Сия Колковска, Ванина Сумрова
Институт за български език „Проф. Любомир Андрейчин“ – БАН

ЛЕКСИКОГРАФСКОТО ПРИЛОЖЕНИЕ НА БЪЛГАРСКИЯ НАЦИОНАЛЕН КОРПУС

Българският национален корпус (http://search.dcl.bas.bg/) е създаден през 2009 г. в Института за български език „Проф. Любомир Андрейчин“ при БАН като едноезиков корпус, обединяващ колекциите от текстове на Секцията за компютърна лингвистика и Секцията за българска лексикология и лексикография. Впоследствие *Корпусът* прераства в многоезиков, включващ паралелни корпуси от 47 езика, в резултат на което обемът му нараства значително, като в момента българската част съдържа около 1,2 милиарда думи, включени в около 240 000 документа. Той представлява „голям (според съвременните разбирания), небалансиран, динамично развиващ се корпус“ с развита анотационна схема и „таксономично организиран класификационен модел на метаданните за описание на текстовете“ (Коева 2014: 47). Тези особености на Българския национален корпус (нататък – БНК) са резултат от прилагането на съвременните подходи в корпусната лингвистика, насочени според Св. Коева към „динамично събиране и съставяне на големи по обем многоезикови корпуси, характеризиращи се с разширени категоризационни и анотационни данни, обединени от обща класификационна схема“ (Коева 2014: 44). Системата за разширено търсене в БНК дава възможност за извличане на разнообразна лингвистична информация чрез различни по сложност заявки[1].

БНК се използва широко в различни научни и научно-приложни области, като тук се разглеждат някои аспекти от неговото приложение в лексикографската работа, в която използването на корпуси вече е стандартна практика. Проблемите, свързани с използването на БНК в българската лексикография, са разглеждани многократно[2]. Тук се представят някои резултати от приложението на БНК при изработването на многотомния академичен *Речник на българския език* (по-нататък РБЕ), като целта е да се открояят тези особености на БНК, които са особено полезни с оглед на усъвършенстването на определени показатели на РБЕ.

РБЕ е най-значимият и представителен тълковен речник на българския език с публикуваните досега 15 тома (букви А–Р) и преиздадените преработени (осъвременени и допълнени) първи четири тома, чието първо издание е преди преломната (включително в езиков аспект) 1989 г. До момента РБЕ включва повече от 119 000 заглавни думи, като вече е осигурен и онлайн достъп до неговото съдържание (http://ibl.bas.bg/rbe/).

---

[1] Характеристиките на БНК са разглеждани в редица публикации, срв. Коева и др. 2010; Коева и др. 2011.

[2] Корпуснобазираният подход се прилага в българската академична лексикография от 2005 г., като разнообразните аспекти на това приложение са разглеждани в поредица от разработки (вж. Благоева, Колковска 2011 и посочената там литература).

Какво отличава РБЕ от останалите тълковни речници на българския език? Две са неговите най-важни особености: *пълнотата* в отразяване на речниковото богатството на българския език и *детайлността* в семантичното описание на заглавните думи. Тези два показателя са залегнали в концепцията на РБЕ и определят в голяма степен спецификата му.

Показателят *пълнота* се отнася до редица елементи на макро- и микроструктурата на РБЕ – и словника, и различни зони на речниковата статия (граматична, семантична, фразеологична, справочна и др.). Този показател е свързан, на първо място, с широкия хронологичен и функционален диапазон на лексикалните единици, представени в РБЕ (при обхващане на българската лексика от последните близо 200 години), с включването на думи, значения и употреби извън активния речников фонд (остарели, нови, от субстандартните регистри или срещащи се сравнително рядко), каквито по правило не присъстват в по-малките по обем тълковни речници. На второ място, показателят *пълнота* е свързан с изчерпателното отразяване на характеристиките на заглавните думи (граматични особености на заглавката или на отделни значения, съчетаемост, произход на заемките, фонетични или словообразувателни варианти и др.) и с по-пълното представяне на поликомпонентните лексикални единици във фразеологичния блок (отразяващ участието на заглавната лексема във фразеологизми, в съставни наименования или в сложни съюзи и пр.).

