Български език 62 (2015), 3, 102-119
ISSN 0005-4283

Светла Коева, Диана Благоева, Сия Колковска, Цветана Димитрова, Ивелина Стоянова, Светлозара Лесева
Институт за български език „Проф. Л. Андрейчин“
при Българската академия на науките
София, България

БЪЛГАРСКИЯТ НАЦИОНАЛЕН КОРПУС В КОНТЕКСТА НА СЪВРЕМЕННАТА ЛИНГВИСТИКА

(Резюме)

В статията се представя Българският национален корпус (БНК) и са изложени принципите и практиките, възприети при създаването му. Българският национален корпус е голям съвременен корпус, който се състои от ядро от текстове на български език и 47 паралелни корпуса с общ обем 5,4 милиарда думи. Ядрото на БНК съдържа около 1,2 милиарда думи в над 240 000 текста, отразяващи състоянието на българския език от средата на 20. век до днес. Чуждоезиковите текстове са част от паралелните корпуси и имат обем от около 4,2 милиарда думи. В статията са представени структурата и съставът на Българския национален корпус, като се обръща внимание на богатото метаописание на корпусните документи, което позволява търсенето и извличането на информация според разнообразни езикови и извънезикови критерии. Авторите се спират и на принципите за едноезикова и многоезикова (паралелна) анотация на корпуса, основаващи се на единна рамка за описание, която интегрира различните равнища на анотация. Специално внимание е обърнато на разнообразните приложения на корпуса както в областта на обработката на естествения език (най-вече при обучението и тестването на програми за езикова анотация), така и в областта на компютърната лексикография и други лингвистични дисциплини.

Ключови думи: Български национален корпус, паралелни корпуси, дизайн на корпуси, анотация на корпуси, компютърна лингвистика, компютърна лексикография

✉ Светла Коева
svetla@dcl.bas.bg
✉ Диана Благоева
d.blagoeva@ibl.bas.bg
✉ Сия Колковска
sia_btb@yahoo.com
✉ Цветана Димитрова
cvetana@dcl.bas.bg
✉ Ивелина Стоянова
iva@dcl.bas.bg
✉ Светлозара Лесева
zarka@dcl.bas.bg

Публикувано: 19 октомври 2015