Семантично анотирани ресурси за българския език – БулСемКор

Мария Тодорова, Христина Кукова, Светлозара Лесева

Институт за български език – БАН

Abstract: This paper describes the Bulgarian Sense-Annotated Corpus (BulSemCor) – a balanced semantically annotated corpus of 99,480 lexical units in which each lexeme is assigned a unique sense according to the Bulgarian WordNet (BulNet). After an outline of the good practices in the compilation of sense-tagged corpora in Section 1, the paper goes on to discuss the general principles underlying the annotation in BulSemCor in light of generally-recognised recommendations regarding corpus structure, data representation, annotation schema and annotation principles (Section 2). Section 3 is centred on the structure and data representation of the corpus. The main part of the paper (Section 4) deals with semantic annotation and addresses its relationship to other levels of linguistic processing in the corpus, such as lemmatisation, normalisation, morpho-syntactic tagging. Part of the discussion is focused on the novel features of BulSemCor in comparison with other SemCor-modelled corpora, such as full MWE annotation and semantic tagging of closed-class words. Section 5 sums up the actual and possible applications of BulSemCor. We show that the mapping to the rich semantic representations of the relevant BulNet synonym sets supplies the annotated units with abundant linguistic information. This multi-level annotation augments the applicability of BulSemCor to various tasks besides the primary goals (WSD and MT), including semantic and syntactic analysis, information extraction, clause segmentation and alignment, discourse analysis, among others.

1. УВОД

Семантично анотираните корпуси представляват колекции от структурирани езикови данни, в които на всички или на определени лексикални единици е приписан таг със семантична информация. Той може да обозначава семантична релация между конкретната дума и други думи в корпуса във вид на семантична роля, съотнасяне със семантичен фрейм и др. или да съдържа информация за семантичните свойства на думата – лексикалното значение, реализирано в конкретния контекст, семантичната ѝ съчетаемост и т.н. (Макенери, Уилсън 2001: 61–62). В класическия случай семантичната информация, която се приписва, се основава на анотационна схема от предварително дефинирано (и по възможност разширяемо) множество от лексикални значения, семантични роли и т.н., изработена на базата на съществуващ лексикографски ресурс – тълковен речник, семантико-синтактичен речник, лексикалносемантична мрежа и др.

Тук под семантично анотирани корпуси ще имаме предвид корпуси, в които на всички думи или на част от тях е приписано еднозначно лексикално или граматикализирано значение. Известни са два основни подхода към този тип семантична анотация (Коева 2010б: 11) – чрез избор на значения от зададена анотационна схема или чрез превод на семантично анотиран корпус от един език на друг.

В рамките на първия подход, на който ще се спрем по-подробно, съществуват две основни тенденции: а) анотация на всички думи (от определени части на речта) в даден корпус и б) анотация на срещанията на конкретни многозначни думи. В първия случай обикновено се обхождат и анотират срещанията на думи от една или повече пълнозначни части на речта, най-често съществителни и глаголи, а във втория – думи от отворените класове, подбрани така, че да се срещат в корпуса с честота над определен праг, да имат над определен брой значения или по други критерии.

Най-известният семантичен корпус СемКор (SemCor) (Милър 1995) се състои от 352 текста с общ обем от около 700 000 думи, част от които са подбрани от *Корпуса на съвременния американски английски език на Университета „Браун“*[1] (Brown Corpus) (Франсис, Кучера 1979). Езиковият ресурс, използван при анотацията, е Принстънският WordNet (Princeton WordNet) (Ландс и др. 1998). В около половината от текстовете са лематизирани и ръчно анотирани значенията на пълнозначните думи – съществителни, глаголи, прилагателни и наречия, а в останалата част от корпуса – само глаголите. СемКор е модел за разработване на семантично анотирани корпуси за редица други езици – испански, каталонски и баски (Агире и др. 2006; Наваро и др. 2003, цит. по Коева 2010б), италиански (Италианската синтактико-семантична банка – ISST, вж. Монтемани и др. 2000), нидерландски (Восен и др. 2012) и т.н.

По-различен подход е възприет при семантичната анотация на част от синтактично анотирания корпус Пен Трийбанк (Penn Treebank). Тагирани са около 5000 думи свързан текст (Палмър и др. 2000), като са избрани глаголите и опорите на аргументите и адюнктите им, както и собствени имена, на които са приписани обобщени тагове.

Според втората методология е анотиран корпусът DSO (Нг, Лий 1996), който съдържа анотация на срещанията на 121 съществителни и 70 глагола в корпус от близо 193 000 изречения от Корпуса „Браун“ и „Уолстрийт джърнъл“, като са използвани значенията от Принстънския WordNet.

Друг такъв корпус е TüBa-D/Z (Телйохан и др. 2004), в който обект на анотация са глаголи и съществителни с повече от едно значение в GermaNet и поне 20 срещания в корпуса. Тагирани са около 16 000 срещания на 80 думи, като липсващите значения са попълвани в GermaNet паралелно с анотацията.

Ръчно анотираният семантичен корпус MASC (Пасоно и др. 2010) включва 500 000 думи, част от които са анотирани с морфосинтактична, синтактична и семантична информация. По-специфичното за този корпус е, че една част от него, която съдържа по 1000 срещания на 100 високочестотни многозначни думи, се анотира не само със значенията от Принстънския WordNet, но и със семантичните фреймове и елементи от ФреймНет (FrameNet).

DutchSemCor (Восен и др. 2012) с обем около 1 000 000 думи съдържа примери за значенията на 3000-те най-често срещани многозначни съществителни, глаголи и прилагателни. Характерно за него е, че около 250 000 от думите са ръчно анотирани, а на останалата част автоматично е приписано значение с

[1] За краткост нататък този корпус ще бъде наричан с името, с което е популярен – Корпус „Браун“.

помощта на системи за отстраняване на семантична многозначност, след което резултатите са проверени от експерти.

Известни са и други инициативи за автоматично тагиране, които целят да облекчат и, доколкото е възможно, да заменят ръчната анотация. Сантамария и др. (2003) представят алгоритъм за автоматично съотнасяне между уебдиректории от Open Directory Project и синонимни множества в WordNet, който служи за обогатяване на съответните значения със семантични описания, тематично групиране на значения и конструиране на семантично анотирани корпуси. Хенрих и др. (2012) описват ексцерпиране на корпус (WebCAGe) с примерни употреби на многозначни съществителни, прилагателни и глаголи от интернет, анотиран автоматично със значения от GermaNet на базата на съществуващото съотнасяне между GermaNet и немския Wiktionary. Броят на анотираните думи е 10 750.

Автоматичната семантична анотация е свързана с осигуряване на допълнителни програмни или езикови ресурси като системи за автоматично определяне на значението и надеждни алгоритми за съотнасяне между лексикални и други бази от данни. Ръчно анотираните корпуси продължават да са необходими за трениране на методи за машинно обучение, както и за тестване и оценка на системите за автоматично отстраняване на многозначност. Направеният преглед показва, че създаването на семантични корпуси, тагирани от експерти, продължава да е актуално, като стремежът е да се осигурява богата анотация на морфосинтактично, синтактично и друго равнище, включително многопластов семантичен анализ.

2. БулСемКор И СЪВРЕМЕННИТЕ СЕМАНТИЧНО АНОТИРАНИ КОРПУСИ

Българският семантично анотиран корпус (БулСемКор) следва методологията, използвана при създаването на СемКор, в съчетание с някои специфични принципи (Коева 2010а). Подобно на други корпуси, конструирани по същия модел, БулСемКор е извадка от аналогичен на Корпуса „Браун“ ресурс – Българския „Браун“ корпус (Коева и др. 2006) и е анотиран ръчно със значения от Българския WordNet. Обемът му е значително по-малък от този на СемКор, но е съпоставим с този на много от съществуващите семантично анотирани корпуси. Важна характеристика на БулСемКор е, че при подбора на текстовете са приложени евристични методи, които осигуряват оптимално за обема и структурата на корпуса покритие на разнообразна многозначна лексика (вж. 3).

В световната практика доминиращият ресурс за семантична анотация е WordNet. Българският WordNet – БулНет, също е избран за анотацията на БулСемКор по следните съображения (Коева 2010б: 24–32): гранулираността и изчерпателността на дефинираните значения; комплексната релационна структура на WordNet, която прави възможни редица приложения, свързани с обработката на естествения език; съотнасянето на БулНет с Принстънския WordNet, а чрез него и с други уърднети, което осигурява достъп до еквивалентите на съответните значения в голям брой езици; разширяемата анотационна схема, която позволява паралелно с процеса на анотация да се добавят и редактират значения в съответствие с корпусните данни.

