ПРОГЛАС

Издание на Филологическия факултет
при Великотърновския университет „Св. св. Кирил и Методий“

кн. 2, 2017 (год. XXVI), ISSN 0861-7902

ТЕМА НА БРОЯ: ЛИНГВИСТИКА И ИНТЕРНЕТ

Петя Осенова[1]

БЪЛГАРСКИТЕ ЕЛЕКТРОННИ ГРАМАТИЧЕСКИ РЕСУРСИ В ПАРАМЕТРИТЕ НА УНИВЕРСАЛНОСТТА

Petya Osenova

BULGARIAN ELECTRONIC GRAMMAR RESOURCES IN THE PARAMETERS OF UNIVERSALITY

The paper introduces two approaches to the universal presentation of Bulgarian language resources. The first one is based on the creation of precise resource grammars for various languages in accordance with a particular linguistic theory. The second one is the conversion of an existing language resource in accordance with a new annotation scheme. The paper discusses the two approaches within the context of Bulgarian language, pointing out their strengths and weaknesses.
Keywords: *language resources, precision grammars, universal scheme, Bulgarian, linguistic modeling, dependency theory*

Статията описва два подхода към универсалното представяне на езикови ресурси за българския език. Първият се основава на създаването на прецизни ресурсни граматики за различните езици спрямо определена лингвистична теория. Вторият представлява трансформирането на вече съществуващ езиков ресурс спрямо нова анотационна схема. Двата подхода се разглеждат в контекста на българския език, като се посочват техните силни и слаби страни.
Ключови думи: *езикови ресурси, прецизни граматики, универсална схема, български език, езиково моделиране, дependентна теория*

Увод

От доста време насам се търсят начини за обективно сравнение и сравнимост между различните езици: както от гледна точка на езиковата теория, така и от гледна точка на електронните езикови ресурси. Обръщам внимание на факта, разбира се, че двете гледни точки не само не си противоречат, но и често взаимно се подпомагат.

В тази статия бих искала да споделя своя опит от участието си в две начинания, които целят изграждането на обща езикова парадигма при анализа на различните езици. Едното е изграждането на ресурсна граматика на даден език, а другото – конвертиране на съществуващи синтактично анотирани ресурси (т.нар. *трибанки* от англ. *treebanks*) в обща езикова и репрезентационна рамка.

Изграждането на ресурсна граматика е трудоемка задача, тъй като изисква кодиране на различни езикови явления в тяхното взаимодействие, както и кодирането на информация от различни езикови равнища. Освен това подобни граматики обикновено разчитат на наличието на богати речници и формални описания за даден език. Конвертирането на един съществуващ езиков ресурс, който е следвал определени правила, спрямо друга схема също се оказва нелека задача. Смяната на модела може да

[1] **Петя Осенова** (Petya Osenova) – проф. д-р в кат. „Български език“, Факултет по славянски филологии на СУ „Ñâ. Êëèì åí òÎ õðèñòîôúð“, petyaosenova@slav.uni-sofia.bg

доведе до нарушаване на системността на анотациите, т.е. на кодираното лингвистично знание, или до загубата на ценна лингвистична информация.

От друга страна обаче, изглежда няма по-добър начин за постигане на сравнимост между езиците освен опитите да се приложи обща рамка към тях. Ясно е, че общата рамка е само някакво приближение до универсалността. То може да е по-голямо, когато се работи с по-общи характеристики, и по-малко, когато се отчита по-голяма детайлност на специфичната езикова действителност. Но единствено по този начин параметрите на общо споделените езикови характеристики стават видими и наблюдаеми. Това се отнася и до различията, които възникват поради особеностите на самия език, лингвистичната традиция при този език или несъвършенството на самата обща схема. Ценното в случая е, че колкото повече езици се включват в подобни начинания, толкова повече общата схема (която често бива наричана 'универсална' въпреки явните ограничения по отношение на постигане на универсалност) се коригира и отчита разнообразието.

