DEPARTMENT OF
COMPUTATIONAL
LINGUISTICS

Bulgarian National Corpus: modern trends in computational linguistics

Institute for Bulgarian Language,
Bulgarian Academy of Sciences

19 Nov 2013

Introduction

- BulNC consists of:
  - a monolingual (Bulgarian) part

    240,000 documents, 1.2 billion words
  - 47 parallel corpora

    4.2 billion words not equally distributed among languages
- Mainly written language
- Bulgarian part reflects the state of Bulgarian from the middle of 20th century (1945) until present

Principles I

- Task-independent design and uniform approach with respect to language, modality and classification
- Extensibility of the corpus through inclusion of new categories
- Flexibility and robustness of the design allowing reconsideration and restructuring of classificatory information

Principles II

- Accommodating texts belonging to multiple categories
- Easy access to the relevant documents, including simple and efficient extraction of information

Classification I

Classification is based on:

1. Style - general text category combining register, mode, and discourse
    - Administrative,
    - Science,
    - Journalism,
    - Fiction,
    - Informal,
    - Informal/Fiction (film subtitles),
    - Popular science,
    - Popular

Classification II

2. Domain - style-dependent, although sometimes found across styles
3. Genre - style-dependent, associated with the internal formal features of the text

Classification - principles I

Main principles of classification:

- explicit definition of categories,
- clear-cut structure,
- structure flexibility - no rigid predefined structure,
- extensive metadata

Classification - principles II

Flexible structure

The flexible corpus structure is able to accommodate various types of texts and facilitates restructuring and extraction of subcorpora with specific structure and features.

Monolingual Bulgarian kernel

- Size

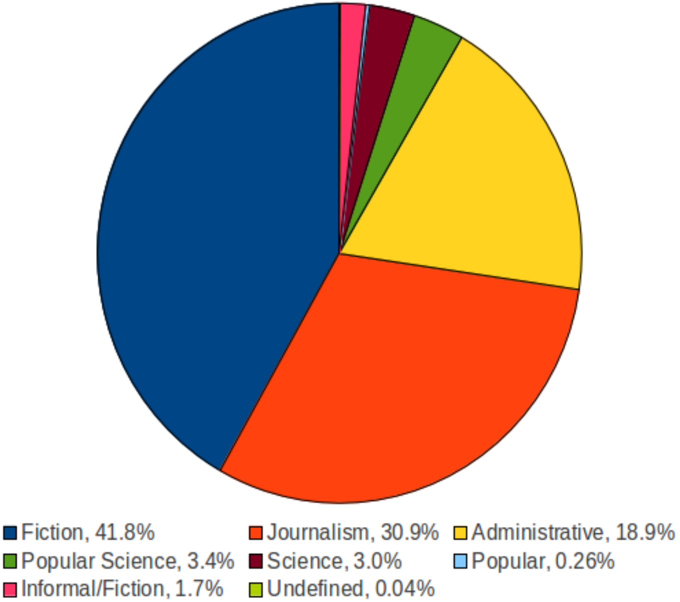
  240 000 text samples, 1.2 billion tokens
- Originality

  37.1% original, 40.5% translated, 22.4% unknown
- Modality

  97.4% written,
  2.6% spoken (lectures, proceedings, subtitles)
- Source

  97.5% from internet, 2.5% from authors/publishers

Bulgarian kernel - by style



Parallel corpora

- Languages

  47 parallel corpora

- Size

  Total 4.2 billion tokens for all foreign languages

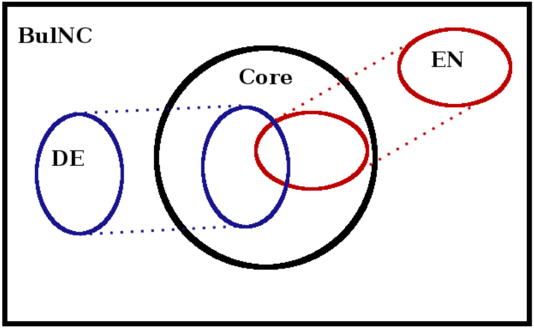
- Structure

  Each foreign language corpus repeats the structure of the Bulgarian kernel

Parallel corpora

- Bulgarian texts are stored once for storage efficiency and linked to parallel equivalents by filename and language code

