Мария Тодорова, Ивелина Стоянова, Светлозара Лесева
Институт за български език – БАН, София

ЛИНГВИСТИЧНО ОПИСАНИЕ И КОМПЮТЪРНА ОБРАБОТКА НА СЪСТАВНИ ИМЕНУВАНИ СЪЩНОСТИ В БЪЛГАРСКИЯ ЕЗИК

Abstract: We present an automatic approach for the creation of a Bulgarian Morphological Dictionary of Named Entities consisting of 21,161 named entities (NEs). Each NE is assigned an inflectional type which reflects its part-of-speech, lexico-grammatical and morphological properties, including the morphonemic alternations. NEs are classified into semantic categories according to an extended hierarchical system. The wordforms of each entry are generated automatically from its inflectional type, and the relevant grammatical characteristics are assigned. The result is validated by experts.

Keywords: multiword expressions, named entities, Bulgarian language, Natural language processing

1. Увод

В компютърната обработка на естествените езици се налага нов подход към съставните думи: всяка от тях трябва да бъде разпознавана като една лексикална единица, състояща се от няколко графични думи и характеризираща се със собствено формообразуване.

Статията представя морфологичен речник от 21 161 единици, който включва географски названия, имена на личности, наименования на събития и други, класифицирани по флективни типове. Ресурсът е приложим при анотацията на корпуси на български език, в системи за автоматичен превод, в компютърната лексикография и в други области.

2. Характеристики на съставните именувани същности

В статията се разглеждат *съставните именувани същности (СИС*), които са подклас както на съставните единици (англ. *multiword expressions*), така и на именуваните същности (англ. *named entities*). СИС са прагматично обусловена категория и се причисляват към институционализираните съставни единици (Саг и др. 2002), наричани още прагматеми (Мелчук 1995). Приемаме, че СИС обхващат дефинираните от Крипке (Крипке 1980: 48 – 49) еднозначни названия, които означават едни и същи обекти във всички възможни светове. На практика, според определението на Секине (Секине 2007), те отговарят на въпроса „Какво е **името** на …?“: *Какво е **името** на това животно? – Брумбар рогач*; *Какво е **името** на този празник? – Ден на независимостта на България*.

Представеният Речник включва СИС, които обозначават: а) уникални референти, които може да са единични – *Черно море*, или множествени – *Се-*

демте рилски езера; б) терминологични названия – *бяла мечка*, *блатен хвощ*. Наименованията на единични обекти в определен контекст може да обозначават и два или повече различни, но еднотипни референта и тогава се срещат и във форма за множествено число – *Министерствата на здравеопазването на България и Гърция*. Същевременно, тъй като терминологичните названия може да се изменят по число, когато означават представители на даден клас обекти, в Речника те са въведени като СИС с пълна парадигма.

СИС се характеризират със структурни и парадигматични ограничения, свойствени на съставните лексеми изобщо, но и допълнително обусловени от силната институционализираност на разглежданата категория:

а) Броят на формите на компонентите е ограничен спрямо тези на съответстващите им самостоятелни думи и се определя от лексикално-граматичните категории на дадено СИС и от парадигмата му; например в СИС със структура AN прилагателното винаги се съгласува със съществителното (*Чешка република*);

б) Не допускат замени на компоненти с друга форма или със синоним, срв. *Мало село* (собствено име) и *малко село* (съществително нарицателно име в свободно словосъчетание);

в) Обикновено не може да бъдат разделяни от външни елементи – например *В Стара ли Загора отиваш?* се среща само инцидентно;

г) Налагат строги ограничения в линейната последователност на компонентите – *Национален съвет за радио и телевизия*, но не *Национален съвет за телевизия и радио*;

д) Много от тях се характеризират с парадигматични ограничения, включващи дефективност на формите за число и/или определеност. Някои типове, основно лични имена и топоними, са неизменяеми, а други позволяват членуване: *Море на спокойствието – Морето на спокойствието*. В такива случаи реализацията на дефективни форми сигнализира, че словосъчетанието е свободно, срв. *Южен парк* и *южни паркове*, *Сейшелски острови* и *сейшелски остров*, *Ново село* и *новото село*.

СИС са групирани в семантични класове, които описват типа на назованите обекти. Тук са възприети основните категории от Разширената йерархична система на Секине (Секине 2007), които включват лични имена, топоними, организации, събития, периоди, продукти, биологични видове и други.

