TOWARDS LINGUISTICS ANALYSIS OF THE BULGARIAN FOLKLORE DOMAIN

Galina Bogdanova, Konstantin Rangochev,
Desislava Paneva-Marinova, Nikolay Noev

Abstract: *This paper presents an investigation of the lexical structure of the Bulgarian folklore, made during the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage"*[1] *project. This is the first attempt for computational lexical analysis of the Bulgarian folklore and its constituents. Based on this research some linguistic components, aiming to realize different types of analysis of text folk objects are implemented in the Bulgarian folklore digital library. Thus, we lay the foundation of the linguistic analysis services in digital libraries aiding the research of kinds, number and frequency of the lexical units that constitute various folk objects.*

Keywords*: multimedia digital libraries, frequency and concordance dictionaries, systems issues, user issues, online information services, folklore rubrics.*

ACM Classification Keywords*: H.3.5 Online Information Services – Web-based services, H.3.7 Digital Libraries – Collection, Dissemination, System issues*

Linguistics Research and Analysis of the Bulgarian Folklore

The research of the lexical structure of the Bulgarian folklore is very important task for different science domains such as folkloristic, ethnology, linguistics, computational linguistics, etc. Until today, such a linguistic analysis hasn't been made; it is unclear what the lexical structure of Bulgarian folklore works is. During the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" project [Bogdanova et al., 2006] [Paneva-Marinova et al., 2010] [Todorov, 2007] we lay the foundation of the computational lexical analysis of the Bulgarian folklore and its constituents. Our attention was directed to these researches in order to enrich both the content and functionality of the developed multimedia digital library of Bulgarian folklore[2] (also called Bulgarian Folklore Digital Library or BFDL, http://folknow.cc.bas.bg/)

[1] The "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" is a national research project of the Institute of Mathematics and Informatics, supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant No IO-03/2006. Its main goal is to build a multimedia digital library with a set of various objects/collections (homogeneous and heterogeneous), selected from the fund of the Institute for Folklore of the Bulgarian Academy of Science. This research aims to correspond to the European and world requirements for such activities, and to be consistent with the specifics of the presented artefacts [Bogdanova et al., 2008][Berger et al., 2008].

[2] The Bulgarian folklore digital library is built during the "Development of Digital Libraries and Information Portal with Virtual Exposition 'Bulgarian Folklore Heritage'" module of the national research project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage". This

[Pavlov et al., 2010] [Paneva-Marinova et al., 2010] [Rangochev et al., 2007]. Thus we aim to expand the target group of potential users of the library, covering not only those who are interested in Bulgarian folk music, but also narrow specialists in different fields of humanities (folklore, ethnology, linguistics, text linguistics, structural linguistics, etc.). The Bulgarian folklore digital library has a flexible structure that involves additional linguistic components in order to provide real observation and analysis of text folk objects. Digital library with similar analyzing services are presented at [Pavlov et al., 2007] [Pavlova-Draganova et al., 2007] [Pavlov et al., 2006].

As a basis of our research we took the analysis of folklore lexical structure so called main component of the linguistic research of the Bulgarian folklore. We try to answer to the questions: How many and what token it contains? Is there and what is the domination or the lack of some groups of tokens, etc. Until today, such a linguistic analysis hasn't been made; it is unclear what the real lexical structure of Bulgarian folklore works is. With a few exception (for Bulgarian heroic epoc [Rangochev, 1994] and for "Veda Slovena", See http://www.bultreebank.org/veda/index.html) lexical analysis for the Bulgarian folklore and its constituents is missing, the regional characteristics of the folklore lexical structure is unknown. Unfortunately, in 2011 the Bulgarian linguistics, folklore, ethnology, etc. cannot answer the question what are the lexical components of Bulgarian folklore (number, frequency, word forms, etc.) and so far, this type of research is carried out systematically and with a purpose.

This paper presents the basic components of the linguistics research – the different types of dictionaries, frequency dictionaries, concordance dictionaries, terminological dictionaries, valence dictionaries, etc. The paper also describes the BDFL linguistics components for frequency analysis that manipulate the sets of folklore objects of text media type. Finally, the project of a dictionary – concordances of songs, prose, interviews, etc. is outlined.

