Bulgarian National Corpus Project

Svetla Koeva, Diana Blagoeva, Siya Kolkovska
Institute for Bulgarian Language – Bulgarian Academy of Sciences
52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria
E-mail: svetla@dcl.bas.bg, diablag@mail.bg, sia_btb@yahoo.com

Abstract

The paper presents Bulgarian National Corpus project (BulNC) - a large-scale, representative, online available corpus of Bulgarian. The BulNC is also a monolingual general corpus, fully morpho-syntactically (and partially semantically) annotated, and manually provided with detailed meta-data descriptions. Presently the Bulgarian National corpus consists of about 320 000 000 graphical words and includes more than 10 000 samples. Briefly the corpus structure and the accepted criteria for representativeness and well-balancing are presented. The query language for advance search of collocations and concordances is demonstrated with some examples - it allows to retrieve word combinations, ordered queries, inflexionally and semantically related words, part-of-speech tags, utilising Boolean operations and grouping as well. The BulNC already plays a significant role in natural language processing of Bulgarian contributing to scientific advances in spelling and grammar checking, word sense disambiguation, speech recognition, text categorisation, topic extraction and machine translation. The BulNC can also be used in different investigations going beyond the linguistics: library studies, social sciences research, teaching methods studies, etc.

1. Introduction

National electronic corpora are representative of the state of a particular language at a certain period of its development. During the last decades, national electronic corpora have been developed for most of the Slavic languages – Croatian, Czech, Polish, Russian, Slovak. The Bulgarian National Corpus – BulNC, (Български национален корпус) is another large Slavic corpus project. It is being developed at the Institute for Bulgarian Language *Prof. L. Andreychin* by researchers from the Department of Computational Linguistics and the Department of Bulgarian Lexicology and Lexicography. The access to The Bulgarian National Corpus is free and available at: http://search.dcl.bas.bg/.

2. General description of the BulNC

The Bulgarian National Corpus project is building a large-scale, representative, publicly available corpus of Bulgarian. The BulNC can also be defined as a monolingual general corpus, fully morpho-syntactically (and partially semantically) annotated. Presently, the Bulgarian National corpus consists of about 320 000 000 graphical words and includes more than 10 000 samples (a detailed description can be found in the publications posted at the Bulgarian National Corpus web site: http://ibl.bas.bg/en/BGNC_en.htm – Figure 1).

Figure 1: Bulgarian National Corpus web site

The BulNC incorporates different types of materials drawn from varied sources: digitalised printed documents and manuscripts, documents that exist in an electronic form only, and transcripts of spoken data. The main characteristics of the BulNC are:

- The corpus represents *only one language* – Bulgarian. It includes modern original and translated Bulgarian texts that contain a limited number of foreign words, be they written in Cyrillic (Russian) or Latin (mainly English).
- The corpus mirrors the *synchronic state* of the language. It covers texts from the middle of the XX century until the present. Since this period is known for a high dynamic in the social and political life of Bulgaria, the corpus offers some opportunities for diachronic investigations as well.
- The corpus is *general*. It includes texts from different styles, thematic domains, genres, etc. In other words, it reflects language diversity and is consequently a representative resource for the analysis of modern Bulgarian.
- The corpus is *large* enough (even according to modern views) collection of written Bulgarian electronic texts. It gives an opportunity for various representative corpus studies in a wide range of branches in applied and theoretical linguistics.

The samples that go into the corpus should meet the following criteria (with some exceptions allowed):

- Originality.
- Recentness – the corpus excerpts are taken from texts created and published after 1945, and if possible – after 2000.
- Domain association – classification into categories and subcategories should correspond to the Brown Corpus[1] classification, which was chosen as a primary model.
- Availability of the source of the text.
- The sample should be an excerpt from a text or texts written by a single author or one team of authors (there are corpus samples accepted composed of more than one

[1] The Brown Corpus model was drawn up in 1963 and its appropriateness can be analyzed on a number of criteria: with respect to the current state of a language; the specifics of Bulgarian fiction and informative prose; etc.

author and with no author specified).

- The corpus sample should be an excerpt from a single text (there are corpus samples accepted composed of more than one text).
- Length between 2 000 and 50 000 words.