В статията се представят две инициативи (Осенова 2016), свързани с поставянето на българския език в типологически универсален контекст. Едната е изграждането на българска ресурсна граматика (Osenova 2011), а другата е конвертирането на синтактично анализирания ресурс Бултрибанк (Осенова и Симов 2007) спрямо схемата на универсалните зависимости[2] (Osenova and Simov 2015).

1. Ресурсната граматика за българския език[3]

Ресурсната граматика за българския език беше създадена в своя първи вариант през 2010 г. благодарение на Фулбрайтовата стипендия, чрез която имах възможността да бъда в Станфордския университет за 5 месеца. Тази граматика е част от инициативата DELPH-IN[4] за създаване на прецизни лингвистични граматики за различните езици в рамките на Опорната фразова граматика (ОФГ)[5]. Имплементацията се извършва в системата с отворен код LKB (Linguistic Knowledge Builder)[6]. Освен това се предлага една обща архитектура за писане на граматики, която се нарича Граматическа матрица (Grammar Matrix[7]). Тази граматическа матрица задава основните параметри, свързани с кодирането на различни явления, като например субкатегоризацията, построяването на видовете фрази, координацията и др. Казано по друг начин, матрицата задава скелета на граматиката чрез определяне на йерархията от типове и характеристики, както и посоките на наследяване на информацията. Целта ѝ е, от една страна, да създаде основа за типологическо сравнение между различни езици, а от друга, да ускори процеса на създаване на граматики за нови езици. На практика обаче не се получава точно така. Първо, защото матрицата трябва да се променя и развива заедно с прибавянето на всеки нов език, за да отрази неговите особености. Второ, защото в граматиката често се кодират решения, които са извън матрицата. Това се налага поради съществуването на езикови специфики, които не са били предвидени в матрицата, или поради различните предпочитания на авторите по отношение на модела за кодиране на езиковия анализ. Например в португалската граматика колегите са избрали да работят с фрази от типа *опора-функтор* (според разработките на Van Eynde (2006)) вместо с фрази от типа *опора-спецификатор* или *опора-адюнкт*. Трето, независимо от подпомагащата роля на матрицата, изработването на граматика с добро покритие за даден език остава доста трудоемка работа. Така например, английската граматика (ERG[8]), която в момента е с най-добро покритие, се разработва от началото на 90-те години на XX век. Граматиките за португалския, испанския и норвежкия език се разработват повече от 10 години. Четвърто, трябва да се има предвид ролята на различните

[2] Схемата на универсалните зависимости следва идеологията на Граматиките на зависимостите, или още – Депендентните граматики. В този тип граматики се отчитат зависимостите между думи, а не между фрази, както е положението в конституентните граматики.

[3] Повече информация за граматиката може да се види на следния адрес: http://www.bultreebank.org/BURGER/index.html

[4] http://www.delph-in.net/wiki/index.php/Home

[5] Head-driven phrase structure grammar (HPSG)

[6] http://moin.delph-in.net/LkbTop

[7] http://www.delph-in.net/matrix/

[8] http://www.delph-in.net/erg/

специализирани области. Ако се напише и тества добра граматика в една област (например медии), то в друга област тя няма да анализира добре текстовете (например медицина или финанси).

Тъй като ОФГ е теория, която организира езиковите явления на едно равнище (т.нар monostratal theory), граматическата матрица предлага и семантично моделиране на езика. Тя въвежда референти и събития, както и семантични релации, използвайки инструментите на т.нар. Минимална рекурсивна семантика[9].

При локализирането на граматическата матрица за българския език основното предизвикателство беше да се намери балансът между голямото количество граматикализирани явления и представянето им на семантично равнище. Така например, проблем възникна при семантичното кодиране на категорията *определеност/неопределеност* при имената и на категорията *вид* при глагола. В първия случай предизвикателството е адекватното кодиране на факта, че в именната фраза от типа на прилагателно и съществително (напр. *високата маса*) морфосинтактичната информация за определителния член на нивото на фразата се наследява от прилагателното, което е събитие, а семантичната информация за определеността – от опорното съществително, което е референт. Неслучайно определителният член често се разглежда като фразов афикс, а не като част от думата. Във втория случай проблем възниква при необходимостта от определяне на ‘основната форма’ измежду двата вида на глагола – свършен и несвършен на семантично равнище. Разбира се, това е проблем само ако се следва идеята, че двата вида се разглеждат като форми на един и същи глагол.