Classification of Basic Components of the Linguistics Research

The basic components of the linguistics research are the dictionaries. According to accepted definitions every dictionary is a list of words and their meanings in alphabetical order. It is also an alphabetically arranged publication containing information about words, meanings, derivations, spelling, pronunciation, syllabication and usage. (See http://www.web-ezy.com/cit/main/webzglos.htm). The dictionary could give information for pronunciation, grammar, derivatives, history and etymology of the basic word, as well as recommendations for usage, examples, phraseological expressions, examples. Dictionaries are usually in the form of books, but recently electronic dictionaries are more and more recognized.

Qualification of dictionaries is based on different criteria. Many qualifications exist in different lexicographic and lexicological papers [Hartmann, 1993][Svensen, 1993]. Usually dictionaries are combined, which makes them more effective, but this makes their differentiation in categories more difficult.

- By form
    - Traditional dictionaries – they are made with the help of a computer, but their end form is on a paper.

Internet-based environment is a place where folklore objects (mainly from the Funds of the Institute for Folklore at the Bulgarian Academy of Sciences) of different kinds and origins were documented, classified, and „exhibited" in order to be widely accessible to both professional researchers and the wide audience.

  - Digital dictionaries – online (web-based) or local (desktop) dictionaries.
- By their purpose
  - Descriptive dictionaries – descriptive dictionary for the meaning of words according to a common convention;
  - Grammar dictionaries – includes definitions and grammar rules;
  - Dictionary of synonyms – unilingual, includes words with similar meanings;
  - Valence dictionary – dictionaries for the variations of one language, for example – British, Bulgarian, American;
  - Dictionaries of etymology – they trace the development of a language's words in time, giving historical examples, to show the origin and the changes afterwards;
  - Phrase logical dictionaries – dictionaries that present phraseological units of one language. They contain: the most used phraseologies in colloquial speech besides literary units, jargon units, folklore units, vulgarisms and civisms, which are typical for the speech of young people;
  - Frequency dictionaries – gives information on how frequently a word, phrase is used in a particular corpus of texts;
  - Translation dictionaries – bilingual dictionaries, which are used for translation from one language to another;
  - Concordance dictionaries – dictionaries that shows the lexeme with/ in her context.
  - Specialized dictionaries – contain words (terms), that are used by a particular group of people in a professional environment;
  - Terminological dictionaries – contain the most frequently used words with detailed description for each of them;
- By the number and type of languages
  - Unilingual (mono-lingual) dictionaries – unilingual is the dictionary, in which words are described in the same language;
  - Bilingual – dictionaries which contain translation of words in two languages;
  - Multilanguage dictionaries – dictionaries, which contain translations of words in more than two languages.

Frequency Dictionaries and Concordance Dictionaries for Bulgarian Folklore

For the folklore domain more suitable dictionaries are the frequency dictionary and the concordance dictionary [Rangochev et al., 2010]. The frequency dictionary presents the frequency of the lexemes in a definite corpus of texts. It is considered that the facts in one frequency dictionary are reliable enough if there are minimum 20 000 lexical units in it. The frequency dictionaries gave versatile information: presence/ absence of definite lexemes or group of lexemes in comparison with a standard frequency dictionary of the Bulgarian speech [Radovanova, 1968]; frequency of verbs (the so called "verb temperature" [Gerganov et al., 1978] (for the Bulgarian speech at least 21 % verbs in the examined corpus of texts); investigating of the paradigmatic relations in the vocabulary of the text corpus (river- stream- brook- rill…). The domination of group lexemes and respectively small number or absence of other group reveals the constituent characteristics of the text type and its originators [Rangochev, 1994].

- A general frequency dictionary – it contains the all lexical units which are in the BFDL (songs, proverb and descriptions of the rites…);
- A regional frequency dictionary – it contains all the text units which come of a definite folklore region or of a concrete settlement (if there are enough texts). Practically, this is a dialect dictionary of the region/ settlement as far as the folklore regions coincides with the dialect areas.
- A functional frequency dictionary – it contains all the text units which have identical functions: descriptions of the rites, various types of songs, narratives etc. This kind of dictionary would describe some genre specifics of the different parts of the Bulgarian folklore;
- Another dictionary – by user's wish.

The advantage of creating of frequency dictionaries is the possibility to make comparisons between the different types of texts and it can be also followed the tendencies in the dynamics of the lexis – presence/absence of various group of lexemes, etc.

The following table 1 illustrates the comparison of the Bulgarian folklore and spoken languages based on data available in frequency dictionaries.

