

# Електронни езикови ресурси и уеб приложения за редактиране на текстове

---

Електронните езикови ресурси на Института по български език към БАН могат да бъдат в полза на езиковедските и литературоведските изследвания по български език, както и в сферата на обучението.

Към **електронните речници** спадат Граматичен речник на българския език и редица терминологични речници на българския език.

**Граматичният речник на българския език** съдържа около 85 000 думи и позволява автоматично генериране, анализиране и синтезиране на словоформите – около 1 140 000 на брой. Граматичният речник представя в пълна степен формообразуването на синтетични словоформи. Освен като помощно средство при създаването на различни програми за обработка на естествения език (за корекция на правописа, за автоматизирано търсене на всички словоформи на зададена дума), речникът осигурява на изследователите възможност лесно и бързо да се построи парадигмата на произволна дума, включена в него, да се разпознае определена форма като част от парадигмата на съответната дума и да се припишат граматичните характеристики. Речникът е достъпен на адрес: <http://dcl.bas.bg/est/dict.php>.

**Терминологичните речници на българския език** описват думите в определена тематична област. Получени са чрез извличане на специализирана терминология от множество речници – Български тълковен речник, специализирани тематични речници; впоследствие са допълнени чрез извличане на терминология от специализиран тематичен корпус. За всяка словоформа, включена в речниците, се съдържа информация за лемата и граматическите характеристики.

- **Медицинският речник** обхваща специализирана лексика от областта на медицината. Към настоящия момент съдържа 2 875 лема и над 15 000 словоформи.
- **Политическият речник** включва термини от областта на политиката. Към настоящия момент съдържа приблизително 500 лема и над 3 000 словоформи).
- **Икономическият речник** обхваща терминология от областта на икономиката. Към настоящия момент съдържа приблизително 700 лема и около 5 500 словоформи.
- **Речникът с правни и съдебни термини** включва специализирана лексика от областта на правото и съдебната дейност. Към настоящия момент съдържа приблизително 900 лема и около 6 500 словоформи.
- **Спортният речник** обхваща терминология от областта на спорта. Към настоящия момент съдържа приблизително 700 лема и около 5 000 словоформи.

- **Военният речник** включва специализирана военна терминология. Към настоящия момент съдържа приблизително 400 леми и около 3 000 словоформи.

Всички терминологични речници са в процес на обработка, но могат да бъдат използвани по заявка, изпратена до Секцията по компютърна лингвистика към Института за български език на БАН. Речниците осигуряват възможност за анализ на специални корпуси, както и за количествено измерване на специализирана лексика в общи корпуси и различни видове текстове.

### **Корпуси**

**Българският национален корпус (БНК)** е голям представителен корпус за български език, който е свободно достъпен за научноизследователски, образователни и други цели, като достъпът се осъществява с помощта на уеб приложение за разширено търсене. БНК е създаден в Института за български език „Проф. Любомир Андрейчин“ от сътрудници от Секцията по компютърна лингвистика и Секцията за българска лексикология и лексикография. Корпусът постоянно се развива и обогатява с нови текстове, като се цели постигането на представителност и балансираност чрез включването на текстове от различна модалност (писмени и устни), различни периоди на създаване, разнообразни тематични области и жанрове.

Ядрото на БНК е съставено от текстове на български език – оригинални и преводни. Към октомври 2012 българската част на БНК обхваща над 200,000 текстови единици, наброяващи общо над 950 милиона думи. Включени са текстове, публикувани в периода от 1945 до 2011 година, като преобладаващата част са издадени след 2000 година. Всеки текст е снабден с издателски метаданни (име на автор, заглавие на текст, източник и др.) и класификационни метаданни (стил, тематична област, жанр).

Българският национален корпус може да бъде използван както за теоретични изследвания на определени лингвистични явления с цел лингвистично 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.

### **Паралелни корпуси**

В състава на БНК са включени също 33 паралелни чуждоезикови корпуса – английски, немски, френски, славянските и балканските езици, както и други европейски и неевропейски езици. Паралелните корпуси включват само текстове, които имат българско съответствие – оригиналът е на български или имаме български превод на съответния текст, или двата текста са превод от трети език. Българските текстове принадлежат към ядрото на БНК.

Отделните езици не са равномерно представени. Най-голям е английският корпус с 280 млн. думи, има 4 корпуса с големина 100-200 милиона думи, 16 корпуса с 30-52 млн. думи, 7 корпуса с 1-10 млн. думи и останалите с под 1 млн. думи, като най-малките за китайски, японски и исландски съдържат под 50 хил. думи (данните са валидни към октомври 2012).

Основните приложения на паралелните корпуси са в областта на компютърната лингвистика – автоматичен превод, изработване на двуезични лексикални ресурси (речници) и други.

БНК 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.

Достъпът до Българския национален корпус е безплатен и включва две нива:

1. Ограничен достъп – за нерегистрирани потребители, който включва достъп до системата за търсене. Броят на извежданите за всяка заявка срещания е ограничен до 30. Няма възможност за преглед на разширен контекст.