В повечето случаи семантичната анотация се комбинира с други равнища на обработка: морфологична (определяне на основната форма, или лематизация) и морфосинтактична (определяне на частта на речта), а нерядко и синтактична, дискурсна и/или прагматична анотация. В БулСемКор наред със семантичното тагиране, при което се извършва съотнасяне на конкретната контекстуална употреба на всяка лексикална единица от корпуса със съответното (точно едно[2]) семантично множество в Българския WordNet, е извършено пълно морфологично, частично морфосинтактично и частично синтактично анотиране.

Както беше посочено, семантичната анотация обикновено обхваща пълнозначните думи (или част от тях), като се отдава предпочитание на съществителните имена и глаголите и по-рядко на прилагателните и наречията. Въпреки че в последните години интересът към затворените класове думи става все по-актуален, обикновено такъв тип анотация се извършва само за определени значения на някои служебни думи. Същото важи и за несвободните фрази[3] (различни езикови конструкции от две или повече думи, които изразяват единно понятие), от които обикновено се анотират само някои типове. В това отношение важни отличителни черти на БулСемКор са единният подход към различните лексеми и възприетият принцип за последователна и изчерпателна анотация. Според тази методология всички лексеми независимо от строежа (несъставни или съставни) или функцията си (пълнозначни или служебни) се смятат за равноправни и се анотират според дефинираните общи критерии. По този начин работата по БулСемКор поставя въпроси и осигурява среда за теоретично и практическо изучаване на различни проблеми, които като цяло са слабо застъпени в научните изследвания. Такива са например проблемите за многозначността при представителите на затворените класове думи и при несвободните фрази. В по-общ план семантичната анотация и разширяването на БулНет със синонимни множества въз основа на засвидетелстваните в БулСемКор значения поставят редица научни задачи, свързани едновременно с разпознаването на езиковите единици и с лексикографското им описание.

Анотираните единици наследяват цялата лингвистична информация, асоциирана с даденото синонимно множество, която освен задължителния морфологичен и семантичен таг може да включва характеризиране на едно или повече от следните допълнителни нива:

(а) частична информация за синтактичната структура на определени типове несвободни фрази – определяне на главната им част и на подчинените им части;

---

[2] Обикновено при семантичната анотацията се избира само едно подходящо значение, но в някои подходи, включително при създаването на СемКор (Милър 1995: 93), се допуска приписването на повече от един семантичен таг, когато значенията на синонимните множества са дефинирани твърде гранулирано или контекстът позволява повече от една интерпретация. При анотацията на БулСемКор е възприето да се приписва точно едно значение на всяка лексикална единица.

[3] Терминът *несвободна фраза* е въведен от Мелчук (Мелчук 1995); повече за мотивацията му в българската лингвистична литература вж. тук статията на Ив. Стоянова и М. Тодорова.

[4] Терминът е зает от областта търсене и извличане на информация и служи за назоваване на информационно значими единици, включващи имена, числени означения, времеви означения и под.

(б) информация за категорията (име, място, организация, дата, число и т.н.), означавана от именуваните същности[4];

(в) информация за онтологичната категория (време, място, начин, степен, количество и т.н.) на наречията;

(г) информация за типа на изразяваното от съюзите синтактично отношение (съчинително или подчинително);

(д) информация за изходната част на речта или форма при субстантивите;

(е) стилистична, граматична и друга информация за синонимните множества или отделни техни членове.

Възприетите принципи осигуряват методологичната база за създаване на висококачествен езиков ресурс, обогатен с многопластова и детайлна лингвистична информация, което го прави приложим за различни изследователски цели.

3. СТРУКТУРА И ФОРМАТ НА КОРПУСА

Българският „Браун“ корпус, от който е ексцерпиран изходният корпус за семантична анотация, се състои от 500 корпусни единици, всяка от които включва около 2000 думи. Текстовете са подбрани така, че корпусът да отразява оптимално синхронното състояние на езика в съответствие с възприетите критерии за балансираност и представителност. Разпределени са в 15 категории, обобщени в 2 групи: художествени и информативни текстове. Представителността на корпуса за семантична анотация се гарантира чрез наследяването на структурата на Българския „Браун“ корпус, като в БулСемКор са включени извадки от приблизително 100 думи (разширени в посока наляво и надясно до край на изречение) от всеки от 500-те текста на Българския „Браун“ корпус (Коева и др. 2006; Коева 2010б).

Изходният корпус се състои от две части, при които са приложени различни стратегии за подбор, описани подробно у Коева (2010б: 15–16). За основен критерий в първата част служи концентрацията на пълнозначни думи с висока честота. За целта са избрани извадки, които съдържат най-голям брой пълнозначни основни форми от честотен речник, съставен от два корпуса, в които е отстранена граматичната многозначност на словоформите: българския превод на „1984“ от Дж. Оруел и корпус с текстове от три тематични области – икономика, право и политика. За постигане на оптимално балансиран подбор, на думите от отделните части на речта се приписва различно тегло: 0,4 на съществителните, 0,3 на глаголите, 0,2 на прилагателните, 0,1 на наречията. Приложени са и допълнителни процедури за осигуряване на оптимално лексикално покритие, в резултат на което са избрани извадките, които съдържат най-голям брой различни леми в комбинация с най-голям брой съответстващи значения в БулНет. Втората част на корпуса за семантична анотация също запазва структурата на Българския „Браун“ корпус. Критерий за подбор на извадките е максимален брой думи от тях да не се срещат в БулНет или ако се срещат, да участват в максимален брой синонимни множества. Така конструираният изходен корпус съдържа 811 извадки с общ обем от 101 791 токъна.

Файловете в БулСемКор са представени в xml формат. Всеки текст е кодиран под формата на поредица от xml тагове с наименование word (Ризов 2010: 43), както е илюстрирано в опростен вид в (1); l и w са атрибути на word, които съхраняват съответно лемата и словоформата:

(1) <*word* l=″документ″ w=″документа″/>

4. ПРИНЦИПИ НА АНОТАЦИЯ

Анотационната схема е изработена в съответствие с посочените у Коева (2010б: 32–35) анотационни стандарти и добри практики: (а) възстановимост на оригиналния текст; (б) създаване на документация за анотирания текст; (в) възможност за извличане на информация за различни равнища на лингвистичен анализ; (г) лингвистична и когнитивна мотивираност на анотацията; (д) адекватно покритие на анотирания текст; (е) оптимална степен на гранулираност; (ж) непротиворечивост на анотацията; (з) гъвкавост и разширяемост на анотационната схема; (и) проверимост на анотацията.

4.1. Лематизация

Преди да се пристъпи към семантичната анотация, се извършва приписване на основна форма на всяка графична дума, за която това е възможно, последвано от нормализация[5].

Тъй като всяка анотирана единица в БулСемКор се съотнася посредством лемата си с всички синонимни множества от БулНет, които съдържат литерал[6] с идентична лема, правилната и еднозначна лематизация обуславя коректното приписване на синонимни множества, а оттам и правилния избор на значение.

Възможно е автоматично приписаната основна форма да не е подходяща в конкретния контекст. Например в (2) формата *млади*, лематизирана първоначално като прилагателното *млад*, в конкретния контекст е субстантивираната форма *млади*, плуралия тантум. За да може да бъде асоциирана правилно със синонимното множество: *{младеж:3, млади:1, /млади/ хора:1}*, се налага корекцията на лемата от *млад* в *млади.*

(2) *Младите обичат живота.*

Друга основна предпоставка за правилен избор на подходящо значение е коректното определяне на границите на анотираните единици. Те могат да бъдат: (а) прости или сложни графични думи; (б) несвободни фрази; (в) (последователности от) букви, цифри и/или специални символи, които могат да се

[5] Под нормализация в БулСемКор се разбира корекция на формите, срещащи се в корпуса, и на техните леми в съответствие с правописните норми. Освен типографските грешки тук се включва заместване на думи с остарял правопис със съвременни варианти (напр. *чувствувам* с *чувствам*), заместване на русизми, архаизми, остарели думи, оказионализми, авторови неологизми и др. Разновидност на нормализацията представлява и определянето на границите на несвободните фрази.

[6] Всеки от синонимите в дадено синонимно множество.

състоят от един или повече токъна. На този етап се извършват и допълнителни операции, като сливане и разделяне на думи според кодифицираните правила в съвременния български книжовен език, когато в корпуса се наблюдава разминаване с нормата.

