**УДК 004.94**

**БАЯЧОРОВА Б.Ж., БАТЫРОВ Ж.**
*КНУ им. Ж. Баласагына, Бишкек*
**BAYACHOROVA B.J., BATYROV J.**
*J. Balasagyn KNU, Bishkek*

# КОМПЬЮТЕРНЫЙ АНАЛИЗ ЧАСТОТНОСТИ БУКВ В КЫРГЫЗСКОМ ТЕКСТЕ

## Кыргызча текстерде тамгалардын жыштыгын компьютерде изилдөө

## Computer analysis of the frequency of letters in the Kyrgyz text

***Аннотация:*** *Описываются программные технологии реализации анализа частотности букв в кыргызском тексте. Приводится структура программы и результаты обработки текстов различного жанра художественной (проза, лирика) и учебной литературы. Результаты обработки различных текстовых данных представлены в виде таблицы и осредненные результаты отражены в виде гистограммы частотности букв.*

***Аннотация:*** *Кыргызча тексттердеги тамгалардын жыштыгын изилдөөнү ишке ашырууга тийиштүү программалык технологиялар сүрөттөлдү. Компьютердик программанын түзмөгү берилип, программалык ишке ашырылышы менен ар кандай жанрдагы көркөм адабияттардагы (проза, лирика), окуу куралдарындагы тексттер боюнча алынган жыйынтыктар келтирилди. Ар түрдүү тексттерди иштеп чыгуу жыйынтыктары таблица түрүндө берилип жана орточо жыйынтыктар гистограмма түрүндө сүрөттөлдү.*

***Annotation:*** *The software technology describes the analysis of the frequency of letters in the Kyrgyz text. The structure of the program and the results of processing texts of various genres of fiction (prose, lyrics) and educational literature are given. The results of processing various text data are presented in tabular form and the averaged results are reflected as a histogram of the frequency of letters.*

***Ключевые слова:*** *частотность букв в тексте, кыргызский алфавит, компьютерная лингвистика.* ***Урунттуу сөздөр:*** *текстеги тамгалардын жыштыгы, кыргыз алфавити, компьютердик лингвистика.*

***Keywords****: frequency of letters in the text, Kyrgyz alphabet, computer linguistics.*

### Введение

Для любого национального языка анализ частотности букв в тексте является одной их важных задач не только в области компьютерной лингвистики, но и в вопросах обеспечения защиты информации, шифрования, дешифрования. В Веб-ресурсах опубликованы информации об исследованиях частотности вхождения букв в текстах, связанные древнеанглийским латинским русским, казахским алфавитами и др. [1, 2, 3], но относительно соответствующих программных разработок нет опубликованы информации. Относительно исследования частотности букв кыргызского, узбекского алфавитов пока не имеются публикации.

Следует отметить, что кыргызский алфавит, основанный на кирилице, был утверждён Указом Президиума Верховного Совета Киргизской ССР 12 сентября 1941 года и состоит из 36 букв, дополнением 3 букв: Ө, ө; Ү, ү; Ң, ң к буквам кирилицы (руского алфавита). В настоящее время в мире кыргызской письменности используются две официальные письменности: одна из них, основанная на кириллице, действует на территории Кыргызстана и в других странах бывшего СССР, а вторая письменность, основанная на арабском алфавите(персидская версия), используется на территории Китая.

В данной статье сделан анализ данных об исследованиях частотности вхождения букв в текстах на разных национальных языках мира. Далее описывается структура разработанной нами компьютерной программы и результаты обработки различного вида текстовых данных.

Об используемых программных технологиях

Для разработки эффективной и достаточно универсальной компьютерной программы анализа частотности букв в кыргызском тексте использованы новые программные технологии, описанные ниже.

- Для разработки программы анализа частотности букв в тексте использованы языки вебпрограммирования **HTML**и**JavaScript.**
- Для создания десктоп программы, не зависящее от браузера, используется фреймворк **Electronjs,** благодаря которому создается возможность собрать .exe файл для запуска приложения на компьютере, не зависимо от браузера.

-Для выполнения **JavaScript**кода в программе вне браузера, использовалась программная платформа **Nodejs**, с помощью которой обеспечивается возможность разработки десктоп программы, и она работает в паре с фреймворк **Electronjs**.

- Для ввода данных, в виде текста, используется тег textarea языка **HTML**.

- Операции выбора отдельных букв в тексте, подсчет частотности и их сортировка описываются на языке **JavaScript.**
- Для выводарезультатов обработки данных в виде графика используется библиотека **Chartjs,** которая отличается своей простотой, и она является бесплатной.

