# Българският национален корпус в контекста на съвременната лингвистика

*Светла Коева, Диана Благоева, Сия Колковска, Ивелина Стоянова, Светлозара Лесева, Цветана Димитрова*


*Европейски социален фонд - ОП „Развитие на човешките ресурси“ 2007-2013*
BG051PO001-3.3.06-0022
„Интегриране на нови практики и знания в обучението по компютърна лингвситика“


***Трети конгрес по българистика, 23-26 май 2013 г., София***

# Какво е БНК?

- Най-големият паралелен корпус с фокус върху български език;
- обем: над 5,4 млрд. думи;
- ядро – български текстове: над 1,2 млрд. думи;
- паралелни корпуси: 47 езика; над 4,2 млрд. думи;
- преобладават на текстове, събирани автоматично от интернет.

# Какви са проблемите?

- Критериите за оценка на корпуса:

  (i) ***големина*** на корпуса и на текстовете в него;

  (ii) ***представителност*** на текстовете от гледна точка на езиковата продукция;

  (iii) ***балансираност*** – мотивирано от езиковата продукция съотношение между текстовете.

- по-трудно постижими при многоезиковите корпуси.

# Какви са практиките?

- Корпуси, създадени според предзададен модел (*BNC*, *HNK*, *COCA*, *НКРЯ*, *SNK*, *OEC*);
- големи корпуси с балансирани подкорпуси (ČNK, NKJP);
- небалансирани корпуси с подробно метаописание, за лесно извличането на подкорпуси (DeReKo);
- много големи небалансирани корпуси от интернет Google Books Corpora;

# Общи положения, възприети в БНК

- По-големи корпуси – повече данни, по-разнообразни приложения;
- обогатяване с допълнително съдържание (метаданни и анотация);
- ***представителност*** – покритие на разнообразни категории текстове;
- ***балансираност*** – определено от целите съотношение между текстовете.

# Принципи при създаване на БНК

- ✔ Унифициран модел на събиране и обработка;
- ✔ Таксономично организиран класификационен модел за описание на документите;
- ✔ Автоматично идентифициране и събиране на документи от интернет;
- ✔ Автоматично извличане на метаданни;
- ✔ Натрупване на анотационни равнища;
- ✔ Автоматична лингвистична анотация.

# Как се събират текстове?

- Готови текстови колекции: **49,2%** от текстовете на български език;

  *Български лексикографски архив*, *Архив на писмени текстове на български език*, *OPUS:*

- ръчно събиране на текстове от интернет;
- автоматично събиране на текстове от интернет: **40%** от българския корпус и **90%** от паралелните.

# Състав на БНК (1)

- Ядро от български текстове:

  над **1,2 млрд. думи**; **240 000 текста**

- източник на текстовете:

  **98,9%** от интернет **1,1%** от автори и издатели

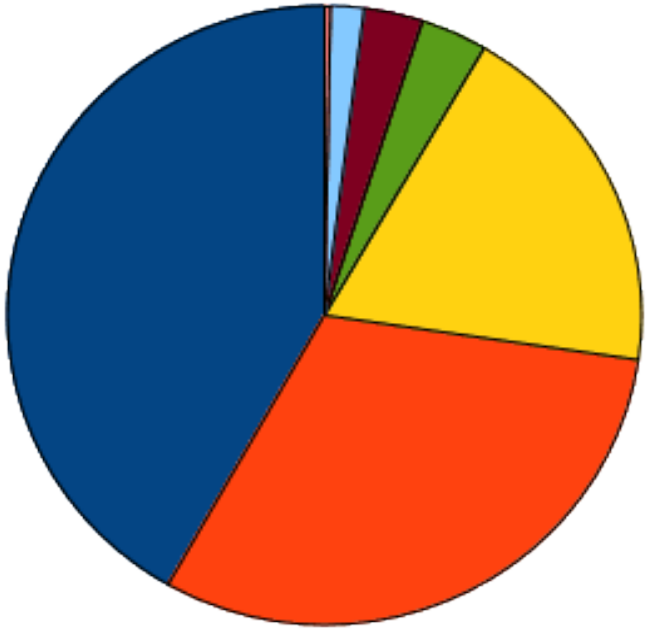
- произход (22,4% с неизвестен произход):