Тъй като лематизацията се извършва на ниво графична дума, и несвободните фрази не се разпознават като такива. Именно в процеса на нормализация и анотация се решава дали даден израз да се лематизира като последователност от свободни думи, или като несвободна фраза. Така например при предварителната обработка думите *учебна* и *година*, съставящи несвободната фраза *учебна година*, автоматично са лематизирани поотделно (3):

(3) *<word l="**учебен**" w="**учебна**"/>*
*<word l="**година**" w="**година**"/>*

За да бъде изразът разпознат като съставно съществително, компонентите му трябва да се свържат и да получат обща лема. Често граматичната форма на някои от компонентите в лемите на несвободните фрази не съвпада с основната форма на съответстващата им свободна дума. Затова на този тип словосъчетания се приписва абстрактна лема (Савъри 2008), в която конституентите се поставят в съответната форма. Например в *учебна година* опората изисква съгласуване по род на прилагателното със съществителното, поради което е необходимо то да се постави във формата за женски род в лемата на несвободната фраза. Самата лема има следния вид (4):

(4) *<word l="учебна" p="140425607156816:0" w="учебна"/><word l="година" p="140425607156816:1" w="година"/>*

Съотнасянето със синонимните множества, съдържащи литерала *учебна година*, се извършва чрез еднаквата стойност на атрибута l в БулСемКор (4) и тага <LEMMA></LEMMA> в БулНет (5):

(5) *<LITERAL>/учебна/ година<LEMMA>**учебна година**</LEMMA> </LITERAL>*

При определяне на основната форма на несвободните фрази в БулСемКор и при въвеждането им в БулНет се спазва принципът за минималност и неутралност на съставната лема (Тодорова 2010). Например някои съставни предлози, като *от страна на*, *в полза на*, следвани от име, длъжност и под. на лице, се появяват и във форма с притежателно местоимение – *от негова страна*, *в негова полза*. Съгласно посочения принцип лемата на съответното съчетание в корпуса се привежда към неутралната форма.

Именувани същности от типовете *наименование*, *дата* и *число* се лематизират по един от двата възприети начина: (а) с обобщена лема в зависимост от категорията, която обозначават – *дата*, *число*, *идентификационен номер*, *буква* и т.н., когато в WordNet не съществува съответното синонимно множество и преценката на анотатора е, че такова множество не следва да се създаде в БулНет (6); (б) с лема, която осигурява връзка с подходящо синонимно множество, когато именуваната същност е въведена или следва да се въведе в БулНет (7):

(6) *<word l="**фамилия**" w="**Петров**"/>*
*<word l="**дата**" w="**20.10.2010**"/>*

(7) *<word l="**Дикенс**" w="**Дикенс**"/>*
*<word l="**11 септември 2001**" w="**11 септември 2001**"/>*

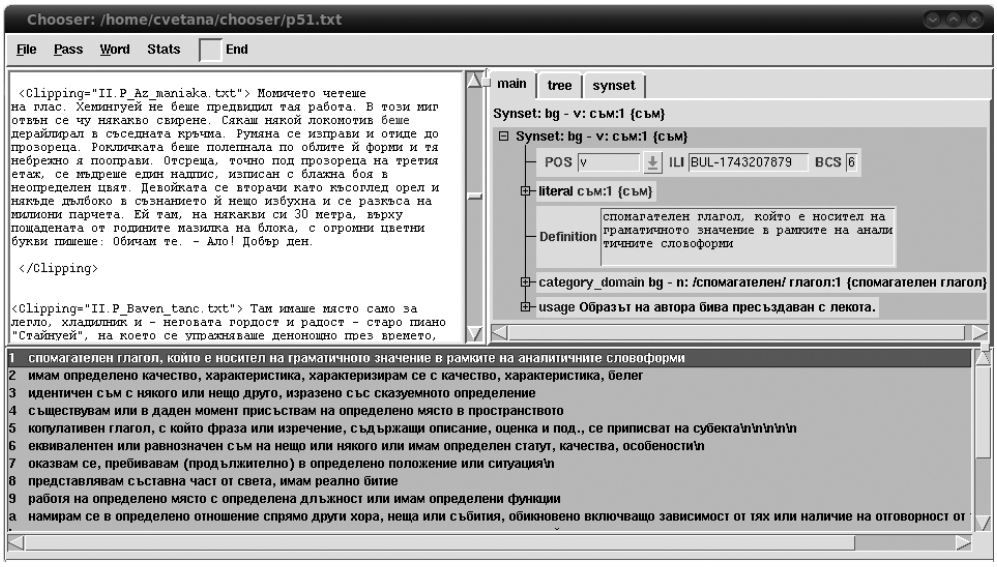
В резултат от коректната и еднозначна лематизация всяка анотирана единица в БулСемКор се асоциира с канонична основна форма, проверена от експерт.

4.2. Семантична анотация

Анотацията на БулСемКор се осъществява с помощта на версия на системата за анотиране на корпуси Chooser (Ризов 2010: 43–50), разработена за целите на отстраняването на семантичната многозначност. Програмата е независима от операционната система на потребителя и хардуера на компютъра. Тя е мултифункционална и лесно приложима за различни нива на езикова анотация, като предлага бърз и лесен достъп до информацията, необходима на анотатора при избора от кандидати, налични в базата от данни. Chooser интегрира различни стратегии за навигация по текста, който се анотира: обхождане на (а) всички единици; (б) на неанотираните единици; (в) на всички срещания на конкретна дума. Основните функции на системата включват: (а) бърз и лесен избор на значение; (б) лингвистична информация за множеството от кандидати, от които анотаторът избира; (в) свързване на съседни или дистантно разположени компоненти на несвободни фрази. Системата Chooser използва и редица операции за редактиране на корпуса: (а) редакция на словоформа; (б) редакция на лема; (в) изтриване и добавяне на думи в самия текст, например при корекция на правописни и печатни грешки, включително слято, полуслято и разделно писане; (г) редакция на неправилно форматиран текст.

Процесът на семантична анотация се състои в свързването на всяка коректно лематизирана лексикална единица или друг токън (буква, цифра, специален символ) в БулСемКор със синонимно множество в БулНет, което според преценката на анотатора най-добре съответства на значението, употребено в съответния контекст, при това определено еднозначно. Значенията, кандидати за анотация на дадена единица в БулСемКор, се представят в програмата за анотация под формата на списък от тълковните дефиниции на съответните синонимни множества в БулНет, измежду които анотаторът може да избира.

На фиг. 1 е представен основният изглед на програмата. Текстът за анотация е зареден в горния ляв прозорец. Използвайки някоя от стратегиите за навигация, анотаторът обхожда корпуса. При селектиране на дадена единица (маркираната форма *беше*) в долния прозорец се визуализира списъкът от тълковни дефиниции, съответстващи на кандидатите за анотация в БулНет, в случая значенията на глагола *съм*. След извършен избор съответното значение се маркира в друг цвят. Списъкът от значения за избор е синхронизиран с изгледа на съответните синонимни множества в БулНет, които се визуализират в горния десен прозорец. Унифицираният изглед на данните улеснява анотацията, тъй като осигурява достъп до пълната лингвистична информация за дадено синонимно множество, а оттам – позволява на анотаторите да прилагат процедури за избор на значение, базирани на различни типове информация, съдържаща се в БулНет.



Фиг. 1. Системата за семантична анотация Chooser

Когато в списъка за избор няма подходящ кандидат, анотаторът проверява дали синонимно множество, изразяващо засвидетелстваното в корпуса значение, съществува в Принстънския WordNet, който представлява основен източник за обогатяване на БулНет. Ако бъде намерено такова значение, то се въвежда в БулНет, в противен случай се създава ново синонимно множество без съответствие в Принстънския WordNet, което се включва в структурата на лексикалносемантичната мрежа в съответствие с възприетите принципи. Веднага след добавяне на ново множество в БулНет списъкът с кандидати за анотация в Chooser се актуализира и съответното значение може да бъде избрано.