Показателят *детайлност* се отнася до равнището на представяне на значенията на заглавните думи, свързано с по-голяма конкретност и диференциация, вследствие на което при редица заглавни думи в РБЕ се отделят повече значения в сравнение с другите речници. В много случаи се разграничават и нюанси към отделните значения, както и специфични типични словосъчетания (наричани според концепцията на РБЕ употреби), а също и т.нар. образни употреби – все елементи, които не намират място в по-малките тълковни речници.

БНК се използва активно при съставянето на *Речник на българския език* от том 13 нататък. Вече се очертават ясно резултатите от неговото приложение, особено в най-новия публикуван том 15 (буква Р), при изработването на който корпусните данни от БНК са използвани много широко и последователно. Резултат от това е усъвършенстването на изданието по отношение на най-важните му показатели – пълнота и детайлност на семантичното представяне[3].

Утвърдена практика при съставянето на най-новите томове на РБЕ, върху които се работи през последните години, е използването на автоматично генерирани от БНК списъци с леми или словоформи, подредени по честота. Благодарение на големия обем на *Корпуса* тези списъци включват не само думите от активния речников фонд, но и множество по-редки или по-малко употребими думи, което гарантира в значителна степен това да не бъдат пропускани лексеми, които присъстват в съвременния български език, но липсват в лексикалните картотеки и в предходни речници. Така например лексеми или дори цели словообразувателни гнезда като *разгазирам*, *разгазиране*, *разгазиран*,

[3] Ролята на корпуснобазирания подход за подобряване на такива важни характеристики на всеки речник като точност, системност и пълнота на представянето на лексикалните единици е посочена от редица изследователи (вж. например Аткинс 2002).

равнопоставям, *равнопоставя*, *равнопоставяне*, *равнопоставено*, *рампирам*, *рампиране*, *рампов*, *развъдчик*, *разностилен*, *разностилие*, *разнотипен*, *разнотипност*, *разобличаващ*, *рутинирам се*, *рутинирано*, *рутинираност*, *различност*, *рангов*, *ранговост*, *раблезиански* и много други са регистрирани само в БНК и включването им в *Речник на българския език* е основано на корпусни данни.

Подобрение на показателя *пълнота* се наблюдава и при отразяване на фразеологичните единици, в които участват заглавни думи в РБЕ. Наблюденията върху поведението на тези думи в големи масиви от текстове в БНК дават възможност за установяване на фразеологични съчетания, които липсват в други речници. Така например въз основа на корпусни данни при заглавните думи *размахвам* и *разменям* са установени фразеологичните съчетания, съответно *размахвам пръст* (*на някого*) и *разменяме си ролите* (*с някого*), които са включени в речниковите статии на тези глаголи. Същото се отнася и за съставното наименование *развален телефон* (название на детска игра) и за фразеологизма *развален телефон* (с вариант *като развален телефон*) със значение ‘ситуация и др., при която информация, факти и под. се предават, представят неточно, изопачено’.

Трябва да се подчертае, че подобряването на показателя *пълнота* на РБЕ в резултат от приложението на БНК се дължи на такава важна особеност на този корпус, каквато е големият му обем. Този съществен резултат от лексикографското приложение на БНК е потвърждение за полезността от създаването на колкото се може по-големи по обем корпуси, посочена от редица изследователи (Килгариф, Грефенстет 2003; Мейер 2004: 14; Аткинс, Ръндел 2008: 61; Чермак 2010; Коева 2014: 39). Според Св. Коева „по-големият обем на корпусите предполага по-достоверна илюстрация на по-широк кръг езикови явления (с по-висока честота на срещане и разнообразна дистрибуция в различни тематични области, стилове и жанрове)“ и същевременно е предпоставка за това корпусите да съдържат „достатъчно на брой срещания дори за рядко употребими думи, рядко употребими колокации и рядко употребими съставни лексикални единици“ (Коева 2014: 39).

Важен резултат от възможността за наблюдения на по-разнообразни контексти на думите в *Корпуса* е по-пълното и детайлно семантично представяне на заглавните лексеми в том 15 на РБЕ. Това се отнася до:

Значенията на заглавните думи и на фразеологичните единици

Например благодарение на наблюдавани в БНК примери се установява развитието на преносно значение на съществителното име *рунд* ‘поредна фаза от продължителен и многоетапен конфликт, съдебен спор, противопоставяне или надпревара’, с което то се употребява извън спортната сфера (пример: *Загубихме първия рунд в борбата с наркоманията*), на преносно значение на прилагателното име *разноглед* ‘за човек – силно затормозен или смутен, объркан, обикн. поради умора, претоварване и под.’, използвано в разговорния език (напр. *ставам разноглед*, *правя някого разноглед*) и т.н. При наблюденията върху срещанията на глагола *развеждам се* в *Корпуса* се установява новото

му значение ‘прекратявам отношенията си с някого, интереса си към някого, нещо’.

