ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 49, КН. 1, СБ. А, 2011 – ФИЛОЛОГИЯ,
PAISIY HILENDARSKY UNIVERSITY OF PLOVDIV – BULGARIA
RESEARCH PAPERS, VOL. 49, BOOK 1, PART A, 2011 – LANGUAGES AND LITERATURE

БЪЛГАРСКОТО ЕЗИКОЗНАНИЕ В СЪВРЕМЕННИЯ СВЯТ – БЪЛГАРСКИЯТ НАЦИОНАЛЕН КОРПУС[1]

Росица Декова
Институт за български език – Българска академия на науките

The paper presents a very concise overview of some electronic language resources as an illustration of the achievements of the Bulgarian linguistic research community in the contemporary scientific world. Special attention is paid to those resources that can be used by researchers, and most of all to the Bulgarian National Corpus. The presentation is organized with respect to the possible uses of the available language resources and technologies to the linguistic community and research in Bulgaria.

Key words: language resources, searches in the Bulgarian National Corpus, linguistic research

В съвременния свят на бързоразвиващи се информационни и комуникационни технологии българското езикознание продължава да се доказва като научна област с непрекъснат напредък и стремеж за следване на европейските и световните стандарти. За това важна роля играе и компютърната лингвистика, чиято главна задача е формалното описание на естествения език с цел създаване на електронни езикови ресурси и разработване и прилагане на компютърни технологии при статистическото и логическото анализиране и моделиране на езика. С тази си роля компютърната лингвистика се явява своеобразен мост между два „свята“ – традиционната лингвистична школа и съвременните компютърни технологии. Със съжаление трябва да отбележим обаче, че не са много езиковедите у нас, които минават по този „мост“ и използват пълноценно предоставените възможности. Ето

[1] Искам да изкажа благодарност на присъствалите на презентацията на тази статия, представена в секция *Науките за езика и тяхното място в съвременния свят* на Юбилейните Паисиеви четения, Пловдив, 2011 г. Всички коментари и предложения бяха от изключителна полза за оформянето ѝ, за да изпълни тя своето предназначение да бъде в помощ на съвременните лингвистични изследвания по български език.

защо настоящата статия си поставя за цел да представи накратко и в достъпен вид някои от съвременните електронни езикови ресурси с оглед на възможностите, които предлагат те в помощ на лингвистичните изследвания по български език.

В България електронни езикови ресурси се създават най-вече в изследователските центрове на Българската академия на науките – Секцията по компютърна лингвистика към Института за български език, Лабораторията за лингвистично моделиране към Института по информационни и комуникационни технологии, Института по математика и информатика, както и в университетите – СУ „Св. Климент Охридски“, ПУ „Паисий Хилендарски“ и др. Поради ограничения обем на статията ще представим единствено някои от ресурсите, създадени в ИБЕ или с участието на изследователи от академичната структура с оглед на възможностите за употребата им в различни видове лингвистични изследвания.

В началото трябва да поясним, че електронни езикови ресурси са тези, които могат да бъдат разчетени от компютърни програми за обработка на езика, но това не означава, че намират приложение единствено за целите на компютърната лингвистика. Електронните езикови ресурси биха могли да бъдат в полза на езиковедските и литературоведските изследвания по български език, както и в сферата на обучението.

Електронните езикови ресурси се делят на електронни речници, корпуси (структурирана съвкупност от текстове, към която има допълнителна информация като например автор и година на създаване, ако са известни, или жанр и тематична принадлежност на текста и др.), анотирани корпуси (включващи морфо-синтактична и/или семантична информация), паралелни корпуси (текстове и техните преводни съответствия на друг език/езици), лексикално-семантични бази от данни и др.

Към електронните речници спада Граматичният речник на българския език, който е достъпен на адрес: http://dcl.bas.bg/est/dict.php.

Към лексикално-семантичните бази от данни спадат Българският ФреймНет и Българският WordNet.

Българският ФреймНет (Коева 2008, Коева, съст. 2008, Коева, Декова 2008, Коева 2010а) представлява семантико-синтактичен речник на българския език и понастоящем включва най-често срещаните български глаголи с всички техни значения и съответстващото им формално семантично и синтактично описание. Всяко едно значение е свързано чрез идентификацинен номер със съответния литерал в

Българския WordNet и е илюстрирано с примери, взети от Българския национален корпус, които са семантично и синтактично анотирани. Информацията в Българския ФреймНет може да бъде използвана за изследването на различни лингвистични явления, свързани със семантиката и/или синтаксиса на български глаголи. Това включва както изследвания върху аргументната структура на конкретни глаголи, например аргументната структура на глагола *мигрирам* (вж. Несторова 2010), така и за сравнение на синтактико-семантичните характеристики на групи глаголи (вж. Декова, Несторова 2011). Българският ФеймНет се използва успешно и в обучението по български език на чужденци (вж. Несторова 2011).

Потребителският интерфейс към Българския ФреймНет е в процес на обработка, като търсенето ще бъде възможно чрез интернет страницата на Секцията по компютърна лингвистика към Института за български език на БАН (http://dcl.bas.bg).