2. Пълнен достъп – за регистрирани потребители. Достъпът е безплатен за некомерсиални цели - научноизследователски, учебни или др. Пълният достъп включва следните функционалности:

- [Изтегляне на части от Българския национален корпус](#)
- [Пълнен достъп до системата за търсене](#)
- [Заявки за колокации](#)
- [Заявки за извличане на подкорпуси](#)
- [Заявки за изготвяне на честотни речници от подкорпуси](#)

**Българският Браун корпус**, създаден съобразно методологията, разработена в университета Браун (Brown university, Providence, Rhode Island, USA) и приложена в създаването на известния Brown Corpus, е част от БНК. Българският Браун корпус включва 500 текста (общо 1 001 286 думи), разпределени в 15 категории от 2 типа – художествени и информативни. Дължината на текстовете е приблизително фиксирана на 2 000 думи като броят на думите варира с оглед на запазване на границите на началното и крайно изречение на всяка извадка. Корпусните единици са части от текстове, създадени или публикувани като първо издание в периода 1990-2005, основната част – след 2000 година. Корпусът е документиран и нормализиран, направени са проверки за погрешна замяна на кирилски с латински букви и за правописни и пунктуационни грешки. Търсене в Българския Браун Корпус може да се прави от страницата за търсене в Българския национален корпус (<http://search.dcl.bas.bg/>).

**Семантичният корпус** е част от Българския Браун корпус. Съдържа 95119 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 може да бъде използван за различни лингвистични изследвания, при които е необходимо разграничаване на значенията на срещнатите в корпуса думи или форми. Корпусът е свободно достъпен на адрес: <http://dcl.bas.bg/semtcor/bg/>.

**Българският PoS аотиран корпус** (Коева 2010а) също е част от Българския Браун корпус. Съдържа 174697 лексикални единици, аотирани с подходящата граматична информация от Българския граматичен речник. Граматическите характеристики на всяка словоформа са определени коректно и еднозначно. Предлага възможности за ефективно търсене на езикови модели и форми в текста. Корпусът е свободно достъпен на адрес: <http://dcl.bas.bg/poscor/bg/>.

### **Лексикални бази данни**

**Българският ФреймНет** представлява семантико-синтактичен речник на българския език, чиито речникови статии съдържат информация за семантичните, синтактичните и лексикално-семантичните ограничения при съчетаемостта на езиковите единици. Изграждането на Българския ФреймНет следва модел за описание (Коева 2008:6), в който се дефинират класове български глаголи, които принадлежат към даден семантичен фрейм на междуезиково равнище; изразяват се с общи морфо-синтактични характеристики (парадигма за лице и число; инхерентна транзитивност, вид на глагола и флективен тип); и към тях могат да се присъединят еквивалентни (комбинации от) задължителни и допустими обкръжения, характеризирани посредством тяхната синтактична категория (включително предлози и въвеждащи думи), синтактична функция и възможност за имплицитна или (зависима) експлицитна употреба. Към настоящия момент Българският ФреймНет включва най-често срещаните български глаголи с всички техни значения и съответстващото им формално семантично и синтактично описание. Всяко едно значение е свързано чрез идентификационен номер със съответния литерал в Българския WordNet и е илюстрирано с примери, взети от Българския национален корпус, които са семантично и синтактично аотирани.

Информацията в Българския Фреймнет може да бъде използвана за изследването на различни лингвистични явления, свързани със семантиката и/или синтаксиса на български глаголи. 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>).

**Уеб приложенията за редактиране на български текстове** включват Граматичния речник на българския език, представен по-рано, и системите за автоматична проверка и корекция на българския правопис WinEst и WinEst+ за Microsoft Office и MacEst за Mac OS.

Разпознават се едновременно различни типове грешки като се генерират предложения за думи с повече от една грешка и предложенията за замяна се подреждат с голяма точност по отношение на вероятността. Системата за автоматична граматична обработка на текстове използва граматичен речник и система за определяне вероятните характеристики на непознати думи, след което определянето на характеристиките на думата за дадено срещане в текста става чрез прилагането на стотици контекстни граматични правила.

Уеб приложенията за редактиране на български текстове са достъпни на интернет страницата <http://dcl.bas.bg/est/>.

Студентите имат възможност за достъп и до редица книги, списания, и други издания по специалността, които могат да се считат за съвременна "виртуална библиотека". Най-известните сред тях са:

- *The Modern Journal of applied Linguistics* (<http://www.mjal.org/index.htm>) – достъпен по интернет
- *Journal of Applied Linguistics* (<http://www.equinoxpub.com/JAL/index>) – платен абонамент
- *Annual Review of Applied Linguistics* – достъпен по интернет (<http://journals.cambridge.org/action/displayJournal?jid=APL>)
- *ITL - International Journal of Applied Linguistics* ([http://poj.peeters-leuven.be/content.php?url=journal&journal\\_code=ITL](http://poj.peeters-leuven.be/content.php?url=journal&journal_code=ITL)) - достъпен по интернет
- *Canadian Journal of Applied Linguistics* (<http://www.aclacal.org/RevueAn.htm>) – двуезичен, достъпен по интернет
- <http://www.ub.edu/filoan/ALJournals.html> – сайт с информация за различни списания в областта на приложната лингвистика