The samples of the BulNC are provided manually with detailed meta-data descriptions in XML format. The meta-data description of each corpus sample (following established standards (Atkins et al., 1992; Burnard, 2007); Lee, 2001) includes general and classificatory information: name of the file; information about the author; information about the text – whether there are one or more texts, title(s); form of the text – written, transcribed, etc.; length of the sample in words; date on which the sample was included in the corpus; date (year) of production or first publication/edition; date (year) of publication/edition included in the corpus; information about the source; additional notes. The samples in the Bulgarian National Corpus are also classified according to their *type*: informative or imaginative; *category*: style and/or genre; *subcategory*: it depends on the particular category, thematic domain as well as on the kind of source.

3. BulNC structure in brief

The BulNC incorporates four general sub-corpora provided with uniform meta-data description and morpho-syntactic annotation that facilitates their processing and grouping according to different criteria. These sub-corpora are the Bulgarian Brown Corpus, the Structural Corpus of Bulgarian Electronic Documents 2001–2009, the Structural Corpus of Bulgarian Printed Editions 1945–2009, and transcripts of spoken data.

3.1. Bulgarian Brown Corpus

Bulgarian Brown Corpus[2] (Figure 2) is a structured general corpus of texts in contemporary Bulgarian, published electronically on the Internet in the period 1990–2005.

Random 150 results from 248 occurrences in "Brown" corpus of Bulgarian, DCL.

Figure 2: Bulgarian Brown Corpus

The methodology for the development of the Bulgarian Brown Corpus was as close as possible to the original Brown Corpus in terms of structure and content, but still it differs in some respects: some categories that are either partially or not at all represented in contemporary Bulgarian usage were replaced by more appropriate ones – i.e. besides Mystery and Detective Fiction gangsters' stories and pulp fiction were included. The Bulgarian Brown Corpus consists of 500 samples (each of approximately 2 000 words) forming a total of 1,001,286 words.

The Bulgarian Brown Corpus with full texts comprises samples of original length and amounts to 4.8 million words.

3.2. Structured Corpus of Bulgarian Electronic Documents (2001–2009)

This sub-corpus includes 2,948 samples and amounts to 292 million words. Non-fiction constitutes 78% (2,318 samples) of the total number of samples and 67% (approx. 20 million words) of the corpus content. Fiction is 22% (630 documents) of the total number of samples and 33% (approx. 10 million words) of the corpus content. Of these 62% (1,831 samples and approx. 18.6 million words) are original Bulgarian-language texts and the remaining 38% (1,117 documents and approx. 10.6 million words) – translations from other languages. The meta-data description of each text includes title, author, thematic domain (the domain distribution is given in Table 1), genre (if such can be defined), source, and number of words.

Domain	Number of electronic documents	Number of words (in millions)
Economy	45	0.8
Politics	207	2.2
Law	268	2.6
Medicine	116	0.7
Sport	28	0.4
Military	28	0.2
Society	8	0.1
Administration	963	4.6
Journalism	350	7.3
Fiction	630	9.1
Other	305	1.2

Table 1: Samples distribution in domains

3.3. Structured Corpus of Bulgarian Printed Editions (1945–2009)

This sub-corpus contains over 6,700 samples and amounts to more than 28.5 million words. The corpus includes electronic versions of books (2,110 documents and 16,4 million words, of these 54% originals and 46% translations) and periodicals (4,522 documents comprising 12.1 million words). In chronological terms 842 samples comprising 62 million words are published between 1945 and 1989, and 5689 samples comprising 223 million words between 1990 and 2009.

[2] http://dcl.bas.bg/Corpus/home_en.html

Fiction (Table 2) constitutes 17% (1,132 samples) of the total number and 44% (approx. 12,5 million words) of the corpus content. Non-fiction (the domain distribution is given in Table 3) constitutes 83% (5,600 samples) of the total number of texts and 56% (approx. 160 million words) of the corpus content.

Genre	Number of electronic documents	Number of words (in millions)
Prose	1152	110,37
Poetry	262	1,63
Children's literature	100	3,82
Plays	25	0,9
Fables	3	0,06
Folklore	43	3,28

Table 2: Fiction samples distribution in genres

The informative samples (Table 3) constitute respectively 83% (5 600 documents) from the total number of texts and comprise 56% (app. 160 million words) from the corpus content.

