БЪЛГАРО-АНГЛИЙСКИ ПАРАЛЕЛЕН КОРПУС СЪС СЪОТНЕСЕНИ ГЛАГОЛНИ ФОРМИ[1]

Тодор Лазаров
Институт за български език, БАН

Статья представляет текущую работу по болгарско-английскому параллельному корпусу с выровненными формами глагола. Рассматривается процесс выбора ресурсов, предварительная обработка ресурсов, аннотации и программные приложения, используемые в этом процессе. Дается описание текущего состояния корпуса.

This article presents the current work on the Bulgarian-English parallel corpus with aligned verb forms. We present the process of resource selection, pre-processing of the resources, the annotation and the software applications used in the process. There is also a description of the current state of the corpus.

Ключови думи: глаголни форми, езиков корпус, анотация, съвременни езикови ресурси

Keywords: verb forms, language corpus, annotation, contemporary language resources

I. Предназначение и цели на корпуса

Настоящата работа по българо-английския паралелен корпус със съотнесени глаголни форми представлява продължение на работата по темата за граматичните паралели при превод на глаголните форми от български на английски. Както е посочвано в предишни изследвания по проблема, преводът на глаголните форми е труден процес дори за превод, извършван от хора. Това се дължи на факта, че българският и английският се различават както по начина, по който изразяват на семантично ниво отношенията между ориентационните моменти и динамичните признаци по темпоралната ос, така и по начина на формообразуване и инвентара на глаголните форми.

Както е известно, граматичната категория време в двата езика може да бъде разглеждана като хиперкатегория, обединяваща граматически-

[1] Тази статия е част от проект № 72-00-40-221 / 10.05.2017: „Българо-английски граматични паралели с оглед на машинния превод. Обогатяване на статистически модел за превод от български на английски с лингвистична информация“, осъществен с финансовата подкрепа на Програма за подпомагане на млади учени и докторанти на БАН – 2017.

те стойности на няколко подкатегории. Вече са посочвани причините (Lazarov/Лазаров 2017), които правят това схващане продуктивно за целите на изследването на граматическите паралели между български и английски, и преимуществата му за формален анализ на особеностите на глаголните системи. Споделяйки множество общи характеристики, хиперкатегориите в двата езика предоставят възможността за паралелен формален анализ на особеностите на формообразуването на глаголните форми и извличането на надеждни данни както за конструиране на езиков модел, така и за конструирането на модел за превод.

Трябва да отбележим, че концепцията за изследването във формален план на особеностите на формообразуването в български и английски (както отделно за двата езика, така и в съпоставителен план) не е нова, съществуват изследвания акцентиращи както върху по-общи проблеми – специално за български вж. Осенова и Симов (Osenova, Simov 2007), така и върху по-специфични. Известни са и опити за представянето на трансферни правила за превод между български и английски. При избора на подход при работата върху българо-английския паралелен корпус със съотнесени глаголни форми сме възприели концепциите за статистическа машинна обработка на естествения език. При статистически машинен превод се разчита основно на достатъчно представителни паралелни езикови корпуси с достатъчен обем, които да представят лингвистичните феномени в техните различни проявления. За създаването на статистически езиков модел за превод е необходимо да знаем каква е вероятността дадена езикова единица (дума, фраза) *e* в целевия език да бъде превод на езикова единица *f* от изходния – $p(e|f)$. За целта са необходими алгоритми, които да обработват информацията, и езикови ресурси, от които да се извлекат лингвистичните данни. И двата избрани от нас езика позволяват подобна работа с тях, тъй като са налични различни по вид свободно достъпни езикови ресурси и паралелни езикови корпуси, които „представляват надежден източник за наблюдение, анализ и изводи (подкрепени от обективни количествени и дистрибутивни данни) за […] автоматично извличане на езикови данни, езикови отношения и модели“ (Koeva/Коева2014: 49).

В контекста на изследванията по темата за граматическите сходства и различия между българската и английската глаголни системи предназначението на българо-английският паралелен корпус със съотнесени глаголни форми е да бъде практически приложим ресурс за създаването на статистически модел за превод на глаголните форми.