Concordance dictionaries are these which show the lexeme with/ in her context – it is present the previous one (or more than one) lexeme and the following lexeme according to the examined lexeme. Example: "Fifty heroes are drinking wine" – the underlined lexeme is the examined and the lexemes in italic are her context. Of course, about the songs this could be concordance dictionary of their verses, about the narrative texts (descriptions of the rituals, etc.) – sentences in which they are contained (from point to point…). The creating and using of concordance dictionaries of the texts from BFDL would give good possibilities for folklorists and ethnologists to solve a series of problematic areas as presence/ absence of formulas in the folklore songs and epics, the structure of the folklore text, etc.

Rank list			
Bulgarian spoken language[3]		Bulgarian heroic epoc[4]	
1. съм – 4 041	14. си – 1065	1. съм – 1342	14. го – 338
2. и – 3764	15. казвам – 1 045	2. да – 1 247	15. му – 320
3. да – 3 148	16. тя – 1044	3. си – 548	16. че – 318
4. аз – 2 433	17. викам – 1 031	4. Марко – 1 036	17. а – 286
5. той – 2 288	18. те – 1014	5. се – 828	18. кон – 276
6. не – 1 956	19. какъв – 938	6. на – 801	19. от – 272
7. се – 1 928	20. за – 913	7. и – 796	20. ми – 233
8. този – 1 701	21. че – 874	8. па – 657	21. ти – 225
9. на – 1 669	22. с – 809	9. у – 582	22. що – 222
10. ти – 1 249	23. имам – 768	10. я – 553	23. по – 218
11. ще – 1 183	24. така – 742	11. та – 526	24. добър – 201
12. един – 1 131	25. от – 731	12. не – 412	25. три – 201
13. в – 1 099		13. юнак – 396	

Table 1: Comparison of the Bulgarian folklore and spoken languages

[3] The frequency dictionary is made of texts of the Bulgarian spoken language and the corpus contains 100000 lexemes [Nikolova, 1987].

[4] The frequency dictionary is made of 100 song from [Romanska, 1971] and the texts of the songs contains 7871 verses while there are in it 40042 lexemes.

A Conceptual Framework of a Linguistics Components in the Bulgarian Folklore Digital Library

In the process of the primary testing of BFDL come into being the necessity of insurance of resources for linguistic analysis of the folklore knowledge [Rangochev et al., 2010]. For this aim it was projected and worked out a frequency dictionary with the following functional specification:

- Linguistic analysis of the available multitude of folklore objects of text media type in BFDB;
- Determination of the frequency of meeting the lexemes in text folklore objects;
- Creating of lists of the lexemes,
  - in frequency order
  - in alphabetical order.
- Taking the number of the lexical units;
- Taking the number of the repeats of the lexical units.

Figure 1 depicts the sequence of actions that has to be executed in order to be generated a frequency dictionary. Standard step is the passing through BFDL search service and its sub-functions: 1) user searches by some criteria; 1.1) service performs search in metadata repository, 1.1.1) service gets media data for the found objects, 1.1.1.1) service returns all found media objects by the search criteria, and 1.1.1.1.1) result sent to user. When the result set is generated the user could choose to generate a functional dictionary (step 2). Dictionary generation is performed and the result could be shown by frequency or alphabetically.

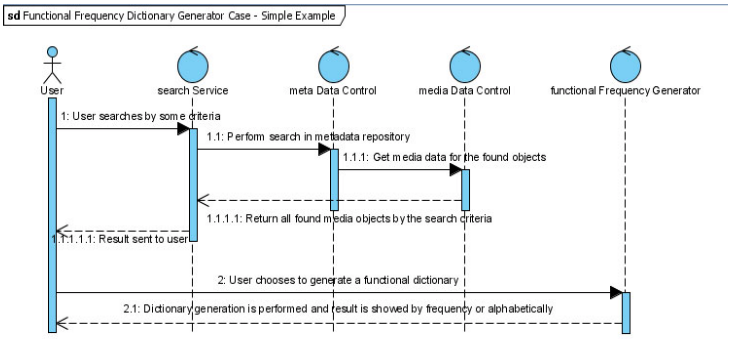
sd Functional Frequency Dictionary Generator Case - Simple Example
User
search Service
meta Data Control
media Data Control
functional Frequency Generator
1: User searches by some criteria
1.1: Perform search in metadata repository
1.1.1: Get media data for the found objects
1.1.1.1: Return all found media objects by the search criteria
1.1.1.1.1: Result sent to user
2: User chooses to generate a functional dictionary
2.1: Dictionary generation is performed and result is showed by frequency or alphabetically

Figure 1: Sequence Diagram

Figure 2 depicts analysis class diagram for the BFDL linguistic component.