Thematic domain	Number of electronic documents	Number of words (in millions)
Journalism	2064	47
Science	380	23,17
Popular science	16	1,4
Documents/memoirs	81	6,8
Economy & trading	641	45,8
Arts & culture	450	7,14
Military science	148	1,7
Computers	85	0,8
Health	137	0,9
Religion	29	1,1
Other	539	18,65

Table 3: Non-fiction samples distribution in domains

3.4. Transcripts of spoken data

The transcripts of spoken data constitute only 1% of the total number of texts in the BulNC and represent two domains – Bulgarian parliament debates and lectures on homeopathy.

4. Representativeness and balance

It has been widely acknowledged that the reliability of the data and results obtained from corpora depends heavily on corpus parameters such as size, diversity and balance (Biber, 1993; Č ermák et al., 2006; Kennedy, 1998; Sinclair, 1995; among many others). The requirements for corpus representativeness for a given language are:

- a corpus with a given size should contain highly varied texts (different types and chronological periods);
- a corpus with a given content should be large enough to serve for statistically reliable analyses of different language phenomena.

The well established balance should ensure the adequate performance of language diversity in all its functional varieties (Kennedy, 1998; among many others).

Although different models for the compilation of representative general corpora are suggested (some of which have become established standards – e.g. the British National Corpus and the Brown Corpus of Standard American English), the task to build a representative well-balanced general corpus remains a challenge. Establishing an appropriate methodology for the selection of text samples is a central issue in ensuring the representativeness of a corpus, as defined in terms of its ability to represent language phenomena in an adequate way and to ensure unbiased results.

At the first stage of the development of the BulNC a pilot corpus, large in volume and showing diversity with respect to style, genre, thematic domains, and usage domains, has been compiled. In the selection process the EAGLES requirements for text typology selection (Sinclair, 1996) were followed, with some adjustments. A set of thematic domain criteria aiming at the inclusion of a maximum diversity of thematic domains has also been followed. As a result the BulNC illustrates a large number of the styles, genres and thematic domains typical for modern Bulgarian.

Major concerns in building a corpus are representativeness (an optimal relation between the components of the corpus) and diversity of the lexis. More particularly, we measure consistency in terms of the relative share of the core lexis of the language; diversity by the amount of recently coined words, specificity in terms of coverage of special domain vocabulary, etc.

The mutual related task is the definition of indicators showing the correlation between texts distribution and lexis diversity in the BulNC and the actual population of Bulgarian (printed and electronic) texts for a given period – direct correlation with the population of texts is not appropriate; due to the big percentage of media texts the representativeness will predominate well-balancing (Przepiórkowski et al., 2008).

In view of the present structure and diversity of the corpus there are several areas for improvement. The basic problem is the small share of spoken data and that they are limited to two narrow domains. It is necessary to increase the number and variety of scientific texts and other non-fiction, as well as of poetry, drama, and other fiction samples. The enrichment of domain-specific subparts of the corpus (namely some important thematic domains such as economy, society, politics, medicine, sport, etc.) would be very useful having in mind the insufficient number of thematic corpuses for Bulgarian.

5. Annotation

The whole corpus is automatically marked up for sentence and word boundaries and annotated for parts of speech, detailed grammatical information and Bulgarian WordNet word senses. Because of the size of the corpus, only parts of it are manually annotated: 300,000-plus words for parts of speech and 150,000-plus words for word senses. For tagsets and annotation schemata see Koeva et al., 2006a and Koeva et al., 2006b.

We share the understanding that standardisation in morpho-syntactic annotation has to cover both correspondences between different languages as well as language specific features (Ide et al., 2003). In various kinds of annotation we follow established standards to the extent that they are compatible with the language-specific features of Bulgarian. In particular, we have taken into consideration the recommendations of the ISO/TC 37/SC 4 subcommittee, the TEI guidelines (Sperberg-McQueen & Burnard, 2007), the Linguistic Annotation Framework instructions (Ide & Romary, 2007) as well as the principle of two-level structuring for tokens and word forms involving the use of feature structures for morpho-syntactic content (Clément et al., 2005). Basically, during the annotation process (either automatic or manual) we observe the following principles: the input text remains unchanged (normalisation is not done); the annotation is performed consecutively; the annotation data are represented as attribute value pairs in feature-like structures and the annotation data are accumulated rather than overwritten. Thus even the annotation from the lowest level is retrievable, the annotation data are separable (only some parts of annotation can be accessible) and mergeable (some parts of annotation can be combined) to facilitate corpus searches (Ide & Romary, 2007). XML output of the annotated corpus is supported.