`0001tABC.txt` `0001tABCen.txt` `0001tABCde.txt`

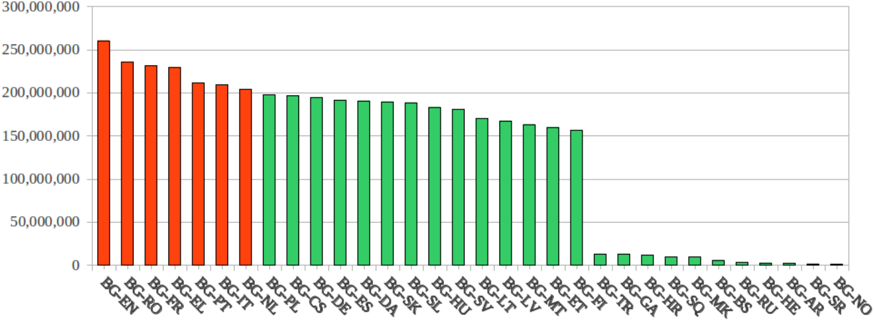


Parallel corpora

Number of corpora of various size:

>260mln.	200-250mln.	150-200mln.	1-15mln.	<1mln.
1	6	14	11	15

Largest parallel corpora in number of tokens:



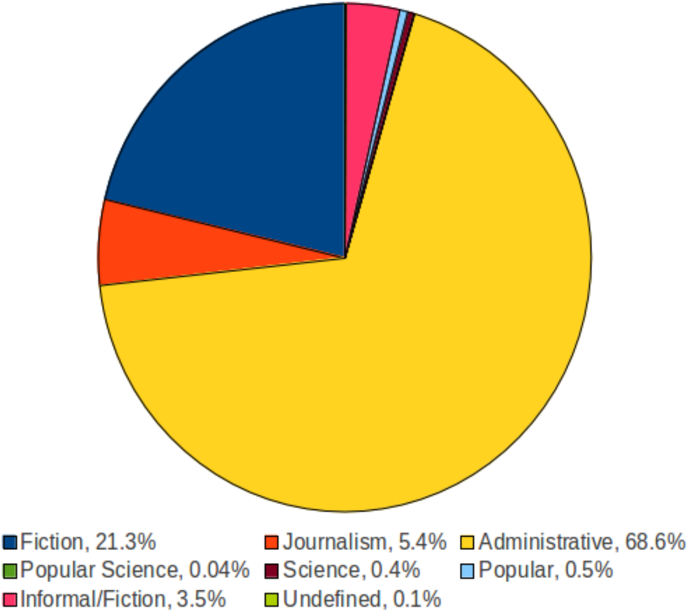
Bulgarian-English parallel corpus

- The largest parallel corpus within BulNC

  260.7 million tokens for English

  263.1 million tokens for Bulgarian

BG-EN parallel corpus - by style



Compilation of BulNC

Three basic approaches:

1. Using readily available text collections:
   - Initial corpora and text archives (55.95% of the corpus)
   - OPUS collection (`http://opus.lingfil.uu.se/`)
2. Manual compilation – browsing and downloading; limited use for a small number of large documents;
3. Automatic compilation – web crawling.

Copyright

Public domain documents

- source acknowledgement (if possible);
- copyright notice or disclaimer acknowledgment (e.g. European Union, `http://eur-lex.europa.eu/`);

Copyright

General Copyright Law (OJ, 2002)

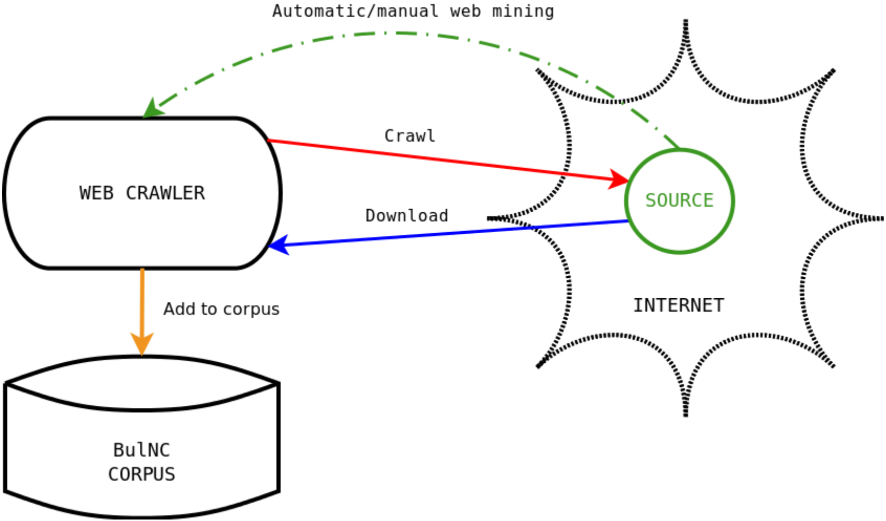
- 3. To use parts of published texts or a relatively small number of texts in other products (texts, collections, etc.) in amount which enables analysis, review or other scientific research; this use is permissible only for scientific and educational purposes with proper citation of the source and the authors name if possible.

Copyright

- redistribution of small portions of text (the context in a search query);
- solely for research and academic purposes;
- extensive metadata, including editorial description (author, text title, source, translator, etc.) whenever these are available.