Резултат от използването на БНК е и по-точното и пълно отразяване на фразеологичните единици в РБЕ. Например, ако в том 3 на РБЕ фразеологизмът *голяма работа*[1] (посочен във фразеологичния блок на прилагателното *голям*) е представен само със значението ‘човек, издигнал се обществено или служебно’, в том 15 (при съществителното *работа*) неговата семантика е представена значително по-пълно, като са отбелязани и значенията ‘човек, който превъзхожда другите със своите качества, способности, възможности’ и ‘много симпатичен, добър, услужлив и под. човек’. Тези значения са извлечени от примери от *Корпуса.* Аналогичен е случаят с фразеологизма *като мехлем на (за) рана*, при който (в сравнение с други речници, както и с другите томове на РБЕ) е прецизирана формата и са отделени четири значения въз основа на примери от БНК.

Така корпусните данни дават основания за идентифициране на значения, които липсват в други тълковни речници на българския език и не са застъпени в лексикалните картотеки.

Семантичните нюанси

Въз основа на корпусни данни е отделен например нюансът на прилагателното име *разноезичен* ‘който е свързан с говор, общуване, разговори на различни езици’ (напр. *разноезична глъч*, *разноезични крясъци*), отбелязан при значението ‘за говор, общуване, разговор и под. – който се извършва на различни езици’. Наблюдавани в *Корпуса* примери дават основание за разграничаване и на нюанса ‘мивка с форма на черупка от мида’, посочен при второто значение на съществителното *раковина* ‘предмет, изделие с форма на черупка на мида’.

Образните употреби

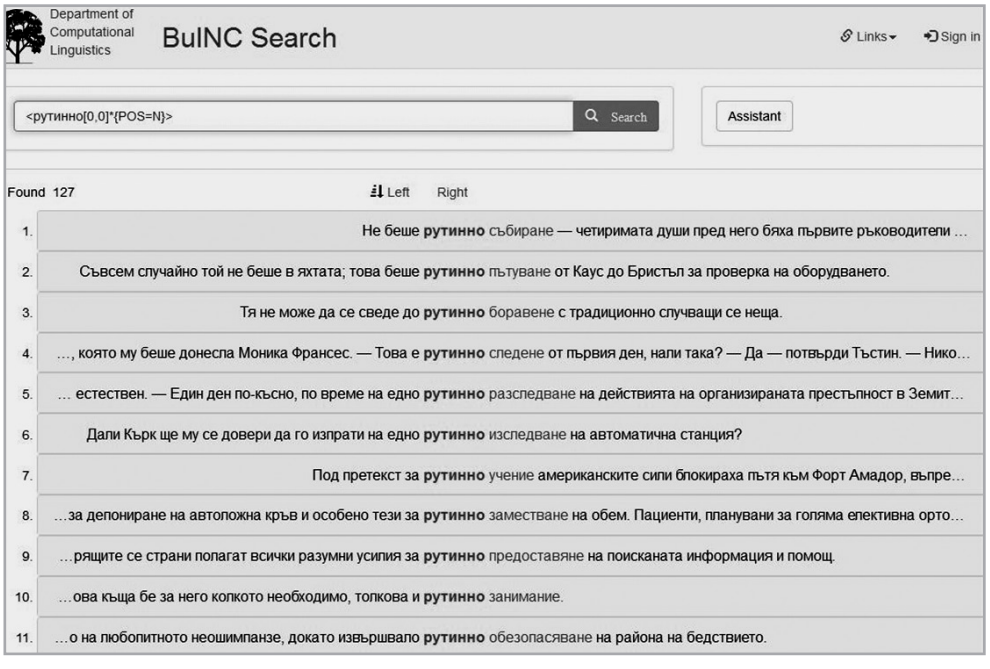
БНК подпомага лексикографската работа и при откриването на образни употреби. Примери като *Облеклото е раковина*, *към която тялото се приспособява; разтворената звездна раковина / на утринта над морския простор* насочват именно към такава употреба на думата *раковина* в основното ѝ значение ‘черупка на мекотело (мида, охлюв, рапан)’.

