Information System for Investigation on the Modern Bulgarian Language

Tihomir Trifonov and Tsvetanka Georgieva-Trifonova
"St. Cyril and St. Methodius" University of Veliko Tarnovo, Bulgaria
tihomirtrifonov@ieee.org, cv.georgieva@uni-vt.bg

***Abstract*—In this paper, an information system for the management of the electronic archive with texts in Bulgarian language is proposed. The developed information system provides the possibility for processing the collected text information. The detailed and comprehensive researches on the letter and the word frequency in the modern Bulgarian language from varied sources (fiction, scientific and popular science literature, press, legal texts, government bulletins, etc.) are performed and the obtained results are represented. They can be utilized by different specialists – computer scientists, linguists, cryptanalysts and others.**

I. Introduction

The frequencies of the letters in the text have often been studied for use in the cryptography [16, 18, 32]. Although the modern ciphers work on bits instead on letters, the values of the frequencies for a given language are still an important tool for the cryptanalysts [7, 27]. More recent analyses show that the letter frequency and the word frequency are distinguishable by the author and by the subject of the examined text. The frequencies of the letters, the bigrams, the trigrams, the words, the lengths of the words and the sentences can be calculated for the particular authors and utilized for confirmation or disproof of the authorship of texts, even for authors whose styles are not so divergent.

The results from researches on the frequencies of the letters and the words in the text are also performed for solving the problems related to keyboard layouts [1, 6, 25, 29, 30, 13].

The accurate relative frequencies of the letters can be gleaned by analyzing a large amount of representative text. The availabilities of modern technologies and tools for computer calculation, as well as for collection and storage of large amount of text corpora, facilitate the accomplishment of such computations. The applications for databases related to the domain of text mining [2, 3, 26, 34] acquire an increasing interest. In these types of applications the frequencies of the words occurring in the texts have significant importance.

The present research is motivated by the lack of detailed and comprehensive results from computations of the frequencies of the letters and the words in the Bulgarian language. In this paper, a realized information system for maintaining and processing texts in Bulgarian language is represented. The results from the accomplished researches on the frequencies of the letters and the words in the modern Bulgarian language are given from various sources (fiction, scientific and popular scientific literature, newspapers, legal texts, governmental bulletins and other genres). The relative frequencies of the letters, the bigrams, the trigrams, the words, the lengths of the words, the first and the last letters in the words, the words with equal lengths, the average length of the words are computed, as well as the index of coincidence.

The rest of the paper is organized as follows. Section 2 contains historical notes for the Bulgarian alphabet and a survey of the existing researches on the frequencies of the letters and the words in the Bulgarian language. In Section 3, a realized information system for maintaining an electronic archive with texts in Bulgarian language is described. In Section 4, the obtained results for the frequencies of the letters and the words in the modern Bulgarian language are represented.

II. Researches on the Frequencies of the Letters and the Words in the Bulgarian Language

The frequency of the occurring the letters, the bigrams, the trigrams, the first and the last letters in the words, the average length of the words, the frequencies of the words reflect the way by which the people use their own language and determine unique characteristics of this language.

Detailed and comprehensive researches on the frequencies of the letters and the words in the English language are already published. The relative frequencies of the letters in the English alphabet are represented in [18]. The first twelve most frequent letters in the English alphabet comprise about 80% of the total used letters. LetterFrequency.org [17] provides detailed information on frequencies of the letters and the words in the English language obtained from various sources (press reporting, religious texts, scientific texts and general fiction). The represented results show differences especially for the general fiction with the position of the letters 'h' and 'i'. Besides the frequencies of the letters in the English language, in [16] the frequencies of the space and others non-alphabetic characters such as digits, punctuation, etc., are computed.

As has be mentioned before, this paper aims to provide detailed and extensive researches on the frequencies of the letters and the words in the modern Bulgarian language. In the present section, a historical information is included, about the origin and the development of the Bulgarian alphabet, as well as a survey of the calculations for the frequencies of the letters and the words in the Bulgarian language performed until now.

