Description of corpus/subcorpus within the Bulgarian National Corpus

Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

BulNC@dcl.bas.bg

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

Contents

1 General information

Name	News corpus
Based on version of BulNC (date)	27.07.2012
Source	http://setimes.com/
Short description	Corpus of journalistic texts about South-East Europe.

Avaiability of multilingual data

Number of tokens is only approximate. Check ISO 639-1 (2-letter) language codes here:
http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes.

Language	# files	# tokens
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 Metadata specification

The description of BulNC consists of individual records for each file. Each record is placed on a separate line and comprises 25 fields separated by `TAB`. The fields are represented in the table below.

Metadata	**Description**
`filename`	File name
`path`	Local path to the file within the BulNC
`date_added_to_corpus`	Date of adding the file into the BulNC
`author_info`	Author (one, many, unknown)
`author`	Author's name
`translator_info`	Translator (one, many, unknown)
`translator`	Name of translator
`text_info`	Text (one, many)
`title`	Title
`year_of_creation`	Year of creation
`publishing_date`	Publishing date
`source_type`	Source type (Internet, scanning, from the author/publisher)
`source`	Source
`translated`	Translated
`medium`	Medium (written, oral)
`number_of_words`	Number of words
`style`	Style
`genre`	Genre
`genre_info`	Additions to genre
`domain1`	Domain (primary)
`domain2`	Domain (secondary)
`domain_info`	Additions to the domain
`notes`	Notes
`keywords`	Key words
`langs`	Languages

3 Copyright

Copyright of the corpus

The BulNC and its subcorpora available for download, are distributed as collections of documents where each document is supplied with extensive metadata – author, title, source, etc. (if available). The permitted uses of the corpus include: restructuring, subcorpora extraction, annotation, metadata modification. The downloadable parts of the BulNC are distributed under the following license: *Creative Commons Attribution-NonCommercial 3.0 Unported License.*

License terms in English are available here:
`http://creativecommons.org/licenses/by-nc/3.0/`
License terms in Bulgarian are available here:
`http://creativecommons.org/licenses/by-nc/2.5/bg/`

Copyright of texts

Texts in the public domain are generally not copyrightable.

EUR-LEX – legislation of the EU

We have 50,000+ documents in Bulgarian and large parallel corpora in 5 other langs: EN, DE, PL, RO, EL

Copyright notice
©European Union, 1998-2012

Except where otherwise stated, downloading and reproduction, for personal use or for further non-commercial or commercial dissemination, of legal texts and other documents publicly available on the EUR-Lex website are authorised provided appropriate acknowledgement is given as follows:

©European Union, `http://eur-lex.europa.eu/`

When legislation proper is reused, the following disclaimer shall be added:

Only European Union legislation printed in the paper edition of the Official Journal of the European Union is deemed authentic.

Translations of texts or documents into languages other than the official language editions displayed on the EUR-Lex website are authorised subject to the condition that due acknowledgement is given at a suitably prominent place, followed by an appropriate disclaimer, both translated into the relevant language:

Translated from the original [specify language] edition published by the Publications Office of the European Union on the EUR-Lex website: ©European Union, `http://eur-lex.europa.eu/`, [full title in source language] Responsibility for the translation into [specify language] lies entirely with [name of the copyright holder of the translation].

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

SETimes – news

We have 30,000+ documens and about 7.5 mln. words for Bulgarian; parallels in 9 languages: EN, HR, TR, RO, SQ, BS, EL, MK, SR

Copyright information. Unless a copyright is indicated, information on the site is in the public domain and may be copied and distributed without permission. Citation of the original source of the information is appreciated. If a copyright is indicated on a photo, graphic or other material, permission to copy these materials must be obtained from the original source.

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

EMEA – Administrative corpus of medical documents

We have ... documents totalling to ... words in Bulgarian and parallel texts in 20+ languages. We have taken the raw texts from OPUS - reorganised them, supplied metadata wherever possible.

European Medicines Agency copyright and limited reproduction notices.

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 – Popular Science

We have 100,000+ articles, 40 mln. words in Bulgarian.

There is no agreement or understanding between you and Wikipedia regarding your use or modification of this information beyond the Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA) and the GNU Free Documentation License (GFDL); neither is anyone at Wikipedia responsible

should someone change, edit, modify or remove any information that you may post on Wikipedia or any of its associated projects.

`http://en.wikipedia.org/wiki/Wikipedia:General_disclaimer`