6. Tools

Two sets of tools are used within the Bulgarian National Corpus project: linguistic annotation tools and an efficient corpus search engine.

6.1. Annotation Tools

The part-of-speech tagger (Koeva, 2007) used is developed utilizing a large manually annotated corpus consisting of 197,000 tokens (150,000 words) randomly extracted from the Bulgarian Brown corpus (Koeva et al., 2006a). The tagger has been developed as a modified version of the Brill tagger. A sophisticated tokenizer that recognises sentence boundaries and categorises tokens as words, abbreviations, punctuation, numerical expressions, hours, dates and URLs has been built as part of the tagger. For each word in the text the initial (most probable) part of speech from the ambiguity set is assigned from a large inflectional dictionary. The words that are not recognised by the dictionary are handled by the guesser, which analyses the suffixes of the unrecognised words and assigns the initial part of speech from the ambiguity set. The part-of-speech ambiguity ratio calculated over the annotated corpus is 1.51 tags per word, which means that on average every second word is ambiguous. For solving the ambiguity, 144 contextual rules are implemented Some additional techniques for the optimisation are implemented: application of dictionaries of abbreviations, proper nouns, grammatically unambiguous words, etc. After POS tagging, the text remains unchanged, and the additional information is added in XML format. A large inflectional dictionary of Bulgarian is used both for lemmatisation and stemming. Bulgarian WordNet is utilised to assign all available senses to a particular word form, as well as the existing semantic relations such as hypernymy, synonymy, etc.

6.2. Search Tool

The online search tool provides selection of sub-corpora, viewing of the nearest context and information for the examples source (Figure 3).

Bulgarian National Corpus
Department of Computational Linguistics
Corpora Search System
Search
Search
Results

Figure 3: BulNC search system

The corpus is kept in a MySQL data base as a set of objects of the type *sentence* (Tinchev et al., 2007). Each *sentence* object corresponds to a real sentence from the corpus and contains a unique identifier. Each sentence, on the other hand, is represented as a list of the indexes of the words that constitute it. For each syntactic term from the user's query, an object of the type *term* is constructed, consisting of the following elements:

- Word w ∈ W – a word from the natural language;
- Relation r ∈ R – binary lexical relation, which correlates w with a set of words (at the moment the system allows the following relations: word form /F/, synonymy /S/, hypernymy /H/, and similar to/L/);
- Feature structures f ∈ F – grammatical and semantic attributes and their values (for example the category number and its values: singular, plural and quantified plural, e.g. {N=s}). The symbol ***** notates any word characterised by a particular set of grammatical features. The attributes which can be searched for at present are: Part of speech POS, Noun gender G, Noun type NT, Numeral type NUMT, Verb aspect VA, Verb transitivity VT, Pronoun type PT (personal and possessive), Number N, Person P, Gender FG, Definiteness D, Time T, Impersonal verb form IVF.
- Interval i ∈ I. This field specifies the number of arbitrary words in the query. Random words in the ordered search are notated by square brackets **[]**, and their number from – until: with numbers.

The query syntax allows to specify word combinations, ordered queries, inflexionally related words, semantically related words, part-of-speech tags, arbitrary words, etc.

Boolean operations are used to combine the elementary queries with the logical operators: *negation*, *disjunction*, *conjunction, implication,* and *equivalence.* The priorities for action of the operations are expressed through grouping.

7. Applications in linguistics

The Bulgarian National Corpus enables a number of applications in various areas of linguistics: in computational linguistics; in lexicography; within theoretical studies of specific linguistic phenomena; for observations of the characteristics of individual language domains; for extracting example sentences for teaching Bulgarian, etc.

The Bulgarian National Corpus serves as an empirical basis for another national project of the Institute for Bulgarian Language, dealing with the development of a large explanatory dictionary of Bulgarian. The collaboration between corpus compilers and lexicographers ensures professional feedback aiming at the improvement of corpus representativeness and well-balancing.

Some of the more specific applications of the corpus are listed bellow.