Българският WordNet, известен още като БулНет (Коева 2007), представлява лексикално-семантична мрежа за български език, която съдържа над 53 000 синонимни множества, разпределени в четири части на речта – глаголи, съществителни имена, прилагателни имена и наречия. Всяко синонимно множество (синсет) кодира релация на еквивалентност между няколко единици (литерали), които имат уникално лексемно значение (sense), принадлежат към една и съща част на речта и изразяват еднакво значение, описано в дефиницията на синсета. Всеки синсет се свързва с кореспондиращ синсет в Принстънския WordNet (PWN2.0) чрез идентификационен номер. Синонимните множества са свързани помежду си посредством два вида релации: семантични (*каузация*, *хиперонимия*, *хипонимия*, *холонимия*, *подобен на*, *също така*, *субсъбитие*, *глаголна груп*а и др.) и екстралингвистични (*тематична област*, *регион*, *обичайна употреба*). Релациите между самите литерали също се делят на две групи: семантични (*синонимия* и *антонимия*) и словообразувателни (*произхождащ*, *причастие*, *производен* и *принадлежащ*).

Изграден по този начин, Българският WordNet осигурява успешно приложение в задачите за търсене и извличане на информация от документи, за автоматична категоризация и анотация на документи, за автоматичен превод и др. Наред с това ресурсът предлага редица възможности в помощ на лингвистичите изследвания по български език: търсене и избор на синоними, справка за семантичните релации на дадена дума по отношение на системата от останали думи в езика; справка за тълкуваното лексикално значение на думата и паралелни

предложения за съответната дума на други езици (сред които английски, немски, френски, испански, италиански, холандски, чешки, естонски, гръцки, румънски, турски и сръбски). Българският WordNet се разпространява от ELDA (Агенция за оценка и разпространение на езикови ресурси). Информация от Българския WordNet е интегрирана в системите за търсене на Българския национален корпус и на Семантичния корпус, които предстои да опишем по-детайлно.

Поради ограничения обем на статията избрахме да се спрем най-подробно на Българския национален корпус, който отговаря на съвременните изисквания за пълнота и балансирано отразяване на българския език и е свободно достъпен по интернет, като предлага голямо разнообразие от възможности за търсене и предоставя широкообхватна лингвистична информация в помощ на изследователите на българския език.

Българският национален корпус (Коева, Стоянова 2009, Коева и др. 2010) е създаден в Института за български език „Проф. Любомир Андрейчин“ от сътрудници в Секцията по компютърна лингвистика и Секцията за българска лексикология и лексикография. Към настоящия момент Българският национален корпус (БгНК) съдържа над 470 милиона думи и включва над 10 000 текста. Материалите, включени в Корпуса, отразяват състоянието на българския език (предимно в неговата писмена форма) от средата на XX в. (1945 г.) до наши дни. Българският национален корпус може да бъде използван както за теоретични изследвания на определени лингвистични явления с цел езиковедско описание или лексикографско отразяване (например наблюдения върху честотата на употреба на думи или езикови конструкции), за наблюдения върху особеностите на отделни области на езика и за извличане на примери за демонстрация при обучението по български език, така и за създаване на приложения в различни области на езикознанието като компютърната лингвистика, лексикографията и др.

Корпусът е свободно достъпен на адрес: http://search.dcl.bas.bg/bg/. Търсенето може да се направи в целия корпус или да се изберат общи или специализирани подкорпуси по определени критерии (тематика, автор, година / период на издаване, източник и др.) в зависимост от спецификата на изследователските цели. Това става чрез избиране на бутона „Корпуси“ и селектиране (поставяне на отметка) само на този корпус или корпуси, в които желаем да търсим. Подкорпусите, от които можем да избираме, са следните:

- **1945–2010** – предлага възможност за тематично търсене на литература, преводи, периодика и фолклор за посочения период, като могат да се селектират една или повече подтеми;
- **Българският Браун корпус** (BulBrown) предлага възможност за търсене, като могат да се селектират една или повече специализирани области – A – административни, B – научни, C – преса, D – литература;
- **Българският паралелен корпус** (Коева и др. 2011), който включва близо 47 000 текста от различни области, предлага възможност за тематично търсене в избрани паралелни корпуси – *Закони*, *Медицина*, *Новини* (разделени по години от 2003 до 2011 год. вкл.), *Субтитри* и *Художествена литература* (където могат да се селектират един или повече жанрове – *детективска литература, детска литература, класическа литература, научна фантастика, приключенски, ужаси, фентъзи*);
- **2001–2010** – предлага възможност за тематично търсене за посочения период, като могат да се селектират една или повече от следните специализирани области – *военно дело, ежедневие, закони (право), икономика, литература, медицина, преса, ръководства, спорт*, както и смесени – *правно-икономически, правно-спортни, правно-медицински*;

Търсенето може да включва една или повече думи, като по подразбиране заявката е ненаредена, т.е. думите се търсят във всяка възможна поредност. Например заявката „*човекът ли е*“ връща като резултат всички срещания на трите думи, независимо в каква последователност се намират те една спрямо друга (примери от (1) до (4)). В началото е даден общият брой на намерените резултати, както и броят на страниците, на които са показани. Във всеки един от резултатите, показани на страницата на корпуса, търсените от нас думи са маркирани за удобство с червен цвят и удебелен шрифт (в настоящата статия те са маркирани единствено с удебелен шрифт).

1) ***Човекът ли е*** *божия грешка?* (БгНК, *файл:* 1945-2010/Translations/след 1990/наука/филос/GKajtazovZNB(prevod).xml).

