>те биха били пеене</bg></ph>
 <ph>
  <en>They would actually have been singing with all their friends last year.</en><bg>Те биха били
пеене с всички техни приятели миналата година.</bg></ph>
 <ph>
  <en>they would have actually been singing</en><bg>те ще са действително пее</bg></ph>
 <ph>
  <en>They would have actually been singing with all their friends last year.</en><bg>Те ще са
действително пее с всички техни приятели миналата година.</bg></ph>
```

За обработка на корпуса се създава отделен файл с име **tokens.xsl** – това е документ, на който чрез езика XSLT се пишат трансформациите, необходими за извличане на нужната информация от другите файлове. Извършва се токънизация на корпуса. Всяка дума от корпуса се загражда с таг **<word>** и по този начин се осигурява лесен достъп до токъните с цел изграждане на речник. Резултатът се записва във файл **tokens.xml** и има следния вид за едно изречение:

```
<ph>
    <en>
      <word>They</word>
      <word>would</word>
      <word>be</word>
      <word>singing</word>
      <word>tonight</word>
      <punct>.</punct>
    </en>
    <bg>
      <word>Те</word>
      <word>ще</word>
      <word>се</word>
      <word>пее</word>
      <word>тази</word>
      <word>вечер</word>
      <punct>.</punct>
    </bg>
  </ph>
```

4.3. Изграждане на морфологични речници за българските и английските думи

За определяне на правилната форма на всяка лексема е необходим морфологичен речник както за английските, така и за българските глаголи. Такива компютърни речници описват цялата граматична информация, носена от словоформата, и я представят в удобен за обработка вид. Двата направени работни речника са малки по обем, защото включват само думите от корпуса с цел практическа демонстрация на избрания подход. Намират се на файловете **dict.xml** (за българските думи) и **english.xml** (за английските). Изградени са като XML документи. Състоят се от елемента **<word>** с атрибут **id=" "**, който получава като стойност конкретната словоформа от корпуса, и по този начин за всяка дума от примерите се образува еднотипна структура. Отделно

<word> има дъщерни елементи **<gram>** и **<form>.** Атрибутите на <gram> съдържат като стойности характеристиките на лемата, а тези на <form> - на словоформата. Символите, използвани за имена и стойности на атрибутите, са избрани така, че да са уникални. За да няма дублиране, имената на елементите и атрибутите са на латиница, а на стойностите – на кирилица. Ето пример за описаната структура в речника на английските словоформи *english.xml*:

```
<word id="I">
   <gram lemma="I" pos="M">
   <form v="л" c="и" p="1" n="е"/></gram>
 </word>
```

Тези четири реда съдържат граматичните характеристики на словоформата *I*: *id=”I”* има за стойност словоформата такава, каквата е извлечена автоматично от корпуса; стойността на атрибута *lemma* показва лемата, в случая тя е *I*; атрибутът *pos* показва каква част на речта е думата като *M* означава местоимение; *v=”л”* посочва вида на местоимението (лично), *c=”и”* – падежа (именителен), *p=”1”* – лицето (първо), *n=”е”* – числото (единствено). Атрибутите на елемента <form> се различават в зависимост от конкретната дума, защото различните части на речта имат различни категории. За елемента <gram> задължителните атрибути са *lemma* и *pos*. Ето и представяне на възможните стойности (определени са единствено спрямо думите в английски текст, защото изработката на цялостен морфологичен речник е извън предмета на изследване; обаче изграждането на примерен работен речник със срещнатите лексеми е необходимо условие за изпълнение на алгоритъма за разпознаване на глаголни форми в езика).

Атрибути на елемента <gram> - граматични характеристики на лемата:

- **lemma** – има за стойност словоформата
- **pos** – има за стойност означение за част на речта на лемата, възможните варианти са **М** – местоимение, **Г** – пълнозначен глагол, **СГ** – спомагателен глагол, **П** – прилагателно име, **С** – съществително име, **Р** – предлог, **Ч** – числително име, **Н** – наречие, **О** – определителен член (article) за *a* / *the*.
- **n** - при местоименията *all* и *every* има за стойност „е” за ед. ч. и „мн” за мн. ч.

Атрибути на елемента <form> - конкретни граматични особености на словоформата

- **p** – има за стойност категорията лице, възможните варианти са 1 – първо, 2 – второ, 3 – трето лице.
- **n** – има за стойност категорията число, възможните варианти са „е” за ед. ч. и „мн” за мн. ч.
- **g** – има за стойност категорията род, възможните варианти за „ж” – женски, „м” – мъжки, „с” – среден род.
- **c** – има за стойност категорията падеж, възможните варианти са „и” – именителен, „в” – винителен, „д” – дателен падеж.
- **t** – има за стойност формален показател на времето или на причастието в английската глаголна форма, възможните варианти са „и” – инфинитив, „с” – сегашно време, „сп” – сегашно причастие (singing), „п” – минало причастие (sang), „мн” – перфектно причастие (sung).
- **a** – има за стойност категорията вид на глагола само при сегашно причастие, единствен възможен вариант е „нсв” – несвършен вид.
- **v** – има за стойност вида на местоимението, възможни вариант са „л” – лично, „п” – притежателно.

Ето част от документа на морфологичния речник за някои думи от корпуса:

