Български език 62 (2015), 3, 102-119

* * *

СВЕТЛА КОЕВА, ДИАНА БЛАГОЕВА, СИЯ КОЛКОВСКА,
ЦВЕТАНА ДИМИТРОВА, ИВЕЛИНА СТОЯНОВА,
СВЕТЛОЗАРА ЛЕСЕВА

БЪЛГАРСКИЯТ НАЦИОНАЛЕН КОРПУС В КОНТЕКСТА НА СЪВРЕМЕННАТА ЛИНГВИСТИКА[1]

SVETLA KOEVA, DIANA BLAGOEVA, SIA KOLKOVSKA,
TSVETANA DIMITROVA, IVELINA STOYANOVA,
SVETLOZARA LESEVA

THE BULGARIAN NATIONAL CORPUS IN THE CONTEXT OF CONTEMPORARY LINGUISTICS

(Abstract)

The paper offers an overview of the methodology adopted in the development of the Bulgarian National Corpus (BulNC), its structure, linguistic annotation and applications with a focus on contemporary lexicography. BulNC is a dynamic corpus of over 5.4 billion words consisting of a monolingual (Bulgarian) part and 47 parallel corpora. The Bulgarian part includes about 1.2 billion words. The foreign language texts (originals or translations of texts in the Bulgarian part) total about 4.2 billion words, which makes BulNC the largest parallel corpus with a focus on Bulgarian. The corpus's structure is extendable and allows enrichment with new texts and categories in a way that ensures the corpus's representativeness and balance with respect to various features. The annotation approach adopted in the creation of the corpus involves automatic linguistic annotation on different levels: lemmatisation and PoS-tagging for Bulgarian and English, and sentence- and clause-alignment for parallel texts. BulNC can be accessed via a web application with advanced options for search using lemmas, wordforms and other linguistic (and extralinguistic) criteria, and functionalities for extraction of collocations and concordances. BulNC has been used for computational and theoretical linguistic research, in comparative studies and other areas. Special attention is paid to the corpus's applications in contemporary computational lexicography, with a focus on its use in the development of various dictionaries such as the comprehensive Dictionary of the Bulgarian Language, the Bulgarian FrameNet, the Bulgarian WordNet.

Keywords: Bulgarian National Corpus, parallel corpora, corpus design, corpus annotation, computational linguistics, computational lexicography

1. Общо представяне на *Българския национален корпус*

Българският национален корпус (*БНК*) е динамичен корпус, съизмерим по обем със световните стандарти. Структурата му позволява не само обогатяване с нови текстове, но и разширяване на модела с нови категории, така че да е възможно извличането на подкорпуси, отговарящи на критериите за представителност и балансираност по отношение на различни показатели. Част от текстовете имат преводни еквиваленти на други езици, в резултат на което *БНК* представлява най-големият паралелен корпус с фокус върху българския език. Общият му обем надхвърля 5,4 млрд. думи, като българската част от корпуса съдържа над 1,2 млрд., а паралелните корпуси – около 4,2 млрд. думи, разпределени между 47 езика.

1.1. Теоретична обосновка

Съставянето на едноезикови и многоезикови корпуси има дългогодишна традиция и намира множество приложения. В последните години се наблюдава благоприятно съчетаване на възможности: все по-голямо количество достъпни и разнообразни по вид документи в интернет и все по-развити технологии за извличане и обработване на информацията, което позволява автоматично или полуавтоматично конструиране на езикови корпуси със значителна големина.

1.1.1. Принципи на създаване на едноезикови и многоезикови корпуси

Актуална практика и при едноезиковите, и при многоезиковите корпуси е стремежът към създаване на ресурси с все по-голям обем, които обхващат разнообразие от стилове, тематични области и жанрове. Тази тенденция е мотивирана от съображението, че колкото по-голям е един корпус, толкова по-голяма е вероятността той да съдържа достатъчно емпирични данни, така че да се илюстрират специфични или редки за даден език явления, на базата на които да се правят статистически валидни заключения (Banko, Brill 2001; Keller, Lapata 2003; Kilgarriff, Grefenstette 2003).

Съвременните едноезикови корпуси съдържат от няколкостотин милиона до няколко милиарда думи. В Таблица 1. са представени количествени данни за някои от по-известните корпуси.

Корпус	**Обем**
Британски национален корпус (BNC) http://www.natcorp.ox.ac.uk/	100 млн.
Хърватски национален корпус (HNK) http://www.hnk.ffzg.hr/cnc.htm	101,3 млн.

Корпус на съвременния американски английски (COCA) http://corpus.byu.edu/coca/	над 450 млн.
Национален корпус на руския език (НКРЯ) http://ruscorpora.ru/index.html	над 500 млн.
Национален корпус на полския език (NKJP) http://nkjp.pl/ (Bański, Przepiórkowski 2010)	около 1 млрд.
Словашки национален корпус (SNK) http://korpus.juls.savba.sk/index_en.html	1,155 млрд. (януари 2013 г.)
Чешки национален корпус (ČNK) http://ucnk.ff.cuni.cz	1,3 млрд. (2010 г.)
Оксфордски корпус на английския език (OEC) http://oxforddictionaries.com/words/the-oxford-english-corpus	2,3 млрд.
Германски референтен корпус (DeReKo) http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html (архив)	5,4 млрд.

Таблица 1. Едноезикови корпуси за различни езици (в брой думи)

В последните години чрез автоматично извличане на текстове от интернет бяха създадени редица корпуси, чийто обем в отделни случаи достига стотици милиарди думи. Пример за това е *Google Books Corpus*, чиято англоезична част надвишава 200 млрд. думи[2].

