Български език 63 (2016), 4, 95-101

ТОДОР ЛАЗАРОВ

АНАЛИЗ НА РЕСУРСИТЕ ЗА СТАТИСТИЧЕСКИ МОДЕЛ ЗА ПРЕВОД НА ГЛАГОЛНИТЕ ФОРМИ ОТ БЪЛГАРСКИ НА АНГЛИЙСКИ

TODOR LAZAOV

ANALYSIS OF THE RESOURCES FOR STATISTICAL TRANSLATION MODEL OF THE VERB FORMS FROM BULGARIAN TO ENGLISH

(Abstract)

The text briefly presents the idea of creating a statistical language model of the verb forms for Bulgarian and English. We present the main difficulties caused by the grammatical features of both languages which hinder the formal description of the verb forms for the purposes of developing transfer-based rules for machine translation. In the article are presented the possible resources for constructing the model with their quantitative and qualitative characteristics. In brief we present the disadvantages of the resources and the difficulties that come up with the future work on the statistical model.

Keywords: statistical language model, verb system, verb forms, tenses, transfer rules, machine translation, formal description of natural language

Идеята за създаване на статистически езиков модел, който да бъде помощно средство към трансферните правила за превод на глаголните форми от български на английски, е породена от проблемите и трудностите, до които води самостоятелното използване на трансферните правила при автоматичен превод на глаголните форми от български на английски.

Вече сме разглеждали особеностите на глаголните системи на българския и на английския и начините за изразяване на категорията време в двата езика (Лазаров 2015). Без да навлизаме в подробности, тук ще отбележим, че основните разлики между двата езика, които също така представляват и трудност при автоматичен превод, са:

– наличието на морфологичната категория род[1] в български език, която отсъства в английски, от което следва, че граматикализираната информация за род на глаголното лице се губи при превод на английски;

– наличието на лексикално-граматичната категория *вид на глагола* (Куцаров 2007: 551) в български език, която изразява завършеността/незавършеността на действието спрямо границата на самия процес (вътрешна (не)завършеност, цялостност), и наличието на морфологичната категория *вид на действието* в английски, която изразява завършеността на действието спрямо друг ориентационен момент по темпоралната ос (външна незавършеност, продължителност). Разликата между значението на двете категории, макар често пъти да е трудно доловима, не може да бъде пренебрегната и тя води след себе си обстоятелството, че някои глаголни форми в български при превод на английски имат две съответствия;

– наличието на категориите *умозаключително наклонение*, *преизказно наклонение, вид на изказването*, хиперкатегорията *характеристика на предаваната информация от говорещия* и т.н. (Ницолова 2008: 248), за които българският глагол има силно развита парадигма от граматикализирани значения, които изразяват в пълна степен различните отношения, докато в английски специфичната семантика на тези форми е невъзможно да се предаде без допълнителни лексикални уточнения.

Изброените качествени и количествени различия между категориите, които глаголите в двата езика притежават, водят до трудности при формалното описание и съставянето на напълно функционални и изчерпателни трансферни правила за превод. Към недостатъците на машинния превод чрез трансферни правила в този случай трябва да добавим и несиметричността между възможните реализации на граматикализирани значения на глаголите в двата езика. Броят на глаголните форми в български е в пъти по-голям от английски, което означава, че при посока на превода от български на английски (част от) граматикализираната информация може да се загуби, а при превод от английски на български информацията, която е нужна за конструирането на съответната форма на български, не е налична в рамките на глаголната словоформа на английски.

Независимо от множеството си недостатъци трансферните правила позволяват да представим избраните от нас езици в тяхното (граматично) съотношение и илюстрират подробно лексикалния и семантичния трансфер, който се осъществява при превода. Въпреки това при превод от български на английски особеностите на двата езика не позволяват да приписваме трансферните правила категорично за всеки отделен случай на превод, тъй като възможните форми в български са в пъти повече отколкото в английски, а граматичните значения на българските глаголни форми позволяват превеждането им на английски с повече от една глаголна форма.

Тъй като по същество трансферните правила представляват вид граматика, която отразява зависимости между двата езика, е възможно такава граматика да бъде допълнена с информация за актуалната употреба на глаголните форми – да бъде създаден статистически езиков модел за превод на глаголните форми.

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 дума *e* в целевия език да бъде превод на дума *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 достатъчно представителни и различни по вид 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“ (Коева 2014: 49).