2) *Наистина* ***ли човекът е*** *по-несъвършен от машината?* (БгНК, *файл:* 1945-2010/Periodicals/списания и годишници/1945-1989/косм/Kosmos1971-2.xml, *сигнатура*: К, 1971, кн. 2).

3) *Познавал* ***ли е човекът*** *от древните времена скоростта на светлината?* (БгНК, *файл:* 1945-2010/Translations/след 1990/наука/-истор/SvKoevZE(prevod).xml).

4) *Не **е ли човекът** това, което прави, а не, което мисли, защото делата му остават, а те могат да не са в съзвучие с ума и душата му.* (БгНК, *файл:* 1945-2010/Literature/1945–1989/проза/-роман/AStojnevENB.xml).

Резултатите включват и случаи, в които търсените думи са разделени от препинателни знаци, които не сигнализират край на изречение, каквито са примерите в (5) и (6).

5) *Човекът не е виновен, разбираш **ли**, **човекът е** виновен само когато е бездарен и некадърен.* (БгНК, 1945-2010/Literature/1945–1989/проза/фантастика/ LDilovPO.xml).

6) – *Основният въпрос на философията **е**: **човекът ли е** за устава, или уставът за човека?* (БгНК, *файл*: 1945–2010/Translations/-след 1990/проза/фантаст/ MStoevTTz(prevod).xml).

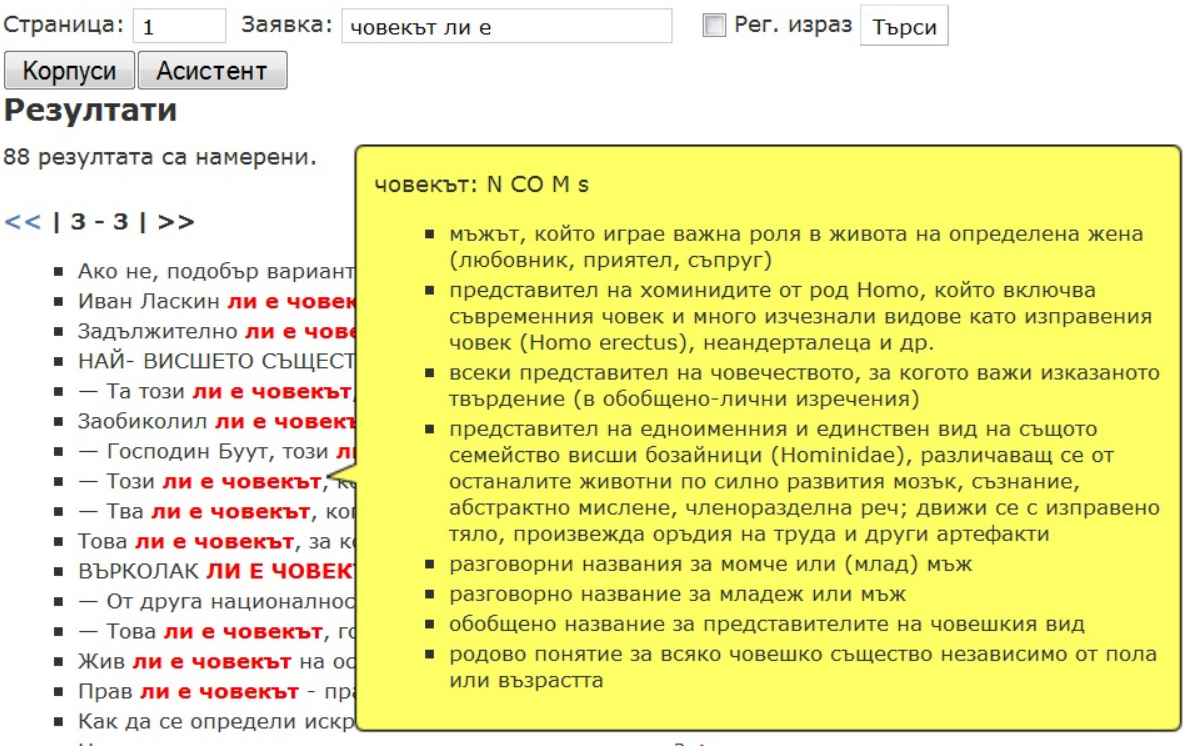
Ако искаме да търсим наредена заявка, то трябва да изпишем думите в желаната от нас последователност и да оградим заявката в ъглови скоби < >. Една такава примерна заявка <*човекът е*> връща като резултат всички срещания в корпуса, в които двете думи се намират точно в този ред и няма да покаже резултати, в които думите са с разменени места (т.е. изречения, които съдържат „*е човекът“,* няма да бъдат върнати като резултат от нашето търсене*)*.

Всяко срещане на търсената дума/думи се вижда на екрана в обкръжението на най-близкия му контекст (обикновено в рамките на едно изречение). Ако искаме да видим резултата в по-широк контекст, натискаме знака +, който се намира в края на всеки от изведените резултати. Освен контекста, ограничен до две изречения преди и две изречения след изречението, което е върнато като резултат, виждаме изписано и името на файла, откъдето са взети, както и данни за автора (ако са налични). Върнатото като резултат изречение е отбелязано в зелен цвят, а думите от заявката са дадени отново в червен цвят и с удебелен шрифт.

Когато цитираме изречение от Корпуса, единственото задължение, което имаме като потребители, е да посочим като източник Българския национален корпус и да приведем данни за автора на текста и за изданието, от което е взет съответният пример.