В процеса на семантична анотация БулНет бе обогатен с редица езиково и/или културно специфични синонимни множества, както и със семантични множества от служебните части на речта. Примери за езиково специфични значения са относителните прилагателни в български, на които в английски съответства атрибутивна употреба на съществителното: *тухлен*, *студентски*; сложни прилагателни и наречия без съответствие в английски: *културно-етнически*, *социокултурно*; някои абстрактни съществителни и субстантиви, като *образност* и *духовно*; деминутиви и аугментативи; редица класове глаголи – ингресивни[7]: *заговорвам – заговоря*; финални: *дописвам – допиша*; атенуативни: *пооздравявам – пооздравея*; делимитативни: *помълчавам – помълча*, и др. Методологията и принципите за обогатяване на БулНет със синонимни множества без съответствия в Принстънския WordNet са представени от: Е. Търпоманова за съществителните имена (Търпоманова 2010а) и местоименията (Търпоманова 2010б), Хр. Кукова за прилагателните имена (Кукова 2010а) и наречията (Кукова 2010б), Св. Лесева за глаголите (Лесева 2010), М. Тодорова за несвободните фрази (Тодорова 2010), Цв. Димитрова за служебните думи (Димитрова 2010).

[7] Термините следват Иванова (1974).

Процедурата за избор на най-подходящото измежду наличните значения се основава на няколко критерия: (а) заменимост на дадената лексема с един или повече синоними от определено синонимно множество; (б) съответствие на значението в контекста с тълковната дефиниция на синонимното множество; (в) позиция на синонимното множество в структурата на лексикалносемантичната мрежа; (г) съответствие на значението с илюстрираното значение в примерите за употреба и др.

Тестът на замяна със синоними е един от най-важните за разграничаване на едно значение от друго, но невинаги е достатъчен критерий, тъй като може да се наблюдава пълно съвпадение на членовете на различни множества, непълнота на синонимното множество и т.н.

Принципите, следвани при съставянето на тълковните дефиниции в БулНет, са изложени подробно у Коева (2010б: 25). Предназначението на дефинициите в лексикалносемантичната мрежа е не само да характеризират изчерпателно значението на конкретно синонимно множество, но и ясно да разграничат отделните значения на дадена дума.

Описателните дефиниции се структурират на базата на генерализиращи и диференциращи признаци. Като генерализиращ признак най-често служи непосредствен или опосредстван (отдалечен на повече от една стъпки) хипероним на множеството, по-рядко холоним (дума, изразяваща цялост или клас, като разглежданото значение обозначава негова част/член) (8):

(8) <LITERAL>*сокол*</LITERAL><LEMMA>*сокол*</LEMMA>
<DEF>*представител на едноименния род (Falco) космополитни дневни грабливи птици от семейство Соколови, от дребни до средно едри, с разнообразна окраска; ловуват насекоми, птици, дребни бозайници и др.*</DEF>

Дефинициите може да съдържат селективни ограничения (9), (10):

(9) <LITERAL>*мъркам*<LITERAL><*LEMMA*>*мъркам*</LEMMA>
<LITERAL>*преда*<LITERAL><*LEMMA*>*преда*</LEMMA>
<DEF>*за котка – издавам гърлени вибриращи звуци, подобни на звука на вретено, обикновено в израз на удоволствие*</DEF>

(10) <LITERAL>*чудесен*<LITERAL><*LEMMA*>*чудесен*</LEMMA>
<LITERAL>*прекрасен*<LITERAL><*LEMMA*>*прекрасен*</LEMMA>
<DEF>*за време (метеорологични условия) – който е много приятен, или за ден и под. – който се характеризира с приятни за сетивата или подходящи за определена цел атмосферни условия*</DEF>

Важна особеност на БулНет в съпоставка с лексикографските ресурси е, че се отдава предпочитание на описателните и на структурно-описателните дефиниции. Поради организацията на лексикалната информация в синонимни множества е прието да не се използват синонимни дефиниции. Спазването на тези принципи гарантира информативността на дефинициите и улеснява семантичната анотация.

Синонимните множества, с които даден кандидат за анотация влиза в определени релации, както и самият тип на релациите в редица случаи улесняват избора. Отчетлив пример за това представлява принадлежността на синонимни множества със съвпадащи членове към различни таксономични класове като

естествени обекти и артефакти: *чело* (част от лицето) и *чело* (музикален инструмент); артефакти от различен тип, например части от конструкцията на сгради и на превозни средства: *прозорец*, *покрив*, *врата*; институции и сградите, в които се помещават: *училище*, *университет*, *болница*; организации, населени места, помещения и т.н., както и съвкупността от хора, които ги обитават или се намират в тях: *село*, *стая*, *офис*; съдове и тяхната вместимост: *бутилка*, *чаша*, *кутия* и др. На свой ред различната позиция на синонимните възли в структурата се отразява в дефиницията.

При избора между различни значения примерите за употребата на членовете на дадено синонимно множество имат помощна функция, като илюстрират и поясняват допълнително конкретното значение. В процеса на работа стремежът бе към наличните примери в БулНет, създадени чрез превод от Принстънския WordNet, конструирани от анотаторите или ексцерпирани от корпуси, да се добавят и засвидетелствани в БулСемКор употреби. Съвкупността от примерни употреби сама по себе си представлява корпус от изречения и изрази, в който значението на илюстрираните думи е еднозначно определено.

В таблица 1 са представени общият брой на токъните в БулСемКор след нормализацията и броят на анотираните токъни и разпределението им между несъставни единици: прости и сложни лексеми и други токъни ((поредици от) букви, цифри и/или специални символи и под.), от една страна, и несвободни фрази (включително поредици от токъни, съдържащи комбинации от букви и/или цифри и/или специални символи и под.), от друга. Разликата между броя на анотираните токъни и общия брой на токъните се дължи на това, че част от тях са пунктуационни знаци (например тире) или имат служебна функция. Несвободните фрази (вж. 4.4) представляват 5,83 % от общия брой анотирани токъни, а средната им дължина е 2,18 токъна.

Таблица 1
Брой на токъните и анотираните единици

Брой токъни	Анотирани токъни	Несъставни единици	Несвободни фрази (НФ)	Токъни в състава на НФ
101 062	99 480	86 842	5797	12 638

Анотираните xml файлове имат вида, представен в (11). Значението е еднозначно определено чрез стойността на атрибута ‘s’. Всички компоненти в състава на несвободна фраза получават една и съща стойност за атрибута ‘p’ (parent). Анотираните думи съдържат още информация за анотатора (атрибута ‘u’) и за времето, когато е извършена анотацията (‘t’). Информация за края на изречението се съдържа в атрибута ‘e’.

(11) *<word l="затова" p="-1529023764" s="1100001720" t="1298483182" u="*zarka@192.168.22.3*" w="затова"/>*

Както беше посочено, анотираната единица наследява цялата езикова информация, асоциирана със съответното синонимно множество: частта на речта (вж. 4.3); описателната дефиниция; примерите за употреба; бележките за граматични, семантични и прагматични ограничения при един или повече членове на

синонимното множество или на синонимното множество като цяло; множество от семантични, морфосемантични и извънезикови релации, принадлежащи към синонимното множество; множество от семантични и деривационни релации, свързани с даден литерал. Чрез уникален идентификатор (ID) всяко синонимно множество е съотнесено със съответствието му в Принстънския WordNet и езиковата информация, асоциирана с него, а опосредствано – и със съответните синонимни множества в други езици.

4.3. Морфосинтактична анотация

Заедно със семантичната анотация на всяка лексема се приписва и морфосинтактична информация чрез тага за част на речта, наследена от задължителния за всяко синонимно множество атрибут POS. Стойността на този атрибут се приписва автоматично при лематизацията или от анотаторите при нормализацията на корпуса. По този начин частта на речта на всяка лексема в БулСемКор е еднозначно определена. За разлика от други корпуси, в които морфосинтактичната многозначност е отстранена на ниво графична дума, какъвто е БулПосКор (вж. тук статията на М. Тодорова и Р. Декова), в БулСемКор наред с таговете, които се приписват на компонентите на несвободните фрази, цялата фраза също получава еднозначно характеризирана за част на речта лема в зависимост от семантиката и функцията си. Коректността на морфосинтактичната анотация е допълнително проверена чрез семантичната, тъй като погрешно определяне на частта на речта на дума в корпуса би довело до съотнасянето ѝ със синонимни множества от друга част на речта. Тъй като това не е представлявало цел на работата по БулСемКор, морфосинтактичната анотация не съдържа информация за граматичните характеристики на словоформите и в това отношение е по-малко детайлна от морфосинтактичното тагиране в БулПосКор, но е съвместима с него.

При разширяването на БулНет се допускат някои случаи на системна асиметрия по отношение на категоризацията на синонимните множества в Българския и Принстънския WordNet, които се отразяват в морфосинтактичната анотация.

