Описание на корпус/подкорпус в рамките на Българския национален корпус

Секция по компютърна лингвистика
Институт за български език
Българска академия на науките

BulNC@dcl.bas.bg

http://ibl.bas.bg/en/BGNC_en.htm

Съдържание

1 Обща информация

Име	Корпус с новини
Базиран на версия на БНК (дата)	27.07.2012
Източник	`http://setimes.com/`
Кратко описание	Корпус от публицистични текстове за Югоизточна Европа.

Данни за езиците в корпуса

Броят думи е приблизителен. Можете да проверите кодовете на езиците по стандарт ISO 639-1 (2-буквени кодове) тук:
`http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes`.

Език	# **файлове**	# **думи**
BG	35337	7882021
BS	20077	4779983
EL	35252	8689232
EN	34104	7282813
HR	33104	7234621
MK	35220	8205944
RO	35187	8584477
SQ	35227	8762032
TR	35245	6826628

2 Спецификация на метаданните

Описанието на БНК се състои от отделни записи за всеки файл. Всеки запис заема един ред и съдържа 25 полета, разделени с `TAB`. Полетата са описани в таблицата.

Метаданни	**Описание**
`filename`	Име на файла
`path`	Локален път до файла в структурата на БНК
`date_added_to_corpus`	Дата на добавяне
`author_info`	Автор (един, много, неизвестен)
`author`	Име на автора
`translator_info`	Преводач (един, много, неизвестен)
`translator`	Име на преводача
`text_info`	Текст (един, много)
`title`	Заглавие
`year_of_creation`	Година на създаване
`publishing_date`	Дата на публикуване
`source_type`	Вид източник (интернет, сканиран, от автора/издателя)
`source`	Източник
`translated`	Преведен
`medium`	Медиум (писмен, устен)
`number_of_words`	Брой думи
`style`	Стил
`genre`	Жанр
`genre_info`	Допълнения към жанра
`domain1`	Тематична област (първостепенна)
`domain2`	Тематична област (второстепенна)
`domain_info`	Допълнения към тематичната област
`notes`	Бележки
`keywords`	Ключови думи
`langs`	Езици

3 Авторски права

Авторски права върху корпуса

Българският национален корпус и подкорпусите, предоставени за ползване, се разпространяват в качеството си на колекции от документи, като всеки документ е снабден с подробно описание – автор, заглавие, източник и др. (където са известни). Разрешените употреби на корпусите включват: промяна на структурата, извличане на подкорпуси, анотация, промяна на описанието и др. Предоставените части от Българския национален корпус се разпространяват със следния лиценз: *Creative Commons Attribution-NonCommercial 3.0 Unported License.*

Условия на лиценза на английски език:
`http://creativecommons.org/licenses/by-nc/3.0/`
Условия на лиценза на български език:
`http://creativecommons.org/licenses/by-nc/2.5/bg/`

Авторски права върху отделните текстове

Текстовете в публичната сфера са посочени като изключения в Закона. Информация нза конкретните условия за отделните корпуси.

EUR-LEX – законодателство на Европейския съюз

Езици: EN, DE, PL, RO, EL

Авторски права

©Европейски съюз, 1998-2012

Освен ако не е посочено друго, изтеглянето и възпроизводството за лично ползване или за по-нататъшно разпространение с нетърговски или търговски цели на законови текстове или други документи, които са публично достъпни на уебсайта EUR-Lex, се разрешават при условие, че това бъде отбелязано по следния начин:

'©Европейски съюз, `http://eur-lex.europa.eu/`'

Когато обект на повторно използване е точно цитиран законодателен текст, трябва да се добави следният отказ от отговорност:

'За автентично се счита само законодателството на Европейския съюз, отпечатано в хартиеното издание на Официален вестник на Европейския съюз'.

Писмени преводи на текстове или документи на езици, различни от тези, на които са официалните издания, публикувани на уебсайта EUR-Lex, се разрешават при условие, че това се отбележи на видно място на съответния език, последвано от подходящ отказ от отговорност, също преведен на съответния език:

‘Превод от оригинално издание на [посочете език], публикувано от Службата за публикации на Европейския съюз на уебсайта EUR-Lex: ©Европейски съюз, http://eur-lex.europa.eu/, [пълно заглавие на оригиналния език] Отговорността за превода на[посочете език] е изцяло на [имената на притежателя на авторските права върху превода].’

http://eur-lex.europa.eu/en/editorial/legal_notice.htm

SETimes – новини

Езици: EN, HR, TR, RO, SQ, BS, EL, MK, SR

Информация относно авторските права. Ако не е посочено авторското право, информацията в този сайт е обществено притежание и може да бъде копирана и разпространявана, без да е необходимо изрично разрешение. Цитирането на оригиналния източник на информацията се приветства. Ако на снимка, графика или друг материал е посочено авторско право, за копирането на тези материали трябва да се получи разрешение от оригиналния източник.

http://setimes.com/cocoon/setimes/xhtml/bg/document/setimes/footer/disclaimer/disclaimer

EMEA – Административен корпус от медицински документи

Европейска агенция по лекарствата

The contents of these webpages are ©EMA [1995-2012].

In particular, unless otherwise stated, the Agency, according to current European Union and international legislation1, is the owner of copyright and database rights of this website and its contents.

Information and documents made available on the Agency's webpages are public and may be reproduced and/or distributed, totally or in part, irrespective of the means and/or the formats used, for non-commercial and commercial purposes, provided that the Agency is always acknowledged as the source of the material. Such acknowledgement must be included in each copy of the material.

Citations may be made from such material without prior permission, provided the source is always acknowledged.

The above-mentioned permissions do not apply to content supplied by third parties. Therefore, for documents where the copyright vests in a third party, permission for reproduction must be obtained from this copyright holder.

http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000178.jsp&mid=

Wikipedia – Научно-популярен корпус

100,000+ статии, 40 млн. думи на български език.

Можете да разпространявате текстове под всякаква форма, ако направите признание на авторството по някой от следните начини: а) чрез хипервръзка (където е възможно) или уникален ресурсен локатор (URL) към статията или статиите, от които черпите материали, б) чрез хипервръзка (където е възможно) или уникален ресурсен локатор към алтернативно, устойчиво онлайн копие, което е свободно достъпно и съблюдава Условията ни за ползване и което прави признание на авторството по начин, еквивалентен на признанието, което предоставя този сайт, или в) чрез списъка на всички съавтори. (Всеки списък от съавтори може да бъде филтриран така, че да изключи много малките или несъществени приноси.) Тези правила важат за текстове, оригинално създадени от общността на Уикимедия. По отношение на текстовете, привнесени от външни източници, може да важат и други допълнителни изисквания за признание на авторството, за които ще направим всичко възможно да ви укажем ясно. Например, на страницата може да има шаблон или бележка под линия, указващи, че съдържанието или част от него са оригинално публикувани някъде другаде. Ако самата страница съдържа такива означения, в общия случай е редно потребителите да ги запазват, когато използват съдържание повторно за свои цели.

Ако правите промени или допълнения към съдържанието на страницата, която на свой ред използвате, трябва да лицензирате така получената производна работа под лиценза Криейтив Комънс Признание - Споделяне на споделеното, версия 3.0 или по-късна.

`http://bg.wikipedia.org/wiki/Уикипедия:Условия_за_ползване`