Целите на корпуса могат да бъдат разделени според предназначението му в две основни насоки – теоретични и практични. Теоретичните цели на корпуса са да даде обективни количествени и дистрибутивни езикови данни за целите на изследванията на особеностите на глаголните системи на български и английски и да представи изчерпателно (за обема

си) качествени и количествени езикови проявления за надежден анализ. Практическите цели на корпуса са да предостави достатъчно надеждни лингвистични данни за създаването на различни езикови приложения за компютърна обработка на естествения език като статистически модел за анотиране и статистически модел за превод на глаголните форми.

II. Подбор и оценка на включените ресурси

Стъпката, която предхожда създаването на самия корпус, е подборът и оценката на подходящи и надеждни езикови ресурси, които отговарят на целите на корпуса. Важно е да бъде направено уточнението, че българо-английският паралелен корпус със съотнесени глаголни форми инкорпорира в себе си различни видове езикови ресурси. Оценката на всеки един подходящ ресурс е направена според няколко критерия. Критериите за подбор на езиковите ресурси могат да бъдат разделени според няколко характеристики:

- според предварително дефинираните езикови особености на самия българо-английски паралелен корпус – според характеристиките, които са възприети за езиков корпус, от една страна, поставената цел пред корпуса е той да бъде представителен по отношение на езиковия феномен, който представя. От друга страна, спецификите на глаголната употреба предопределят небалансираността на текстовите типове, които да бъдат включени в корпуса, и екстралингвистичните характеристики на ресурсите;
- според имплицитните особености на наличните ресурси – подходящите ресурси трябва да представляват паралелни българо-английски сдвоени текстове на определено ниво – изречения или параграфи. Също така целите предполагат наличието на анотационен слой с морфологична информация – PoS тагове;
- според екстралингвистичните характеристики на ресурсите – в метаинформацията на ресурса трябва да бъде посочено кой е езикът на оригинала и езикът на преводния текст (посоката на превод в ресурса). Някои от включените ресурси не съдържат информация за морфологичните характеристики на думите, но все пак са включени (след предварителната им обработка) като ресурс за наблюдение и описване езиковите явления. Другата основна характеристика е ресурсът да представлява актуалната употреба на езика. Естеството на подбора предопределя и статута на авторските права на ресурсите – те трябва да бъдат освободени от авторски права или да бъдат предоставени такива за целите на работата.

В хода на подбора и оценката на езиковите ресурси някои от дефинираните критерии са модифицирани заради особеностите на самите ресур-

си и целите на корпуса. Ще опишем модификациите при представянето на съответните ресурси по-долу.

След като са определени критериите за езиковите ресурси, които ще бъдат включени в корпуса, изборът е ограничен до Българско-английския паралелен корпус със съотнесени изречения (БАПКСИ) (Koeva et al. 2012) и събиране на освободени от авторски права двуезични документи от интернет.

БАПКСИ (Tarpomanova, Dimitrova/Търпоманова, Димитрова 2014) е съставна част от Българско-английският паралелен корпус (БАПК), който от своя страна е част от Българския национален корпус (БНК). Обемът на БАПКСИ е около 367 000 думи, разпределени неравномерно между български и английски. Корпусът включва различни нива на едноезикова и многоезикова анотация. Ключовите характеристики на БАПКСИ заедно с БАПК, които го правят подходящ за включване в Българо-английския паралелен корпус със съотнесени глаголни форми, са:

– подробна морфологична едноезикова и многоезикова анотация, направена с помощта на Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове (Koeva, Genov 2011);
– многоезикова анотация, която включва съотнасяне на изреченията в двата езика (и на простите изречения в състава на сложното), като съотнасянето е проверено или направено от човек, а проверката и корекцията на автоматичното съотнасяне са извършени със специално разработена за целта програма.