1. Функционална подялба на числителните, които в българското езикознание традиционно представляват отделна част на речта, на прилагателни (POS=a) и съществителни (POS=n) според семантиката и функцията им, наложена от практиката в Принстънския WordNet.

2. Въвеждане на адвербиални изрази с част на речта ‘прилагателно’ (POS=a) или ‘наречие’ (POS=b) в зависимост от вида на фразата, която модифицират – NP или VP, заради различното им дефиниране в Принстънския WordNet. Например изразът *на живо* се анотира със семантично множество с POS=a в *предаване на живо* и с POS=b в *предават на живо*. За да се запази коректността към българската традиция, в забележка към множеството се посочва частта на речта според българското езикознание.

3. Въвеждане на местоименията, които са включени в Принстънския WordNet, като прилагателни, съществителни или наречия в зависимост от семантиката и функцията им. Местоименията, които нямат съответствие в Принстънския WordNet, се въвеждат като нови синонимни множества с част на речта ‘местоимение’ (POS=pron).

4. Въвеждане в БулНет на субстантивираните прилагателни имена и причастия без еквивалент в Принстънския WordNet като синонимни множества с част на речта ‘съществително’, като в забележка към литерала или множеството се посочват типът и спецификата им, съответно subst.[a] и subst.[participle].

5. Отнасяне на несвободните фрази към една или друга част на речта на базата на преводните им еквиваленти и синтактико-семантичната им функция при употреба в контекст (а не непременно на базата на опората на израза). Например фразеологизмът *вир вода* се състои от две съществителни, но е включен в синонимно множество с част на речта ‘прилагателно’ заедно с литералите *{мокър:1}* и *{прогизнал:1}*.

4.4. Анотация на несвободни фрази

Степента на композираност е основният маркер, който определя дали дадена комбинация от думи да се анотира като несвободна фраза и да се добави в БулНет, или не. Композираността се изразява в различни по характер ограничения върху морфосинтактичното поведение: ограничения върху заменимостта на всеки елемент; синтактична нерегулярност; изискване за предварително знание; ограничения върху перифразиране с една дума или с определена конструкция; словоредни разместванията и ограничена парадигматична продуктивност.

За разлика от частичната анотация на несвободните фрази по подвидове (идиоми, съставни думи и др.) според категориалната им принадлежност (именни, глаголни, адвербиални), компонентен състав и други критерии, характерна за съществуващите корпуси, предлагащи това ниво на анотация, БулСемКор осигурява пълно и последователно семантично и морфосинтактично тагиране на всички несвободни фрази. Следвани са основни критерии за: (а) определяне на степента на композираност между компонентите на несвободните фрази; (б) разпознаване на компонентите им независимо от възможните вариации в структурата и формите им в контекста; (в) идентифициране на семантичната им стойност, което предопределя установяването на преводните им еквиваленти и релациите, в които участват с други лексикални единици (а оттам и позицията им в структурата на лексикалносемантичната мрежа); (г) определяне на основната им форма и на морфосинтактичната им принадлежност при въвеждане в БулНет.

Семантичната анотация на несвободните фрази се извършва след приписването на абстрактна лема, отчитаща формалните ограничения, на които се подчиняват (вж. 4.1), и абстрактна част на речта, която взема предвид семантиката и функцията на съчетанието. При съотнасянето с подходящо синонимно множество в БулНет заедно с останалата лингвистична информация повечето несвободни фрази наследяват информация за това коя е главната и кои са зависимите части (подчинените думи се ограждат с наклонени черти) (12):

(12) *{цветарница:1; /цветарски/ магазин:1; магазин /за/ /цветя/:1}*

Несвободните фрази позволяват различни изменения в структурата, формите, словореда и компонентния си състав. Характерен случай представлява елипсата. Чрез нея ще илюстрираме необходимостта от прилагане на различен подход към анотацията на фразите в зависимост от това, дали са свободни или несвободни.

В свободните словосъчетания всяка дума се анотира индивидуално. Това важи и при координация (13), така че в съчетанието *добро и весело дете* се анотират самостоятелно думите *добро*, *и*, *весело*, *дете*:

(13) <word l="добър" w="добро"/><word l="и" w="и"/><word l="весел" w="весело"/><word l="дете" w="дете"/>

При несвободните фрази, когато се наблюдава изпускане на опора или друг конституент, в лемата се възстановява пълната (нередуцирана) форма. Ето защо в координационната фраза *прясно и кисело мляко* елементът *прясно* се лематизира като *прясно мляко*, а съчетанието *кисело мляко* – като *кисело мляко* (14):

(14) <word l="прясно мляко" w="прясно"/><word l="и" w="и"/><word l="кисел" p="-1525323956" pl="кисело мляко" w="кисело"/><word l="мляко" p="-1525323956" w="мляко"/>

Мотивацията за възприемане на този подход е различното поведение на координираните елементи на свободните и несвободните фрази. Докато първите могат да бъдат свободно заместени с несъставни или съставни синоними или близки по значение думи (например *добро и весело дете* може да се перифразира в *послушно и радостно дете*) и се превеждат самостоятелно (например *a nice and happy child*), то при несвободните фрази важат следните ограничения: (а) в някои случаи в същия контекст е възможна замяна на несвободната фраза или неин елемент с несъставен синоним, например *цветарски и месарски магазин* > *цветарница и месарница*; (б) преводът на несвободната фраза в други езици може да е несъставна дума, включително несродна, например: *прясно и кисело мляко* > *milk and yoghurt*, или дори да е несвободна фраза с друга опора и да не позволява елипса. Тези съображения налагат различното анотиране на елиптичните конструкции, съдържащи несвободни или свободни фрази.

За несвободните фрази е характерно развиването на самостойно лексикално значение от опората или от подчинена част, което е еквивалентно на значението на несвободната фраза. Примери за това са генерализацията/специализацията (15, 16)[8] и субстантивацията (17):

(15) *{операция:1; /хирургична/ операция:1;...}*
{операция:4; /военна/ операция:1}
{операция:6; /търговска/ операция:1}
{операция:7; /математическа/ операция:1}

(16) *{/духов/ /музикален/ инструмент:1; /духов/ инструмент:1}*

(17) *{мобилен телефон:1; мобилен:1}*
{детска стая:1; детска:1}

Тези особености на несвободните фрази поставят пред анотаторите два главни въпроса: (а) да идентифицират коректно границите на несвободната фраза в контекста и да я съотнесат с подходящото значение; (б) да въведат съответните лексеми в БулНет, когато това е необходимо. Малко засяган, но актуа-

[8] В конкретните примери не засягаме въпроса чрез кой от двата процеса са получени съответните значения.

Фиг. 2. Класификация на несвободните фрази в БулСемКор

лен както за разпознаването и анотацията, така и за лексикографското описание на несвободните фрази, е и въпросът за синонимията, илюстриран в (18):

(18) *{операция:1; /хирургична/ операция:1; /хирургическа/ операция:1; /хирургична/ намеса:1; /хирургическа/ намеса:1; /хирургична/ интервенция:1; /хирургическа/ интервенция:1; /оперативна/ намеса:1; /оперативна/ интервенция:1}*

С оглед на изследването на проблемите, свързани с функциите и степента на семантична, парадигматична и синтактична композираност на несвободните фрази, е извлечен списък на всички несвободни фрази в БулСемКор, от които към момента една трета са класифицирани ръчно.

Фигура 2 представя обобщена класификация (представена по-подробно у Тодорова 2010), основана на изследване на композираността на компонентите на съставните единици в корпуса. Делението на подкласове се основава на типа референция, като на тази основа се различават граматикализирани и лексикализирани несвободни фрази. Показани са и текущите резултати под формата на честотно разпределение на подкатегориите в БулСемКор, които ясно демонстрират необходимостта от последователна семантична анотация на граматикализираните и лексикализираните несвободни фрази, в които се включват и част от затворените класове думи.

4.5. Анотация на затворени класове думи

В последните години бе демонстрирано, че разнообразната многопластова лингвистична анотация има голямо значение в най-различни приложения в компютърната лингвистика, което налага разработване на нови или обогатяване на вече съществуващите езикови ресурси. Например при анализа на отношението на автора (Sentiment analysis) висока информационна стойност имат прилагателните, наречията и частиците, а при разрешаването на анафорите (Anaphora resolution) – местоименията. В дискурсния анализ при интерпретирането на из-

казванията се използват също периферни езикови изрази, каквито са частиците и междуметията. В приложенията, свързани с разбирането и генерирането на естествен език, важен аспект е дискурсивната структура, която включва установяването на релациите между простите изречения в състава на сложните, както и между самостойните изречения.