Automatic compilation

Focused web crawling



Metadata

filename	path_to_file	date_added_to_corpus
author_info	author	translator_info
translator	text_info	title
year_of_creation	publishing_date	source_type
source	translated	medium
number_of_words	style	genre
genre_info	domain1	domain2
domain_info	notes	keywords
languages	quality	accessibility

Linguistic annotation I

Criteria for quality annotation:

- Multi-layered - to cover and accumulate as many levels of linguistic annotation as possible
- Compliance with standards in data formatting and representation of annotation - unification of various tagsets and data formats

Linguistic annotation II

- Uniformity - a common set of attributes and values for different languages, media types, etc. to allow application of language-independent tools
- Consistency - standartisation, validation and evaluation.

Linguistic annotation - monolingual

Bulgarian texts are annotated using the Bulgarian language processing chain:

- Sentence splitter and tokeniser - based on regular expressions
- POS-tagger using SVM
- lemmatiser using a dictionary
- finite-state chunker
- wordnet sense annotation tool

`http://dcl.bas.bg/en/DCLservices.html`

Linguistic annotation - monolingual

English texts are annotated using the following:

- Apache OpenNLP with pre-trained models - sentence segmentation, tokenisation, and POS tagging
- Stanford CoreNLP - sentence segmentation, tokenisation, and POS tagging
- Stanford CoreNLP - lemmatisation
- RASP - lemmatisation

OpenNLP can also be used for other languages.

Linguistic annotation - parallel

Alignment of parallel texts at sentence level is applied using:

- HunAlign
- Maligna

(Both use Gale-Church algorithm)

Automatic alignment at subsentential level is performed for a small part of the Bulgarian-English corpus (BulEnAC).

BulPosCor

- A structured sample from the reference Bulgarian Brown Corpus (174,697 wordforms);
- POS-annotated, with grammatical features disambiguated;
- training and test corpus for POS taggers;
- BulNC is automatically POS-tagged with the BgTagger.

BulSemCor

- A structured sample from the reference Bulgarian Brown Corpus – 95,119 lexical units and 99,480 wordforms;
- POS-annotated, manually sense disambiguated according to the Bulgarian WordNet;
- training and test corpus for a WSD system.

BulEnAC

- A Bulgarian-English Sentence- and Clause-aligned Parallel Corpus – 366,865 tokens altogether;
- automatically sentence-aligned;
- manually clause-aligned;
- training corpus for MT enhancement.

Access to BulNC

DCL Corpora Search

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

- Two languages with uniform result handling
  - Extended Search
  - Regular Expressions
- Corpora Selection
- Metadata Filtering
- Query Assistant
- Alignment Filtering
- Result Details with Parallel Corpora Support

Access to BulNC

Extended Search Queries

- Unordered Sequences

  here comes she

- Ordered Sequences

  $< He * \{POS = V\} chess >$

- Boolean combinations

  $!p, \quad p\&q, \quad p|q, \quad p \Rightarrow q, \quad p \Leftrightarrow q$

Access to BulNC

The plaform

- multiple servers;
- RESTful webservices;
- dynamically added / removed.

Access to BulNC

Download of public-domain subcorpora

- distributed as collections; each document is supplied with extensive metadata - author, title, source, etc. (if available);
- uses: restructuring, subcorpora extraction, annotation, metadata modification;
- distributed under the Creative Commons Attribution-NonCommercial 3.0 Unported License.

Access to BulNC

EUR-LEX – legislation of the EU

- 50,000+ documents in Bulgarian and large parallel corpora in 5+ other languages: at least EN, DE, PL, RO, EL;
- Copyright notice European Union, 1998-2012.

Access to BulNC

SETimes (news in Balkan languages + English)

- 30,000+ documens and about 7.5 mln. words for Bulgarian;
- parallels in 8 Balkan languages and English: EN, HR, TR, RO, SQ, BS, EL, MK, SR.

Access to BulNC

EMEA – Administrative corpus of medical documents

- 18,000 documents in Bulgarian and parallel texts in 20+ languages;
- the raw texts taken from OPUS are reorganised and supplied with metadata wherever possible.

Access to BulNC

Wikipedia – Popular science

- 100,000+ articles, 40 mln. words in Bulgarian;
- Downloaded in XML format via `Special:Export`; extensive metadata are extracted, including domain, keywords, etc.

Access to BulNC

- Frequency dictionaries compiled from the BulNC and some of its subcorpora;
- Collocation webservice – a RESTful web service which supports queries through `http`; the queries return the collocations of a given word in the NoSketchEngine format

Future work

- Extending the corpus by adding new texts and new categories in the classification
- In particular, spoken texts and informal texts (from forums, blogs, etc.)
- Extending metadata description
- Enhancing monolingual and parallel annotation - improving quality and applying on more languages

Thank you!

BulNC@dcl.bas.bg