A. Origin and Development of the Bulgarian Language

The modern Bulgarian alphabet [8, 11, 21] is a modernized version of the Cyrillic alphabet, developed in the 9th century A.D. The basis of the early Cyrillic is

founded on the Glagolhic alphabet from the 9th century in which some letters are borrowed from the Greek alphabet. The Saints brothers Cyril and Methodius, monks from Thessalonica, are the creators of the Glagolhic alphabet. This alphabet has significant importance not only for the Slavonic nations, but for entire Europe, because by its usage the dissemination of the Christian culture becomes comprehensible for these nations languages [22]. Non fortuitously the device on the highest Bulgarian award after the creation of the Third Bulgarian state "St. St. Equal to the Apostles Cyril and Methodius" is "Light from the East – Ex Oriente Lux". The early Cyrillic is developed from Saint Clement from Ohrid, a follower of the Saint Cyril and Saint Methodius, as well as others their followers which are worked at Preslav school in the north-eastern Bulgaria.

The modern languages based on the Cyrillic [12], use alphabet, obtained by removing some superfluous letters. The Bulgarian alphabet consists of 30 letters, the Russian has 32 letters (33, with inclusion of the soft sign), the Serbian has 30 letters, the Ukrainian – 32 (33). The modern Cyrillic is adapted for many non-Slavonic languages, in some cases with addition of the special letters. The languages utilizing Cyrillic are six Slavonic languages – Bulgarian, Russian, Ukrainian, Belarusian, Serbian, Macedonian; one Persian language – Tadjik; three Turkic languages – Kazakh, Uzbek, Kyrgyz; one Altaic language – Mongolian.

Today the nations writing by using Cyrillic alphabet are more than 60 nations living in seven countries – Bulgaria, Macedonia, Russia, Ukraine, Belarus, Serbia and Mongolia, composing about 10% from the population on the Earth. Together with the Latin and the Arabian alphabets, the Cyrillic is one of the three most used alphabets, and with the Latin alphabet – one of two alphabets of the world, on whose basis new nations build their scripts.

With the membership of Bulgaria to the European Union on 1 January 2007, Cyrillic alphabet became the third official alphabet of the European Union, along with Latin and Greek.

B. Frequencies of the Letters and the Words in the Bulgarian Language

[24] provides information about the arrangement of the Bulgarian letters by descending order of their frequencies in a fragment of the novel „Under the yoke". The authors are examined a fragment containing 131050 letters and the spaces between words.

In [29] the electronic issue of the novel „Under the yoke" is used [33] and the frequencies of the letters in the entire novel are found.

In the rubric "The frequency of the letters in texts in Bulgarian and English language" of the paper [30] the results in percentage are given from examinations of the authors on the frequencies of the letters in some Bulgarian texts.

The problems related to the usage of the Cyrillic alphabet in the electronic communication are discussed in [13]. The possibilities for improving the keyboard layout and proposing a new standard for keyboard layout are considered. The results from the study of the ANABELA (Association for the National Bulgarian Electronic Archive) are obtained on two text corpora – the first is from the 50 million running words general corpus of the Bulgarian language, the second is from a 1 million running words corpus of texts from the register of state administration.

On [5] Bulgarian texts are available. They are collected from Internet and contain more than 72000000 running words, from which 15% of the texts belong to the fiction, 78% come from newspapers and about 7% are form legal texts, government bulletins and other genres. Besides, a list with the first 100000 most frequent words in the archive are accessible.

A demo version of the system developed by Ognyan Chernokozhev and Atanas Kiryakov morphological analyzer is accessible. The system recognizes the word forms of more than 110000 Bulgarian lexemes and assigns to them the appropriate morphosyntactic characteristics.

The XML-based software system CLaRK for management, storage and extraction of the text documents is proposed in [28]. The main purpose of this system is to support the linguists in their work during building the text corpora.

The performed survey indicates the necessity of the detailed and comprehensive researches on the frequencies of the letters and the words in the Bulgarian language, what already exist for the English and other languages.