От една страна, е добре да има общо тестово множество от изречения за различните езици, които да демонстрират различни явления. От друга страна, подобни изречения не изчерпват всички особености на различните езици, защото първоначално се създават спрямо един определен език (най-често английския). Така е и в този случай. Но те все пак помагат за постигане на начално състояние на сравнимост, когато се преведат на другите езици. След това, разбира се, могат да се добавят и още изречения със специфични езикови явления към тестовото множество.

От съществено значение е също каква част от лингвистичното знание се кодира в самата автоматична граматика и каква част – извън нея. Това означава, че в първия случай създателят на граматиката трябва да моделира всички равнища в тази граматика – сегментиране, морфологично анотиране, синтаксис, семантика и др. Във втория случай част от обработките (например сегментация и морфология) могат да се направят чрез компоненти извън граматиката, като резултатът се подава като вход към граматиката преди синтактичния анализ. При кодирането на цялото лингвистично знание вътре в граматиката възниква въпросът как да се разпредели това знание между различните модули. Например каква част да се представи като типове в йерархията от езикови явления и каква - като декларативни правила. В много от случаите моделирането може да се извърши по поне два начина, като всеки от тях има своите силни и слаби страни. Тук вече думата има разработчикът на граматиката (т.нар. grammar writer).

Заради типологическите особености на отделни групи езици, каквато е групата на славянските езици, в Avgustinova and Zhang (2009) се представя идеята за базисна славянска граматика. Нейната функция е да детайлизира общата граматическата матрица, като предоставя повече информация за явленията, специфични за славянските езици.

Българската ресурсна граматика в настоящия си вариант следва модела на общата граматическа матрица. Тя започна развитието си с покриване на явленията в 178 изречения, които се получиха при превода на 100 изречения от многоезиковото множество изречения към граматическата матрица. Тъй като бяха добавени още някои изречения, сред които и 20 тенденциозно грешни (за целите на оценката по-късно), тестовото множество нарасна на 213 изречения. Това множество илюстрира основни явления като: представяне на комплементи и модификатори, координация, съгласуване, контрол, пасивизация, номинализация, релативни изречения, негация, квантификация и др. Освен тях са представени езиковоспецифични явления за българския език в сравнение с английската граматика, сред които: нулевата субектност, лексикално изразената категория *вид*, клитичните местоимения и дублирането на местоименията, по-свободният словоред и др.

[9] Minimal Recursion Semantics. Повече информация може да се намери на http://lingo.stanford.edu/sag/papers/copestake.pdf

За българския език беше взето решение морфологията да се кодира като част от самата автоматична граматика, за да може с тази граматика да се прави не само *анализ на текст*, но и *генерация на текст*. Генерацията е много важна стъпка например при машинния превод. Съществуваха два възможни подхода. Единият беше да се прекодира ръчно цялата морфологична система заедно с правилата и изключенията, които са доста за флективния ни език. Това, разбира се, е лингвистично издържан подход, но е доста трудоемък. Вторият подход беше да се пренесат морфологичните класове (особено на глаголите) от вече съществуващ електронен морфологичен речник за българския език в граматиката. Избран беше именно този втори подход с оглед на осигуряване на време за работа по синтактичните правила. Но пренасянето на толкова много информация стана на цената на получаването на голям брой морфологични правила – 2600 само за личните глаголи. Този подход, макар че осигури пълнота за синтактичните анализи, предизвика проблеми в системата за създаване на граматика от гледна точка на бързината на зареждане на граматиката. Този проблем още не е решен. Един възможен подход е да се направи версия на граматиката с определен брой най-чести глаголи в българския език. Когато граматиката се развие достатъчно на синтактично равнище, ще се добавят и по-рядко срещаните глаголи или техни форми в лексикона.