3. Съставните именувани същности в компютърната лингвистика

Правилното тагиране (определяне на частта на речта и граматичните характеристики) и лематизиране (определяне на основната форма) на съставните лексикални единици спомага за усъвършенстването на езиковата анотация (тагиране, синтактичен анализ, семантична анотация), което може да допринесе и за подобряване на разработваните езикови технологии, включително машинния превод, търсенето и извличането на информация, автоматичното отговаряне на въпроси и др.

Всяко СИС трябва да бъде разпознавано като цялостна лексикална единица, съставена от две или повече графични думи. Необходимостта от правилно лематизиране на съставните лексикални единици и на СИС в частност се налага най-вече в случаите, когато съдържат пълнозначни думи, например *Черно море*, за да се разграничи свободното значение (на прилагателното *черен* и съществителното *море*) от значението на СИС. Много от съществуващите програми разпознават и тагират пълнозначните компоненти на СИС като отделни единици: например *Черно*{*черен*, прил.} *море*{*море*, същ. нариц.}.

Правилното определяне на морфосинтактичните характеристики на СИС също представлява предизвикателство в морфологично богати езици като българския. Преобладават два основни подхода към компютърното представяне на съставните единици: а) чрез крайни автомати, които генерират всички форми, използвайки речници, в които единиците са групирани във флективни типове (Зилберщайн 2005; Савари 2005); и б) чрез формализма на унификационните граматики в съчетание с речници, като чрез променливи се задават структурни зависимости и ограничения (Саг и др. 2002). Тук е възприет първият подход.

4. Речник на съставните единици за целите на компютърната обработка

4.1. Структура на речниковото описание

Описанието на СИС в Речника се основава на принципите и означенията, възприети в рамките на граматичното описание на простите думи в Граматичния речник на българския език (Коева 1998) и предложената концепция за граматично описание на съставните лексеми (Коева 2006).

Всяко СИС е представено в речниковата си статия с основна форма (лема) и граматично описание. За лема на СИС е използвана максимално неутралната граматически правилна форма (например *Варненско езеро*). Формалното лексикално-граматично описание на единиците във възприетия речников формат включва:

а) категориална информация, която характеризира лемата и служи за групиране на думите в класове според частта на речта, към която принадлежат – СИС принадлежат към класа на съществителните имена (N);

б) лексикално-семантичен разред в рамките на дадената част на речта – при съществителните имена лексикално-граматичните разреди се поделят според това, дали съществителните назовават понятия, явления и под. (нарицателни имена, означени със С) или уникални обекти (собствени имена, означени с Н);

в) лексикално-категориална граматична информация, групираща думите в граматични подкласове – при СИС, като част от съществителните имена, такава характеристика е родът, за който се използват означенията – M (мъжки род), F (женски род), N (среден род);

г) допълнителна парадигматична характеристика, отразяваща определени особености на парадигмата – при СИС такава е дефективността на парадигмата по отношение на категорията число, тъй като често те са или само единични, или само множествени обекти: (S) сингулария тантум, (P) плуралия тантум;

д) семантичен клас: географски названия (G), събития (E), лица (R), организации (A), продукти на човешката дейност (J), небесни обекти (C), растителни видове (B), животни (Z);

е) номер на флективния тип, който еднозначно определя флективната парадигма и процеса на формообразуване – при СИС парадигмата се определя от броя и реда на компонентите, от гореописаните характеристики (а-д), от фонетичните изменения и др.

В системата на Речника характеристиките от а) до д) са зададени с еднобуквени означения, а номерът на флективната парадигма – с число, като комбинацията им в строго определен ред (с изключение на семантичния клас) представлява флективният тип, който характеризира СИС с еднакви лексикално-граматични, морфосинтактични и други особености.

4.2. Дефиниции на флективните типове

Всеки флективен тип се идентифицира от уникален код (съставен от компонентите, описани в 4.1.) и дефинира начина, по който от лемата се генерират формите. Пример 1 представя формалното описание на флективния тип NHMS2 за съществителни собствени имена от мъжки род, сингулария тантум, състоящи се от два компонента – прилагателно и съществително име, като: 1) по форма се изменя само първият компонент (*Южен парк*, *Южния парк*); 2) словоформите се образуват от основната по следния начин: предпоследният символ отпада (напр. „е“ в „южен“, „ъ“ в „малък“) и окончанието се конкатенира към остатъка.