Допълнителна информация за всяка от думите в намерените резултати може да бъде получена, когато посочим желаната дума с курсора. Тогава се появява прозорец, в който е изписана посочената от нас дума, нейната част на речта, както и граматичните ѝ характерис-

тики. В случая сме посочили една от въведените при заявката думи, а именно *човекът*. В прозореца се визуализира: *човекът* N CO M s, което означава съществително нарицателно от мъжки род, единствено число. Освен граматичната информация ни е представена и семантична информация чрез изписването на всички възможни значения на думата (които са въведени в Българския WordNet), както е илюстрирано на изображението на Фиг. 1 по-долу.



Фиг. 1. *Възможност за показване на допълнителна семантична информация и граматични характеристики на избрана дума от върнатите при търсенето резултати*

Търсенето може да бъде конкретизирано чрез добавяне на различни символи, като символите за релация (семантична или граматична свързаност между думите) се ограждат в наклонени скоби //, а символите за граматични характеристики във фигурни скоби {}.

Така например можем да търсим определена дума заедно с всичките ѝ словоформи, като изпишем /F/ след думата[2]. Заявката *вървя* /F/ връща всички срещания в корпуса на която и да е от синтетичните форми на глагола *вървя*. За сравнение заявката *вървим* /F/ ще

[2] Символът за словоформи се огражда в наклонени скоби /F/, защото се разглежда като вид граматична релация между основната форма и словоформите.

върне същите резултати, т.е. думата, чиито форми търсим, може да не е в основна форма.

По подобен начин можем да търсим синоними /S/ на съществителни, глаголи, прилагателни и наречия, както и хипероними /H/ на съществителни и глаголи и релацията *подобен на* /L/ за прилагателни. Например заявката *вървя* /S/ извежда като резултат изречения с глаголите *ходя, мина, минавам, отида, отивам* и др. И тук с посочване на курсора върху избрана дума могат да се видят нейните граматични характеристики, както и всички възможни значения в съответствие с Българския WordNet. Заявката *дете* /H/ извежда като резултат всички хипероними[3] на думата *дете*, срещнати в корпуса: *малолетен, непълнолетен, потомък, потомство*. Заявката *хубав* /L/ връща като резултат всички литерали и техните форми, които се срещат в Българския WordNet и в корпуса, свързани с релацията *подобен на* с думата *хубав*: *красив, прекрасен, приятен, лош, изключителен, превъзходен, пълен, задоволителен* и т.н. Виждаме, че заявката *подобен на* /L/ може да включва както синоними, така и антоними, и хипоними или хипероними на търсената дума.

Когато искаме да търсим дума с конкретни граматични характеристики, то заявките имат следния вид {атрибут=стойност}. Това е така, защото граматичните характеристики се разглеждат като признак, на който се приписват определени стойности. Възможностите за търсене по атрибути и стойностите, които могат им бъдат приписани, както и символите[4], с които се записват, са описани в Таблица 1.

[3] Според йерархията на думите в Българския WordNet и посочените там семантични релации.

[4] Всички символи в заявките се изписват задължително с латински букви. Символ, написан на кирилица, дори да има същата графична визуализация (например кирилско А и латинско А), ще доведе до грешка в търсенето.

Атрибут	**Символ на атрибута**	**Стойности**	**Символ на стойността**
Род на съществително име	**G**	мъжки	**M**
		женски	**F**
		среден	**NE**
Тип на съществителното име	**NT**	нарицателно	**CO**
		собствено	**PR**
Тип на числителното име	**NUMT**	бройно	**C**
		редно	**O**
Вид на глагола	**VA**	свършен	**PE**
		несвършен	**IM**
Преходност на глагола	**VT**	преходен	**T**
		непреходен	**IN**
Тип на местоимението	**PT**	лично	**L**
		притежателно	**POSS**
Число	**N**	единствено	**s**
		множествено	**pf**
		бройна форма	**cf**
Лице	**P**	първо	**1**
		второ	**2**
		трето	**3**
Род	**FG**	мъжки	**mf**
		женски	**ff**
		среден	**nf**
Определеност	**D**	нечленувана форма	**0**
		членувана форма	**df**
Време	**T**	сегашно	**r**
		минало свършено	**e**
		минало несвършено	**j**
Нелична глаголна форма	**IVF**	сегашно деятелно	**y**
		минало свършено	**x**
		минало несвършено	**q**
		страдателно причастие	**w**
		деепричастие	**z**

Таблица 1. *Описание на атрибутите и стойностите, които могат да им бъдат приписани, както и символите, с които се означават при търсене в корпуса*

Символът *звезда* [*] означава произволна дума, характеризирана с определено множество граматични характеристики. Например заявката *{POS=PRON} намира всички местоимения в корпуса.

Ето и записването на някои конкретни заявки. Например със заявката *скачам* /F/ *{POS=PREP} търсим всички възможни предлози, които са атестирани в Българския национален корпус след глагола *скачам* (независимо от неговата словоформа). Част от получените резултати са: *през, от, извън, пред, покрай, около, във, към, с, на* и т.н. По този начин можем да получим статистически резултати за възможната аргументна структура на глагола и за семантичните ограничения на неговите аргументи.