III. Information System for Maintaining fnd Processing the Texts of Bulgarian Language

In the present section, the architecture of the developed client/server system is described, as well as the services included in its realization. The structure of the created database, which stores the needed information, is represented. A client application ADP (*access data project*) [23], whose purpose is to provide possibilities for insertion, edition and searching the data, is proposed.

A. Architecture of the System

The architecture of the developed client/server system is based on the two-layer information model (fig. 1).

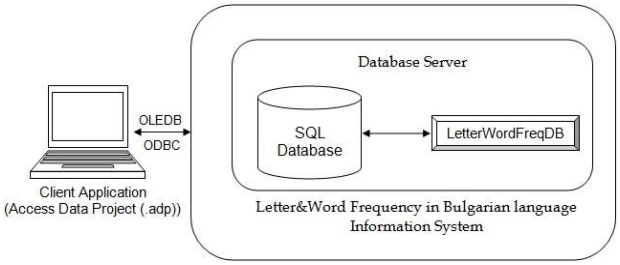

Figure 1. Architecture of the system

The layer for data processing is realized by using the database management system. For the present system we use Microsoft SQL Server, which allows efficient storage of large databases and provides functionality for accessing the data [4, 9, 10, 14, 15, 19, 20].

The client part consists of an ADP application, providing a convenient interface for insertion, updating and searching the data.

B. Database on SQL Server for Data Storage

The database is realized by means of the database management system Microsoft SQL Server. The relevant relational tables are shown in Figure 2.

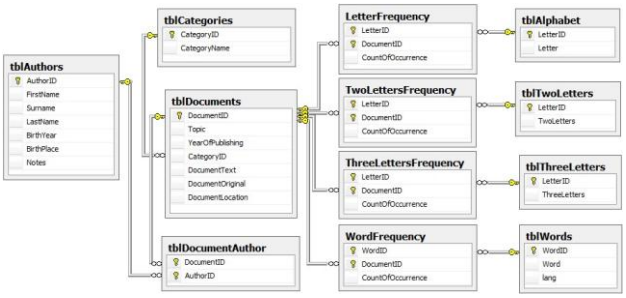

Figure 2. Relational model of the database `LetterWordFreqDB`

The structure of the database is defined to provide the best efficiency of the most frequently used operations – insertion, updating, searching the data.

The `LetterWordFreqDB` database serves for storage and processing the data for the text documents. Information on the topic of the document, the year of publishing, the category (genre), the author(s), the text contained in the document, the location of the file with the document is maintained. For each document the possibility for storage the computed frequencies of the letters and the words in it, is provided.

The basic functions of the database include:

- addition of a new document in the database;
- editing of the data in the documents;
- deletion of documents from the database;
- browsing the data in the documents;
- searching the documents by applying different criteria;
- computing the frequencies of the letters, the bigrams, the trigrams, the words in one or more documents;
- encrypting and decrypting the text in a chosen document with the Vigenere cipher [32].

The `LetterWordFreqDB` database of SQL Server contains the created views for extracting the data from several related tables, as well as the stored procedures for computing the frequencies of the letters and the words in the documents; obtaining the information about the documents published in a fixed year or period; the documents from a chosen category; the documents from a given author. The stored procedures provide a better performance of the client/server system because they decrease the exchange of data between the client and the server. Besides the stored procedures can accept parameters and therefore they can be executed from multiple client applications by applying different input data.

C. Client Application ADP for Insertion and Searching of the Data

Microsoft Access allows the establishment of a connection between the current database and tables from other databases of Microsoft SQL Server and other data sources. ADP is connected with a database of SQL Server and provides an access to the objects created in that database (such as tables, views, stored procedures, triggers, etc.). The data are stored in the database of SQL Server. ADP does not contain any data and tables, but it can be used for easy creation of forms, reports, or macros. As a result of that, the end user has the opportunity for insertion, editing, and deletion of the data by means of a comfortable interface. Forms for insertion and updating the data are realized. Their purpose is to facilitate actualization of the information.