2. Универсализирането на българския синтактично анализиран корпус – Бултрибанк[10]

Паралелно с опитите за създаване на прецизни лингвистични граматики с голям обхват на действие за различните езици доста усилия са насочени и към създаването на сравними анотирани данни. За разлика от създаването на граматики, където процесът е насочен към влагане на лингвистично знание в даден теоретичн модел, тук идеята е да се извлича знание от вече анализирани по определен начин езикови ресурси. От една страна, тези усилия са свързани с прехвърляне на вече анализираните данни към обща анотационна схема. Тук попадат различните състезания и хакатони. Едно такова състезание беше проведено през 2006 г. с цел сравняване на синтактични дependентни парсери. От друга страна, при създаването на нов езиков ресурс различните групи следват установени добри практики. Пример в това отношение са синтактично анотираните корпуси, които следват анотационната схема на пражкия синтактично анотиран корпус Prague Dependency Treebank[11] (създаден по дependентна теория) или на американския синтактично анотиран корпус PennTreebank[12] (създаден по конституентна теория). Следването на добрите практики е стъпка, подобна на създаването на ресурсни граматики по определена матрица, тъй като изначално се следва някакъв общ модел на представяне. Но разликата е, че при създаването на граматика обикновено един или малък брой разработчици директно кодират знание в теоретичен модел за автоматичен синтактичен анализ, а при втория случай група анотатори анализират лингвистично даден корпус спрямо определена анотационна схема. Тази схема всъщност играе ролята на приближение към даден теоретичен модел и се усъвършенства непрекъснато. По-късно анотираният ръчно или полуавтоматично корпус може да послужи за обучение на морфологични, синтактични и семантични анализатори.

В последните няколко години се появи една по-съвременна инициатива за ‘универсализиране’ на синтактично анотирани корпуси. Тя се нарича Универсални зависимости (Universal Dependencies)[13], защото в основата на лингвистичното представяне е дependентната теория. Анотационната схема обаче включва всички езикови равнища: сегментиране на думи и символи, разпознаване на многокомпонентни думи, части на речта, граматически характеристики, дependентни релации. Тази инициатива също има своя история. От една страна, в Станфорд е разработена първоначална схема за универсални дependентни релации (т.нар. Stanford Type Dependencies), които са около 50 (De Marneffe and Manning 2008). От друга страна, от Гугъл излизат с предложение за представяне на многоезиковите данни в универсален тагсет[14], който включва 12 тага[15] (Petrov et al. 2012): съществително, глагол, прилагателно, наречие, местоимение, детерминатор, предлози (или следлози), числителни, съюзи, частици, пунктуация и таг за всичко (например чужди думи или абревиатури). Този тагсет се основава на сравнението на

[10] http://bultreebank.org/Resources.html

[11] https://ufal.mff.cuni.cz/pdt3.0

[12] https://www.cis.upenn.edu/~treebank/

[13] http://universaldependencies.github.io/docs/#language-u

[14] Списък с етикети, които кодират част на речта и граматичните ѝ характеристики, когато ги има.

[15] Конкретен етикет от списъка с етикети. Напр. *Amsi* означава следното: A=adjective; m=masculine; s=singular; i=indefinite.