  **37,1%** оригинални **40,5%** преводни

- модалност:

  **2,7%** устни **97,3%** писмени

# Състав на БНК (2)



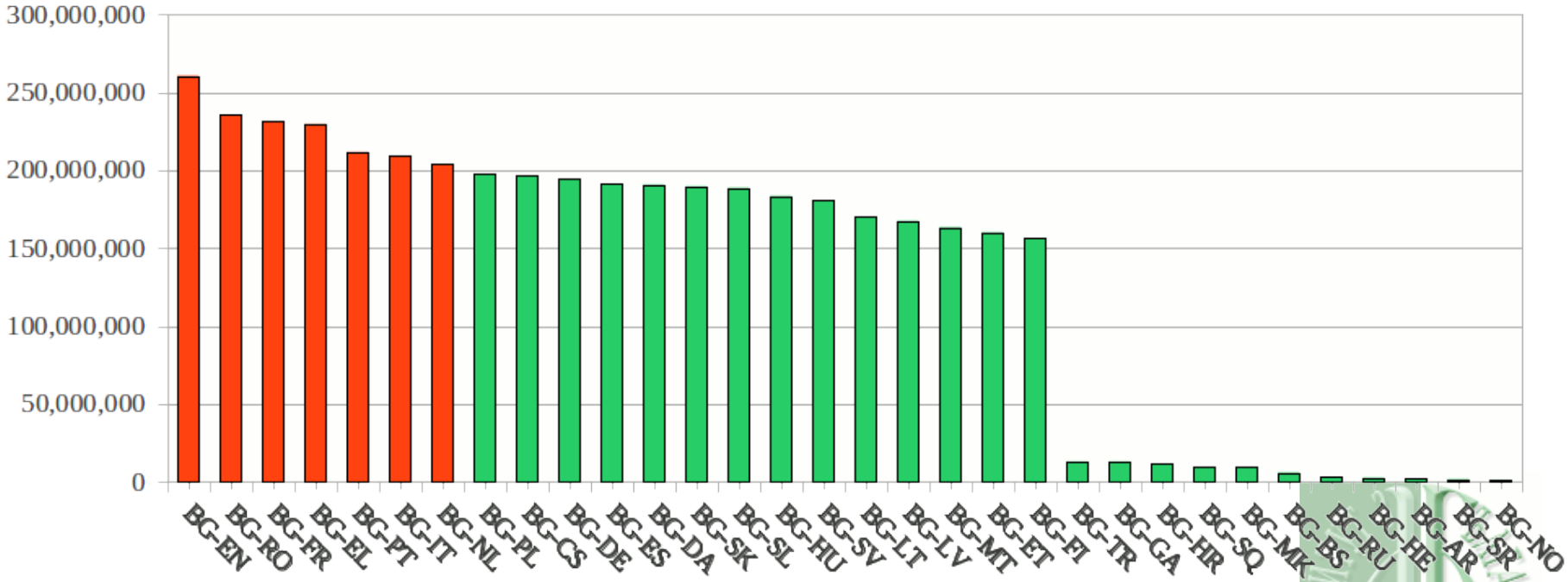

**Разпределение на стиловете в ядрото на БНК**

# Състав на БНК (3)

Bul-X-Cor – 47 паралелни корпуса:

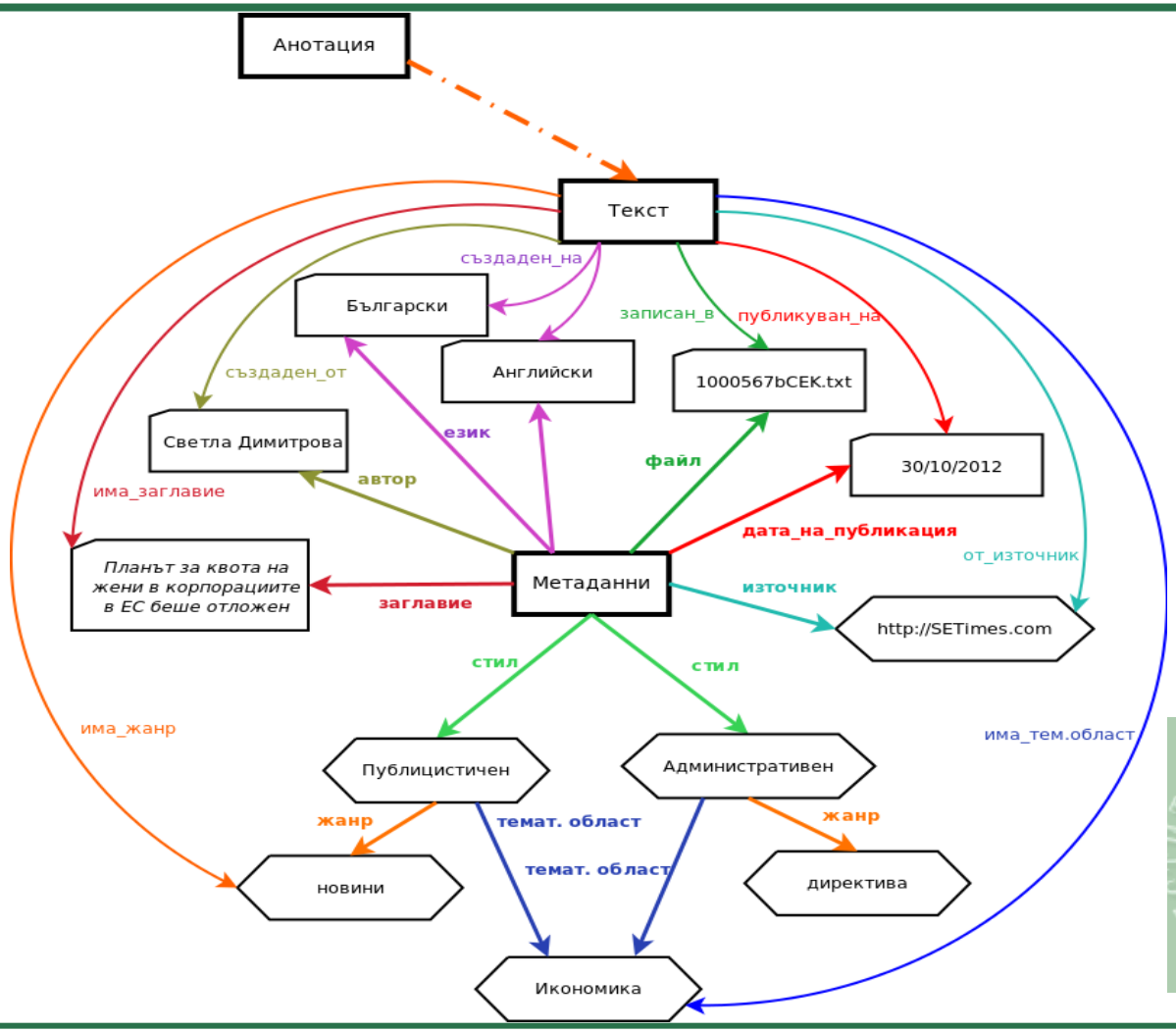
- *Българско-английски паралелен корпус* - над **260 млн. думи** за език;
- **200-250 млн. думи** за език 6 корпуса
- **100-200 млн.** думи за език 14 корпуса
- **1-15 млн. думи** за език 11 корпуса
- под **1 млн. думи** за език 15 корпуса
- *Българско-японски корпус* - **50 хил. думи** за език

# Състав на БНК (4)




**Паралелни корпуси с обем над 1 млн. думи**

# Метаданни


Анотация
Текст
създаден_на
Български
Английски
записан_в
публикуван_на
1000567bCEK.txt
създаден_от
Светла Димитрова
език
файл
30/10/2012
автор
има_заглавие
дата_на_публикация
от_източник
Планът за квота на жени в корпорациите в ЕС беше отложен
Метаданни
заглавие
източник
http://SETimes.com
стил
стил
има_жанр
има_тем.област
Публицистичен
Административен
жанр
темат. област
темат. област
жанр
новини
директива
Икономика

# Анотация

- **Едноезикови корпуси** – български и английски:

  токънизация, лематизация, определяне на част на речта, граматични характеристики, (частична) синтактична и семантична анотация; и др.

