МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2017
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2017
Proceedings of the Forty-sixth Spring Conference
of the Union of Bulgarian Mathematicians
Borovets, April 9–13, 2017

TEXTS, LETTER FREQUENCIES AND VARIATIONS – A METHOD FOR OPTIMIZATION OF THE RESEARCH[*]

Svilena Hristova

The topic of this study is the following problem: Let T be some text and $T1$ be an excerpt from that text with a length of n. Can a number m $(n > m)$ be found such that the letter frequencies in $T1$ are a good approximation of the letter frequencies in T, i.e. every excerpt $T1$ in T with "length" equal to the number of letters n $(n > m)$ to characterize the letter frequencies in the whole text well enough? We propose the following hypothesizes:

1. $m = 4000$ for texts written in Bulgarian language.
2. The number 4000 does not depend on the language in which the text is written.

In this paper we work with a book written in Bulgarian language.

1. Introduction. There are large differences between the letter frequencies in different texts, depending on the topic, author, etc. [1]. When we study the letter frequencies, we usually work with excerpts of length 1000 letters [2]. It gives a good idea of the letter frequencies in the text. But there are often great deviations from their middle values. The question is whether we can find a number of letters, which guarantees a good approximation of the letter frequencies in the text. This research gives an answer:

Yes, it is enough to count an excerpt of length 4000 letters and we'll have a very good idea of the letter frequencies in the whole text.

In other words, for all excerpts of length 4000 letters the deviations of letter frequencies from the frequencies in the whole text are *little*.

2. The method. We focus on 4 letters – "a", "e", "o" and "r" and the text "Little money bible" in Bulgarian [3]. Let us take an excerpt of length 6000 letters. We divide it into 6 parts, each of length 1000 letters and then we calculate the letter frequencies for each part. After that we calculate the middle values for the excerpt of length 6000 letters. Now we remove the last part and then we calculate the middle values of the frequencies for the rest (5) parts. We do the same for 4 parts and for 3 parts. The results can be seen in the following table:

[*] **2010 Mathematics Subject Classification:** 62-07.

Key words: letter frequencies, optimization.

	a	**e**	**o**	**r**
6 parts	11.08	10.72	9.25	1.48
5 parts	11.2	10.51	9.08	1.6
4 parts	11.28	10.68	8.95	1.73
3 parts	11.73	10.08	9.25	1.87

Let us see the letters “a” and “e”. The deviations of the values for 3 parts from those for 6 parts are much greater than those for 4 parts and 5 parts. Now, we make the assumption that 4 parts is the optimal number which gives quite good information about the letter frequencies.

Now, let us see if this is true for other excerpts of length 4000 letters. We take a random excerpt of length 4000 letters. The results for the letter frequencies are the following:

a	**e**	**o**	**r**
10.78	11.13	10.13	1.13

We see that they are similar to the results for 6 parts, that we obtained above.

We do the same with other random excerpt of length 4000 letters. The result is the following:

a	**e**	**o**	**r**
11	11.25	8.95	1.35

The deviations are little again.

Now, let’s calculate the middle values of letter frequencies and the standard deviations for all 14 parts of length 1000 letters. When we compare them with the results for the 3 excerpts of length 4000 letters, we obtain the following table:

	a	**e**	**o**	**r**
middle value	10.97	10.99	9.41	1.34
standard deviation	0.91	1.19	1.01	0.38
1st excerpt	11.28	10.68	8.95	1.73
2nd excerpt	10.78	11.13	10.13	1.13
3rd excerpt	11	11.25	8.95	1.35

All letter frequencies in the excerpts of length 4000 letters (with one small exception) deviate from the middle values for 14 excerpts of length 1000 letters in the frames of the standard deviation for text length 14 000 letters.

3. The result. We claim that the following is true:

- It is enough to have an excerpt of length 4000 letters to get a good approximation of the middle values for the whole text.
- And we suppose that it is true for any text written in any language.
- Naturally, for some specialized texts, if there are some words in the text, which are very often repeated, then we can remove them. This depends on the aims of the research.

REFERENCES

[1] Б. Стефанов, В. Бирданова. Честота на буквите в текстовете на български и английски език. *Computer*, No 2 (1997), 56–62.

[2] М. Добрева. Моделиране на вариативност в старобългарски текстове. Дисертация, ИМИ – БАН, София, 1999.

[3] С. Уайлд. Малка библия на парите. София, Анхира, 2009

Svilena Hristova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev St, bl. 8
1113 Sofia, Bulgaria
e-mail: svilenajh@abv.bg

ТЕКСТОВЕ, ЧЕСТОТИ НА БУКВИТЕ И ВАРИАЦИИ – МЕТОД ЗА ОПТИМИЗАЦИЯ НА ИЗСЛЕДВАНИЯТА

Свилена Йорданова Христова

Обект на изследване в тази статия е следната задача: Нека T е произволен текст с букви, а $T1$ е част от него с n букви. Може ли да се посочи число m ($m > n$), такова че буквените честоти на $T1$ са добро приближение на буквените честоти на T, т.е. всеки откъс $T1$ от T с „дължина“, равна на броя на буквите n ($n > m$), да характеризира добре буквените честоти на целия текст? Предлагаме следните хипотези:

1. $m = 4000$ за текстове на български език.
2. числото 4000 не зависи от езика, на който е написан текстът.

В статията работим с книга на български език.