Продължава да е актуален и въпросът за оптималното съотношение между големината на корпуса и размера на отделните текстове, които го съставят. Няма създадена методология, която да осигурява предписания за оптималния размер и адекватната структура на корпусите – очевидно е, че тези характеристики са пряко следствие от предназначението на всеки конкретен корпус. Оптималната големина на корпуса би трябвало да осигурява лексикално, граматично и стилистично разнообразие по отношение на тематични области, жанрове или езикови явления, докато критерият за размера на отделните текстови единици взема под внимание необходимостта от мотивирано съотношение между текстовете.

Представителността (достоверна илюстрация на езиковата употреба) и балансираността (съотношението между текстовете, които принадлежат към различни стилове, жанрове, тематични области) на корпусите са пряко свързани с тяхната големина. Въпреки опитите за обективно дефиниране на тези категории (Atkins 1992; Biber 1993; Sinclair 2005), все още не са установени надеждни критерии за оценката им, а някои автори дори ос-

порват основателността им (Kilgarriff, Grefenstette 2003; Kupietz et al. 2010). Обикновено при компилацията на даден корпус се следва специално създадена класификация, която включва различни категории и различно съотношение на текстовете в дадена категория. Не съществува консенсус и по отношение на начина, по който се дефинират характеристиките на корпуса. Например балансираността се определя като равномерно количествено застъпване на предварително избрани категории (Davies 2010), като пропорционално разпределение на текстовете според стилистични (Przepiórkowski et al. 2010), социологически (Čermak, Schmiedtová 2003), маркетингови (Tadić 2002) и други характеристики, или по други критерии. Остава открит и въпросът какво е взаимодействието между отделните характеристики и как да се осигури оптимално отношение между тях.

Тези въпроси са особено актуални при многоезиковите корпуси поради ограниченото количество паралелни текстове за много двойки езици. Голяма част от най-широко използваните многоезикови корпуси представляват колекции от текстове от определен жанр или тематична област, създадени за конкретни цели. Такива са корпуси като *EuroParl* (Koehn 2005), съдържащ стенограми от заседанията на Европейския парламент; редица корпуси от административно-юридически текстове като *JRC-Acquis* (Steinberger et al. 2006) и *Административния корпус на SEE-ERA.NET* (Tufiş et al. 2009), съдържащи текстове от Acquis communautaire.

При създаването на други корпуси се използват преводи на специално подбрани текстове – най-често това са корпуси с художествена литература. Такива са *Multext-East* (Dimitrova et al. 1998), който съдържа романа на Джордж Оруел „1984“ на шест езика, както и *Литературният корпус на SEE-ERA.net* (Tufiş et al. 2009) с романа на Жул Верн „80 дни около света“ в превод на шестнайсет езика.

През последните години, поради улеснения достъп до все повече и поразнообразни данни в интернет, се създават големи корпуси, които същевременно се стремят към представителност и балансираност. *Чешко-английският паралелен корпус* (*CzEng*) съдържа над 400 млн. токъна, разпределени в текстове от седем тематични области – художествена литература, законодателство на ЕС, субтитри на игрални филми, уебстраници с паралелно съдържание, техническа документация, новини и др. (Bojar et al. 2012). *Hunglish* е корпус, съставен от паралелни унгарски и английски текстове от областите литература, религия, международно право, субтитри, софтуерна документация, списания и отчети на компании (Varga et al. 2005). *Полско-руският паралелен корпус*[3] съдържа класическа и съвременна литература, журналистически и юридически текстове.

В общия случай както едноезиковите, така и многоезиковите корпуси са анотирани по част на речта, а в някои се приписва и синтактична, семантична и друга анотация. Достъпът се осъществява чрез системи за търсене, които в зависимост от нивата на анотация и детайлността на мета-

данните, от една страна, и от езика за търсене, от друга, позволяват изпълнение на различни по сложност заявки.

Придържането към традиционния подход при създаването на корпуси, според който се предпоставя модел, изграден на базата на обективни и/или субективни фактори, поставя редица проблеми. Тъй като не съществува единство по въпроса какъв да е моделът, а различните изследователски задачи може да изискват различни модели, обикновено се налагат практически съображения за структурата и състава на корпуса. В контекста на динамично развиващите се езикови технологии и огромния обем от достъпни текстове в интернет изискванията към основните характеристики на корпуса – размер, балансираност и представителност – могат да бъдат преразгледани. Подробен критичен преглед на актуалната проблематика при съставянето на езикови корпуси и богата библиография по въпроса са представени у Коева и др. (Koeva et al. 2012c).

1.1.2. Принципи за създаване на *БНК*

Съвременните технологии предоставят огромни възможности за откриване на информация; обхождане на интернет; фокусирано автоматично събиране на документи; класифициране на текстовете към определен стил, жанр, тематична област; автоматично извличане на разнообразна информация (автор, заглавие и др.), както и на информация, която не е така видима – например кои са документите с по-голяма честота на съставни думи или неправилно употребени транскрибирани имена и т.н.; автоматично обогатяване на документите с лингвистична анотация – част на реч-та, основна форма, значение, в което е употребена думата, фраза и изречение, в които се съдържа думата, преводни еквиваленти, съотносима употреба на думи, фрази и изречения в паралелни документи на други езици. Досега не ни е известна практика, която да обединява всички възможности.

Подходът, който приемаме (Koeva et al. 2012b; Коева и др. 2012), се основава на следните положения: 1) по-големите корпуси предоставят по-големи възможности за езиков анализ, независимо от спецификата на поставената задача; 2) обогатяването на документите с допълнително съдържание (подробни метаданни и разширена лингвистична анотация) позволява извличане на разнообразни подкорпуси, представителни за илюстрация на определено явление и балансирани по определен начин; 3) представителността не може да бъде универсално определена, а се основава на обхвата и разнообразието на категориите, спрямо които се класифицират текстовете, което предпоставя изработването на методология за постигане на гъвкава йерархична структура на метаданните; 4) балансираността не може да бъде универсално дефинирана, а се определя от предназначението на корпуса, което предпоставя създаването на методология за извличане на балансирани подкорпуси на базата на подробната категоризация с метаданни и лингвистичната анотация.