Осъзнаването на значимостта на служебните думи води до създаването на езикови ресурси, включващи съответната анотация. Такъв например е Пен Трийбанк (Палмър и др. 2000), в който семантичното анотиране обхваща не само пълнозначните думи, но също и: а) местоимения, на които е приписано значението на референта там, където е възможно; б) някои типове предложни фрази (въвеждани от предлозите *in* в значение за време, място и начин и *to* за посока). В Дискурсивната банка, част от Пен Трийбанк, е добавена анотация на дискурсивни конектори, които включват съчинителните и подчинителните съюзи и адвербиални изреченски модификатори (Милцакаки и др. 2004).

По-долу ще се спрем по-подробно на предлозите и съюзите в БулСемКор.

4.5.1. Анотация на предлозите

Предлозите изразяват релации между части на изречението. Абстрактността и комплексността на тези отношения и високата степен на полисемия обуславят трудностите при анализа им, което отчасти обяснява относително слабото изучаване и приложение на предлозите в компютърната лингвистика. Наред с това изследванията най-вече в областта на Семантиката на фреймовете разкриват богатата семантична съчетаемост на предлозите при въвеждането на аргументи и адюнкти и способността на различни предлози да свързват елементи с една и съща семантична роля (Литковски, Харгрейвз 2007). В резултат от това все повече се осъзнава важността на тази част на речта както при семантичния анализ – при автоматичното приписване на семантична роля (semantic role labeling), така и при синтактичния анализ, например при определянето на мястото на присъединяване на предложните фрази (PP attachment).

Целта на проекта за описание на предлозите Preposition Project (Литковски, Харгрейвз 2005) е да се дефинират значенията на английските предлози с оглед на обработката на естествения език. Всяко значение се характеризира от гледна точка на семантичната роля и на синтактичните и семантичните характеристики на комплемента. Друг езиков ресурс, който предлага описание, макар и на част от предлозите, е ФреймНет (Рупенхофър и др. 2010). В Българския ФреймНет (Коева 2010в) е възприето да се посочва предлогът или множеството от предлози, които въвеждат аргумент или адюнкт на даден предикат, като се използва наборът от синонимни множества на предлозите в БулНет.

При анотацията в БулСемКор значенията на предлозите, а и на останалите служебни думи, са дефинирани основно на базата на срещанията им в корпуса. Значения, незасвидетелствани в него, са въвеждани с оглед на последователното разграничаване на значенията на даден предлог и на коректната анотация. В резултат от приложения подход наборът от използвани значения на предлозите се характеризира с по-висока степен на гранулираност, отколкото в тълковните речници. В таблица 2 е представен броят на значенията в БулНет в съпоставка с няколко лексикографски източника.

Таблица 2
Брой значения за най-многозначните предлози в БулНет в съпоставка с лексикографски източници и граматики

Предлог	**на**	**с**	**по**	**за**	**в**	**от**
БулНет	54	42	41	40	35	35
РБЕ (нови и редактирани томове)	38	–	29	–	12	28
БТР	17	11	9	9	7	8
ГСБКЕ	11 (18)	5 (18)	6 (13)	11 (24)	8 (15)	9 (16)

Наблюдава се значителна разлика между броя на значенията в различните ресурси. В „Български тълковен речник“ от Л. Андрейчин и колектив (БТР 2002) семантиката на предлозите е представена в по-обобщен вид, като значенията са в голяма степен съотносими с дефинираните в т. 2 на „Граматика на съвременния български книжовен език“ (ГСБКЕ 1983) по-общи значения, например пространствени, темпорални, причинни и др. (извън скобите), или с техните по-частни разновидности (в скобите) в зависимост от степента на гранулираност. Характерно е, че в новоизлезлите и редактираните томове на „Речник на българския език“ (РБЕ, т. 1–14, 1977–2012), където също се прилага корпуснобазиран подход, броят на значенията е значително по-голям, отколкото в БТР (2002) и ГСБКЕ (1983). При въвеждането на предлозите в БулНет са анализирани значенията в посочените ресурси, както и в различни преводни и тълковни речници на английския език.

Наблюдаваните разлики при дефинирането на набора от значения произтичат в голяма степен и от спецификата на тази част на речта. Предлозите означават релации между предикат и аргумент(и) или между предикат и адюнкт(и) (Ницолова 2005: 151; Ницолова 2008: 457–459). Много от тях са развили както по-тесни и специфични, така и високоабстрактни значения, обобщаващи различни по обхват класове от аргументи, много от които съотносими с определени семантични роли (19). Наблюдавана е и тенденция някои предлози да се превръщат в универсални маркери за аргументи, които се проектират в определена синтактична позиция независимо от конкретната семантична роля. Пример за това е *на*, използван за въвеждане на непряк обект (Ницолова 2005: 144–151; Ницолова 2008: 457–459, примерите в (20) са аналогични на представените у Ницолова 2005; Ницолова 2008):

(19) *Получих писмо* ***от Иван***. (адресант)
Жените са ***от Венера***. (източник)
Книгата е написана ***от Маркес***. (агенс в пасив)
Тя е по-голяма ***от сестра си***. (екватив)
Детето се страхува ***от паяци***. (стимул)

(20) *Купиха подарък* ***на детето***. (реципиент)
Заявих ***на полицая****, че това не е редно*. (адресат)
Подаръкът се хареса ***на детето***. (експериенцер)
Любуват се ***на природата***. (стимул)

Освен това предлозите въвеждат адюнкти и/или аргументи с темпорална, локативна и друга адвербиална семантика (21):

(21) *Работиха* ***на нивата***. (пространство)
Книгата е ***на масата***. (разположение спрямо повърхност)
Животните мигрират ***на юг***. (направление на движение)
Корабът акостира ***на десет мили от брега***. (разстояние)
Хапнахме ***на крак***. (начин)
Заминахме ***на сутринта***. (време)

Някои значения на предлозите въвеждат тясно дефинирани класове от аргументи, селектирани от глаголи с близка семантика (22, 23):

(22) *Минава* ***за шпионин***.
Представя се ***за шпионин***.
Мислят го ***за шпионин***.

(23) *Пада си* ***по авантюрите***.
Луда е ***по филмите на ужасите***.
Не съм ***по тази част***.

В определени случаи комбинацията от глагол и предлог може да се разглежда като идиоматизирана, тъй като съставките ѝ не могат да се интерпретират индивидуално. Това се наблюдава най-добре при съпоставка със свободни съчетания от глагол и предлог (24, 25):

(24) *Икономиката* ***се радва на*** *бум.* – ‘изживявам благоприятен развой’
Радвахме се *на хубавото време.* – ‘наслаждавам се’

(25) *Нищо* ***не разбра от*** *живота.* – ‘извличам полза или наслада от нещо’
Какво ***разбра*** *от урока?* – ‘проумявам, схващам’

Не на последно място предлозите свързват и аргументи на имплицитни предикати (Ницолова 2005: 144–146) – *трудът на родителите му*, *мнението на сестра му*, *книга на Набоков*, *шофьор на такси* (примерите са от Лесева 2010).

Отделните предлози са развили в различна степен абстрактни и пълнозначни значения и по различен начин концептуализират дадени отношения. Таблица 3, таблица 4 и таблица 5 показват данни за част от значенията на *на* и *в(ъв)* и тяхната честота в БулСемКор. Например първите пет значения на предлога *на* свързват аргументи с техния предикат или аргументи с имплицитен предикат в номиналната фраза. Значенията, изразяващи типични адвербиални отношения, имат многократно по-ниска честота. Обратно, петте най-често срещани значения на *в(ъв)* изразяват пространствени или абстрактни пространствени отношения.