```
<word id="you">
    <gram lemma="I" pos="M">
    <form v="л" c="и" p="2" n="е"/>
    <form v="л" c="и" p="2" n="мн"/>
    <form v="л" c="в" p="2" n="е"/>
    <form v="л" c="в" p="2" n="мн"/>
    <form v="л" c="д" p="2" n="е"/>
    <form v="л" c="д" p="2" n="мн"/>
</gram>
</word>
<word id="we">
    <gram lemma="I" pos="M">
    <form v="л" c="и" p="1" n="мн"/></gram>
</word>
<word id="her">
    <gram lemma="my" pos="M" >
    <form v="п" g="ж" p="3" n="е"/></gram>
    <gram lemma="I" pos="M">
```

```
    <form v="л" g="ж" c="в" p="3" n="e"/>
    <form v="л" g="ж" c="д" p="3" n="e"/></gram>
</word>
<word id="sing">
    <gram lemma="sing" pos="Г">
    <form t="c"/> </gram></word>
<word id="sings">
    <gram lemma="sing" pos="Г">
    <form t="c" p="3" n="e"/></gram>
</word>
<word id="sang">
    <gram lemma="sing" pos="Г">
    <form t="п"/></gram>
</word>
  <word id="sung">
    <gram lemma="sing" pos="Г">
    <form t="мп"/></gram>
  </word>
  <word id="singing">
    <gram lemma="sing" pos="Г">
    <form t="сп" a="нсв"/></gram>
</word>
<word id="am">
    <gram lemma="be" pos="СГ"/>
    <form t="c" p="1" n="e"/>
</word>
```

Морфологичният речник на българските думи е конструиран по същия принцип, има малки разлики в тагирането, свързани с граматичните специфики в двата езици. Например и в двата речника има добавени атрибути при лемата, характерни само за нея – при *child* е за означаване на рода, при *пеене* – за посочване на вида на съществителното като отглаголно и основната форма на глагола. Ето как изглежда информацията за думата *пеене*, пълните речници са намират на приложения диск:

```
<word id="пеене">
    <gram lemma="пеене" type="отгл" pos="С" verb="пея">
    <from n="e" a="н"/>
    </gram>
 </word>
```

Консултирането с информацията в морфологичния речник позволява безпроблемното откриване и позициониране на английското сказуемо. Разграничаването между спомагателен и пълнозначен глагол не само е лингвистично правилно, а и предполага ефективното разпознаване на аналитичните форми.

Информацията за съществителните имена и местоименията ще помогне за правилното съгласуване на формата на сказуемото с подлога в изречението.

4.4. Създаване на алгоритъм за откриване на времето на английския глагол и конструиране на правила за превода му със съответната българска форма

Необходимият алгоритъм трябва да изпълнява следните стъпки:

1) намиране на пълнозначния глагол в изречението;
2) разрешаване на многозначности, резултат на граматическа омография;
3) определяне на темпоралната характеристика на сказуемото;
4) намиране на подлога в английското изречение;
5) проверка на формата на преведеното английско сказуемо от системата Google Translate;
6) генериране на граматически правилна форма на български глагол (извършва се и съгласуване с подлога);

Първо следва да се определят границите на сказуемото в английското изречение, което вече е преструктурирано като елемент <en> с деца <word>. Начална стъпка на този процес е откриване на пълнозначния глагол чрез езика XSLT. Създава се нов файл с име **predicate.xsl** и се прави връзка между него и морфологичния речник *english.xml*. Чрез XSLT команда се търси думата в изречението, за която записът в речника съдържа таг за пълнозначен глагол. Т.е. намират се стойностите на атрибута *lemma=””*, за които е вярно *pos=”Г”*. Ето и реда от кода на командата, поставящ конкретните условия:

```
select="word[.=document('english.xml')//word[gram[@pos='Г']]/@id and not(preceding-sibling::*[1][self::word[.=document('english.xml')//word[gram[@pos='О' or @pos='ПТ']]/@id]])]"/>
```

При проверка на генерирания списък с пънозначните глаголи в корпуса се намират лексеми, на които при едно от значенията има елемент с атрибут *pos=”Г”*, но в текста те не функционират като глаголи. Примери за този вид многозначност са думите *tones* и *rising*. Проблемът при първата идва от речника поради възможността лемата *tone* да е съществително име и глагол. При втората дума сегашното причастие изпълнява функция на прилагателно име. Откриването и отстраняването на подобни употреби от

желания списък с глаголни форми става чрез още един *template* на езика XML. Командата изключва автоматично думите, пред които непосредствено намира прилагателно име или определителен член – това са позиции, блокиращи позицията на сказуемо – *the rising sun, high tones*. Ето как изглеждат представянето на лексемата *tones* в морфологичния речник и реда от инструкцията в XSLT, съдържащ указанията за търсене:

```
<word id="tones">
    <gram lemma="tone" pos="С">
    <form n="мн"/>  </gram>
    <gram lemma="tone" pos="Г">
    <form t="с" p="3" n="е"/> </gram>
```

```
select="word[.=document('english.xml')//word[gram[@pos='Г']]/@id and not(preceding-
sibling::*[1][self::word[.=document('english.xml')//word[gram[@pos='О' or @pos='П']]/@id]])]"/>
```

След откриване на всички форми на английските пълнозначни глаголи в корпуса следва да се приложи локален алгоритъм за разпознаване на времето на сказуемото. Този алгоритъм трябва да изпълнява главно две неща – да се обръща към морфологичния речник за определяне на вида на спомагателните глаголи и да формира само едно глаголно време като резултат. Командите са написани на езика XSLT чрез инструкцията *choose – when – otherwise*. Основните стъпки на процеса са следните:

1) ако за формата на пълнозначния глагол е вярно t=”с” → отива на 2, ако не е вярно – на 4.

2) ако няма спомагателен глагол преди формата → избира сегашно просто време (СВ), иначе отива на 3.

3) ако за първия спомагателен глагол е вярно t=”с” → избира бъдеще просто време (БВ), иначе избира бъдеще в миналото просто време (БМВ)

4) ако за формата на пълнозначния глагол е вярно t=”п” → избира минало просто време (МВ), иначе отива на 5.

5) ако за формата на пълнозначния глагол е вярно t=”мп” → отива на 6, иначе – на 9.

6) ако има два спомагателни глагола преди формата → отива на 7, иначе отива на 8.

7) ако за първия спомагателен глагол е вярно t=”c” → избира бъдеще перфектно време (БПВ), иначе избира бъдеще в миналото перфектно време (БМПВ).

8) ако за спомагателния глагол е вярно t=”c” → избира сегашно перфектно време (СПВ), иначе избира минало перфектно време (МПВ).

9) ако има един спомагателен глагол преди формата → отива на 10, иначе – на 11.

10) ако за спомагателния глагол е вярно t=”c” → избира сегашно прогресивно време (СПрВ), иначе – минало прогресивно време (МПрВ).

11) когато има два спомагателни глагола преди формата → отива на 12, иначе – на 15.

12) ако за втория спомагателен глагол е вярно t=”и” → отива на стъпка 13, иначе – на 14.

13) ако за първия спомагателен глагол е вярно t=”c” → избира бъдеще прогресивно време (БПрВ), иначе – бъдеще в миналото прогресивно време (БМПрВ).

14) ако за първия спомагателен глагол е вярно t=”c” → избира сегашно перфектно-прогресивно време (СППрВ), иначе – минало перфектно-прогресивно време (МППрВ).

15) ако за първия спомагателен глагол е вярно t=”c” → избира бъдеще перфектно-прогресивно време (БППрВ), иначе – бъдеще в миналото перфектно-прогресивно време (БМППрВ).

Ето част от алгоритъма, написан на езика XSLT, целият код се намира на файла *predicate.xsl* на приложения диск. Долните редове съответстват на условията от пета до осма инструкция включително и показват алгоритъма за определяне на перфектните времена.