7.1. Sub-corpora extraction

Extraction of specific or general sub-corpora according to particular criteria (subject, author, year / period of publication, source, etc.) can be exploited as training corpora for a number of applications – grammatical and semantic tagging, among others, as well as for other research purposes. The following elements from the meta-data description may be used to extract sub-corpora for special research purposes:

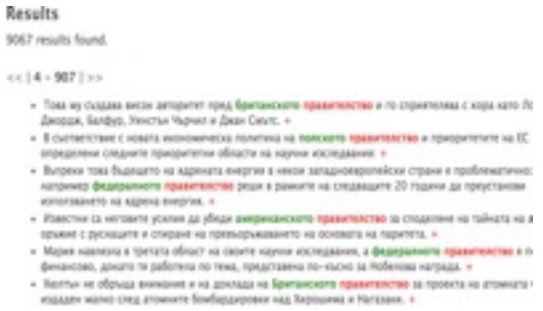
- author (one or several);
- text type (written and/or transcribed);
- originality (originals and/or translations);
- date (year or period of production or first publication);
- source (Internet, electronic or OCR-ed from author or publisher; or transcript of spoken data);
- style (administrative, journalistic, fiction, scientific, colloquial);
- genre (novels, short stories, essays, etc.);
- thematic domain (economy, politics, social science, education, medicine, sport, etc.).

As an example, a sub-corpus created from samples later than 2000 is actively used in the compilation of a dictionary of new words in Bulgarian. Experience show that thematic domain sub-corpora are among the most often used. Thematic domain and genre-based sub-corpora can be also applied in different investigations going beyond the pure linguistic studies. Criteria such as representativeness and well-balancedness apply to special sub-corpora as well.

7.2. Collocations

Investigation of the frequency of words or collocations (word sequences that occur with a certain frequency) and generation of frequency lists are other specialized application of the BulNC. Collocation studies can show how the words combine *lexicall*y (which words are parts of compound words or of multi-word expressions); *syntactically* (which prepositions are used with a given verb), etc. For example, the ordered query **<твърде естраден/F/>**, that matches the word **твърде** (‘too’), immediately followed by any form of the word **естраден** (‘popular’) is not found in BulNC, while the ordered query **<твърде естествен/F/>**, that matches the word **твърде** immediately followed by any form of the word **естествен** (‘natural’), gets 54 hits. The query **<*{POS=A} *{POS=A} акцент>** searching for two adjectives immediately followed by the word **акцент** (**‘**accent’) retrieves 155 hits; there are 14 matches when the first adjective is **подчертан** (‘pronounced’) – **<подчертан [1,1] акцент>** and 0 matches when the first adjective is **небрежен** (‘careless’) **<небрежен [1,1] акцент>**.

The ordered query ***<министерски съвет>&президент*** retrieves 5 sentences where the compound word ***министерски съвет*** (‘council of ministers’) occurs together with the word ***президент*** (‘president’). The query **!български/F/&<*{POS=A} правителство>** retrieves 9,067 sentences where the word ***правителство*** (‘government’) occurs with any other adjective than **български** (‘Bulgarian’), i.e. **британско правителство** (‘British government’**), новото правителство** (‘the new government’), etc. – Figure 4. Some collocation parameters, such as frequency and statistical importance (showing whether the collocation is casual) are not displayed in this version of the search tool interface (although these parameters are retrievable).



Results

9067 results found.

<< | 4 - 907 | >>

- Това му създава висок авторитет пред британското правителство и го сприятелява с хора като Ло Джордж, Балфур, Уинстън Чърчил и Джан Смътс. +
- В съответствие с новата икономическа политика на полското правителство и приоритетите на ЕС определени следните приоритетни области на научни изследвания. +
- Въпреки това бъдещето на ядрената енергия в някои западноевропейски страни е проблематично: например федералното правителство реши в рамките на следващите 20 години да преустанови използването на ядрена енергия. +
- Известни са неговите усилия да убеди американското правителство за споделяне на тайната на а оръжие с руснаците и спиране на превъоръжаването на основата на паритета. +
- Мария навлиза в третата област на своите научни изследвания, а федералното правителство я ф финансово, докато тя работела по тема, представена по-късно за Нобелова награда. +
- Хюлтън не обръща внимание и на доклада на Британското правителство за проекта на атомната издаден малко след атомните бомбардировки над Хирошима и Нагасаки. +