Заявката *{POS=A} *кафе* /F/ извежда като резултат всички възможни прилагателни, които се срещат пред думата *кафе* (и нейните словоформи) в корпуса: *горещо, разтворимо, ароматно, насипно, разлятото, бразилско, истинското* и т.н.

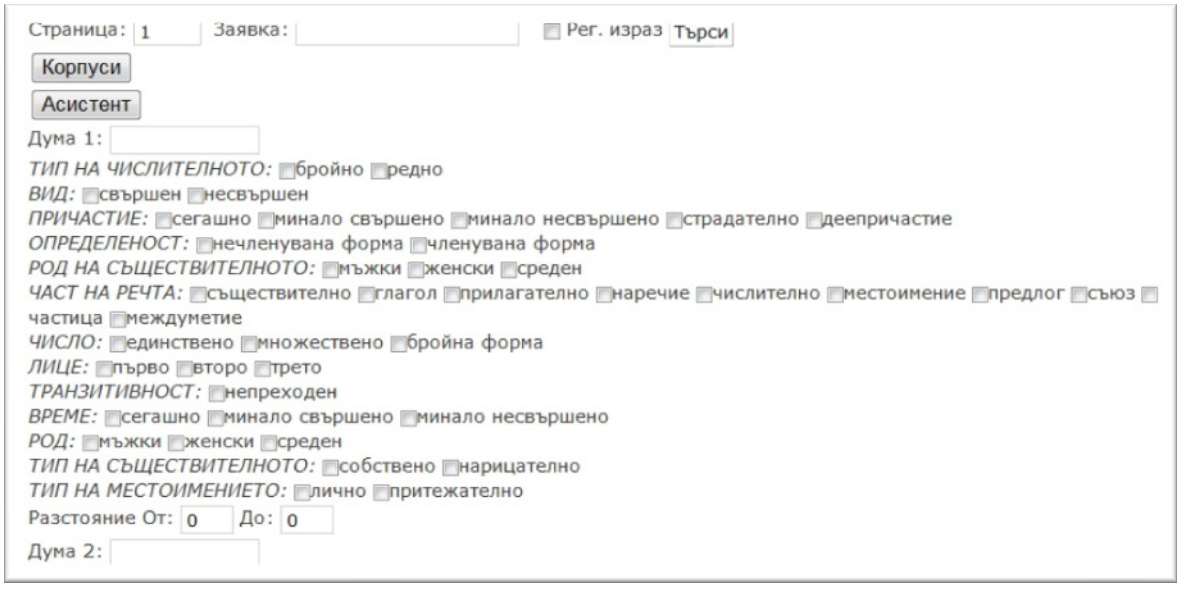
Корпусът предлага търсене на поредици от думи, които съдържат произволни думи. Произволните думи в наредената заявка се отбелязват в правоъгълни скоби [], а броят им се конкретизира с цифри, разделени със запетая [от, до]. Заявката <ударих [1,2] *{POS=N}> намира последователностите от глагола *ударих*, най-малко една и най-много две произволни думи след думата *ударих* и произволно съществително име. Тази заявка ще върне резлутати от вида *ударих на камък*, *ударих с камък*, *ударих се в камък* и т. н., но не и срещания от вида *ударих камъка*, тъй като в последния пример думата *ударих* не е отделена от поне една дума от съществителното.

В търсене с произволни думи можем също да включим и конкретна дума или думи. Така например заявката <*{POS=V} или *{POS=V}> извежда всички срещания на два глагола, свързани със съюза *или: срещам или застигам, излъчва или предава, потопи или взе* и др.; а заявката <един *{POS=N} и една *{POS=N}> извежда срещания от словосъчетания като *един ден и една нощ, един мъж и една жена, един режим и една система, един котак и една жълтица* и т.н.

В търсенето могат да се използват също и символите за конюнкция &, дизюнкция |, отрицание !, импликация => и еквивалентност <=>. Например заявката *черно&бяло* намира едновременно срещане на думите *черно* и *бяло* в рамките на едно изречение. Заявката !*черно*/F/&*бяло*/F/намира всички форми на *бяло* в изречения, в които не се среща никоя от формите на *черно*. Групирането става с употребата на кръгли скоби (). Например заявката !(*цвят*=>*бял*/F/) връща всички срещания на *цвят*, където не се среща някоя от формите на *бял*.

Подробни инструкции за търсене, атрибутите и стойностите, които могат да бъдат търсени, както и символите, с които се записват, са дадени на интернет страницата на корпуса: http://www.ibl.bas.bg/-BGNC_search_bg.htm.

Друг вариант за конкретизиране на търсенето е опцията, която се предлага с бутона „Асистент“, където можем да посочим до две думи с възможност за отбелязване на брой произволни думи между тях. Всички изброени атрибути и стойности са изписани заедно с квадратчета за селектиране, както се вижда на Фиг. 2.



Фиг. 2. *Визуализация на опцията „Асистент“ за подготвяне на заявки за търсене в Българския национален корпус*[5]

С тази опция не е необходимо да помним символите на атрибутите и техните стойности, както и правилата за тяхното изписване. Трябва обаче ръчно да селектираме всички желани от нас характеристики на думата. Стойност, която не е селектирана, няма да бъде взета под внимание при съставянето на заявката. Резултатите, които ще получим, ще включват единствено думи със селектираните от нас стойности.