Предварителният критерий, по който сме избрали езикови корпуси за целите на нашето изследване, е наличието на качествена морфологична анотация, тъй като тя може да представлява основен етап в обработката на лингвистичните ресурси за целите на по-късното извличане на статистически данни и конструирането на езиков модел. Необходимо е да направим уточнението, че трите основни характеристики за качествен езиков корпус – илюстративност, представителност и балансираност (Коева 2014: 30), са присъщи за избраните от нас езикови ресурси. В този текст ще разгледаме Българския национален корпус (заедно с българските паралелни корпуси), Българския корпус, анотиран с части на речта и граматични характеристики, и Британския национален корпус. Ще представим накратко и Корпуса с документи на Европейския парламент.

Българският национален корпус (БНК) е най-големият представителен корпус за български език. БНК се състои от ядрен корпус, включващ текстове само на български език, и 47 паралелни корпуса. Общият обем на корпуса е около 5,4 милиарда думи[2]. Различните нива на анотация в БНК – подробни метаданни, едноезикова и частична многоезикова анотация, позволяват работата с него за различни цели. С оглед на нашето проучване изборът на БНК като източник на лингвистична информация е продиктуван от следните негови характеристики:

– подробна морфологична анотация на българските текстове, направена с помощта на Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове[3], и последващо ръчно разрешаване на случаите на многозначност (за част от корпуса);

– наличието на паралелни корпуси на 47 езика (в това число и английски), което позволява използването му за конструирането на статистически модели за превод.

Част от 47-те паралелни корпуса на БНК е и Българско-английският паралелен корпус (БАПК). С големината си от около 260 милиона думи за език той представлява и най-големият паралелен корпус в състава на БНК. Подобно на останалите паралелни корпуси и БАПК се състои от текстове, които имат българско съответствие – като българският текст може да бъде оригинал или превод, както и превод от трети език. Съставна част от БАПК е Българско-английският паралелен корпус със съотнесени изречения (и прости изречения в състава на сложното) (БАПКСИ). Неговият обем е от около 367 000 думи, разпределени неравномерно между български и английски. Корпусът включва различни нива на едноезикова и многоезикова анотация. Ключовите характеристики на БАПК заедно с БАПКСИ, които го правят подходящ за целите на създаването на статистически езиков модел за превод на глаголни форми, са:

– подробна морфологична едноезикова и многоезикова анотация, направена с помощта на Българската многокомпонента система за първична обработка и лингвистична анотация на текстове;

– 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 програма[4].

БАПК и БАПКСИ са особено ценни езикови ресурси за създаването на статистически лингвистичен модел, тъй като лингвистичните данни са снабдени с достатъчно метаданни за осъществяване на по-нататъшна статистическа обработка.

Българският корпус, анотиран с части на речта и граматични характеристики (БулПосКор), е извлечен от Българския Браун корпус (ББК). Неговата големина е около 175 000 думи, а структурата му се базира на тази на ББК. БулПосКор представлява ценен ресурс при конструирането на лингвистичен модел на глаголите за български език, тъй като съдържа информация за граматичните характеристики на думите. Морфологичното анотиране на БулПосКор се състои от първичен етап на автоматично приписване на граматични характеристики от Българския граматичен речник (Коева 1998) и последващ етап на ръчно разрешаване на случаите на многозначност.

Британският национален корпус, подобно на Българския национален корпус, представлява най-големият представителен корпус за британския английски език. Британският национален корпус е едноезиков корпус с

големина от около 100 милиона думи. Британският национален корпус съдържа едноезикова анотация на различни нива. Морфологичната анотация е направена с помощта на системата CLAWS4 и последващо ръчно снемане на многозначност. За целите на съставянето на статистически езиков модел на глаголите за английски Британският национален корпус представлява ценен ресурс именно със системата си за морфологична анотация, тъй като граматичната информация е приписана с прецизност до 98% (Leech 1994).

Наличието на подробна морфологична анотация на езиковия материал в корпусите е предпоставка за по-нататъшната им обработка за различни цели. Една от основните характеристики на избраните от нас корпуси е именно тази – подробна едноезикова анотация за части на речта и граматични характеристики, на която да бъдат базирани статистическите езикови модели. Разбира се, при създаване на статистически модел от съществено значение са количеството данни, на които този модел ще бъде базиран. Различните видове езикови корпуси съдържат в себе си метаданни от различен характер. Някои корпуси не съдържат информация за морфологичните характеристики на думите, но въпреки това те представляват ценен ресурс за наблюдение и описание на лингвистичните явления и тяхната верификация.