```
<xsl:when test="$W[descendant::form[@t='мп']]">
        <xsl:choose>
          <xsl:when test="$AUX=2">
            <xsl:choose>
              <xsl:when test="$A2[descendant::form[@t='c']]">
                <xsl:text>БПВ</xsl:text>
              </xsl:when>
              <xsl:otherwise>
```

```
            <xsl:text>БМПВ</xsl:text>
          </xsl:otherwise>
        </xsl:choose>
      </xsl:when>
      <xsl:otherwise>
        <xsl:choose>
          <xsl:when test="$A1[descendant::form[@t='c']]">
            <xsl:text>СПВ</xsl:text>
          </xsl:when>
          <xsl:otherwise>
            <xsl:text>МПВ</xsl:text>
          </xsl:otherwise>
        </xsl:choose>
      </xsl:otherwise>
    </xsl:choose>
  </xsl:when>
```

По този начин се разпознават формите на английските времена. На елемента <en>, ограждащ английското изречение, се слага атрибут *tense=” ”*, на който автоматично се приписва като стойност резултата на алгоритъма за откриване на времето.

Следващата стъпка е намиране на подлога в английското изречение с цел извличане на граматичните му характеристики и генериране на правилната българска глаголна форма спрямо тях. Типичният английски словоред SVO предопределя мястото на субекта. Чрез XSLT команда се първата намира дума пред пълнозначния глагол, която е съществително име или лично местоимение в именителен падеж. Инструкцията изглежда така:

```
<xsl:variable name="subj" select="$V/preceding-
sibling::word[.=document('english.xml')//word[gram[@pos='C' or (@pos='M' and form[@v='л' and
@c='и'])]]/@id]"/> .
```

От речника се вземат морфологичните характеристики на думата, избрана за подлог, и са слагат като стойности на атрибутите *n=””* и *p=””* за категориите число и лице на елемента <word>. Той вече е получил атрибут *func=”subj”*, който показва на системата, че думата, заградена с този таг, изпълнява функцията на подлог в изречението на английски език. Заедно с атрибута *tense=” ”* на елемента <*en*> новата информация се записва във файла **tagged-corpus.xml**, към който ще се приложат правила за генериране на правилната форма на българското сказуемо. Пример за едно тагирано английско изречение е следната извадка от корпуса, времето на сказуемото е правилно разпознато като бъдеще в миналото перфектно-прогресивно време:

```
<ph>
<en tense="БМППрВ">
  <word func="subj" n="e" p="3">He</word>
  <word>would</word>
  <word>have</word>
  <word>been</word>
  <word>singing</word>
  <word>with</word>
  <word>all</word>
  <word>my</word>
  <word>friends</word>
  <word>last</word>
  <word>year</word>
  <punct>.</punct>
</en>
<bg>
  <word>Той</word>
  <word>щеше</word>
  <word>да</word>
  <word>пее</word>
  <word>с</word>
  <word>всички</word>
  <word>си</word>
  <word>приятели</word>
  <word>миналата</word>
  <word>година</word>
  <punct>.</punct>
</bg>
</ph>
```

Следващата стъпка в алгоритъма за превод е генериране на граматически правилната форма на българското сказуемо. Създава се файл **translate.xsl** – той съдържа всички инструкции, които сега се задават към новосъздадения корпус с добавена информация за подлога и времето в английското изречение *tagged-corpus.xml*. Проверява се дали резултатът от машинния превод на английското сказуемо е верен. Намира се пълнозначният глагол в българското изречение, преведено от Google Translare. Това става с команда на XSLT, търсеща тази дума в изречението, която има

таг за пълнозначен глагол в морфологичния речник. За разрешаване на многозначности от вида на възникналите при формите *има* и *изгряващо*, при които резултатът би бил два пълнозначни глагола в едно изречение, се поставят ограничения при търсенето – думата да е последният елемент в редицата с атрибут *pos=”Г”*[4], за който е изпълнено едно от условията: 1) съдържа атрибути за лице *p=” ”* с цифрова стойност и за сегашно или минало свършено време *t=”c”* или *t=”a”*; 2) съдържа атрибутите *p=”n”*, *t=”a”*, *v=”a”* и не съдържа атрибут *a=””* - думата е нечленувано аористно причастие. Ако няма резултат от това търсене, следва команда за намиране на отглаголното съществително име, с което е преведена английската форма, и по този начин се прави връзка с българския глагол. Случаите на превод на конструкцията *–ing* със съществително име са често срещани, броят на грешните преводи по тази причина е 122 от общо 598 израза. Точната инструкция за намиране на думата, заемаща позиция на сказуемо, има следния вид:

```
<xsl:template match="bg">
<xsl:variable name="V">
  <xsl:choose>
    <xsl:when test="word[.=document('dict.xml')//word[descendant::*[@pos='Г'] and
(descendant::form[@p!='п'] or (descendant::form[@p='п' and @t='a' and @v='a' and
not(@a)]))]/@id]">
      <xsl:value-of select="word[.=document('dict.xml')//word[descendant::*[@pos='Г'] and
(descendant::form[@p!='п'] or (descendant::form[@p='п' and @t='a' and @v='a' and
not(@a)]))]/@id][position()=last()]"/>
    </xsl:when>
    <xsl:otherwise>
      <xsl:value-of select="word[.=document('dict.xml')//word[descendant::*[@pos='C' and
@type='отгл']]/@id]"/>
    </xsl:otherwise>
  </xsl:choose>
</xsl:variable>
```

[4] Статистическият машинен превод на Google, използващ подход, основан на биграми, има предимството да запазва английския словоред в българския превод. Въпреки че понякога резултатът е синтактично неиздържан, тази особеност предполага с висока степен на сигурност възможните позиции на генерираните български думи в израза.

След намиране на пълнозначен глагол в българското изречение следващата стъпка е определяне на граматичните характеристики на сказуемото и сравнение с информацията за подлога в английското изречение. Ако преведената от Google форма е грешна, се генерира правилната по следния начин. Прави се връзка с файла *tagged-corpus.xml* с помощта на променливите V, VL, T, P, N, G. Във V се записва информация за конкретната глаголна форма на сказуемото, във VL – за лемата, в T – за времето в английското изречение, в P – за лицето на думата, заемаща позицията на подлог, в N – за числото на същата дума, в G – за нейния род (който оказва влияние при резултата на превода на българските времена, образуващи се с минало свършено причастие). Командата за трансформации в XSLT задейства локален алгоритъм за генериране на българската глаголна форма, която съответства на английската. Той се състои от следните стъпки:

1) Обръща се към променливата T и открива времето.

2) Отива в българския морфологичен речник и намира тази форма на лема, съвпадаща със стойността на VL, за която са верни фактите:

   - притежава атрибут t=”c”, t=”a” или t=”н” за сегашно, минало свършено или минало несвършено време в зависимост от търсения изход
   - притежава атрибут p=”” със стойност, съвпадаща с информацията на променливата P
   - притежава атрибут n=”” със стойност, съвпадаща с информацията на променливата N
   - притежава атрибут g=”” със стойност, съвпадаща с информацията на променливата G, или изобщо няма такъв атрибут

По този начин се намират точните форми на спомагателните и пълнозначните глаголи. После се изпълняват инструкции с правила за генериране на аналитичните форми. Процесът се състои в правилното позициониране на вече извлечените от речника конкретни думи и добавяне на частици на определени места. Ето някои редове от алгоритъма, който успява да конструира правилната глаголна форма, целият код се намира на файла *translate.xsl* на диска:

команда за генериране на бъдеще време в миналото като превод на английските форми за бъдеще в миналото неопределено време:

```
<xsl:when test="$T='БМВ'">
        <xsl:value-of select="document('dict.xml')//word[descendant::gram[@lemma='ща'] and
descendant::form[@t='н' and @p=$P and @n=$N]]/@id"/>
        <xsl:text>да</xsl:text>
        <xsl:value-of select="document('dict.xml')//word[descendant::gram[@lemma=$VL] and
descendant::form[@t='с' and @p=$P and @n=$N]]/@id"/>
    </xsl:when>
```

команда за генериране на бъдеще в миналото предварително време като превод на английските форми за бъдеще в миналото перфектно време:

```
<xsl:when test="$T='БМПВ'">
        <xsl:value-of select="document('dict.xml')//word[descendant::gram[@lemma='ща'] and
descendant::form[@t='н' and @p=$P and @n=$N]]/@id"/>
        <xsl:text>да</xsl:text>
        <xsl:value-of select="document('dict.xml')//word[descendant::gram[@lemma='съм'] and
descendant::form[@t='с' and @p=$P and @n=$N]]/@id"/>
        <xsl:value-of select="document('dict.xml')//word[descendant::gram[@lemma=$VL] and
descendant::form[@t='а' and @p='п' and @n=$N and (not (@g) or @g=$G)]]/@id"/>
    </xsl:when>
```

Полученият български превод на изречението *By midnight I would have sung all songs* вече е ***До полунощ щях да съм пял вече всички песни*** вместо статистическия вариант на Google *До полунощ бих вече са пели всички песни*. При прилагането на описаните правила за превод няма значение дали има обстоятелствено пояснение между формите на сказуемото, защото инструкциите се задават главно за глаголните форми. Проблемът с превода на английското сегашно причастие като отглаголна съществително име също бива разрешен. Пример за положителния резултат и в тази насока е двойката изречения *I have been sweetly singing this for ten minutes* и генерираното ***Пея сладко това в продължение на десет минути*** (в сравнение с българския превод от корпуса *имам мелодично пеене това в продължение на десет минути*). Като пример за правилното съгласуване с подлога по лице и число при наличието на причастие в българското сказуемо е преводът ***Той беше пял високи тонове*** на *He had been singing the high tones*, който премахва грешките в първоначалния вариант *Той* беше пеене *на високи тонове*.

В приложените файлове на диска може да се види резултата и за останалите глаголни форми и времена.

5. Заключение

В дипломната работа се изработи модел за превод от английски глаголни форми на български език. Приложеният подход представлява комбиниран начин, реализиран на два етапа – статистически машинен превод и машинен превод, базиран на правила относно резултата от действието на първия метод. Целта на настоящата работа е конструиране на правила за оптимизиране на статистически подход, пример за който е системата за превод Google Translate. Комбиниран машинен превод с действащ алгоритъм по описания в тезата начин все още не е приложен на практика в българските научни среди. Постигнатите резултати показват, че това би бил универсален подход към разрешаване на често допускани от статистическите методи грешки, защото правилата относно граматични характеристики са в голяма степен контекстното независими. Изграденият модел разрешава проблемите, свързани с превода на английски форми на български език като се справя с различни видове многозначност, намира лемата на глагола от отглаголното съществително име, определя границите на сказуемото в изречението, разпознава аналитичната английска глаголна форма, правилно генерира съответната българска чрез консултация в морфологичния речник.