Figure 4. Concordances of **!български/F/&<*{POS=A}**

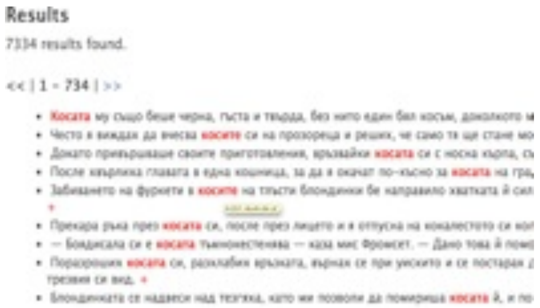
7.3. Concordances

The BulNC search tool excerpts environments (limited to the sentence boundaries) of a given language expression (the concordances). Searches in the BulNC (for instances of particular linguistic phenomena, lexicographic examples or for educational purposes) can be exploited to find all the occurrences of a particular word and its forms; exactly defined forms of a particular word; a word and its synonyms (or other semantic relations from the Bulgarian WordNet).

For instance, the query **хубав/S/** returns concordances of all synonyms of the word **хубав** (‘beautiful’) found in the Bulgarian WordNet together with their inflected forms: **хубав** (‘beautiful’, masculine singular), **хубава** (‘beautiful’, feminine singular), **добър** (‘nice’, masculine singular), **добра** (‘nice’, feminine singular) and so on.

The query **мусака/H/** returns concordances of all hyperonyms of the word **мусака** (‘moussaka’) found in the Bulgarian WordNet and their inflected forms: **ястието** (‘dish’, singular definite), **блюдо** (‘dish’, singular indefinite), **блюда** (‘dish’, plural indefinite) and so on. The query велик/**L**/ returns concordances of all words in the Bulgarian WordNet that are connected with the ‘similar-to’ relation to the adjective **велик** (‘great’) and their inflected forms: **значим** (‘considerable’, masculine), **значима** (‘considerable’, feminine), **голям** (‘great’, masculine, singular), г **олеми** (‘great’, plural), **важен** (‘important’, masculine), **важно** (‘important’, neuter) and so on. The query **рисувам/F/** returns concordances of all synthetic forms of the word **рисувам** (‘to draw’), but the search term does not have to be given in its lemma form – the query **рисуват/F/,** where the verb is 3rd person plural, returns the same results as the query with the lemma.

The query syntax allows the specifying a particular class or category value: noun, common noun, singular, etc. For example, the query **коси/F/{P=2}** matches concordances of the 2nd person forms of the verb **кося** (‘to mow’), while the query **коси/F/{D=df}** matches occurrences of the definite forms of the noun **коса** (‘hair’). Mousing over the retrieved result causes the grammatical annotation for the word in question to pop up (Figure 5).



Results

7334 results found.

<< | 1 - 734 | >>

- Косата му също беше черна, гъста и твърда, без нито един бял косъм, доколкото м
- Често я виждах да вчесва косите си на прозореца и реших, че само тя ще стане мо
- Докато привършваше своите приготовления, връзвайки косата си с носна кърпа, сь
- После кълцаха главата в една кошница, за да я окачат по-късно за косата на гра
- Забиването на фуркети в косите на тлъсти блондинки бе направило хватката й сил
  +
- Прекара ръка през косата си, после през лицето и я отпусна на [illegible]
- — Боядисала си е косата тъмночестенява — каза мис Фроксет. — Дано това й помо
- Поразроших косата си, разхлабих вратовръзката, върнах се при [illegible] и се постарах д трезвия си вид. +
- Блондинката се надвеси над тезгяха, като ми позволи да помириша косата й, и по

Figure 5. Concordances of **коси/F/{D=df}**

Searches can be made for an arbitrary set of words defined by a particular set of grammatical features. For instance, the query ***{POS=A}** returns all sentences where an adjective appears, while the query ***{POS=A D=df}** returns sentences with a definite adjective, and so on.

The advance search option also provides an ordered query that is useful for retrieving concordances of compound words <**машинно** *масло*> (‘engine oil’), analytical word forms <**бил съм отишъл**> (‘I have gone’, non-evidential form), syntactic phrases <**тук не е ходено**> (‘nobody has walked here’), idioms **<една лястовица пролет не прави>** (‘one swallow does not make a summer’) and so on.