[5] Поради липса на пространство Фиг. 2 показва само атрибутите и стойностите, принадлежащи на Дума 1. Същите се повтарят и за Дума 2. За всяка дума могат да се селектират произволен брой стойности. Ограничението може да бъде само лингвистично, като при зададени противоречиви стойности търсенето ще даде грешка.

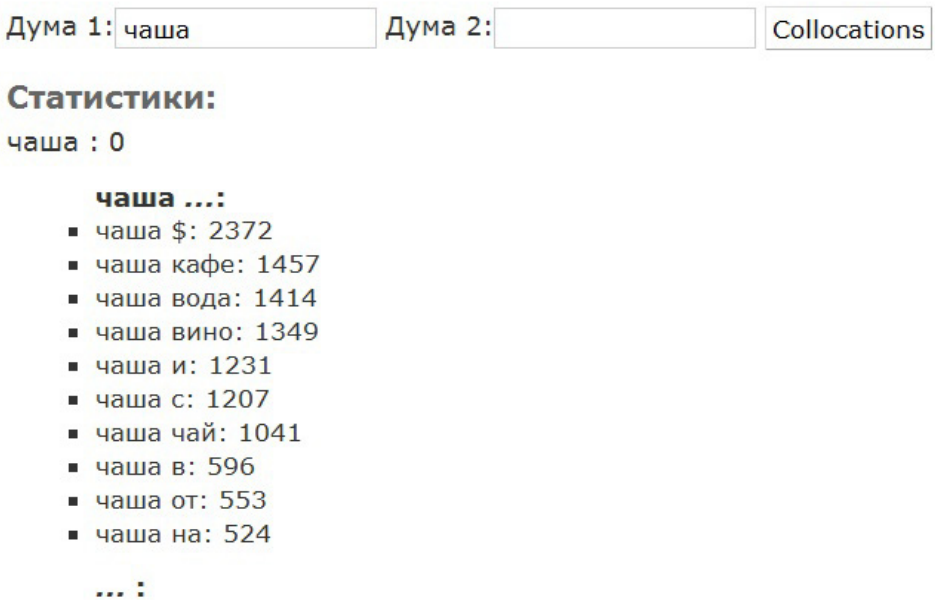
В полето за Дума 1 и/или Дума 2 можем да напишем както конкретна дума, така и знака за произволна дума *, която ще отговаря на зададените граматични критерии.

Когато сме готови с изписването на думите и селектирането на техните характеристики, трябва да натиснем бутона „Подготви“. В полето „Заявка“ се появява израз, еквивалентен на желаната от нас заявка, и едва тогава натискаме бутона „Търси“. Резултатите от заявка, генерирана чрез опцията „Асистент“, се визуализират по същия начин, както и ако сами напишем заявката, и за тях важат същите правила и възможности – изписване на броя на всички намерени срещания, получаване на допълнителна информация за думите (чрез посочване на желана от нас дума с курсора на мишката), визуализиране на допълнителен контекст и информация за избрано от нас срещане чрез натискане на знака +, оцветяване в червено на заявката в резултатите и т.н.

Опцията, която не сме споменали до този момент, е „регулярен израз“ (търсене на текст по даден шаблон). Тази опция се намира вляво от бутона „Търси“ и може да бъде селектирана чрез поставяне на отметка в квадратчето непосредствено до нея. При селектирана опция регулярен израз можем да търсим единствено по правилата на регулярните изрази. Тази опция се използва успешно от компютърни лингвисти, програмисти и математици, които са предварително запознати с шаблоните за търсене на регулярни изрази. Най-тривиалният пример за търсене на регулярни изрази е търсене на единични символи в корпуса и може лесно да бъде възпроизведен от неспециалисти. Например регулярният израз [?] връща като резултат всички срещания на въпросителния знак, т.е. всички въпросителни изречения.

В горния ляв ъгъл на страницата за търсене в Българския национален корпус можем да изберем търсене за *колокации*. В най-общия смисъл колокация е едновременното срещане на две или повече думи. Терминът *колокация* се използва в по-тесен смисъл за означаване на група от думи, които се срещат заедно по-често от статистически случайното и формират общо значение.

Можем да търсим най-често срещаните думи след или преди избрана от нас дума, като я напишем съответно в заявката за Дума 1 или Дума 2. Така например заявката Дума 1: *чаша* извежда десетте думи, които се срещат най-често след думата *чаша* (Фиг. 3), а заявката Дума 2: *чаша* извежда десетте думи, които се срещат най-често пред думата *чаша* (Фиг. 4).



Фиг. 3. *Резултати от заявката за колокации Дума 1:* ***чаша***

Фиг. 4. *Резултати от заявката за колокации Дума 2:* ***чаша***

Ако изпишем една и съща дума и в двете полета за търсене, в резултат ще получим както десетте най-често срещани думи преди търсената дума, така и десетте най-често срещани думи след нея. Например заявката Дума 1: *виждам* и Дума 2: *виждам* ни връща най-често срещаните колокации преди и след думата виждам, както и статистиката *виждам виждам*: 66, т.е. случаите на повторение, които могат да бъдат намерени, ако пуснем търсенето <*виждам виждам*> в

