Chapter 7
Research on Letter and Word Frequency and Mathematical Modeling of Frequency Distributions in the Modern Bulgarian Language

Tihomir Trifonov
St. Cyril and St. Methodius University of Veliko Tarnovo, Bulgaria

Tsvetanka Georgieva-Trifonova
St. Cyril and St. Methodius University of Veliko Tarnovo, Bulgaria

ABSTRACT
The purpose of this chapter is to present current research on the modern Bulgarian language. It is one of the oldest European languages. An information system for the management of the electronic archive with texts in Bulgarian language is described. It 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. The index of coincidence of the Bulgarian language as a whole and for the individual sources is computed. The results can be utilized by different specialists – computer scientists, linguists, cryptanalysts, and others. Furthermore, with mathematical modeling, the authors found the letter and word frequency distributions and their models and they estimated their standard deviations by documents.

INTRODUCTION
The frequencies of the letters in the text have often been studied for use in the cryptography (Lee, 1999; Lewand, 2000; Tilborg, 2000). 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 (Computer news, 1999; Quaresma, 2008). More recent analyses show that the letter frequency and the word frequency are distinguishable by the
Research on Letter and Word Frequency and Mathematical Modeling

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 (Anson et al., 2001; Burrell, 2009; Piepgrass, 2006; Skordev, 2007; Stefanov & Birdanova, 1997; Functional Multilingual Extensions to European Keyboard Layouts, 2008).

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 (Barsegyan, Kupriyanov, Stepanenko, & Holod, 2008; Berry, 2003; Plantevit, Charnois, Klema, Rigotti, & Cremilleux, 2009; Xue & Zhou, 2009) 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 of the words, the words with equal lengths, the average length of the words are computed, as well as the index of coincidence. Moreover, we found the frequency distributions of the length of the distinct words, the length of the words, the letters, the words, the first and the last letters of words, the bigrams, the trigrams and we estimated their standard deviations by documents.

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. In Section 5, the letter and word frequency distributions in the modern Bulgarian language are examined. The results from the computed standard deviations of the letter and word frequencies are summarized in Section 6. The additional results included in Section 7 are evaluating the entropy of the text and the average length of the letter representations in Morse code.

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 of 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 (Lewand, 2000). The first twelve most frequent letters in the English alphabet comprise about 80% of the total used