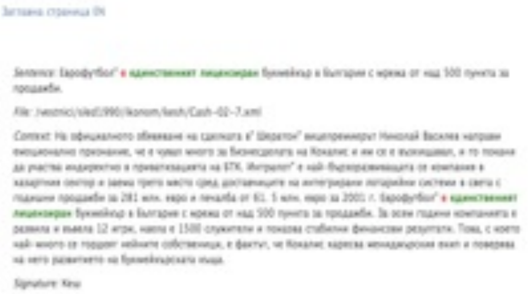
For example the query **<син/F/{D=df} *{POS=N}>** matches concordances of definite forms of the adjective **син** (‘blue’) immediately followed by a noun: **синята вратовръзка** (‘the blue tie’), **синята дама** (‘the blue lady’) and so on. The query **<*{POS=A} както и *{POS=A}>** matches sentences with two adjectives linked with the compound coordinative conjunction **както и** (‘as well as’): **позитивни, както и негативни** (‘positive as well as negative’); **миналият, както и настоящият** (‘the past as well as the present’) and so on.

Ordered queries may include also a random word(s) [at least n, at most n]. For example, the query **<на [1,2] ден>** matches sequences of the preposition **на** (‘of, to’), at least one and at most two random words, followed by the word **ден** (‘day’). The query **<на [2,2] *{POS=N}>** retrieves all sentences where a string consisting of the preposition **на** (‘of, to’), two random words and a noun occurs. The query **<съм/F/ [1,1] *{POS=V IVF=q FG=mf}>** matches sentences where the sequence of a form of the verb ***съм*** (‘to be’), a random word and a past passive participle appear.

Boolean operators and grouping of elements in the queries give more power and efficiency to searches. For example, the query **този&нов** retrieves sentences where both of the words **този** (‘this’) and **нов** (‘new’) occur. The query **този|нов** retrieves sentences where either of the word **този** (‘this’) or the word **нов** (‘new’) occurs. The query **! български/F/&банка** returns sentences where the word **банка** (‘bank’) appears, but none of the forms of the word **български** (‘Bulgarian’). A query with negation of the implication **!(фигура=>шахмат/F/)** retrieves all sentences with the word **фигура** (‘figure’), but without the word **шахмат** (‘chess’).

The query **<Земеделски съюз>|<десен/F/ политически/F/ *{POS=N}>** finds sentences with either the multi-word **Земеделски съюз** (‘Agrarian Union’) or sequences containing any form of the word **десен** (‘right’), any form of the word **политически** (‘political’) plus a noun: **дясното политическо пространство** (‘right-wing political spectrum’), **дясна политическа формация** (‘right-wing political formation’) and so on.

Thus the query system satisfies one of the main aims for external corpus users – to search for word and phrase concordances. The user can also check the nearest context of the retrieved sentences and the meta-data information specifying the author, title, source, and other useful information (Figure 6).



Sentence: Еврофутбол" е единственият лицензиран букмейкър в България с мрежа от над 500 пункта за продажби.

File: /vestnici/sled1990/ikonom/kesh/Cash-02-7.xml

Context: На официалното обявяване на сделката в" Шератон" вицепремиерът Николай Василев направи емоционално признание, че е чувал много за [illegible] на Кокалис и [illegible] [illegible] Еврофутбол" е единственият лицензиран букмейкър в България с мрежа от над 500 пункта за продажби. [illegible]

Signature: [illegible]

Figure 6. The nearest context of the retrieved sentences

8. Conclusion and future work

The Bulgarian National Corpus already plays a significant role in natural language processing of Bulgarian contributing to scientific advances in spelling and grammar checking, word sense disambiguation, speech recognition, text categorisation, topic extraction and machine translation. The BulNC can also be used in different investigations going beyond the linguistics: library studies, social sciences research, teaching methods studies, etc.

The work on the enlargement of the BulNC with respect to its better representativeness and well-balancing is continuing in two directions: 1) further elaboration of criteria and methodology for reliable evaluation, and 2) corpus expansion and reconstruction (mainly trough adding new samples). There are 3,076 samples, approx. 21 million words in total, due to be included in the corpus in the near future. One of the nearest goals is to reach a volume of 1 billion words.

Some of the more concrete tasks are: establishing criteria for spoken data selection, increasing the relative quota of spoken data, adding more samples from thematic domains and genres with low coverage, etc.