Пример 1. Флективен тип NHMS2

NHMS2 = <1> <2>/smo	*основна форма*
+ <1><L2><S><R>ият <2>/sml	*членувана форма – пълен член*
+ <1><L2><S><R>ия <2>/smh	*членувана форма – кратък член*

<1> <2> – компоненти на лемата: ***Южен парк*** (за генериране на формите вж. 5.2.)

/ – разделител, след който към генерираната форма се приписват граматичните ѝ характеристики.

Основните означения, използвани за граматичните характеристики, са: s – единствено число, p – множествено число; m – мъжки род, f – женски род; n – среден род; o – нечленувана форма, l – членувана форма (м. р., пълен член), h – членувана форма (м. р., кратък член), d – обща членувана форма (ср. и ж. р.), c – бройна форма (м. р.).

4.3. Структурни и флективни типове

Към момента Речникът обхваща следните най-често срещани синтактични структури при СИС[1]: а) *AN – Южен парк*; б) *AAN – прав речен рак*; в) NN – *бръмбар рогач*; г) *NPN – Море на спокойствието*; д) *ANPN – Българска академия на науките*; е) *NPAN – Секция по компютърна лингвистика*; ж) *ANPAN – Българска асоциация по информационни технологии*.

Структурните типове се разпределят в съответните флективни типове според броя на членовете в парадигмата:

а) Неизменяеми СИС (с едночленна парадигма), например *Ню Йорк* (флективен тип *NHMS0*), *Нова Зеландия* (*NHFS0*).

б) Изменяеми СИС с парадигматични ограничения по категорията число. Парадигматичната характеристика, сигнализираща ограничението, се реализира с две стойности – сингулария тантум (S), например *Александров архипелаг* (NHMS1), и плуралия тантум (P), например *Сейшелски острови* (NHMP1).

в) Изменяеми СИС с пълна парадигма. Както при останалите съществителни, при СИС парадигмата е шестчленна в мъжки род (основна форма – ед.ч. нечленувана; форма в ед.ч., членувана с пълен член; форма в ед.ч., членувана с кратък член; бройна форма; нечленувана форма за мн.ч.; членувана форма за мн.ч.) – например *див заек* (NHM17); четиричленна в среден и женски род (основна форма – ед.ч., нечленувана; членувана форма за ед.ч.; нечленувана форма за мн.ч., членувана форма за мн.ч.). Към парадигмите на СИС, за които категорията е релевантна, може да бъдат добавени и звателни форми.

4.4. Съдържание на Речника

Към момента Речникът на съставните именувани същности включва: 21 161 СИС, класифицирани в 81 флективни типа и 8 семантични класа, от които се генерират 39 028 словоформи. Броят на флективните типове се обуславя от различната структура на СИС, различния брой на елементите, ограниченията и особеностите на парадигмата, както и от различните съгласувателни условия. Таблица 1 представя разпределението на единиците по семантични класове.

Таблица 1. Разпределение на СИС в Речника по семантични класове

Семантичен клас	Брой единици	Брой форми
Географски имена	11 965	20 923
Събития, периоди	1 947	871
Лица	2 212	1 747

[1] Използвани са следните означения: A – прилагателно име, N – съществително име, P – предлог.

Организации	842	842
Артефакти	989	989
Растения	1 697	7 228
Животни	1 430	6 322
Небесни обекти	79	106
Общо	**21 161**	**39 028**

5. Създаване на Речника

Речникът е създаден с помощта на автоматични методи, подпомогнати от ръчна експертна верификация. Кандидатите за речникови единици са извлечени по два основни начина: а) от общи и специализирани лексикографски ресурси (речници, списъци с термини и под.); и б) с помощта на евристики от Уикипедия и други структурирани и/или частично анотирани ресурси. Възможно е също извличането на СИС чрез статистически методи от свободни текстове от Българския национален корпус.

5.1. Автоматично приписване на флективни типове

При създаването на Речника са приложени различни евристики, позволяващи автоматичното предсказване на формоизменителните характеристики и флективния тип на СИС: а) разпознаване на опората и подчинените компоненти – синтактичната структура в голяма степен предсказва кои компоненти се изменят при формообразуване; б) разпознаване на рода (според рода на опората) и предсказване на броя на формите (например пълен и кратък член при съществителните от мъжки род); в) анализиране на завършека на прилагателните за определяне на формообразуването (например при членуване прилагателните от м.р., завършващи на *-ски*, добавят *-я/-ят*, докато останалите присъединяват *-ия/-ият*); г) търсене на възможните словоформи на компонентите в корпус, за да се потвърди флективният тип (например за прилагателното *червен* в мн.ч. в корпуса се среща формата *червени*, но не и *червни* и така се предсказва правилната парадигма).