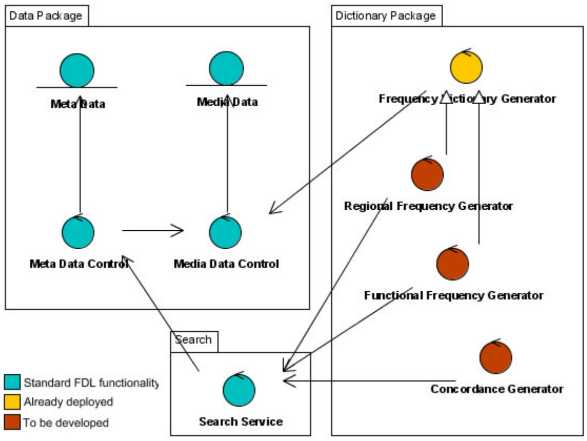

Figure 2: Analysis class diagram

The diagram shows the relations between the data package, the dictionary package and the search service. In the dictionary package there are clearly illustrated different types of generators for frequency dictionary, regional frequency dictionary, functional frequency dictionary and dictionary-concordance [Rangochev et al., 2010].

The Frequency Dictionary Project

The main objective of this project is to build frequency dictionary for texts with folklore themes. The dictionary provides information on how often a particular word or phrase is used in a particular corpus of texts. For the project aims a special hierarchical dataset and WEB interface have been created. The system allows full text search of big corpuses of texts. The dictionary uses rules and concepts in the field of Bulgarian folklore that filter the words/phrases (figure 4). The words/phrases are representatives of 20 different folklore rubrics (thematic headings).

The chosen folklore rubrics are: 1) Village information; 2) Rituals and feasts; 3) Songs; 4) Instrumental music (descriptions); 5) Dance folklore (descriptions); 6) Children folklore; 7) Prose; 8) Proverb, saying; 9) National beliefs and knowledge; 10) National medicine; 11) Magic; 12) Fortune-telling; 13) Dreams; 14) Clothing and adornment; 15) Belongings; 16) National art; 17) Architecture, monuments; 18) Food and feeding; 19) Festivals, gatherings and reviews; 20) Others.

The dictionary serves two types of users: administrator and ordinary user. The administrative area is composed of sections of the main operations of the data modeling. The section for adding of texts allows addition of text and a place for uploading source file. The system has an option for subscription of information from a file. It can upload it to a server as a useful source of reference that can be use by other applications. User area allows the user to search for a word in different sections, the system returns a complete answer on how many times and where the word contains.

The ***administrative part*** contains the following sections for the main operations on data modifying:

- Adding (presented on figure 3):
  - of a text: here the application has a text field, that enables addition of text and a field that enables upload of the source file.
  - of a rubric: the application is simplified to the limit and the administrator chooses the level on which he wants to add a rubric and gives only its name.

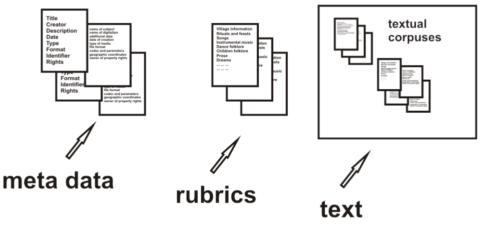

Figure 3: Adding data

Change of a rubric: The application gives an option for a change of the name of the rubric and the unique key is the same. The administrator has an option to choose the rubric, which he wants to change. After the choice is made the text is saved in a field that can be modified. The query the data base is simplified to the highest degree. All needed parameters are given by drop down menus and all rubrics are a part of that menu, which contains their respective identification numbers. There are two types of the query:

"UPDATE rubrics SET rubric_name='" + TextBox1.Text + "' WHERE id = " + Convert.ToInt32(DropDownList1.SelectedValue);.

- ➢ Deletion: After an object is chosen to be deleted at the chosen level, the system deletes cascade all lower levels. The following source code is the query:

```
int sid = Convert.ToInt32(RadioButtonList1.SelectedValue);
if (sid == 1)   {SqlDataSource1.DeleteCommand = "DELETE FROM rubrics WHERE id = " + Convert.ToInt32(DropDownList1.SelectedValue);      SqlDataSource1.Delete();}
```

User part is composed of the search form that allows for selecting a desired item, the level and the corresponding text. The results appeared on the screen in which information rubric, how many files and how the words are distributed.