Другите езикови ресурси, които са избрани за включване в корпуса, са извадки от текстове и техните преводи от интернет. Недостатъкът на тези текстове е, че макар да предоставят информация за посоката на превода, автора и жанра, в тях не се съдържа лингвистична анотация, което води до допълнителен етап на обработка и унифициране между езиковите ресурси.

Трябва да посочим, че и двата вида описани ресурси не отговарят на предварително зададените критерии за тях. От една страна, въпреки че БАПКСИ представлява надежден паралелен българо-английски корпус, съдържащ информация за морфологичните характеристики на думите, той не съдържа информация за източника и целевия език за всеки от съставящите го текстове. От друга страна, избраните документи от интернет съдържат информация за жанра, посоката на превод и други специфики на текста, но събирането им не е подчинено на строго определена таксономия, дефинирана върху текстовите типове и техните характеристики.

За целите на корпуса се налага унификация и всички предимства и недостатъци на описаните езикови ресурси са взети предвид при конструирането на структурата на българо-английския корпус със съотнесени

глаголни форми. Корпусът се състои от избрани части от двата вида ресурси, като за минимална единица, представляваща всеки отделен текст, е възприет параграфът. Инкорпорираните ресурси се разпростират от специално подбрани параграфи от отделни текстове до цели кохерентни текстове от различни източници – новини, художествена литература, филмови субтитри и други произведения. Метаинформацията на всеки отделен документ в корпуса включва данни за източника (име на файла или URL), изходния и целевия език, датата на събиране и датата на вписване в корпуса, наличието на PoS тагове и текстовия тип. В крайния си вид българо-английският корпус със съотнесени глаголни форми се състои от около 84% текстове от БАПКСИ, около 15% текстове от интернет и под 1% от конструирани изречения, извлечени от примери от учебник по граматика за български език.

III. Използвани софтуерни приложения и инвентар от PoS тагове

Както беше посочено по-горе, по време на първичната фаза на подбор и оценка на подходящите езикови ресурси за включване в корпуса възникна проблемът за неговата анотационна структура и унификация на текстовите типове. Използваните езикови ресурси нямат еднакво разпределение на качеството и количеството на анотационните слоеве (или такива липсват), следователно е необходимо структурата на общата корпусно анотация да бъде конструирана на базата на целевите експлицитни характеристики.

Лингвистичните данни в корпуса имат два слоя анотация. Първият слой е слоят на морфологичните характеристики на думите и се състои от PoS-таговете на езиковите единици. Възприета е анотационната структура на БАПКСИ поради факта, че текстовете от него представляват по-голямата част от българо-английския паралелен корпус със съотнесени глаголни форми. Текстовете в БАПКСИ са анотирани с помощта на Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове. Тя включва програми за обработка – токънизатор и разделител на изречения, основани на регулярни изрази, SVM тагер по части на речта, лематизатор на основата на речник, чънкер, програма за семантична анотация от Уърднет, които са адаптирани да работят в единна система, при което се осигурява тяхната свързаност, ефективност и висока точност.

За езиковите ресурси, при които няма налична първична анотация, е избран свободно достъпния инструмент за анотиране TreeTagger. TreeTagger е инструмент за анотиране на текст с информация за морфологичните характеристики на думите. Разработен е от Хелмут Шмид (Schmid 1995) в Института по компютърна лингвистика към университета

в Щутгарт. Поради факта, че TreeTagger е приспособим към други езици, ако има наличен лексикон и ръчно анотирани текстове за обучение, той е избран да бъде използван при работата по създаването на корпуса. Анотацията, използвана от TreeTagger, е взаимствана от BulTreeBank за български (Simov et al. 2004) и от Penn Treebank за английски (Santorini 1991). За да се осигури съвместимостта на анотацията между текстовете от БАПКСИ, които вече притежават анотационен слой, и текстовете, събрани от интернет и тагирани с TreeTagger, след процеса на анотация на текстовете с TreeTagger е приложено конвертиране на таговете с помощта на недетерминиран краен автомат, разработен за целта.