Besides this, the application allows the execution of different queries, which perform finding the specific information, corresponding to the given searching criteria. Each user can search by filling in text boxes and/or list boxes which correspond to the listed characteristics of the documents stored in the database. The results from each query are presented in a format convenient for the end user. The forms and the reports are implemented with the record sources – views and stored procedures designed for:

- extracting the data about documents from a chosen category and/or year (or period) of publishing;
- extracting the data about documents from a chosen author and/or year (or period) of publishing;
- computing the frequencies of the letters from the Bulgarian alphabet in a given document or chosen documents (by category, author, and/or year (or period) of publishing);
- computing the frequencies of the bigrams from the Bulgarian alphabet in a given document or chosen documents;
- computing the frequencies of the trigrams from the Bulgarian alphabet in a given document or chosen documents;
- computing the frequencies of pairs of identical letters from the Bulgarian alphabet in the documents;
- computing the frequencies of the words from the Bulgarian alphabet in a given document or chosen documents;
- computing the frequencies of the words with the same lengths;
- computing the frequency of the serial letter from the beginning or from the end of the words in the documents;
- computing the frequencies of the lengths of the words in the documents;
- computing the frequencies of the letters immediately following the letter 'a';
- computing the frequencies of the letters immediately preceding the letter 'a';
- extracting the data about percentage proportion of the words in Cyrillic alphabet and Latin alphabet in the documents;
- extracting the data about percentage proportion of the words from the different categories of the documents;
- encrypting the text in a chosen document with the Vigenere cipher;
- decrypting the text with the Vigenere cipher.

IV. Relative Frequencies Of the Letters and the Words in the Modern Bulgarian Language

In this section, detailed and comprehensive results from the usage of the realized information system for computing the frequencies of the letters and the words in the collected documents are represented. The calculations are performed on the Bulgarian texts containing more than 1090497 running words, for which 37.14% of the words are found in the texts that belong to fiction, 30.15% come from scientific and popular scientific literature, 13.08% are from newspapers and about 19.63% are from legal texts, government bulletins and other genres.

A. Relative Frequencies of the Letters, Bigrams, Trigrams in the Modern Bulgarian Language

The relative frequencies of the letters in the Bulgarian language for all documents are shown on Table 1, where the letters are arranged in alphabetical order.

TABLE 1.
RELATIVE FREQUENCIES OF THE LETTERS IN THE TEXTS

Letter	Frequency (in percentages)	Letter	Frequency (in percentages)
а	12.323	п	2.928
б	1.533	р	4.920
в	4.449	с	5.081
г	1.567	т	7.604
д	3.526	у	1.312
е	8.972	ф	0.210
ж	0.708	х	0.655
з	2.292	ц	0.595
и	8.875	ч	1.338
й	0.590	ш	0.615
к	3.390	щ	0.612
л	3.284	ъ	1.832
м	2.465	ь	0.014
н	7.136	ю	0.138
о	9.079	я	1.957

Figure 3 also illustrates the relative frequencies of the letters in the Bulgarian language for all documents, but the letters are arranged according to their frequency in the texts.

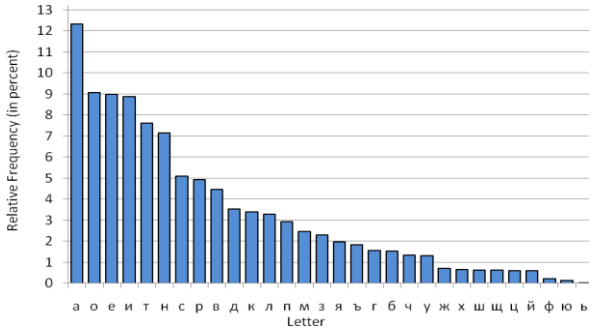

Figure 3. Relative frequencies of the letters in the texts, sorted by their frequencies in descending order

From the obtained results it becomes clear that like the English alphabet the first twelve most frequent letters in the Bulgarian alphabet comprise about 80% of the total used letters.

The relative frequencies of the top thirty most frequent bigrams and trigrams in the Bulgarian language for all documents are presented in Table 2 and Table 3.