25 синтактично анотирани ресурса, сред които и българският ресурс Бултрибанк. И двете инициативи, макар че отчитат ролята на голямо количество ресурси за различни езици, се правят от отделни изследователи и групи. Затова бързо започва да става ясно, че дependентните релации всъщност включват разнородни и разнорангови релации (напр. *комплемент, агент, референт* и под.), а универсалният тагсет е толкова общ, че не може да отчете спецификите на езиците в достатъчна степен. Така се ражда идеята за координирането на подобна инициатива с участието на самите разработчици на ресурсите, които да преценяват доколко техният език и ресурс позволява универсализиране и доколко е необходимо да се отчетат езиковоспецифичните характеристики. Жизненият цикъл на начинанието включва следните стъпки: фиксиране на схема с универсалните принципи и възможностите за кодиране на езиковите особености; кодиране на ресурса спрямо тази схема; дискусия на проблемите между участниците; промяна на анотационната схема; кодиране на ресурса спрямо променената схема. Публикуването на ресурсите става на всеки 6 месеца, като промяната на схемата е по-бавен процес. Ресурсите са свободно достъпни[16].

Процесът на конвертиране на Бултрибанк в рамките на универсалните зависимости не е първият за ресурса спрямо друга схема на представяне. Първо, от оригиналния си вариант, който е ориентиран към ОФГ, той е пренесена в дependентен формат. В него са изключени елипсите като езиково явления поради трудността на кодиране в новата теоретична рамка. Така от 215 000 думи, корпусът намалява на 196 000 думи. Конверсии са правени и в схемата на PennTreebank, както и в първоначалната схема на Станфорд. Подобни прехвърляния на данни се оказват полезни, тъй като посочват проблемите в обяснителната сила на теоретичните рамки, трудностите при формализация на езиковите явления и съответно – предизвикателствата при представяне на лингвистичната информация в по-специфични детайли.

Представянето на Бултрибанк в рамките на универсалните зависимости довежда до минимизиране на йерархизацията, т.е. бинарните структури стават равни. Този въпрос е свързан и с въпроса за опората. При универсалните зависимости е възприето, че опората е пълнозначната дума във всички фрази (именни, глаголни и т.н.). Т.е. подходът е по-скоро семантичен. В оригиналната си версия Бултрибанк възприема спомагателния глагол за опора при глаголните фрази. И двата подхода, разбира се, имат своите силни и слаби страни. При по-семантичния подход се позволява ‘скриването’ на функционални думи като предлозите, копулата и под. Т.е. може да се симулира дълбинно равнище на представяне в смисъла на Чомски или в тектограматичното представяне на чешкия синтактичен ресурс и др. При по-синтактичния подход се дава превес на повърхнинното представяне на синтактичните отношения.

По отношение на морфологията инициативата възприема синтактичен, а не морфологичен подход. Това означава, че функцията доминира над произхода, генеалогията. Затова се наложиха следните действия при трансфера на лингвистична информация: а) директно прехвърляне на частите на речта; б) разделяне на частите на речта на подвидове и в) промяна на частите на речта. В първия случай има пълно съответствие при таговете (етикетите). Затова директно се прехвърлят предлози в предлози; прилагателни в прилагателни и т.н. Във втория случай местоименията се подразделят според функцията си на детерминатори, местоимения и наречия; числителните се подразделят на прилагателни, наречия и числителни; глаголите се подразделят на пълнозначни глаголи, спомагателни глаголи и прилагателни (причастията).

Разбира се, тези класификации по функция не са безпроблемни. Затова непрекъснато се търсят по-добри решения.

Последният случай в) се отнася например за частиците *да* и *не*, които първоначално се кодираха като междуметия по схемата. Това решение изглеждаше доста странно и затова беше променено.

По отношение на синтаксиса, както вече беше споменато, се възприема семантичен подход при представяне на релациите. Тук бяха предприети следните стъпки: а) директен пренос на релации; б) недиректен пренос на релации; в) използване на ‘плаващи’ релации и г) необработване на релации.

Директни са релациите като *dobj* (директен обект), *iobj* (недиректен обект), *nsubj* (номинален подлог), *csubj* (изреченски подлог) и др., които се срещат в много анотационни схеми.

[16] http://universaldependencies.org/

Недиректни са релациите като *да-изреченията* (CLDA) и *че-изреченията* (CLCHE), които трябваше да се трансформират в изречения с контрол (*xcomp*) или липса на контрол (*ccomp*). Т.е. имаше нужда от експлициране на повече информация относно синтактичните връзки в изреченията.