Таблица 3
Петте най-често срещани значения на предлога ***на***

Значение на предлога *на*	**Пример**	**Брой**
Принадлежност, притежание, собственост	*книгата на детето*	1438
Обект на действие, изразено чрез съществително	*четенето на книгата*	904
Агенс, източник или друг субект на действие или състояние	*старанията на учителя*	493
Получател (адресат, бенефициент, реципиент и др.)	*изпрати писмо на Иван*	228
Отношение между вършител, изпълнител, притежател и неговия предмет на дейност, служба, титла, собственост	*шеф на партията* *шофьор на такси*	159

Таблица 4
Характерни обстоятелствени значения за предлога ***на***

Значение на предлога *на*	**Пример**	**Брой**
Място, пространство, в близост до което се извършва или се наблюдава посочената ситуация	*работя на нивата* *стоя на вратата*	117
Положение във времето – част от деня, определена дата или едновременност с протичащо събитие	*заминавам на сутринта* *на зазоряване*	62
Контактно разположение с горната повърхност на нещо	*чашата е на масата*	51
Направление, посока на движение	*Птиците отлитат на юг*	21
Характеризиране според начина на извършване	*свит на кълбо*	17
Разстояние от дадено място или точка	*отдалечен на десет мили*	12

Таблица 5
Петте най-често срещани значения на предлога ***в***

Значение на предлога *в*	**Пример**	**Брой**
Пространствено положение в границите или вътрешността на място, пространство или предмет	*свиреха в клуба* *спеше в леглото*	477
Област на дейност или проява на нещо	*известен в тези среди*	353
Преминаване във вътрешността или границите на дадено пространство, място или предмет	*влезе в стаята* *влетя в кафеза*	180
Положение във времето	*започваме в 10 часа*	104
Състояние или ситуация, в която някой или нещо се намира	*приет в тежко състояние*	78

Представените особености на предлозите са отчетени при определянето на значенията им и последвалото анотиране. В резултат от семантичната анотация в корпуса са анотирани отношенията между предикати и техни аргументи, предикати и техни адюнкти и имплицитни предикати и техни аргументи, когато представляват свободни синтактични съчетания, като дефиницията на тези релации е в голяма степен съотносима с дефиницията на семантичните роли. Когато предлозите са засвидетелствани в състава на несвободни фрази, се следват общите принципи и те не се анотират отделно.

4.5.2. Съюзи

Съюзите изразяват отношения между еднородни части в рамките на простото изречение, между еднородни или разнородни изречения в състава на сложното изречение или между цели изречения (Ницолова 2008: 460) и се характеризират с относително малък брой устойчиви значения. Това наблюдение се подкрепя и от данните в БулСемКор. Най-общо съчинителните съюзи изразяват определен набор от релации между синтактично равноправни единици – съединяване, съпоставяне, разделяне, обобщаване, често съчетани с отношения като едновременност, темпорална последователност (предходност или следходност) или причинно-следственост (Ницолова 2008). Подчинителните съюзи изразяват отношение на зависимост между синтактично главни и синтактично несамостойни единици. Традиционно значенията на съюзите се дефинират от гледна точка на релацията, която изразяват. Тези специфики са залегнали и при дефинирането на съюзите в БулНет, като са взети предвид тълкуванията на значенията им в съществуващите лексикографски ресурси. Не се наблюдават съществени разминавания както между отделните речници, така и между данните от корпуса и речниците.

Таблица 6 демонстрира значителното припокриване на значенията на подчинителните съюзи *да* и *че*, както и сходната подредба на най-често срещаните им значения.

Таблица 6
Значения на подчинителните съюзи ***да*** и ***че***, анотирани в корпуса

Значение на *да*	Брой	**Значение на *че***	**Брой**
допълнителни изречения	559	допълнителни изречения	398
определителни изречения	243	определителни изречения	122
подложни изречения	88	подложни изречения	76
за цел с глаголи за движение	63	сказуемноопределителни изречения	33
сказуемноопределителни изречения	44	за последица и заключение	28
за цел	25	за причина	11
за причина	4	за отстъпка	1

На анотираните съюзи в БулСемКор се приписва имплицитната информация за типа свързване (съчинително или подчинително) чрез релацията между синонимното множество, от което е част съответният съюз, и синонимното множество: *{съчинителен съюз:1}* или *{подчинителен съюз:1}*. Функцията на съюза се конкретизира в тълковната дефиниция.

4.6. Проверка на БулСемКор

Основна предпоставка за коректността и последователността на семантичната анотация е съгласуването на работата между анотаторите и описанието на езиковите явления, проблемите и стандартите за анотация в подробни анотационни конвенции. Изборът на подходящо значение е подпомогнат от информацията за относителната честота на употребата на отделните значения, достъпна посредством програмата Chooser въз основа на анотирания езиков материал.

Коректността, пълнотата и последователността на семантичната анотация се осигуряват чрез комбинирането на различни методи за избор на значение – линеен (текстов) и лексикален (пресичащ се) (Коева 2010б: 35–36). При първия се обхождат последователно всички думи в текстовете, а при втория – само срещанията на конкретна дума или значение. При анотацията на БулСемКор предимствата на линейния и на лексикалния метод са съчетани: (а) чрез линейния метод се извършва отстраняване на многозначността въз основа на по-широк контекст и информация за съчетаемостта на анотираната единица с думите от обкръжението ѝ, което, от една страна, улеснява избора на значение и за предходните и следходните думи, а от друга, спомага за правилното определяне на границите и компонентите на несвободните фрази; (б) чрез лексикалния метод се изследват различните контексти на анотираната дума в по-голям отрязък (част от корпуса или в целия корпус), което осигурява последователност при избора на значение. Вторият метод на обхождане е използван и при дефинирането, допълването и редактирането на множеството от значения на непълнозначните думи в БулНет, тъй като чрез него значително се улеснява разграничаването на значенията.

Комбинираният метод е приложен и при проверката на анотацията, извършена от същия или от друг анотатор в рамките на част от корпуса. Лексикалното обхождане има приоритет при валидирането на отделни лексеми, при които изборът е затруднен от големия брой значения и/или от близостта на (част от) значенията. По същия начин е извършена и проверка на анотацията на служебните думи след окончателното дефиниране на множеството от значенията им.

Коректността на анотацията на несвободните фрази в БулСемКор (описана в 4.4) е проверена в хода на изследването на тяхната композираност и класифицирането им по подтипове въз основа на извлечен от корпуса изчерпателен списък на несвободните фрази. Пропуски при анотацията, включително в случаите, когато са настъпили промени в синонимните множества (например изтриване на синоним от синонимно множество), са отстранени чрез функцията на програмата Chooser за обхождане по неанотираните думи и избиране на подходящо значение.

5. ЗАКЛЮЧЕНИЕ

Структурирането и анотацията на БулСемКор следват добрите примери в световната практика, като същевременно задават нови насоки за работа. Последователното анотиране на всички лексикални единици води както до създаването на анотирани ресурси, конвенции и методологии, така и до натрупването на теоретично знание и практически опит, служещи за разрешаване на редица проблеми в компютърната лингвистика. Изчерпателно анотираните корпуси, в които значенията на езиковите единици са определени в реалното им обкръжение, в широк контекст, са незаменим ресурс както за отстраняване на семантичната многозначност, така и за целите на автоматичния превод, автоматичното извличане на информация, автоматичния синтактичен и семантичен анализ и много други.

Основното предназначение на БулСемКор е като тренировъчен и тестов корпус при автоматичното отстраняване на семантичната многозначност за целите на автоматичния превод.

В резултат от паралелното протичане на анотацията и разширяването на анотационната схема успоредно със създаването на БулСемКор бе обогатяван и друг голям лингвистичен ресурс – БулНет. В хода на семантичното анотиране броят на синонимните множества в него е увеличен с около 50 % – от 21 000 в началото на проекта до 32 000 при завършването му. Същевременно е повишено и качеството му в две основни насоки: (а) синонимните множества са допълнени със синоними, тълковните дефиниции са прецизирани, примерите за употреба са обогатени, извършени са други видове редакция; (б) в резултат от подбора на текстове в корпуса значителна част от нововъведените синонимни множества включват засвидетелствани в реален контекст високочестотни лексеми в съвременния български език, като при това дефинирането на значенията се основава на езиковата употреба.

БулСемКор има голямо приложение за разширяване на различни типове речници с нови значения, регистрирани в корпуса, както и с неописани в лексикографските ресурси несвободни фрази. Изучаването на структурата и особеностите на несвободните фрази би подпомогнало и разработването на модели за автоматичното им разпознаване.

Разпознаването и класифицирането на именувани същности представлява друга важна лингвистична задача, която има широко приложение в областта на автоматичния превод, автоматичното извличане на информация, автоматичното генериране на отговори на въпроси. Такъв тип езикови данни се съдържат и в семантично анотирания корпус, като част от тях са анотирани според категорията, към която принадлежат, чрез обобщена лема.

Анотираните данни са отправна точка за разработването на модели за семантичен анализ. Така например информацията за семантичния клас на анотираните предикати и техните аргументи и адюнкти (наследена от WordNet) в съчетание с релациите, изразявани от предлозите и съюзите, и онтологичния тип на адвербиалните обкръжения на предикатите позволява изследването и формализирането на семантичните релации между участниците в ситуациите и дефинирането на когнитивно валидни селективни ограничения.