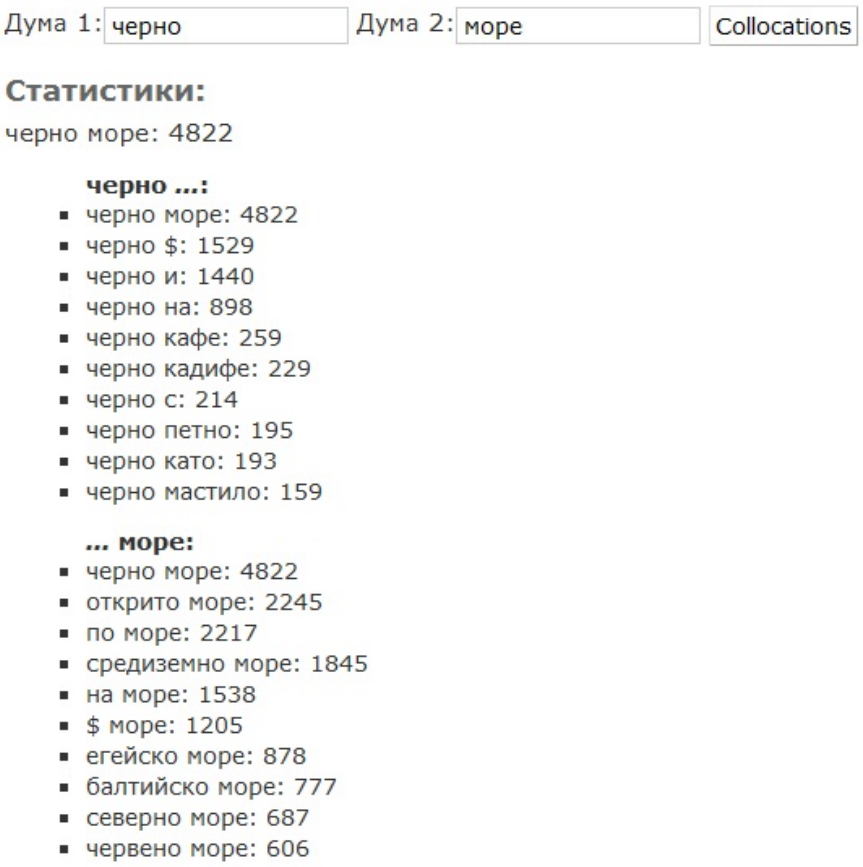
Българския национален корпус. Примерите показват, че повторенията обикновено са стилистични колокации:

(7) – *Ето значи каква била работата… **Виждам, виждам**…* (БгНК, *файл:* 1945-2010/Translations/след 1990/проза/фантаст/-MStoevHPZA(prevod).xml).

(8) – ***Виждам, виждам** – завика вече в изстъпление човекът, – светлина виждам.* (БгНК, *Файл:* 1945-2010/Literature/1945-1989/проза/роман/StZagorchinovDP .xml).

Можем също да напишем различни думи във всяко едно от местата за търсене. Например заявката Дума 1: *черно* и Дума 2: *кафе* връща статистики за думата *черно* на първо място и независими статистики за думата *кафе,* стояща на второ място, от които се вижда, че изразът *черно кафе* има 259 срещания. Думата *кафе* е на пето място в списъка на най-често срещаните думи след думата *черно*, докато думата *черно* липсва изцяло от списъка с най-често срещаните думи пред думата *кафе.*

За сравнение заявката Дума 1: *черно* и Дума 2: *море* извежда статистиките, показани на фиг. 5. В тях се вижда, че изразът *черно море* се среща общо 4822 пъти. Думата *море* е на първо място в списъка на най-често срещаните думи след думата *черно*, както и думата *черно* заема първото място в списъка на най-често срещаните думи пред думата *море*. И наистина *Черно море* е вид колокация, назоваване на обект.



Фиг. 5. *Резултати от заявката за колокации*
Дума 1: ***черно*** *и Дума 2:* ***море***

Както вече споменахме, Българският национален корпус съдържа в себе си няколко подкорпуса. Сред тях е и **Българският Браун корпус**, който е достъпен на адрес: http://dcl.bas.bg/Corpus/home_-bg.html.

Семантичният корпус (Коева, ред. 2010) е част от Българския Браун корпус. Съдържа 95 119 лексикални единици, като значението на всяка лексикална единица е еднозначно определено в зависимост от контекста, в който е употребена. Определянето става чрез приписване на най-подходящото синонимно множество от Българския WordNet. Корпусът е свободно достъпен на адрес: http://dcl.bas.bg/semcor/bg/, както и от страницата за търсене в Българския национален корпус http://search.dcl.bas.bg/bg/, за тази цел най-отгоре на страницата вляво трябва да изберем БулСемКор.

Освен за автоматично отстраняване на семантичната многозначност за целите на автоматичния превод, корпусът може да бъде използван за различни лингвистични изследвания, при които е необходимо разграничаване на значенията на срещнатите в корпуса думи

или форми, като търсенето може да бъде лимитирано до конкретната дума чрез изписване на търсената дума в полето *Дума* или за всички словоформи на думата чрез изписване на основната форма на търсената дума в полето *Основна форма*. Така например заявката Основна форма*: игра* извежда статистически анализ на всички срещания на някоя от формите на лемата *игра* (независимо дали е в единствено или множествено число, членувана или нечленувана), подредени по брой срещания за всяко значение (според Българския WordNet) заедно с едно изречение за пример.

За сравнение заявката Дума*: игра* връща статистически анализ на всички срещания на всички значения на думата *игра*, която може да е както лексема на лемата *игра*, така и словоформа от лексемния ред на лемата *играя*.