БНК е създаден в съгласие със следните основни принципи:

(а) унифициран подход при събирането, класифицирането и обработката на документи на различни езици;

(б) таксономично организиран класификационен модел за описание на документите, който позволява включване на нови категории и лесна реорганизация;

(в) автоматично идентифициране и събиране на подходящи документи от интернет;

(г) автоматично извличане на информация за даден документ в рамките на унифициран класификационен модел;

(д) анотационен модел, основан на принципа за натрупване на лингвистична информация;

(е) автоматично обогатяване на документите с лингвистична анотация.

Прилагането на тези принципи позволява създаване на големи по обем многоезикови корпуси (в рамките на *БНК*), които са класифицирани и анотирани съгласно унифицирани модели, позволяващи акумулиране на нова лингвистична и екстралингвистична информация. *БНК* не е „балансиран“, нито „представителен“ в традиционното разбиране, но дава възможност за генериране на едноезикови, многоезикови, паралелни, тематично ориентирани, синхронни и диахронни, балансирани и представителни за дадено явление корпуси. За разлика от останалите известни ни практики *БНК* използва възможностите, които съвременните технологии предоставят, и поважно – организиран е така, че новите модели могат да се надграждат върху използваните досега.

1.2. Структура и съвременно състояние на *БНК*

БНК е създаден през 2009 г. като едноезиков корпус с текстове на български език за целите на компютърната лингвистика и лексикография, а през 2011 г. в него се включват и редица паралелни корпуси, създадени за целите на задачи, свързани с машинния превод и разработването на програми за обработка на многоезикови ресурси.

1.2.1. Създаване и разширяване на *БНК*

За създаването на *БНК* и допълването на корпуса с нови текстове са използвани три основни метода:

1) Включване на готови текстови колекции. Ядрото от български текстове в *БНК* първоначално е формирано на основата на *Българския лексикографски архив* и на *Архива от писмени текстове на български език*, които в момента представляват 49,2% от корпуса. Включени са и два специализирани корпуса от колекцията *OPUS*[4] – корпусът от текстове от Европейската агенция по лекарствата EMEA (медицински административни текстове) и *OpenSubtitles* – корпус от субтитри на филми.

2) Ръчно събиране на текстове от интернет. В миналото това е бил основен подход за събиране на корпуси, но в момента се прилага в ограни-

чени случаи за малко на брой, но големи по обем документи, предимно художествени текстове. Повечето от старите корпуси в състава на *БНК* са събирани ръчно, като например *Българският „Браун“ корпус* (500 текста, 1 млн. думи).

3) Автоматично събиране на текстове. Използват се добре познати и широко прилагани методи за автоматично събиране на текстове, които са приспособени за специфичните нужди и са оптимизирани с оглед на ефективността на работата и прецизността на резултатите. По този начин в *БНК* са добавени голям брой административни текстове, новини, научни и научно-популярни текстове (напр. статии от Уикипедия) и други, които съставляват около 40% от текстовете в българския корпус и над 90% в паралелните.

1.2.2. Структура на *БНК*

Приемаме следните характеристики на текстовете като основни за тяхното класифициране: функция и роля на участниците в комуникативната ситуация (стил), тематично съдържание (тематична област) и композиционна структура (жанр). Взаимовръзките между тези характеристики са важни за изграждане на добър модел за описание и класификация на корпуса.

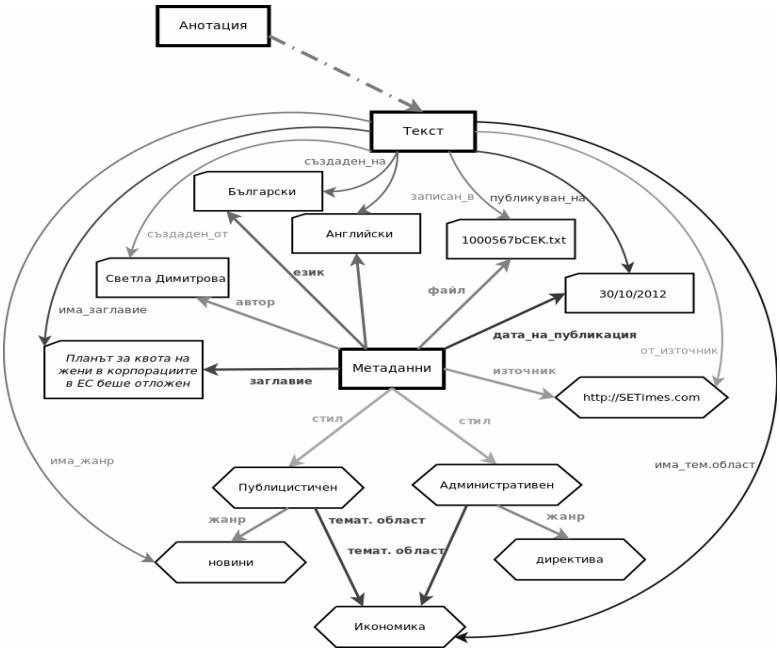
1) Стил. Стилът се дефинира като обща комплексна текстова категория, която комбинира понятията за регистър, модалност и дискурс. В системата на *БНК* са включени следните стилове: административен, научен, публицистичен, художествен, разговорен, художествен/разговорен (субтитри), научно-популярен и популярен.

2) Тематична област. Всеки стил представя набор от тематични области. Тематичните области са обусловени от стила, но някои се срещат в рамките на различни стилове. Например текстове, свързани с икономика или политика, могат да се открият както в научния, така и в публицистичния стил. Поради големия процент текстове с комбинирана тематика се позволява и класифициране по повече от една тематична област.