'Плаващите' релации са тези, които се конкурират при дадено явление, защото все още не е установено на многоезиково равнище коя е най-подходящата. Например въпросителната частица *ли* е кодирана в универсалната версия на Бултрибанк чрез релацията *discourse* (дискурсна релация, която е типична за вметнати части, междуметия и дискурсни частици), но в други езици подобни частици са кодирани като *aux* (спомагателна релация, която е типична за отношението между спомагателен и пълнозначен глагол), *expl* (експлетивна релация, т.е. типична за отношението между нереферентен подлог и сказуемо) или *mark* (маркираща релация, която е типична като връзка между комлементизатора 'че' и опорния глагол на сказуемото в подчиненото изречение).

Релациите, които не са обработени все още, се отнасят най-вече до изреченията с елипси. Макар че в Осенова и Симов (2007) е представена типология на елипсите, автоматичното им пренасяне към универсалната схема се оказва трудно.

Заключение

В тази статия бяха представени два подхода към идеята за сравнение и сравнимост между езиците. Единият е създаването на прецизна лингвистична граматика за даден език с голямо покритие и спрямо определена матрица. Другият е конвертирането на съществуващ езиков ресурс от една схема на представяне към друга. Тези два подхода не си противоречат, а се допълват. И при двата подхода има проблеми. При първия кодирането и тестването на граматиката отнема много време. При втория в процеса на конвертиране може да се изгуби или изкриви лингвистичната информация. Но и двата подхода са необходими за постигането на равнища за сравнение между различните езици. И при двата подхода най-успешна се оказва идеята за привличане на специалисти за различните езици, които да оформят многоезиков екип.

Целта на подобни начинания е не само да се създаде лингвистично по-обективна среда за типологическо сравнение между езиците, но и да се създаде технически по-надеждна среда за приложения като обучение на автоматични синтактични анализатори (парсери), машинен превод, извличане на информация.

ЛИТЕРАТУРА

Осенова 2016: Осенова, П. *Граматическо моделиране на българския език (с оглед на автоматичната обработка на естествен език)*. Парадигма. // **Osenova 2016:** Osenova, P. *Gramatichesko modelirane na balgarskiya ezik (s ogled na avtomatichnata obrabotka na estestven ezik)*. Paradigma.

Осенова и Симов 2007: Осенова, П. и Симов, К. *Формална граматика на българския език*. ИПОИ. // **Osenova i Simov 2007:** Osenova, P. i Simov, K. *Formalna gramatika na balgarskiya ezik*. IPOI.

Avgustinova and Zhang 2009: Avgustinova, T. and Zhang, Yi. Exploiting the Russian National Corpus in the Development of a Russian Resource Grammar. // *Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains*, 2009, Borovets, Bulgaria, 1–11.

De Marneffe and Manning 2008: de Marneffe, M. and Manning, Ch. The Stanford Typed Dependencies Representation. // *Proceedings of the Workshop on Cross-Framework and CrossDomain Parser Evaluation*, 1–8, Stroudsburg, PA, USA. COLING, Association for Computational Linguistics.

Osenova 2011: Osenova, P. Localizing a Core HPSG-based Grammar for Bulgarian. // Hanna Hedeland, Thomas Schmidt, Kai Worner (eds.) *Multilingual Resources and Multilingual Applications*, *Proceedings of GSCL 2011*, ISSN 0176-599X, Hamburg, 175–180.

Osenova and Simov 2015: Osenova, P. and Simov, K. Universalizing BulTreeBank: a Linguistic Tale about Glocalization. // *Proceedings of BSNLP 2015*, Hissar, Bulgaria, 81–89.

Petrov et al. 2012: Petrov, S., Das, D. and McDonald, R. A universal part-of-speech tagset. // Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Van Eynde 2006: Frank van Eynde. NP-internal agreement and the structure of the noun phrase. // *Journal of Linguistics 42 (2006),* Cambridge University Press, 139–186.