Много важен резултат от използването на БНК е улесняването на лексикографа в случаите, когато той трябва да се ориентира в изобилен лексикален материал, който следва да анализира внимателно, за да се постигне необходимото равнище на пълнота и детайлност на лексикографското представяне в РБЕ. Проблемът, свързан с изобилието от лексикален материал, което може да затрудни анализа на лексикалните единици, се отнася както за традиционните лексикографски методи, така и за корпуснобазираните методи. За разлика от традиционната лексикография корпуснобазираните методи обикновено включват различни средства за справяне с този проблем. БНК също предоставя такива възможности, базиращи се на развитата му анотационна схема, на разширената система от метаданни, с която той разполага, и на функционалности-

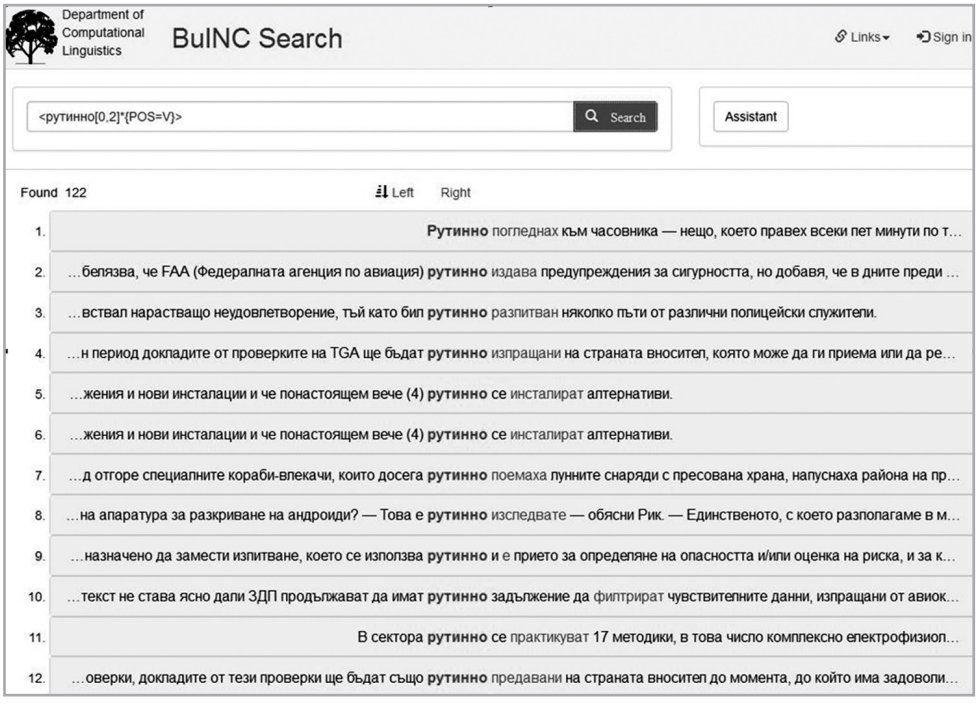
те на системата за разширено търсене в него. Тези особености на БНК дават възможност за филтриране на нерелевантните за дадена цел употреби, което помага на лексикографа да се ориентира в лексикалния материал и съкращава времето за откриване и подбор на подходящи примери.

Много полезна в това отношение е възможността за ограничаване на търсенето в БНК по определен хронологичен, стилов, жанров или стилистичен признак. Например преносното значение на съществителното име *спирачка* ‘дейност, фактор, който е пречка за извършването, протичането на нещо’, за което се предполага, че би трябвало да се среща извън техническата област, може лесно да се идентифицира в текстовете от БНК и да се илюстрира с подходящи примери. Чрез ограничаване на търсенето в подкорпуса MassMedia и чрез избор на стойност на категорията стил „публицистичен“ се откриват редица примери, в които това съществително име е използвано в посоченото преносно значение, срв. *Сериозна спирачка за последващ ръст на акциите на дружеството може да се окаже липсата на каквато и да е информация от управляващите за неговото бъдеще* (в-к „Банкеръ“).