A user-friendly interface for sub-corpora selection and query writing is being developed at the moment. The search tool is going to provide various statistical data, more WordNet relations, and WordNet sense definitions. Enhancement of the corpus with additional annotation (i.e. syntactic) will substantially increase its value as a resource for research and education.

To conclude, the Bulgarian National Corpus is a valuable large online available resource for various applied and theoretical research.

9. Acknowledgements

The Bulgarian National Corpus project is supported by The National Science Fund, Ministry of Education, Youth and Science, project ДТК 02/53.

The compilation of the Structured Corpus with Bulgarian Printed Editions (1945-2009) is partially financed by The National Science Fund, Ministry of Education, Youth and Science, project OXH 1512/2005.

The team express its gratitude to the following publishers and institutions, who have kindly provided texts for inclusion in The Bulgarian National Corpus: Bulgarian parliament, Editorial board of *24 chasa* newspaper, *Trud* Publishing house, *Regalia* Publishing House, *Zlatnata yabalka* Publishing house, among many others.

10. References

Atkins et al. (1992) Atkins, B.T.S., Clear, J., and Ostler, N. , 'Corpus Design Criteria' , *Literary and Linguistic Computing,* 7, pp. 1-16.

Biber, D. (1993) Representativeness in corpus design. *Literary and Linguistic Computing.* 8, pp. 243-257.

Burnard (2007) Burnard L. (ed) Reference Guide for the British National Corpus (XML Edition) http://www.natcorp.ox.ac.uk/docs/URG/

Clément et al. (2005) Clément Lionel and Éric de la Clergerie. MAF: a morphosyntactic annotation framework. In *Proc. of the* Language and Technology Conference, Poznan, Poland, pp. 90-94.

Ide et al. (2003) Ide, N., L. Romary, and E. Villemonte de la Clergerie. International standard for a linguistic annotation framework. In Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology. Edmonton. http://www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf

Ide & Romary (2007) Ide, N. and L. Romary. Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (eds.), Evaluation of Text and Speech Systems, Springer, pp. 263-84.

Kennedy (1998) Kennedy, G. An Introduction to Corpus Linguistics. London-New-York, Longman.

Koeva et al. (2006a) Sv. Koeva, Sv. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova. Bulgarian Tagged Corpora. – In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages. Sofia, pp. 78-86.

Koeva et al. (2006b) Sv. Koeva, Sv. Leseva, M. Todorova, Bulgarian Sense Tagged Corpus. In: Proceedings of the 5th SALTMIL Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages, May 23rd 2006, Genoa, Italy, pp. 79-87.

Koeva (2007) Koeva, Sv. Multi-word Term Extraction for Bulgarian, ACL 2007, In: Proceedings of the Workshop on Balto-Slavic NLP, pp. 59-66.

Lee (2001) Lee, D. 'Genres, registers, text types and styles: clarifying the concepts and navigating a path through the BNC Jungle' in *Language Learning and Technology*, vol 5 no 3, September 2001; http://llt.msu.edu/vol5num3/lee/default.html

Przepiórkowski et al. (2008) Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B., and Łazi´nski, M. Towards the National Corpus of Polish. In Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008, Marrakech. ELRA.

Sinclair (1995) Sinclair, J. Corpus typology: a framework for classification. – In: G. Melchers, B. Warren (eds.). Studies in Anglistics. Stockholm, Almqvist & Wiksell, pp. 17-33.

Sinclair (1996) Sinclair, J. Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P. http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Sperberg-McQueen & Burnard (2007) Sperberg-McQueen, C. M. and L. Burnard (eds) TEI Guidelines, P5. Oxford, Virginia, Brown: Text Encoding Initiative, http://www.tei-c.org/P5/.

Čermák et al. (2006) Čermák, F., M. Křen. Large Corpora, Lexical Frequencies and Coverage of Texts. – Proceedings from the Corpus Linguistics Conference Series, vol. 1, № 1 P. Danielsson, M. Wagenmakers (eds.). Birmingham, http://www.corpus.bham.ac.uk/PCLC.

Tinchev et al. (2007) Tinchev, T., Sv. Koeva, B. Rizov, N. Obreshkov. System for advanced search incorpora. In: Literature and writing in Internet, St. Kliment Ohridski University Press, Sofia, pp. 92-111.