Корпусът съдържа частично синтактично анотирани данни (в рамките на несвободните фрази) и информация за структурата и словоредните им особености, които представляват ценен ресурс както за формалното синтактично описание на несвободните и свободните словосъчетания, така и за автоматичния синтактичен анализ.

Анотацията на съюзите има голямо приложение при подобряване на резултатите от автоматичното сегментиране на простите изречения в състава на сложните, при определяне на релациите между двойки прости изречения и при синтактичния анализ. Друга сфера на употреба е описанието на логическите и дискурсивните отношения между изреченията за целите на разбирането, генерирането и опростеното представяне на езиковите структури.

Посочените приложения на БулСемКор очертават (без изчерпателност) част от областите, в които корпусът се използва или би могъл да намери приложение. Многопластовата и разнообразна анотация на езиковите данни предоставя възможности за разрешаването на много други научни и практически задачи.

ЛИТЕРАТУРА

Агире и др. 2006: *Agirre, E., I. Aldezabal, J. Etxeberria, E. Izagirre, K. Mendizabal, M. Quintian, E. Pociello.* A methodology for the joint development of the Basque wordnet and Semcor. – In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy, 2006.

БТР 2002: *Андрейчин, Л., Л. Георгиев, Ст. Илчев, Н. Николов, Ив. Леков, Ст. Стойков, Цв. Белчев.* Български тълковен речник. Допълнен и преработен от Д. Попов. София: Наука и изкуство, 2002.

Восен и др. 2012: *Vossen, P., A. Görög, R. Izquierdo.* DutchSemCor: targeting the ideal sense-tagged corpus. – In: Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12). Istanbul, Turkey, 2012.

ГСБКЕ 1983: *Андрейчин, Л., П. Асенова, Ел. Георгиева, К. Иванова, Р. Ницолова, П. Пашов, Хр. Първев, Р. Русинов, В. Станков, Ст. Стоянов, Кр. Чолакова.* Граматика на съвременния български книжовен език. Том 2. Морфология. София: БАН, 1983.

Димитрова 2010: *Димитрова, Цв.* Лингвистични конвенции при анотация на затворените класове. – В: *Коева, Св.* (съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 141–165.

Иванова 1974: *Иванова, К.* Начини на глаголното действие в съвременния български език. София: БАН, 1974.

Коева 2010а: *Коева, Св.* (ред. и съст.). Българският семантично анотиран корпус. София, 2010.

Коева 2010б: *Коева, Св.* Българският семантично анотиран корпус – теоретични постановки. – В: *Коева, Св.* (ред. и съст,). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 7–42.

Коева 2010в: *Коева, Св.* Българският ФреймНет 2010. София, 2010.

Коева и др. 2006: *Koeva, S., S. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova.* Bulgarian Tagged Corpora. – In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, 18–20 October 2006, Sofia, Bulgaria. Sofia, 2006, pp. 78–86.

Кукова 2010а: *Кукова, Хр.* Лингвистични конвенции при анотация на прилагателните. – В: *Коева, Св.* (съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 108–118.

Кукова 2010б: *Кукова, Хр.* Лингвистични конвенции при анотация на наречията. – В: *Коева, Св.* (съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 119–125.

Ландс и др. 1998: *Landes, S., C. Leacock, R. I. Tengi.* Building semantic concordances. – In: *Fellbaum, C.* (ed.). WordNet: An Electronic Lexical Database. Cambridge (Mass.): The MIT Press, 1998.

Лесева 2010: *Лесева, Св.* Лингвистични конвенции при анотация на глаголите. – В: *Коева, Св.* (съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 82–107.

Литковски, Харгрейвз 2005: *Litkowski, K., O. Hargraves.* The Preposition Project. – In: ACL-SIGSEM Workshop on "The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications". University of Essex – Colchester, United Kingdom, 2005, pp. 171–179.

Литковски, Харгрейвз 2007: *Litkowski, K., O. Hargraves.* SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. – In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic. Association for Computational Linguistics, 2007, pp. 24–29.

Макенери, Уилсън 2001: *McEnery, A. M., A. Wilson.* Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University, 2001.

Мелчук 1995: Mel'čuk, I. Phrasemes in language and phraseology in linguistics. – In: *Martin Everaert, M., E.-J. van der Linden, A. Schenk, R. Schreuder* (eds.). Idioms: Structural and psychological perspectives. Hillsdale, N.J. and Hove, UK: Lawrence Erlbaum, 1995, pp. 167–232.

Милцакаки и др. 2004: *Miltsakaki, E., R. Prasad, A. Joshi, B. Webber.* Annotating discourse connectives and their arguments. – In: NAACL/HLT Workshop on Frontiers in Corpus Annotation. Boston MA, 2004.

Милър 1995: *Miller, G. A.* Building Semantic Concordances: Disambiguation vs. Annotation. – AAAI Technical Report SS-95-01, 1995, pp. 92–94.

Монтемани и др. 2000: *Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, R. Delmonte.* The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation. – In: Proceedings of the COLING Workshop on "Linguistically Interpreted Corpora (LINC – 2000)". Luxembourg, 2000, pp. 18–27.

Наваро и др. 2003: *Navarro, B., M. Civit, A. Marti, R. Marcos, B. Fernandez.* Syntactic, Semantic and Pragmatic Annotation in Cast3LB. – In: Corpus Linguistics 2003 Workshop on Shallow Processing of Large Corpora. Lancaster, UK, 2003.

Нг, Лий 1996: *Ng, H. T., H. B. Lee.* Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. – In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), 1996, pp. 40–47.

Ницолова 2005: *Ницолова, Р.* За значението на предложните съчетания в българското изречение. – В: *Коева, Св.* (съст.). Аргументна структура. Проблеми на простото и сложното изречение. София: СемаРШ, 2005, с. 141–152.

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

Палмър и др. 2000: *Palmer, M., H. T. Dang, J. Rosenzweig.* Sense Tagging the Penn Treebank. – In: Proceedings of LREC 2000. Athens, Greece, 2000.

Пасоно и др. 2010: *Passonneau, R., C. Baker, C. Fellbaum, N. Ide.* MASC: The Manually Annotated Sub-Corpus of American English. <http://www.icsi.berkeley.edu/pubs/ai/mascword12.pdf> [дата на достъп 7 юли 2013].

РБЕ 1977–2012: Речник на българския език. Т. 1–14. София: АИ „Проф. Марин Дринов“, ЕТ „ЕМАС“, 1977–2012.

Ризов 2010: *Ризов, Б.* Система за анотиране Chooser. – В: *Коева, Св.* (ред. и съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 43–50.

Рупенхофър и др. 2010: *Rupenhofer, J., M. Ellsworth, M. R. L. Petruck, Ch. R. Johnson, J. Scheffczyk.* Framenet II. Extended Theory and Practice. <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf> [дата на достъп 7 юли 2013].

Савъри 2008: *Savary, A.* Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. – Linguistic Issues in Language Technology, 1(2), 2008, pp. 1–53.

Сантамария и др. 2003: *Santamaria, C., J. Gonzalo, F. Verdejo.* Automatic Association of Web Directories with Word Senses. – Computational Linguistics, 29, 2003.

Телйохан и др. 2004: *Telljohann, H., E. Hinrichs, S. Kübler, R. Kübler.* The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. – In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).

Тодорова 2010: Тодорова, М. Съставни единици в Българския семантичен корпус. – В: *Коева, Св.* (съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 166–186.

Търпоманова 2010а: *Търпоманова, Е.* Лингвистични конвенции при анотация на съществителните. – В: *Коева, Св.* (съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 66–81.

Търпоманова 2010б: *Търпоманова, Е.* Лингвистични конвенции при анотация на местоименията. – В: *Коева, Св.* (ред. и съст.). Българският семантично анотиран корпус. София: ИБЕ, 2010, с. 133–140.

Франсис, Кучера 1979: *Francis, N., H. Kucera.* Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers. Department of Linguistics, Brown University, Providence, R. I., U.S.A., original ed. 1964, revised 1971, revised and augmented 1979. <http://icame.uib.no/brown/bcm.html> [дата на достъп 7 юли 2013].

Хенрих и др. 2012: *Henrich, V., E. Hinrichs, T. Vodolazova.* WebCAGe – A Web-Harvested Corpus Annotated with GermaNet Senses. – In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23–27, 2012. Avignon, 2012, pp. 387–396.