3) Жанр. За целите на *БНК* приемаме, че жанрът се определя от вътрешните формални характеристики на текста. За всеки стил е възприета система от жанрове, която се доразвива при добавяне на нови текстове към корпуса.

Текстовите единици в структурата на корпуса са описани с подробни метаданни (Burnard 2005). Екстралингвистичните метаданни съдържат информация за източника на текста, например име на автора (или на преводача), език (на оригинала или на превода, както и посока на превода), издание (източник на превода), дата на публикуване; описателна информация, служеща за категоризиране (например по стил, жанр и др.); административна информация за файла и достъпа до него. Лингвистичните метаданни представляват приписана лингвистична информация за езиковите единици на различни нива на анотация. Статистическите метаданни съ-

държат количествени данни за текста като брой токъни, леми, уникални думи, именни фрази, изречения и др.



Фигура 1. Графично представяне на системата от метаданни

Метаданните, с които се описват текстовите единици в *БНК*, обхващат 27 категории. Фигура 1. илюстрира представянето на метаданните във вид на граф, като върховете означават категориите, а ребрата – отношенията между категориите (стил, тематична област, жанр и др.).

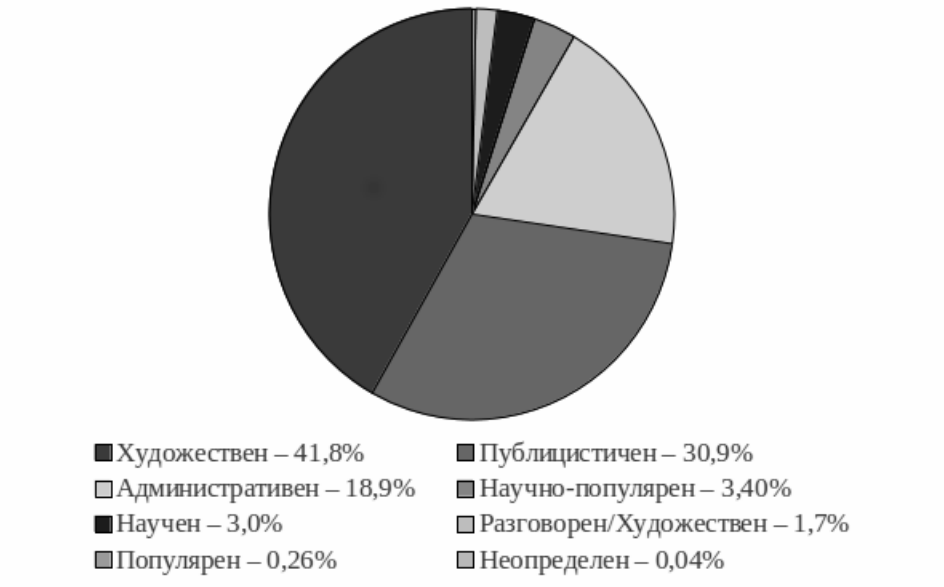
Подробните метаданни позволяват изчерпателно класифициране и лесно подбиране на текстове при създаване на подкорпуси по определени критерии (например тематична област, година на публикуване, авторство, превод).

1.2.3. Корпус на български език

Ядрото на *БНК* се състои от текстове на български език – около 1,2 млрд. думи в над 240 хил. текстови документа. Фигура 2. представя разпределението на текстовете от българския корпус по стилове.

Оригиналните текстове на български език съставляват 37,1% от корпуса, преводните – 40,5%, а за останалите 22,4% липсва информация за източник или посока на превода. В *БНК* са включени и текстове от различна модалност, като преобладават писмените (97,3%), а устните (2,7%) са ограничени по тип – лекции, парламентарни дебати и субтитри. По-голямата

част от текстовете (98,9%) са събрани от интернет, а останалите (1,1%) са предоставени от автори или издатели.



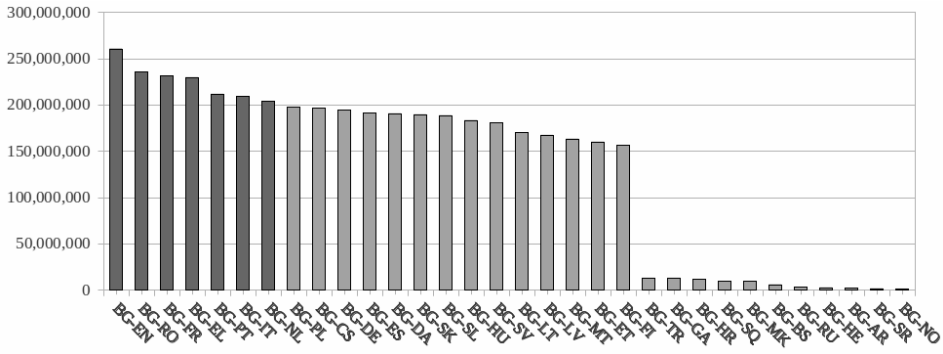
Фигура 2. Разпределение на текстовете на български език в *БНК* по стил

1.2.4. Паралелни корпуси в състава на *БНК*

Обогатяването на *БНК* с паралелни корпуси е свързано с нарастващия интерес на компютърната лингвистика към разработване на многоезикови приложения за машинен превод, извличане на информация от многоезикови ресурси и други.

В състава на *БНК* са включени 47 паралелни корпуса на различни езици, колективно наречени *Bul-X-Cor.* Паралелните корпуси се различават по големина и по покритие на категориите текстове. Разнообразието им се определя от наличните текстове в интернет за дадена двойка езици. *Bul-X-Cor* съдържа текстове на английски, немски, френски, всички славянски и балкански езици, всички езици на държавите от Европейския съюз и някои други европейски и неевропейски езици.

Всеки паралелен корпус се състои от текстове, които имат българско съответствие, като българският текст може да бъде оригинал, превод от другия език или от трети език. Паралелните корпуси са неделима част от *БНК* и следват неговия модел по отношение на структура, формат и описание.