Последният от корпусите, които ще разгледаме, представлява именно корпус, в който отсъства анотация за частите на речта. Корпусът с документи на Европейския парламент (КДЕП) представлява набор от текстове – официалните документи на Европейския парламент от 1996 година насам. Съдържа текстове на всички официални за съюза езици, а обемът му е от около 10 до 50 милиона думи за език. Част от корпуса на КДЕП е и българо-английският паралелен корпус, който съдържа около 10 милиона думи. Въпреки липсата на морфологична анотация, целта на създаването на корпуса е била именно статистическият машинен превод, затова корпусът съдържа изречения с идентификатор, които са автоматично съотнесени с помощта на алгоритъма Чърч–Гейл (Koehn 2005).

Както се вижда, изборът на ресурси за създаване на статистически езиков модел за превод на глаголните форми между български и английски е продиктуван най-вече от наличието на подробна морфологична анотация в описаните корпуси, тъй като нашата цел е да проследим какво се случва с граматикализираната информация при превод. Почти всички изброени ресурси в този текст притежават автоматична анотация за части на речта и граматични характеристики. При ресурса КДЕП липсва подобна анотация, но изборът му е продиктуван от неговия обем и от факта, че съдържа съотнесени изречения, което спомага за по-нататъшна работа с него. Разбира се, нужно е да добавим, че представените от нас корпуси имат различни нива и различни конвенции за анотация, което възпрепятства директната

работа с тях. За целите на конструирането на езиковите и преводните статистически модели ще се наложи унифициране на възприетите анотационни модели на различните корпуси. Също така от представените тук корпуси единствено Българско-английският паралелен корпус със съотнесени изречения (БАПКСИ) съдържа съотнесени езикови единици на равнище прости изречения в състава на сложното изречение, което улеснява до известна степен работата с него за целите на нашето изследване.

Представените тук корпуси са подбрани с оглед на отделни техни качества, които в различна степен оптимизират за работата върху тях. При по-нататъшната работа с изброените корпуси ще бъде необходимо съставянето на единен анотационен модел с оглед на глаголните форми (и на базата на възприетите анотационни модели в различните корпуси) и съотнасянето им (с оглед на семантиката им).

БЕЛЕЖКИ

[1] Категорията род не е еднотипна при съществителните имена и при останалите изменяеми части на речта. Съществителните имена нямат родови словоформи и следователно при тях не може да се говори за морфологична категория. В този смисъл общата морфологична категория род обхваща […] глаголните форми, съдържащи причастия (Куцаров 2007: 184–185).

[2] http://dcl.bas.bg/corpora_bg.html#BNK

[3] http://dcl.bas.bg/DCLservices-bg.html

[4] *http://dcl.bas.bg/clauseAlignedCorpus.html*

ЛИТЕРАТУРА

Коева 1998: *Коева, Св*. Граматичен речник на българския език. Описание на концепцията за организацията на лингвистичните данни. – БЕ, № 6, с. 49–58.

Коева 2014: *Коева, Св.* Българският национален корпус в контекста на световната теория и практика. – В: Езикови ресурси и технологии за български език. София, АИ „Проф. Марин Дринов“, с. 29–53.

Куцаров 2007: *Куцаров, Ив.* Теоретична граматика на българския език. Морфология. Пловдив, УИ „Паисий Хилендарски“.

Лазаров 2015: *Лазаров, Т.* Особености на глаголните системи и начините за изразяване на времето в български и английски. Семантичен трансфер при превод на глаголните форми от български на английски. – В: Littera et lingua – електронно списание за хуманитаристика. Софийски университет

http://slav.uni-sofia.bg/naum/lilijournal/2015/12/1-2/tlazarov

Ницолова 2008: *Ницолова, Р.* Българска граматика. Морфология. София, УИ „Св. Климент Охридски“.

KOEHN, P 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit.

http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf

LEECH 1994: *Leech, G., R. Garside, and M. Bryant.* CLAWS4: The tagging of the British National Corpus. – In: Proceedings of the 15th International Conference on Computational Linguistics, p. 622–628.

✉ *Тодор Лазаров*
Секция по компютърна лингвистика
Институт за български език „Проф. Л. Андрейчин“ при БАН
бул. „Шипченски проход“ 52, бл. 17, 1113 София, България
todorlazarov91@abv.bg

✉ *Todor Lazaov*
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
52 Shipchenski prohod blvd., bl. 17, 1113 Sofia, Bulgaria
todorlazarov91@abv.bg