TABLE 2.
RELATIVE FREQUENCIES OF THE MOST FREQUENT BIGRAMS IN THE TEXTS

Bigram	Frequency (in percentages)	Bigram	Frequency (in percentages)
на	2.990	от	1.196
то	2.015	за	1.192
та	1.830	ия	1.166
ни	1.705	не	1.149
ст	1.649	по	1.134
ат	1.612	да	1.083
ра	1.482	ко	1.077
ва	1.469	ов	0.970
те	1.453	ка	0.955
ен	1.336	ри	0.894
но	1.269	ед	0.886
ка	1.267	ет	0.883
пр	1.251	ос	0.878
от	1.232	ти	0.873
по	1.200	ли	0.760

TABLE 3.
RELATIVE FREQUENCIES OF THE MOST FREQUENT TRIGRAMS IN THE TEXTS

Trigram	Frequency (in percentages)	Trigram	Frequency (in percentages)
ите	1.091	ава	0.364
ата	0.984	при	0.359
пре	0.611	ане	0.352
ето	0.548	ани	0.348
ени	0.546	раз	0.344
ост	0.492	тел	0.341

ото	0.464	**ият**	0.327
ред	0.425	**ния**	0.326
кат	0.395	**ски**	0.317
ств	0.393	**ато**	0.317
про	0.392	**нит**	0.313
нат	0.381	**ван**	0.303
ста	0.372	**ран**	0.299
ние	0.370	**ест**	0.269
ова	0.366	**ава**	0.364

In the documents the following pairs of identical letters are found, ordered by their frequencies: нн (36.212%), тт (24.556%), ии (21.718%), ее (7.183%), оо (5.083%), дд (1.855%).

B. Relative Frequencies of the Words in the Modern Bulgarian Language

The relative frequencies of the top thirty most frequent words in the Bulgarian language for all documents are represented, sorted by the frequency of occurring (Table 4).

TABLE 4.
RELATIVE FREQUENCIES OF THE MOST FREQUENT WORDS IN THE TEXTS

Word	Frequency (in percentages)	Word	Frequency (in percentages)
на	5.190	**като**	0.519
и	3.627	**са**	0.478
да	2.600	**но**	0.407
в	2.333	**а**	0.390
се	2.031	**до**	0.327
за	1.851	**той**	0.320
от	1.790	**или**	0.318
е	1.509	**му**	0.318
не	1.178	**които**	0.316
с	1.133	**година**	0.297
че	0.974	**България**	0.274
си	0.771	**към**	0.274
по	0.626	**ми**	0.267
това	0.586	**го**	0.262
ще	0.579	**при**	0.246

From the found words in the collected documents we obtain the conclusion that the average length of the words in the Bulgarian language is 5.209. The information system provides a possibility for computing the frequencies of the words with the same lengths.

The realized information system allows computing the frequency of the serial letter from the beginning or from the end of the words in the documents.

The calculations of the frequencies of the lengths of the words in the examined texts indicate that the most used words are the two letter words.

Besides, the set of the words, found in the collected documents is explored and it is established that the most frequent length of the different words is 7.

C. *Index* of Coincidence

The index of coincidence, invented by William Friedman, is applied for analyzing the text containing natural language as well as for cryptanalysis [31]. Moreover this index support establishing whether two texts are written in the same language or in different languages using the same alphabet, because the count of the coincidences for texts in the same languages is distinctly higher than the texts in the different languages.

The expected value for the index of coincidence can be computed from the relative frequencies of the letters f_i (i=1,…,c) on the corresponding language:

$$IC_{\text{expected}} = \frac{\sum_{i=1}^{c} f_i^2}{1/c}$$

where c is the number of the letters in the alphabet.

According our computations the value for the index of coincidence for the letters in the Bulgarian language is 1.93389. For comparison this index for Russian language is 1.76.

We can expect that for an arbitrary string of Bulgarian language its index of coincidence will be approximately equal to this obtained value.

Figure 4 represents the computed indexes of coincidence by the different categories of texts sources. The obtained results show that texts, belonging to the fiction, have the lowest value of the index of coincidence 1.88091, and these belonging to the scientific literature and legal texts have the highest values: 2.00328 and 2.02419, respectively. Obviously, this fact is caused by the greater formalization and specialization to the scientific and legal sources. The proposed considerations can be used in the algorithms for text mining and authorization of documents, books, papers, etc.