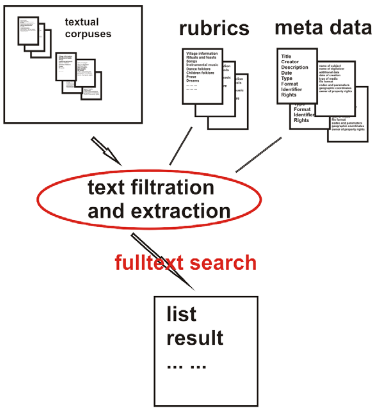

Figure 4: Full-text search

An example of search query is:

```
try  {
string query = "SELECT info_text, path FROM rubric_info WHERE Contains(info_text,@text) AND rid = @rid AND table_id = @tblID";
SqlCommand cmd = new SqlCommand(query, cn);
cmd.Parameters.AddWithValue("@text", TextBox1.Text);
cmd.Parameters.AddWithValue("@tblID", tblId);
cmd.Parameters.AddWithValue("@rid", rid);
cn.Open();
SqlDataReader dr = cmd.ExecuteReader();
int count = 0;
int fileNumber = 1;
while (dr.Read())  {
string text = dr["info_text"].ToString();
string path = dr["path"].ToString();
count += CountWords(text, TextBox1.Text, i, fileNumber,    path);
fileNumber++;
 }
fileNumber -= 1;
Label3.Visible = true;
lbAllFiles.Text = "Number of selected files by this rubric is: " + fileNumber + ".";
lblNumWords.Text = "Number of words is: " + count + ".";
dr.Close();
dr.Dispose();
}
finally     {
cn.Close();
}
```

Because of the nature of the task, the usage of the following additional function is needed. It counts words in the respective texts.

```
private int CountWords(string text, string word, int i, int fNum, string fileName)      {
char[] delims = new char[] { ' ', '.', ',', ';',':',"",'\'', '\t', '\n', '\0' };
foreach(string s in text.Split(delims)){
if (s.ToLower() == word.ToLower()) i++;
}
return i;
}
```

In result there is shown information how many files there are in every rubric, how words are divided, etc.

MsSQL, Visual Studio, HTML, CSS, JavaScript are used for the creation of the dictionary. A hierarchical structure of data (tree) is used for organization of data. The hierarchical structure of data has tables included for administration of rubrics (categories) and growing of the tree structure is allowed in volume and depth.

The system offers uses an easy and fast search system, due to the hierarchy of the data. It enables introduction of many different rubrics and nevertheless they don’t influence the speed of searching. The individual tables contain only the names of rubrics, as well as their keys for organization the hierarchy. The help table contains all texts of all rubrics, organized with the help of indexes, which enables a fast access to the relevant texts and rubrics. There is an option for construction of a dynamic growing of the tree of tables in depth.

Acknowledgements

This work is supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant №IO-03-02/2006 "Development, Annotation and Protection of a Digital Archive "Bulgarian Folklore Heritage"" and №IO-03-03/2006 "Development of Digital Libraries and Information Portal with Virtual Exposition "Bulgarian Folklore Heritage"" from the project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage".

Bibliography

[Berger et al., 2008] Berger, T., Todorov, T., Improving the Watermarking Process with Usage of Block Error-Correcting Codes, Serdica Journal of Computing, 2008, Vol. 2, pp. 163-180.

[Bogdanova et al., 2006] Bogdanova, G., Pavlov, R., Todorov, G., Mateeva, V., Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage, Advances in Bulgarian Science Knowledge, National Center for Information and Documentation, 2006, Vol. 3, pp. 7-15.

[Bogdanova et al., 2008] Bogdanova, G., Todorov, T., Georgieva, Ts., New approaches for development, analyzing and security of multimedia archive of folklore objects, Computer Science Journal of Moldova, 2008, Vol. 16, 2(47), pp.183-208.

[Gerganov et al., 1978] Gerganov, E., Mateeva, A., Experimental Research of the Frequency of the Bulgarian Language, In the Proceedings of the national conference "Contemporary problems of the native language education", Sofia, Bulgaria, 1978.

[Hartmann, 1993] Hartmann, R.R.K., Lexicography. Principles and Practice, Applied Language Studies) London/New York: Academic Press, 1993.