Фигура 3. Паралелни корпуси с обем над 1 млн. думи в *БНК*

Най-големият паралелен корпус в състава на *БНК* е *Българско-английският паралелен корпус*, който съдържа над 260 млн. думи за език (Фигура 3.). Шест от корпусите са с обем 200–250 млн. думи, четиринадесет – 150–200 млн., три – 100–150 млн. Останалите корпуси са сравнително малки: единадесет имат големина 1–15 млн. думи, още петнадесет са под един 1 млн думи. Най-малък е *Българско-японският корпус* с 50 хил. думи за език.

2. Лингвистична анотация

Лингвистичната анотация увеличава използваемостта на корпуса, като позволява извличане на разнообразна информация, разширява неговите функции за различни цели и предлага данни за количествени изследвания върху употребата на езика. Във възприетия от нас подход приемаме следните критерии за качествена анотация:

- Многопластовост – постепенно наслагване на разнообразна анотация.
- Съгласуваност със стандартите за форматиране на данни и представяне на анотацията.
- Последователност – използване на установено множество от атрибути и стойности за различните езици и типове данни, което улеснява съпоставителните изследвания и позволява прилагане на езиково независими програми за обработка.
- Непротиворечивост – приложение на механизми за откриване на несъответствия, както и за проверка и оценка на качеството.

2.2.1. Едноезикова анотация

Българските текстове са анотирани с помощта на *Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове* (Koeva, Genov 2011). Тя включва програми като токънизатор и разделител на изречения, тагер за определяне на част на речта, лематизатор, които са интегрирани в единна система, осигуряваща ефективност и

точност. Разработени са и компютърни програми за разделяне на простите изречения в рамките на сложното, за разпознаване на имена на хора, организации, места и т.н., на съставни лексикални единици и други.

Английските текстове са анотирани (на ниво разделяне на изречения, токънизация, тагиране по част на речта) с *Apache OpenNLP*[5] с предварително тренирани модели и *Stanford CoreNLP*[6]. Модели за *OpenNLP* могат да бъдат тренирани и приложени и за други езици, а за някои като немски и испански са вече достъпни. Лематизацията на английските текстове се извършва с помощта на *Stanford CoreNLP* и *RASP* (Briscoe 2006).

Съвместимостта на анотацията за български и за другите езици е гарантирана чрез съблюдаване на общи стандарти и конвертиране на различните видове анотация към единен формат. Изработената система от тагове за български език осигурява пълноценното морфосинтактично описание на лексикалните единици в български, като е адаптирана и разширена и за другите езици.

2.2.2. Паралелна анотация

Съотнасянето на паралелните текстове по изречения е необходимо при обработката на паралелни ресурси и приложението им за разработване на компютърни програми. В *БНК* части от паралелните корпуси са съотнесени с помощта на системите *HunAlign*[7] и *Maligna*[8], базирани на езиково независимия алгоритъм на Гейл-Чърч (Gale, Church 1993), който използва мярка за подобие между изреченията, основана на дължината им в символи.

Следващ етап в обработването на паралелните корпуси е автоматичното съотнасяне на части от изречения: прости изречения в състава на сложното, фрази и думи. Върху част от *Българско-английския паралелен корпус* е извършено съотнасяне на ниво просто изречение в състава на сложното, в резултат на което е създаден *Българско-английският паралелен корпус със съотнесени (прости) изречения* (*БулЕнСи*).

3. Приложение на *БНК* за различни изследвания

3.1. В компютърната лингвистика

БНК се използва за извличане на специализирани и балансирани по определени критерии едноезикови и многоезикови корпуси за частни задачи в компютърната и традиционната лингвистика.

От самото си създаване корпусът се използва за трениране и тестване на програми за анотация на различни езикови равнища (Koeva et al. 2010; Коева и др. 2010). За тази цел са разработени няколко корпуса по система от критерии.

БулПосКор[9] е структурирана извадка от *Българския „Браун“ корпус* с обем 174 697 словоформи, в която еднозначно са определени частите на речта и граматичните характеристики на всяка словоформа. *БулПосКор* е

използван като тренировъчен и тестов корпус при създаване на приложения за автоматично отстраняване на граматична многозначност и определяне на част на речта (т.нар. тагери) като програмата *BgTagger* (Koeva, Genov 2011), с която е извършено автоматично анотиране на *БНК*.

Българският семантично анотиран корпус (БулСемКор)[10] (Коева и др. 2011) също е извлечен от *Българския „Браун“ корпус* и запазва неговата вътрешна структура. *БулСемКор* включва 95 119 лексикални единици (прости и съставни) и 99 480 словоформи. На всяка проста или съставна единица е приписано еднозначно най-подходящото значение от *Българския WordNet* (*БулНет*)[11]. Семантичният корпус е използван за разработване на програма за автоматично отстраняване на семантичната многозначност.

Българско-английският паралелен корпус със съотнесени (прости) изречения (Koeva et al. 2012a; Koeva et al. 2012b) е с общ обем 366 865 токъна. Върху корпуса е извършена експертна ръчна анотация, включваща: а) определяне на синтактичните отношения (координация и субординация) между простите изречения в състава на сложното; б) анотиране на съюзните връзки между тях; в) съотнасяне на паралелните прости изречения в рамките на съответстващите сложни. *БулЕнСи* се използва за синтактичен анализ и като тренировъчен ресурс в машинния превод (Koeva et al. 2012a).

Wiki1000+ е подкорпус от *БНК*, който включва статии от Уикипедия с общ обем от 13,4 млн. думи. Корпусът е автоматично тагиран с *BgTagger*, след което са анотирани съставните лексикални единици.