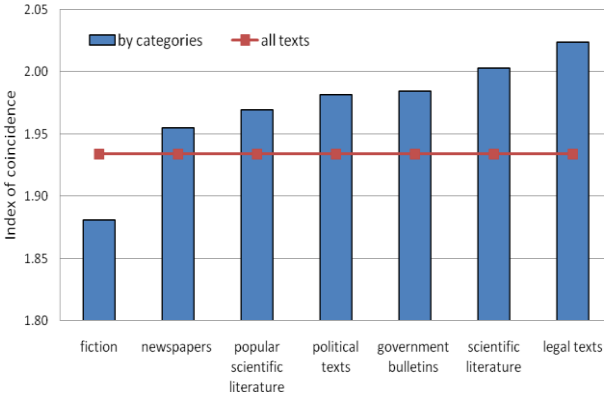

Figure 4. Index of coincidence by categories

V. CONCLUSION

In this paper, the automated system is proposed. It explores a client/server based approach to managing the information on texts in the Bulgarian language. The created database contains information about different characteristics of the documents and it is realized on

Microsoft SQL Server. The interface is developed by means that allow the establishment of a connection with the database of the ADP project. This gives users the possibility of easily accessing detailed information about collected documents.

The realized system allows computing the frequencies of occurring letters, the bigrams, the trigrams, the words in a chosen document or a set of collected documents.

Our future work includes calculating the frequency of the space and other non-alphabetic characters, some phrase, as well as development of the application for mining the constraint-based association rules in the text of the documents, which allows performing the association analysis of the different words from their contents. The improved application will be also utilized for the purposes of the cryptanalysis and the authentication of the text documents in Bulgarian language.

REFERENCES

[1] D. Anson, C. L. Eck, J. King, R. Mooney, C. Sansom, B. Wilkerson, and D. Wychulis, "Efficacy of alternate keyboard configurations: Dvorak vs. Reverse-QWERTY", 2001. Retrieved June 03, 2010, from http://atri.misericordia.edu/Papers/Dvorak.php

[2] A. A. Barsegyan, M. S. Kupriyanov, V. V. Stepanenko, and I. I. Holod, *Technologies for data analysis: data mining, visual mining, text mining, OLAP*. Sankt-Peterburg, Russia: BHV-Peterburg, 2008.

[3] M. W. Berry, *Survey of text mining: clustering, classification, and retrieval*. Springer, 2003.

[4] D. Bieniek, R. Dyess, M. Hotek, J. Loria, A. Machanic, A. Soto, and A. Wiernik, *Microsoft SQL Server 2005 implementation and maintenance – training kit*. Microsoft Press, 2006.

[5] BulTreeBank Group, Available linguistic resources for Bulgarian, 2010. Retrieved June 03, 2010, from http://www.bultreebank.org/Resources.html

[6] M. Burrell, Serbian Dvorak, 2009. Retrieved June 03, 2010, from http://www.csd.uwo.ca/~mburrel/software/serbian-dvorak

[7] Computer news, Computer linguistics, 1999. Retrieved June 03, 2010, from http://kv.minsk.by/index1999262201.htm

[8] F. Curta, *Southeastern Europe in the Middle Ages, 500-1250*. Cambridge University Press, 2006.

[9] H. Garcia-Molina, J. D. Ullman, and J. Widom, *Database systems: The complete book*. Moscow, Russia: Williams, 2002.

[10] M. Gruber, *Mastering SQL*. Sofia, Bulgaria: SoftPress, 2001.

[11] I. Grudev, "The first Slavonic-Bulgarian alphabet – the Cyrillic", 1999. Retrieved June 03, 2010, from http://www.fortunecity.com/victorian/coldwater/293/pyr_k.htm

[12] Encyclopedia Britannica, Cyrillic alphabet, *Encyclopedia Britannica 2009 Student and Home Edition*. Chicago: Encyclopedia Britannica, 2009.