В хода на работата събраните езикови ресурси с унифицирания първи слой анотация са разделени в 22 работни файла, съдържащи неравномерно разпределени сдвоени текстове и морфологичните характеристики на думите в тях, а където липсва сдвояване между текстовете, събрани от интернет, са сдвоени на ниво изречение. Всяка двойка изречения на български и английски получава пореден номер в рамките на работните файлове. След процеса на анотиране, конвертиране и сдвояване, данните са проверени ръчно и са направени корекции, където е необходимо. Анотираните работни файлове са разделени за двата език и към тях е добавена метаинформация. Всеки от работните файлове получава идентификационен номер и се конвертира в триколонен TSV (tab separated values) файл. Всеки ред от файла съдържа по една дума/токен. Първата колона от всеки ред от файла съдържа думата/токена, втората колона представя лемата на думата, а третата колона е съответният PoS-таг. Празният ред сигнализира изреченска граница.

Вторичният анотационен слой съдържа информация за глаголните форми в двата работни езика. При работата върху него е спазен принципът на надграждането между равнищата (Koeva et al. 2010: 3681) – „различните равнища на анотация приписват нова информация към езиковите единици, без да променят анотациите от предходните равнища“. За втория слой на анотация е използван свободно достъпния инструмент за работа с езикови ресурси WebAnno (Yimam et al. 2014). WebAnno е универсален инструмент за анотиране за различни равнища на анотация, включващи различни слоеве от морфологични, синтактични и семантични характеристики. Освен предварително наличните инвентари от тагове, могат да бъдат дефинирани индивидуални равнища от езикови характеристики, което позволява WebAnno да се използва и за структурно недетерминирани цели. Приложението поддържа различни режими на анотация, включително режим на ръчно приписване на характеристики, корекция и самообучение, въз основа на предварително дефинирана таксономия и структурни правила, което позволява приписване на лингвистична информация на базата на вероятностен модел. WebAnno позволява работа с множество файлови

формати, но за целите на корпуса е избран форматът CONLL. WebAnno използва преработена версия на файловия формат CONLL-X. Анотациите характеристики се кодират в нормализирани текстови файлове (UTF-8, с използване на LF символ за прекъсване на реда, включващ LF символ в края на файла/текста) с три типа редове: ред, съдържащи анотацията на думата/токена с 5 колони разделени с табулация; празни редове, маркиращи края на изреченията; и редове с коментари. Изреченията се състоят от един или повече редове с думи, а редовете с думи съдържат следните колони:

- пореден номер на изречението;
- думата/токена;
- PoS таг от първия анотационен слой;
- таг на съставната глаголна форма, ако токенът е част от такава;
- структурно отношение между двата слоя анотация – представлява I/O структура на частите (chunk I/O structure), съдържа атрибутите B (begin/начало), I (in/част от) и O (out/не е част от).

Вторият анотационен слой е добавен ръчно чрез инструмента WebAnno с частично използване на функцията на програмата за приписване на лингвистична информация на базата на вероятностен модел и последваща корекция. Стойностите на езиковите маркери от този слой представляват краен списък от по-малко на брой възможни маркери, отколкото първия слой на анотацията. Те се приписват върху първия PoS анотационен слой. Благодарение на факта, че инструментът WebAnno разглежда глаголните форми (в този случай) като компоненти, съставени от определен брой структурни части, означава, че един от маркерите от втория слой може да бъде приписан на повече от една единица от първия слой.

Инвентарът на маркерите от втория анотационен слой, съдържащ информация за глаголните форми в български и английски, е разработен специално за целите на корпуса и се състои 26 стойности, които представят информация за морфологичната категория време, разглеждана според класическите наименования на членовете ѝ. Към стойностите на втория анотационен слой не е включена информация за лице и число на формите, тъй като тя се унаследява/формира от стойностите на първичния слой. Представени схематично, маркерите от втория анотационен слой изглеждат така:

- За български:
  - Vpraesens – глаголна форма в сегашно време‘
  - Vaor – глаголна форма в минало свършено време;
  - Vimperf – глаголна форма в минало несвършено време;
  - Vfutur – глаголна форма в бъдеще време;
  - Vperfect – глаголна форма в минало неопределено време;
  - Vplusqperf – глаголна форма в минало предварително време;

- Vfutexact – глаголна форма в бъдеще предварително време;
- Vfutexpraet – глаголна форма в бъдеще предварително време в миналото;
- Vfutpraet – глаголна форма в бъдеще време в миналото.