БНК позволява извличане на честотни[12] и специализирани речници, словници и др., подпомага създаването на: компютърни лексикони, съдържащи семантична, синтактична, граматична, прагматична и друга информация; онтологии; лексикално-семантични мрежи. Използван е при разработване на *Българския ФреймНет* – семантико-синтактичен речник, който описва семантичните, синтактичните и лексикално-семантичните ограничения при съчетаемостта на български глаголи (Коева 2008), и *Българския WordNet* – лексикално-семантична мрежа на българския език, която съдържа над 50 хил. синонимни множества, свързани помежду си със семантични, морфосемантични и екстралингвистични релации.

Не по-малко важно приложение на корпуси от мащаба на *БНК* е тренирането на езикови модели, базирани на N-грами, за целите на отстраняването на семантична многозначност, машинния превод, извличането на думи, фразеологизми, фрази и други структури, включително на преводни еквиваленти на различни структурни равнища.

3.2. За решаване на общи и частни лингвистични проблеми

Със своя обем, състав, структура и равнища на лингвистична анотация *БНК* осигурява база за изследвания в областта на морфологията, синтаксиса, семантиката, лексикологията, стилистиката, текстологията и други.

Корпусният подход дава възможност за решаване както на по-общи лингвистични задачи като създаване на граматики, така и на по-частни проблеми като изследване на дистрибуцията и функционирането на определени граматични форми, категории или лексикални единици (Коева и др. 2011). Чрез специално разработената система за търсене[13], която позволява разнообразни и комплексни заявки, потребителите могат да съставят собствени извадки по определени от тях лингвистични и екстралингвистични критерии.

Извличането на подкорпуси по определени параметри като година на създаване, стил, тематична област и други предоставя емпирична база за изучаването на определени езикови явления в различни функционални стилове на българския език, в съответни хронологични отрязъци, в идиолекта на даден автор и т.н.

Подкорпуси, създадени по подходящи критерии, биха подпомогнали проследяването на езиковата динамика в неголеми времеви периоди, т.е. на „микроеволюцията“ на езика (Плунгян 2008: 14). Датирането на текстовете в *БНК* дава възможност да се документира хронологията на промените от такъв тип през последните десетилетия, както и да се идентифицира и извлече нова лексика и фразеология, навлязла в българския език в периода от края на 20. и началото на 21. век. (Kolkovska et al. 2012).

Лингвистичната информация в *БНК* подпомага и кодификаторската практика, тъй като позволява да се проучат узуалните употреби в голям масив от реална съвременна писмена продукция и да се приложат по-обективни критерии при установяване на съвременните езикови норми.

Многоезиковите корпуси, състоящи се от паралелни текстове на български и други европейски и неевропейски езици, са база за различни по характер съпоставителни изследвания.

3.3. В лексикографията

БНК се използва много активно в лексикографията, при работата върху речници на българския език от различен жанр – тълковни, неологични, двуезични (с изходен език български), правописни, синонимни и др. Пример са многотомният академичен *Речник на българския език*, най-значимият лексикографски проект у нас, разработван в Института за български език при БАН, както и академичните речници на новите думи в българския език, синонимните и двуезичните речници и др.

Извлечените от *БНК* данни намират приложение във всички етапи на работата по даден речник – при определянето на словника (списъка от думи, подлежащи на лексикографско описание) и отразяването на вариантността, парадигмата, съчетаемостта и системните релации, при съставянето на тълковни дефиниции и подбирането на илюстративни примери. *БНК* дава възможност както за оптимизиране на речниковата работа в отделните етапи, така и за създаване на лексикографски продукти с високо

качество заради възможността за постигане на по-обективно, пълно и системно представяне на лексикалните единици.

Ексцерпирането на данни за срещанията на отделни лексеми и за честотата им допълва използването на традиционните лексикографски ресурси за изготвяне на словници. Автоматично генерираните списъци с регистрираните в *БНК* основни форми или словоформи, подредени по честота, осигуряват обективни данни за разпространението на лексикалните единици в голямо множество от текстове.

Извадки от *БНК* се използват и при определяне на това, кой от регистрираните фонетични или графични варианти на една лексема следва да бъде посочен като заглавна дума в съответния речник. Например информацията за преобладаващата честота на срещане в корпуса на формата *плейър* (30 срещания) в сравнение с *плейер* (0 срещания) мотивира обособяването на заглавка *плейър* в *Речника на българския език.*

БНК е източник на обективна информация за граматичните особености на представяните в речниците лексеми. Така например извлечената от корпуса информация потвърждава отсъствието на определени форми като формата за множествено число при отглаголни съществителни от среден род като *разменяне* и *размножаване*. Данни от *БНК* се използват и при уточняване на квалификаторите, които характеризират една заглавка в историческа перспектива (например остаряла дума, нова дума) или според честотата на употреба (рядка дума, рядко значение, рядка форма).

Корпусът позволява по-широки наблюдения върху семантиката на лексикалните единици, а оттам и прецизна семантична интерпретация и отделяне на редица значения на тълкувани думи, чието установяване с традиционните методи би било извънредно трудно. Друг тип лексикографски релевантна информация, извличана от корпуса, се отнася до колокациите на търсена дума и до тяхната честота. Тази информация позволява по-пълно и адекватно отразяване на лексикалната съчетаемост на съответната дума. Възможността за извеждане не само на най-честите колокации, но и на по-малко фреквентните и по-нетипичните съчетания, е особено ценна за неологичните речници, защото такива съчетания са показател за оформянето на ново значение в семантичната структура на колокацията.

Използването на *БНК* като източник на илюстративен материал също е в съответствие с актуалните тенденции в лексикографската практика. Възможностите за ограничено търсене в корпуса – в рамките на отделни подкорпуси по хронологичен, стилов, тематичен или друг признак – позволяват наблюдения върху по-тесен кръг от източници според конкретните задачи. Търсенето с регулярни изрази е удобен инструмент за филтриране на нерелевантните резултати от заявката.