[13] Functional Multilingual Extensions to European Keyboard Layouts, "The Bulgarian alphabet and keyboard in the context of EU communications", 2008. Retrieved June 03, 2010, from http://www.csc.fi/english/pages/meek/The-Bulgarian-Alphabet-and-Keyboard

[14] F. Houlette, *SQL: A beginner's guide*. Sofia, Bulgaria: SoftPress, 2001.

[15] D. M. Kroenke, *Database processing*. Moscow, Russia: Piter, 2003.

[16] E. S. Lee, *Essays about computer security*. University of Cambridge Computer Laboratory, 1999.

[17] LetterFrequency, "Letter & Word Frequency in English / Other Language Frequencies", 2010. Retrieved June 03, 2010, from http://www.letterfrequency.org

[18] R. E. Lewand, *Cryptological mathematics*. The Mathematical Association of America, 2000.

[19] Microsoft Corporation, *MCSE training: Microsoft SQL Server 2000 – design and implementing databases*. Sofia, Bulgaria: SoftPress, 2001.

[20] Microsoft Corporation, *Transact SQL*, 2008. Retrieved June 03, 2010, from http://www.microsoft.com/sql

[21] K. Mirchev, "Konstantin-Cyril, the creator of the Old-Bulgarian literary language", *Journal "Bulgarian language"*, vol. 13, no. 3, 1963.

[22] J. Paul II, "Encyclical epistle slavorum apostoli of the supreme pontiff John Paul II to the bishops, priests and religious families and to all the Christian faithful in commemoration of the eleventh centenary of the evangelizing work of SAINTS CYRIL AND METHODIUS", 1985. Retrieved June 03, 2010, from http://www.vatican.va/holy_father/john_paul_ii/encyclicals/documents/hf_jp-ii_enc_19850602_slavorum-apostoli_en.html

[23] W. Pearson, *MS Access for the business environment: stored procedures from the MS Access client*, 2004. Retrieved June 03, 2010, from http://www.databasejournal.com/features/msaccess/article.php/3363511

[24] B. Penkov, A. Obretenov, B. Sendov, T. Kirpikova, and T. Joukanov, "Frequencies of letters in written Bulgarian", *Acad. bulgare Sci.*, vol. 15, pp. 243-244, 1962.

[25] D. Piepgrass, "The Asset Keyboard", 2006. Retrieved June 03, 2010, from http://millikeys.sourceforge.net/asset

[26] M. Plantevit, T. Charnois, J. Klema, C. Rigotti, and B. Cremilleux, "Combining sequence and itemset mining to discover named entities in biomedical texts: a new type of pattern", *International Journal of Data Mining, Modelling and Management*, vol. 1, no. 2, pp. 119 – 148, 2009.

[27] P. Quaresma, *Frequency analysis of the Portuguese language*. Centre for Informatics and Systems of University of Coimbra, TR 2008/003, ISSN 0874-338X, 2008.

[28] K. Simov, Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, and A. Kiryakov, "CLaRK – an XML-based system for corpora development", *In Proceedings of the Conference on Corpus Linguistics*, pp. 558-560, 2001.

[29] D. Skordev, "Some consideration in relation with the Bulgarian keyboard layouts", 2007. Retrieved June 03, 2010, from http://www.fmi.uni-sofia.bg/fmi/logic/skordev/bg_layouts0.htm

[30] B. Stefanov, and V. Birdanova, "Hygienic-ergonomic evaluation of the computer keyboard layouts", *Computer*, vol. 2, pp. 56-62, 1997.

[31] D. Stinson, *Cryptography: theory and practice*. CRC Press, 2006.

[32] H. C. A. Tilborg, *Fundamentals of cryptology*. Kluwer Academy Publisher, 2000.

[33] I. Vazov, *Under the yoke*, Slovoto, 1999. Retrieved June 03, 2010, from http://www.slovo.bg/showwork.php3?AuID=14&WorkID=5778&Level=1

[34] X.-B., Xue, and Z.-H. Zhou, "Distributional features for text categorization", *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 3, pp. 428-442, 2009.