Ако искаме да конкретизираме нашето търсене, можем да дефинираме както търсената от нас дума (лексема), така и нейната основна форма (лема). Например заявката Дума: *игра*, изписана едновременo със заявката Основна форма: *игра,* връща като резултат статистика на всички срещания на всички значения на съществителното нарицателно *игра*.

Българският PoS анотиран корпус (Коева 2010a) също е част от Българския Браун корпус. Съдържа 174 697 лексикални единици, анотирани с подходящата граматична информация от Българския граматичен речник. Граматическите характеристики на всяка словоформа са определени коректно и еднозначно. Той предлага възможности за ефективно търсене на езикови модели и форми в текста.

Корпусът е свободно достъпен на адрес: http://dcl.bas.bg/poscor/bg/, както и от страницата за търсене в Българския национален корпус http://search.dcl.bas.bg/bg/, като най-отгоре на страницата вляво изберем БулПосКор.

Правилата за търсене в Българския PoS анотиран корпус са същите като при Семантичния корпус. Тук например заявката Дума: *игра* връща като резултат статистика на всички срещания на думата *игра,* групирани по морфо-синтактични характеристики, подредени по брой срещания и придружени с по едно изречение за пример. Докато заявката Основна форма: *игра* връща стистически анализ на всички срещания на някоя от формите на лемата *игра*, подредени по брой срещания за всяка словоформа заедно с едно изречение за пример.

Ако искаме да потърсим всички срещания на думата *игра,* като имаме предвид не съществителното, а глагола заедно с останалите му словоформи, то тогава в заявката Основна форма трябва да изпишем лемата *играя.*

Можем също така да ограничим търсенето, като напишем Дума: *игра* и Основна форма: *играя*. В резултат ще получим статистика на всички срещания на глаголната форма *игра*.

Накратко можем да заключим, че съвременните електронни езикови ресурси предлагат много и разнообразни възможности за едно интердисциплинарно развитие на българското езикознание, което допринася за успешното интегриране на лингвистичните изследвания по български език в световната научноизследователска сфера. Качествените езикови ресурси и технологии, създадени в рамките на национални и международни проекти и партньорства, и участието ни в международни научни форуми представят постиженията на българското езикознание като част от съвременния лингвистичен (а и технологичен) свят.

Литература

Декова, Несторова 2011: Декова, Р., П. Несторова. Формално описание на някои непреходни глаголи за движение в Българския ФреймНет. // *Български език*, №1, 31–43.

Коева 2004: Коева, Св. Съвременни езикови технологии – приложения и перспективи. // *Закони на/за езика*, София: Хейзъл, 2004, 111–157.

Коева 2007: Коева, Св. БулНет (лексикално-семантична мрежа на българския език) – част от световната лексикално-семантична мрежа. // *Български език*, 2007, №1, 34–50.

Коева 2008: Коева, Св. Българският ФреймНет. Семантико-синтактичен речник на българския език – концептуален модел. // *Българският ФреймНет. Семантико-синтактичен речник на българския език*, съставител Св. Коева, София: БАН, 2008, 5–57.

Коева 2009: Коева, Св. Езикови ресурси и компютърни програми с приложение в лингвистичните изследвания. // *Приложение на информационните технологии в работата на филолога и при изграждането на езикови ресурси*. София: Архимед, 2009, 54–75.

Коева 2010а: Коева, Св. *Българският ФреймНет 2010*. София: БАН, 2010.

Коева 2010б: Koeva Sv. Syntactic Annotation in Bulgarian National Corpus. // *Proceedings from the seventh international conference Formal Approaches to South Slavic and Balcan Languages*, *Dubrovnik*, 2010, 35–41.

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

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

Коева, Декова 2008: Koeva, S. & R. Dekova, Bulgarian FrameNet. // Tadic, M., M. Vulchanova and Sv. Koeva (eds). *Proceedings from The Sixth*

International Conference Formal Approaches to South Slavic and Balkan Languages, Dubrovnik, Croatia, 2008, 59–67.

Коева, Стоянова 2009: Коева, Св., И. Стоянова. Български национален корпус. // *Български език*, 2009, №3, 137–146.

Коева и др. 2010: Koeva Sv., D. Blagoeva, S. Kolkovska. Bulgarian National Corpus Project. // *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, D. Tapias (eds.), Valletta, European Language Resources Association (ELRA), 2010, 3678–3684.

Коева и др. 2011: Коева, Св., И. Стоянова, Р. Декова. Българо-английски-X+ паралелен корпус. // *Български език*, 2011, №3, 100–118.

Несторова 2010: Несторова, П. Аргументна структура на глагола **мигрирам**. // *Български език*, 2010, №1, 80–87.

Несторова 2011: Несторова, П. Българският ФреймНет и преподаването на български език на чужденци. // *Международна работна среща на преподавателите по български език като чужд и на преподавателите по другите южнославянски езици, „Преподаването на южнославянските езици в съвременна Европа" – по повод на 140 години от смъртта на видния български възрожденец д-р Петър Берон*, 18–19.04.2011 г., Масариков университет – Бърно, Чехия, 148–157.

Несторова, Декова 2011: Несторова, П., Р. Декова. 2011. Опит за класификация на глаголите за движение в Българския ФреймНет. // *Научни трудове на УХТ*, 2011, том 58, Свитък 2, 481–486.