5.2. Автоматично формообразуване

Автоматичното формообразуване се извършва с помощта на флективни граматики, които по зададения флективен тип на лемата генерират всички нейни форми със съответните им морфосинтактични характеристики. Методът е основан на формализма на крайните преобразователи, които разпознават дадена последователност от символи и я трансформират в друга форма (с добавени граматични характеристики). Така морфосинтактичните изменения на СИС в настоящия речник се описват чрез множество от взаимно свързани

трансформации. Работата на автомата за даден флективен тип може да се опише с представените в Пример 2 операции.

Пример 2. *Генериране на определената форма с пълен член и приписване на граматичната информация за флективния тип NHMS2.*

<1><L2><S><R>ият <2>/sml

Символите в триъгълни скоби означават: (а) пореден компонент; или (б) команда с брой повторения. С диез е означена текущата позиция:

<1> – преместване на устройството на автомата след края на първата дума: ***Южен#***

<L2> – преместване вляво (L) от текущата позиция 2 пъти: ***Юж#ен***

<S> – изтриване на дясностоящия символ (е): ***Юж#н***

<R> – преместване вдясно: ***Южн#***

ият – добавяне на окончанието ***ият*** на мястото на текущата позиция: ***Южният***

<2> – преместване след края на втората дума

/sml – приписване на граматичните характеристики

Генерирана форма	Основна форма	Характеристика на лемата	Морфосинтактични характеристики
Южният парк	*Южен парк*	*NHMSG*	*sml*

5.3. Верификация

Флективният тип и генерираните форми са проверени ръчно, а в случаите на неправилна категоризация по една или повече характеристики са нанесени корекции. Средно 4,7% от единиците имат грешно приписан флективен тип, като това засяга най-вече следните категории: а) грешно приписан флективен тип на СИС от мъжки род поради особености във формообразуването на прилагателното име (например с отпадаща гласна или редуване, *Западен парк – Западния парк*); б) грешно приписан флективен тип на СИС поради фонетични промени в съществителното при образуване на множествено число (например *див заек – диви зайци*); в) неустановени ограничения в парадигмата (например флективен тип, допускащ членуване, приписан на неизменяемо СИС – *(връх) Голям Богдан – *Големият Богдан*.

6. Бъдещо развитие и приложения на Речника

Речникът представлява резултат от работата по систематичното описание на формообразувателните специфики на съставните лексикални единици, насочена към създаването на унифициран речник на съставните лексеми от различни части на речта и с различна степен на идиоматичност. Предвижда се разширяване на Речника както с други семантични класове СИС, така и с

нови структурни и флективни типове. По-прецизното автоматично идентифициране на флективния тип може да се постигне чрез по-комплексни методи за морфологичен анализ на компонентите.

Правилното лематизиране и приписване на граматически характеристики на СИС би спомогнало за по-качествена лингвистична анотация на корпуси, което от своя страна може значително да повиши качеството на езиковите технологии, базирани на тези ресурси.

ЛИТЕРАТУРА

Зилберщайн 2005: Silberztein, M. NooJ's dictionaries. – In: *Proceedings of LTC'05*. Poznań: Wydawnictwo Poznańskie, 2006, pp. 291 – 295.

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

Коева 2006: Koeva, S. Inflection Morphology of Bulgarian Multiword Expressions. – In: *Computer applications in Slavic studies. Proceedings of Azbuki@net*, *International Conference and Workshop*, Sofia, 2006, pp. 201 – 216.

Крипке 1980: Kripke, Saul. *Naming and Necessity*. Harvard University Press, Blackwell, 1980.

Мелчук 1995: Melchuk, I. Phrasemes in language and phraseology in linguistics. – In: *Idioms: structural and psychological perspectives. Chap. 8*. Lawrence Erlbaum Associates, 1995, pp. 167 – 232.

Савари 2005: Savary, A. A formalism for the computational morphology of multi-word units. – *Archives of Control Sciences*, 2005, No 15(3), pp. 437 – 449.

Саг и др. 2002: Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., Flickinge, D. Multiword Expressions: A Pain in the Neck for NLP. – In: *CICLing*, 2002, pp. 1 – 15.

Секине 2007: Sekine, S. *The Definition of Sekine's Extended Named Entities. Version 7.1.0*, *09-07-2007*. <http://nlp.cs.nyu.edu/ene/version7_1_0Beng.html> (дата на достъп: 13.11.2015).