Лексикографът е улеснен при ориентацията си в лексикалния материал и в редица други отношения: например когато трябва да намери употреби на омонимни лексеми и форми (по-специално при прилагателни имена в ср. р. ед. ч. и при наречия), на субстантивирани форми на прилагателни имена (отразявани при определени условия в РБЕ) и др. В тези случаи БНК дава възможности за елиминиране на нерелевантните примери чрез търсене с регулярни изрази, отразяващи различия в съчетаемостта на лексемите. На фиг. 1 и 2 са представе-



Фиг. 1. Търсене на словоформата **рутинно**
(нечленувана форма за ср. р. ед. ч. на прилагателното **рутинен**) в БНК



Фиг. 2. Търсене на наречието **рутинно** в БНК

ни резултатите от заявките <рутинно[0,0]*{POS=N}> и <рутинно[0,2]*{POS=V}>, с които се търсят съответно прилагателното *рутинно* в ср. р. ед. ч., нечленувано и наречието *рутинно*. Откритите срещания съдържат достатъчен брой релевантни примери и за двете лексеми.

Чрез наредената заявка <румънски[0,2]*{!POS=N}> се извличат субстантивирани употреби на прилагателното име *румънски*, в които то има значение ‘румънски език’, срв. примерите: *… като емигрант в Румъния публикувал брошура на румънски*; *…аз трябва да науча румънски* и др.

Използването на голям по обем корпус, в който с достатъчна честота присъстват различни форми на лексемите, е предпоставка и за по-точното отразяване на граматичните характеристики на заглавните думи. БНК е източник на обективни данни както за граматичните ограничения в парадигмата на отделни думи (напр. липса на форма за мн. ч. при някои съществителни имена), така и за преобладаващата употреба на определени форми на някои лексеми (напр. на форми за ед. или мн. ч.). Например корпусните данни от БНК за преобладаващи срещания на съществителното имена *руина* в мн. ч. (858 срещания в мн. ч. срещу 118 срещания в ед. ч.) са обективно основание за граматичната бележка *Обикн. мн.* при тази заглавка в РБЕ. Отсъствието в БНК на форми за множествено число на редица съществителни имена от ср. р. (като *разностилие*, *раболепие*, *разбягване*) е основание за бележката *мн. няма* в речниковите статии на тези думи. Както и обратното – наличието в БНК на форми за мно-

жествено число при имена, посочени в по-стари речници като дефективни по число, дава основание за отразяването на тези форми в РБЕ.

В заключение може да се посочи, че отбелязаните насоки в усъвършенстване на *Речник на българския език* по отношение на най-важните му показатели *пълнота* и *детайлност* са от значение не само за представяното издание, но и за българската лексикография като цяло поради активното използване на този лексикографски труд при изработване на множество други речници на българския език (включително и двуезични).

ЛИТЕРАТУРА

Аткинс 2002: Atkins, B. T. S. Then and now: competence and performance in 35 years of lexicography. – In: Proceedings EURALEX. Vol. 1. Copenhagen, 2002, p. 1–28.

Аткинс, Ръндел 2008: Atkins, B. T. S., M. Rundell. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press, 2008.

Благоева, Колковска 2011: Благоева, Д., С. Колковска. Корпусният подход в българската лексикография – практика и перспективи. – В: Съвременни методи и подходи в лексикографската практика. Сборник студии и статии. София: Авангард Прима, 2011, с. 7–45.

Килгариф, Грефенстет 2003: Kilgarriff, A., G. Grefenstette. Introduction to the Special Issue on Web as Corpus. – Computational Linguistics, 2003, 29:3, p. 333–347.

Коева 2014: Коева, С. Българският национален корпус в контекста на световната теория и практика. – В: Езикови ресурси и технологии за българския език. София: АИ „Проф. Марин Дринов“, 2014, с. 29–52.

Коева и др. 2010: Koeva, S., D. Blagoeva, S. Kolkovska. Bulgarian National Corpus Project. – In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, D. Tapias (eds.). Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta: European Language Resources Association (ELRA), pp. 3678–3684. http://www.lrec-conf.org/proceedings/lrec2010/index.html

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

Чермак 2010: Čermák, F. Notes on compiling a corpus-based dictionary. – In: Lexikos 20 (AFRILEX 20: 2010), p. 559–579.