- **Паралелни** – *Българско-английски корпус*:

  съотнесени изречения; (частично) съотнесени прости изречения в състава на сложното; (частично) съотнесени фрази и думи.

## Подкорпуси: *БулПосКор*

- Големина: **174 697 словоформи**;
- анотация: част на речта и граматични характеристики на всички словоформи;
- приложение: тренировъчен и тестов корпус за програми за морфосинтактична анотация (BgTagger).
- *БНК* е автоматично анотиран с BgTagger.

## Подкорпуси: *БулСемКор*

- Големина: **95 119 лексикални единици** (прости и съставни) и **99 480 словоформи**;
- анотация: ръчно приписано значение от *Българския wordnet* (БулНет) на всяка дума;
- приложение: в програма за автоматично отстраняване на семантична многозначност.

# Подкорпуси: *BulEnAC*

- Големина: **366 865 токъна** за двата езика;
- анотация:
  - автоматично съотнесени изречения;
  - ръчно анотирани синтактични отношения и съюзни връзки между простите изречения;
  - ръчно съотнасяне на простите изречения;
- приложение: тренировъчен ресурс при приложения за машинен превод.

# Приложения: лексикални ресурси

- Честотни, специализирани и други речници;
- компютърни лексикони с морфологична, синтактична, семантична, прагматична, др. информация (*Българският ФреймНет);*
- семантични мрежи (*БулНет);*
- онтологии; и др.

# БНК: честотни речници

| Админ. | Научен | Публиц. | Худож. |
|---|---|---|---|
| *член* | *време* | *дейност* | *ръка* |
| *приложение* | *страна* | *дружество* | *време* |
| *продукт* | *година* | *година* | *око* |
| *параграф* | *живот* | *решение* | *човек* |
| *регламент* | *човек* | *съд* | *глава* |

Съществителни с най-висока честота в БНК по стил

## Тренировъчен (и тестов) корпус за

- Езикови модели, базирани на N-грами;
  - автоматично отстраняване на граматична многозначност;
  - автоматично отстраняване на семантична многозначност;
  - приложения за машинен превод;
  - търсене и извличане на информация.

# БНК в езиковедските изследвания

- Система за търсене по комплексни заявки с формули и регулярни изрази:

**http://search.dcl.bas.bg**

  - Наблюдения върху дистрибуцията и функционирането на определени форми;
  - наблюдения с оглед на кодификацията;
  - регистриране на нови думи и значения;
  - извличане на преводни еквиваленти.

# БНК в съвременната лексикография

- определяне на словника;
- извличане на информация за вариантност;
- извличане на информация за съчетаемост;
- извличане на информация за системни релации;
- съставяне на тълковни дефиниции;
- подбор на илюстративни примери.

# БНК в съвременната лексикография

- *Речник на българския език*;
- Речници на новите думи и значения;
- синонимни, антонимни и др. речници;
- двуезични и многоезични речници;
- др.

# 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),*

*финансиран от ЕСФ и РБългария по ОП “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 ГД “Структурни фондове и международни образователни програми” към МОМН.*

*Европейски социален фонд - ОП „Развитие на човешките ресурси“ 2007-2013*
BG051PO001-3.3.06-0022
„Интегриране на нови практики и знания в обучението по компютърна лингвситика“

# Благодарим за вниманието!

svetla@dcl.bas.bg, diablag@mail.bg, sia_btb@yahoo.com,
zarka@dcl.bas.bg, cvetana@dcl.bas.bg, iva@dcl.bas.bg

http://ibl.bas.bg/