Използването на езиков материал от големи корпуси като *БНК* дава възможност за въвеждане на качествено нови методи в лексикографската дейност, като съществено я оптимизира и е предпоставка за създаване на

речници, отразяващи с много по-висока степен на точност, обективност и пълнота особеностите на лексикалните единици.

4. Заключение

Статията представя *Българския национален корпус* като голям по обем корпус за български език, съобразен със съвременните изисквания в компютърната и традиционната лингвистика и съизмерим с големите корпуси за други езици. Създаването и разширяването на *БНК* следва принципи, които осигуряват неговото високо качество и широка приложимост. Корпусът е снабден с подробни метаданни за отделните текстови единици, както и с богата лингвистична анотация за български и за английски език, а лингвистичната обработка за други езици е бъдеща задача. *БНК* намира разнообразни приложения в компютърната лингвистика, лексикологията и лексикографията, за целите на теоретични и приложни езиковедски изследвания, както и в други сфери на хуманитарното познание, в езиковото обучение и в преводаческата практика.

Благодарности

Статията е подготвена в рамките на проекта „Интегриране на нови практики и знания в обучението по компютърна лингвистика“ (Договор BG051PO001-3.3.06-0022), който се финансира от Европейския социален фонд и Република България по Оперативна програма „Развитие на човешките ресурси“ 2007–2013 в рамките на схемата за безвъзмездна финансова помощ „Подкрепа за развитието на докторанти, постдокторанти, специализанти и млади учени“ на Главна дирекция „Структурни фондове и международни образователни програми“ към Министерството на образованието и науката.

БЕЛЕЖКИ

[1] Текстът е представен като доклад на Третия международен конгрес по българистика (София, 23–26 май 2013 г.).

[2] http://googlebooks.byu.edu/

[3] http://www.pol-ros.polon.uw.edu.pl/

[4] http://opus.lingfil.uu.se/

[5] http://incubator.apache.org/opennlp/

[6] http://nlp.stanford.edu/software/

[7] http://mokk.bme.hu/en/resources/hunalign/

[8] http://align.sourceforge.net/

[9] http://dcl.bas.bg/poscor/bg/

[10] http://dcl.bas.bg/semcor/bg/

[11] http://dcl.bas.bg/BulNet/general_bg.html

[12] http://dcl.bas.bg/dictionaries_bg.html

[13] http://search.dcl.bas.bg/

ЛИТЕРАТУРА

Коева 2008: *Коева, С.* (съст.). Българският ФреймНет. Семантико-синтактичен речник на българския език. София, Блаком.

Коева и др. 2010: *Коева, С., Д. Благоева, С. Колковска*. Българският електронен езиков корпус и негови приложения. – Наука, № 5, с. 64–69.

Коева и др. 2011: *Коева, С., Д. Благоева, С. Колковска.* Проектът Български национален корпус – резултати и перспективи. – БЕ, № 3, с. 34–53.

Коева и др. 2012: *Коева, С., И. Стоянова, Ц. Димитрова, С. Лесева*. Традиции и новаторство в корпусната лингвистика: Българският национален корпус. – Списание на Българската академия на науките, кн. 3.

Плунгян 2008: *Плунгян, В. А*. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики. – Русский язык в научном освещении, 2, с. 7–20.

ATKINS B. T. S. 1992. Theoretical lexicography and its relation to dictionary-making. – Dictionaries: Journal of the Dictionary Society of North America 14, pp. 4–43.

BANKO, M., E. BRILL. 2001. Scaling to very very large corpora for natural language disambiguation. – Proceedings of ACL 2001, pp. 26–33.

BAŃSKI, P., A. PRZEPIÓRKOWSKI. 2010. The TEI and the NCP: the model and its application. – Proceedings of LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management (LRSLM 2010), pp. 34–38.

BIBER, D. 1993. Representativeness in corpus design. – Literary and Linguistic Computing 8:4, pp. 243–258.

BOJAR ET AL. 2012: *BOJAR, O., Z. ŽABOKRTSKÝ, O. DUŠEK, P. GALUŠČÁKOVÁ, M. MAJLIŠ, D. MAREČEK, J. MARŠÍK, M. NOVÁK, M. POPEL, A. TAMCHYNA*. The joy of parallelism with CzEng 1.0. – Proceedings of LREC 2012, pp. 3921–3928.

BRISCOE, T. 2006. An Introduction to Tag Sequence Grammars and the RASP System Parser. Computer Laboratory Technical Report. University of Cambridge.

BURNARD, L. 2005. Metadata for corpus work. – Developing Linguistic Corpora: a Guide to Good Practice, Oxford: Oxbow Books, pp. 30–46.

ČERMAK, F., V. SCHMIEDTOVÁ. 2003. The Czech National Corpus Project and lexicography. – Asialex ’03 Tokyo Proceedings: Dictionaries and Language Learning: How Can Dictionaries Help Human and Machine Learning?, pp. 74–80.

DAVIES, M. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. – Literary and Linguistic Computing 25:4, pp. 447–465.

DIMITROVA ET AL. 1998: *DIMITROVA, L., T. ERJAVEC, N. IDE, H.-J. KAALEP, V. PETKEVIC, D. TUFIŞ*. Multext-East: parallel and comparable corpora and lexicons for six Central and Eastern European languages. – Proceedings of COLING-ACL’98, Montréal, pp. 315–319.

GALE, W. A., K. W. CHURCH. 1993. A Program for aligning sentences in bilingual corpora. – Computational Linguistics 19:1, pp. 75–102.

KELLER, F., M. LAPATA 2003. Using the web to obtain frequencies for unseen bigrams. – Computational Linguistics 29:3, pp. 459–484.

KILGARRIFF, A., G. GREFENSTETTE 2003. Introduction to the Special Issue on Web as Corpus. – Computational Linguistics 29:3, pp. 333–347.