[Nikolova, 1987] Nikolova, C., A frequency dictionary of the Bulgarian spoken language, Sofia, Bulgaria, 1987.

[Paneva-Marinova et al., 2010] Paneva-Marinova, D., Pavlov, R., Rangochev, K., Digital Library for Bulgarian Traditional Culture and Folklore, In the Proceedings of the 3-rd International Conference dedicated on Digital Heritage (EuroMed 2010), Lymassol, Cyprus, 2010, Published by ARCHAEOLINGUA, pp. 167-172.

[Pavlov et al., 2006] Pavlov R., Pavlova-Draganova, L., Draganov, L., Paneva, D., e-Presentation of East-Christian Icon Art, In the Proceedings of the Open Workshop "Semantic Web and Knowledge Technologies Applications", Varna, Bulgaria, 2006, pp. 42-48.

[Pavlov et al., 2007] Pavlov R., Paneva, D., Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, Cybernetics and Information Technologies, 2007, Vol. 6, № 3, pp. 51-62.

[Pavlov et al., 2010] Pavlov, R., Paneva-Marinova, D., Rangochev, K., Goynov, M., Luchev, D., Towards Online Accessibility of Valuable Phenomena of the Bulgarian Folklore Heritage, In the Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech'10), Sofia, Bulgaria, 2010, ACM ICPS Vol. 471, pp. 329-334.

[Pavlova-Draganova et al., 2007] Pavlova-Draganova L., Georgiev, V., Draganov, L., Virtual Encyclopaedia of Bulgarian Iconography, Information Technologies and Knowledge, 2007, Vol.1, №3, pp. 267-271.

[Radovanova, 1968] Radovanova, V., „Representative frequency dictionary of text with length 500 000 tokens", Master thesis, University of Sofia 'St. Kl. Ohridski", Sofia, Bulgaria, 1968.

[Rangochev et al., 2007] Rangochev K., Paneva, D., Luchev, D., Bulgarian Folklore Digital Library, In the Proceedings of the International Conference on Mathematical and Computational Linguistics „30 years Department of Mathematical Linguistics", Sofia, Bulgaria, 2007, pp. 119-124.

[Rangochev et al., 2010] Rangochev, K., Goynov, M., Paneva-Marinnova, D., Luchev, D., Linguistics Research and Analysis of the Bulgarian Folklore, Experimental Implementation of Linguistic Components in Bulgarian Folklore Digital Library, In the Proceedings of the International Conference „Classification, Forecasting, Data Mining" (CFMD 2010), Varna, Bulgaria, 2010, pp. 131-137.

[Rangochev, 1994] Rangochev, K., "Structural particularities of the epic text (using material of the Bulgarian heroic epos)", PhD Thesis, Sofia University "St. Kliment Ohridski, Sofia, Bulgaria, 1994.

[Romanska, 1971] Romanska Cv. (Ed.), "Bulgarian heroic epos", Col. 53, Sofia, Bulgaria, 1971.

[Svensen, 1993] Svensen, Bo, Practical Lexicography: Principles and Methods of Dictionary-Making. Oxford: Oxford University Press, 1993.

[Todorov, 2007] Todorov, T., Performance of an error correction scheme for image watermarking, Proc. of the International Workshop on Optimal codes and Related Topics, White Lagoon, Bulgaria, 2007, pp. 233-236.

Authors' Information

Galina Bogdanova *– PhD in Informatics, Associated Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: galina@math.bas.bg*

Major Fields of Scientific Research: Information Society Technologies, Multimedia Digital Archives, Data Mining, Information Society Technologies, Steganographia, Coding Theory, Computer Science, Algorithms, Knowledge Technologies and Applications.

Konstantin Rangochev *– PhD in Philology, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: krangochev@yahoo.com*

Major Fields of Scientific Research: Ethnology, Folklore studies, Culture Anthropology, Linguistics, Computational Linguistics, Digital Libraries.

Desislava Paneva-Marinova *– PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: dessi@cc.bas.bg*

Major Fields of Scientific Research: Multimedia Digital Libraries, Personalization and Content Adaptivity, eLearning Systems and Standards, Knowledge Technologies and Applications.

Nikolay Noev *– PhD student in Informatics, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: nickey.noev@gmail.com*

Major Fields of Scientific Research: Multimedia Digital Archives, Internet Technologies, Computer Science, Knowledge Technologies and Applications.