Конструираните праавила и алгоритми биха могли да се приложат към системи за статистически машинен превод на по-обемни бази данни. За това единствено е необходимо наличието на добре структурирано и граматически правилно формално представяне на езиковите факти. Ефективен начин за подобрение на процеса на машинен превод е добавянето на морфологичен речник и неговото използване в система за проверка на резултата.

Може да се направи заключението, че изграденият модел е сполучлив и може да се приложи и върху други конструкции на простото и сложното изречение. Правилата за превод могат да се допълнят и да се разработят нови, отразяващи повече езикови явления, например категориите вид на изказването и евиденциалност, което е обект на бъдещи изследвания в областта.

Библиография:

Бъркалова 1997: Бъркалова, П. Българският синтаксис - познат и непознат. Увод в курса по синтаксис на СБЕ. Пловдив, Университетско издателство „Паисий Хилендарски", 1997.

Брокет 2002: Brockett, C., Aikawa, T., Aue, A., Menezes, A., Quirk, C. English-Japanese Example-Based Machine Translation Using Abstract Linguistic Representations. *http://www.mt-archive.info/authors-B.htm*

Винивартер 2007: Winiwarter, W. Machine translation using corpus-based acquisition of transfer rules. *http://ieeexplore.ieee.org*

ГСБКЕ 1983: Граматика на СБКЕ.Том 3. Синтаксис. БАН, София, 1983.

ГСБКЕ 1983: Граматика на СБКЕ.Том 2. Морфология. БАН, София, 1983.

Грийнбаум 1995: Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. A Comprehensive Grammar of the English Language. L., Longman, 1995.

Журавски 2005: Jurafsky, D., Martin, J. Speech and Language Processing: An introduction to natural language processing, computational linguistics and speech recognition. Prentice Hall Publishers, 2009.

Иванова 1974: Иванова, К. Начини на глаголното действие. София, БАН, 1974.

Кабакчиев 1992: Кабакчиев, К. Видът в английския език. С., Албо, 1992.

Кабакчиев 1998: Кабакчиев, К. Английска граматика. С., Пенсофт, 1998.

Коева 2005: Коева, С. Аргументна структура. Проблеми на простото и сложното изречение. С. Коева (съст). Сема РШ, София, 2005.

Коен 2008: Koehn, P., Callison-Burch, C. Statistical Machine Translation. *http://www.mt-archive.info/authors-K.htm*

Коен 2010: Koehn, P., Hoang, H. Improved translation with source syntax labels. In *http://www.mt-archive.info/authors-K.htm*

Ницолова 2008: Ницолова, Р. Българска граматика. Морфология. С., Университетско издателство „Св. Климент Охридски”, 2008.

Пенчев 1993: Пенчев, Й. Български синтаксис. Управление и свързване. Пловдив, Университетско издателство „Паисий Хилендарски”, 1993.

Пенчев 1998: Пенчев, Й. Синтаксис на съвременния български книжовен език. Пловдив, 1998.

Помагало по българска морфология 1976: Помагало по българска морфология. Глагол. П. Пашов, Р. Ницолова (съст.). София, Наука и изкуство, 1976.

Осенова 2007: Осенова, П., Симов, К. Формална граматика на българския език. С., Институт за паралелна обработка на информацията, БАН, 2007.

Уейвър 1955**:** Weaver, W. Translation. *http://www.mt-archive.info/Weaver-1949.pdf*