• За английски:
- Vprs – глаголна форма в сегашно просто време;
- Vps – глаголна форма в минало просто време;
- Vfs – глаголна форма в бъдеще просто време;
- Vprp – глаголна форма в сегашно перфектно време;
- Vpp – глаголна форма в минало перфектно време;
- Vfp – глаголна форма в бъдеще перфектно време;
- Vprc – глаголна форма в сегашно продължително време;
- Vpc – глаголна форма в минало продължително време;
- Vfc – глаголна форма в бъдеще продължително време;
- Vprpc – глаголна форма в сегашно перфектно продължително време;
- Vppc – глаголна форма в минало перфектно продължително време;
- Vfpc – глаголна форма в бъдеще перфектно продължително време;
- Vfsp – глаголна форма в бъдеще просто време в миналото;
- Vfpp – глаголна форма в бъдеще перфектно време в миналото;
- Vfcp – глаголна форма в бъдеще продължително време в миналото;
- Vfpcp – глаголна форма в бъдеще перфектно продължително време в миналото‘
- Vinf – инфинитив.

IV. Съвременно състояние на българо-английския паралелен корпус със съотнесени глаголни форми

В текущия си вид българо-английският паралелен корпус със съотнесени глаголни форми представлява неголям, небалансиран, илюстративен (за поставената си цел), двуезичен паралелен корпус с общ обем от 5070 анотирани и съотнесени глаголни форми и техните преводни съответствия. Българските текстове съставляват 45 843 думи (48,12% от общия обем), а английските – 49 432 думи (51,88% от общия обем). Първоначално поставената цел за детерминирана посока на превод – корпусът да представлява изцяло български текстове с техните английски преводи – е изоставена по време на подбора на текстовете, поради екстралингвистичните особености на наличните текстове — недостатъчно свободно достъпни и представителни оригинални български текстове, които да обхващат разглеждания езиков феномен във всичките му проявления. Поради тази причина в корпуса са включени както преводи от английски на български, така и преводи на български и английски от текстове на

трети език. Таксономията на текстовите типове няма единна структура, а е основана на периодичната проверка за езиково насищане на корпуса – при внасянето на поредния обработен текст към него, се проверява дали съотношението на изследваните езикови единици (глаголните форми) ще се промени значително, или броят им ще остане сравнително постоянен.

V. Бъдещи цели и работа

Бъдещата работа по българо-английския паралелен корпус със съотнесени глаголни форми е продиктувана от неговите първоначални цели и причината за конструирането му. Поставените задачи могат да бъдат разделени в две направления – теоретично–практически и приложими. Основните теоретично-практически задачи са две: първата, да бъде конструиран статистически преводен модел за глаголните форми между български и английски, базиран на информацията, извлечена от корпуса – тази задача ще включва моделирането на фразово-структурни, категориални и трансферни правила, създаването на езиков модел и преводен модел и верификацията им; и втората, да бъде конструиран модел за анотиране на съставните глаголни форми за двата езика, който ще допринесе за бъдещото обогатяване на корпуса и осъществяването на дефинираните последващи цели.

Основната приложима задача е качването на корпуса в интернет и осигуряването на свободен достъп до него и създаването или модифицирането на съществуващ инструмент за търсене в корпуса и извличане на информация. Осъществяването на свободен достъп до корпуса ще предостави възможност за бъдещи изследвания върху езиковия материал и информацията, представени в него, които са извън обхвата на първоначално дефинираните за настоящето проучване.