KOEHN, P. 2005. Europarl: A parallel corpus for statistical machine translation. – *Proceedings of MT Summit,* 2005, pp. 79–86.

KOEVA ET AL. 2010: *KOEVA, S., D. BLAGOEVA, S. KOLKOVSKA*. Bulgarian National Corpus Project. – Proceedings of LREC 2010, pp. 3678–3684.

KOEVA ET AL. 2011: *KOEVA, S., S. LESEVA, B. RIZOV, E. TARPOMANOVA, T. DIMITROVA, H. KUKOVA, M. TODOROVA*. Design and Development of the Bulgarian Sense-Annotated Corpus. – Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora. Actas del III Congreso Internacional de Lingüística de Corpus. Valencia: Universitat Politécnica de Valéncia, pp. 143–150.

KOEVA ET AL. 2012A: *KOEVA, S., B. RIZOV, E. TARPOMANOVA, T. DIMITROVA, R. DEKOVA, I. STOYANOVA, S. LESEVA, H. KUKOVA, A. GENOV*. Application of clause alignment for statistical machine translation. – Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), 12 July 2012, Jeju, Korea, pp. 102–110.

KOEVA ET AL. 2012B: *KOEVA, S., B. RIZOV, E. TARPOMANOVA, T. DIMITROVA, R. DEKOVA, I. STOYANOVA, S. LESEVA, H. KUKOVA, A. GENOV*. Bulgarian-English Sentence- and Clause-Aligned Corpus. – Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), Lisbon, 29 November 2012. Lisboa: Colibri, pp. 51–62.

KOEVA ET AL. 2012C: *KOEVA, S., I. STOYANOVA, S. LESEVA, T. DIMITROVA, R. DEKOVA, E. TARPOMANOVA*. The Bulgarian National Corpus: theory and practice in corpus design. – Journal of Language Modelling, 0:1, pp. 65–110.

KOEVA, S., A. GENOV. 2011. Bulgarian language processing chain. – Proceedings of Integration of multilingual resources and tools in Web applications. Workshop in conjunction with GSCL 2011. University of Hamburg.

KOLKOVSKA ET AL. 2012: *KOLKOVSKA, S., D. BLAGOEVA, A. ATANASOVA*. The application of corpus-based approach in the Bulgarian new-word lexicography. – Proceedings of the 15th EURALEX International Congress 2012, 7–11 August 2012, Oslo, pp. 991– 996.

PRZEPIÓRKOWSKI ET AL. 2010: *PRZEPIÓRKOWSKI, A., M. ŁAZIŃSKI, R. L. GÓRSKI, B. LEWANDOWSKA-TOMASZCZYK*. Recent developments in the National Corpus of Polish. – Proceedings of LREC 2010, pp. 994–997.

SINCLAIR, J. 2005. Corpus and text: Basic principles. – Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books, pp. 1–16.

STEINBERGER ET AL. 2006: *STEINBERGER, R., B. POULIQUEN, A. WIDIGER, C. IGNAT, T. ERJAVEC, D. TUFIŞ, D. VARGA*. The JRC-Acquis: A multilingual aligned

parallel corpus with 20+ languages. – Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 2142–2147.

TADIĆ, M. 2002. Building the Croatian National Corpus. – Proceedings of LREC 2002, Canary Islands, Spain, pp. 441–446.

TUFIŞ ET AL. 2009: *TUFIŞ, D., S. KOEVA, T. ERJAVEC, M. GAVRILIDOU, C. KRSTEV*. Building language resources and translation models for machine translation focused on South Slavic and Balkan languages. – Scientific results of the SEE-ERA.NET Pilot Joint Call, Vienna, pp. 37–48.

VARGA ET AL. 2005: *VARGA, D., L. NÉMETH, P. HALÁCSY, A. KORNAI, V. TRÓN, V. NAGY*. Parallel corpora for medium density languages. – Proceedings of RANLP 2005, pp. 590–596.

KUPIETZ ET AL. 2010: *KUPIETZ, M., C. BELICA, H. KEIBEL, A. WITT*. The German Reference Corpus DEREKO: A primordial sample for linguistic research. – Proceedings of LREC 2010, pp. 1848–1854.

✉ *Проф. д-р Светла Коева*; *гл. ас. д-р Цветана Димитрова*;
гл. ас. д-р Ивелина Стоянова; *гл. ас. д-р Светлозара Лесева*
Секция по компютърна лингвистика
Институт за български език „Проф. Л. Андрейчин“ при БАН
бул. „Шипченски проход“ 52, бл. 17, 1113 София, България
svetla@dcl.bas.bg, cvetana@dcl.bas.bg, iva@dcl.bas.bg, zarka@dcl.bas.bg

✉ *Проф. д-р Диана Благоева*, *проф. д-р Сия Колковска*
Секция за българска лексикология и лексикография
Институт за български език „Проф. Л. Андрейчин“ при БАН
бул. „Шипченски проход“ 52, бл. 17, 1113 София, България
d.blagoeva@ibl.bas.bg; *sia_btb@yahoo.com*

✉ *Prof. Svetla Koeva, PhD*; *Assist. Prof. Tsvetana Dimitrova, PhD*;
Assist. Prof. Ivelina Stoyanova, PhD; *Assist. Prof. Svetlozara Leseva, PhD*
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria
svetla@dcl.bas.bg, cvetana@dcl.bas.bg, iva@dcl.bas.bg, zarka@dcl.bas.bg

✉ *Prof. Diana Blagoeva, PhD*; *Prof. Sia Kolkovska, PhD*
Department of Bulgarian Lexicology and Lexicography
Institute for Bulgarian Language, Bulgarian Academy of Sciences
52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria
d.blagoeva@ibl.bas.bg, sia_btb@yahoo.com