REFERENCES/ БИБЛИОГРАФИЯ

Gerdzhikov, G. 2000. Kategoriyata vreme kato hiperkategoriya – Balgarski ezik i literatura, 1. [Герджиков, Г. 2000. Категорията време като хиперкатегория – *Български език и литература,* 1.] http://liternet.bg/publish/ggerdzhikov/hyper.htm [10/12/2018]

Koeva, Sv. 2014. Balgarskiyat natsionalen korpus v konteksta na svetovnata teoriya i praktika – V: Ezikovi resursi i tehnologii za balgarski ezik. Sofiya: AI „Prof. Marin Drinov“, 29–53. [Коева, Св. 2014. Българският национален корпус в контекста на световната теория и практика – В: *Езикови ресурси и технологии за български език.* София: АИ „Проф. Марин Дринов“, 29–53.]

Koeva, S. & A. Genov 2011. *Bulgarian language processing chain*. – Proceedings of Integration of multilingual resources and tools in Web applications. Workshop in conjunction with GSCL 2011, 26 September 2011, University of Hamburg.

Koeva Sv., D. Blagoeva, S. Kolkovska 2010. “Bulgarian National Corpus Project”. – In: *Proceedings of LREC-2010,* Valletta: ELRA, 3678–3684.

Koeva, Sv., B. Rizov, Ek. Tarpomanova, Ts. Dimitrova, R. Dekova, Iv. Stoyanova, Sv. Leseva, Hr. Kukova, and A. Genov 2012. Application of Clause Alignment for Statistical Machine Translation. – In: *Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Jeju, Republic of Korea, 12 July 2012, The Association for Computational Linguistics: ACL 2012 / SIGMT / SIGLEX Workshop, 2012, 102–110.

Lazarov, T. 2017. Funktsionalni gramatichni paraleli pri prevoda na glagolnite formi ot balgarski na angliyski – V: Dokladi ot Mezhdunarodnata yubileyna konferentsiya na Instituta za balgarski ezik „Prof. Lyubomir Andreychin“ (Sofiya, 15–16 may 2017 godina), 321–327. [Лазаров, Т. 2017. Функционални граматични паралели при превода на глаголните форми от български на английски – В: *Доклади от Международната юбилейна конференция на Института за български език „Проф. Любомир Андрейчин“ (София, 15–16 май 2017 година),* 321–327.].

Osenova, P., Simov, K. 2007. Formalna gramatika na balgarskiya ezik. 1-vo izdanie, Sofiya: Institut za paralelna obrabotka na informatsiyata, BAN. [Осенова, П., Симов, К. 2007. *Формална граматика на българския език.* 1-во издание, София: Институт за паралелна обработка на информацията, БАН.].

Santorini 1991: *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. Penn Treebank Project Technical Report № 3. 1991 University of Pennsylvania [https://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports] [10/12/2018].

Schmid H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. – In: Armstrong S., Church K., Isabelle P., Manzi S., Tzoukermann E., Yarowsky D. (eds) *Natural Language Processing Using Very Large Corpora. Text, Speech and Language Technology*, vol 11. Dordrecht: Springer, 13–22.

Simov, K., P. Osenova, M. Slavcheva 2004. *BTB-TR03: BulTreeBank Morphosyntactic Tagset*. BulTreeBank Project Technical Report № 03. 2004 [http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR03.pdf] [10/12/2018].

Tarpomanova, E., Ts. Dimitrova 2014. Balgarsko-angliyski paralelel korpus sas saotneseni (prosti) izrecheniya. – Ezikovi tehnologii i resursi za balgarski ezik. Sofiya: Akademichno izdatelstvo «Marin Drinov», 105–126. [Търпоманова, Е., Цв. Димитрова 2014. *Българско-английски паралелел корпус със съотнесени (прости) изречения.* – Езикови технологии и ресурси за български език. София: Академично издателство «Марин Дринов», 2014, 105–126.].

Yimam, S., Castilho, E., & Gurevych, I. 2014. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. – In: *Proceedings of ACL-2014, demo session*. Baltimore, MD, USA.