Структура программы анализа частотности букв в тексте

С целью выявления частотности букв в кыргызском тексте, основанной на кириллице, нами разработана компьютерная программана языках JavaScript, HTMLсовместно с программной платформой NodeJS и фреймворк Electron JS обработки текстовых данных на кыргызском языке. Структура программы состоит из следующих этапов:

1. Ввод данных в виде текста организован тегом textarea языка **HTML**.
2. Команды программы, выбора отдельных букв в тексте и подсчет, описаны на языке **JavaScript.**
3. Для формирования десктоп программы, не зависящее от браузера, использован фреймворк **Electron js,** благодаря которому создана возможность собрать .exe файл для запуска приложения на компьютере, не зависимо от браузера.
4. Для выполнения **JavaScript** кода в программе вне браузера, использовали программную платформу **Node js**, с помощью которой обеспечили возможность разработки десктоп программы в паре с фреймворк **Electron js**.
5. На основе использования библиотеки **Chart js** осуществляется вывод результатов в виде гистограммы.

Результаты программной реализации обработки данных

По описанной программе обработаны случайные массивы различных текстовых данныхна кыргызском языке, разного жанра (проза, поэзия, учебная литература: (Ч. Айтматов “Белый пароход” [4], С. Эралиев “Ырлар жыйнагы” [5], Кыргыз тарыхы боюнча кыскача энциклопедия [6). В результате обработки текстовых данных [4, 5, 6] получена нижеуказанная Таблица 1. частоты вхождения букв в кыргызских текстах в алфавитном порядке.

*Таблица 1. Результаты обработки текстовых данных на кыргызском языке [4,5,6].*

| **Буквы** | | **А, а** | **Б, б** | **В, в** | **Г, г** | **Д, д** | **Е, е** | **Ё, ё** | **Ж, ж** |
|---|---|---|---|---|---|---|---|---|---|
| **Частотность в %** | **[3]** | 12.48 | 3.23 | 0.04 | 2.87 | 3.98 | 5.95 | 0 | 1.87 |
| | | 13.56 | 3.32 | 0.06 | 2.73 | 4.10 | 4.76 | 0.007 | 2.47 |
| | **[4]** | 13.002 | 3.02 | 0.43 | 3.67 | 4.12 | 3.78 | 0.007 | 1.90 |

| **З, з** | **И, и** | **Й, й** | **К, к** | **Л, л** | **М, м** | **Н, н** | **Ң, ң** | **О, о** |
|---|---|---|---|---|---|---|---|---|
| 1.29 | 3.69 | 2.76 | 6.94 | 4.88 | 2.60 | 6.66 | 0.93 | 4.60 |
| 1.46 | 3.15 | 2.18 | 6.90 | 5.09 | 2.61 | 6.38 | 0.87 | 3.52 |
| 1.39 | 4.29 | 1.40 | 6.34 | 4.73 | 2.75 | 8.54 | 0.24 | 3.71 |

| **Ө, ө** | **П, п** | **Р, р** | **С, с** | **Т, т** | **У, у** | **Ү, ү** | **Ф, ф** | **Х, х** |
|---|---|---|---|---|---|---|---|---|
| 2.38 | 2.98 | 5.43 | 2.29 | 5.51 | 4.67 | 2.46 | 0.02 | 0.04 |
| 2.91 | 2.88 | 6.65 | 2.56 | 5.70 | 4.15 | 3.27 | 0.06 | 0.03 |
| 2.18 | 1.30 | 6.40 | 2.53 | 5.11 | 4.91 | 1.34 | 0.09 | 0.49 |

| **Ц, ц** | **Ч, ч** | **Ш, ш** | **Щ, щ** | **Ъ, ъ** | **Ы, ы** | **Ь, ь** | **Э, э** | **Ю, ю** | **Я,я** |
|---|---|---|---|---|---|---|---|---|---|
| 0.5 | 1.40 | 1.51 | 0 | 0 | 5.06 | 0.02 | 1.15 | 0.14 | 0.19 |
| 0.003 | 1.2 | 1.61 | 0 | 0 | 5.54 | 0.007 | 0.94 | 0.10 | 0.12 |
| 0.05 | 1.46 | 1.75 | 0 | 0.008 | 6.53 | 0.07 | 0.63 | 0.30 | 0.50 |

Из данных в Таблице 1. методом осреднения результатов частотности букв и сортировкой в порядке убывания, получена нижеуказанная гистограмма, отражающая частотность вхождения букв кыргызского алфавита в различных текстах (Рис.1).

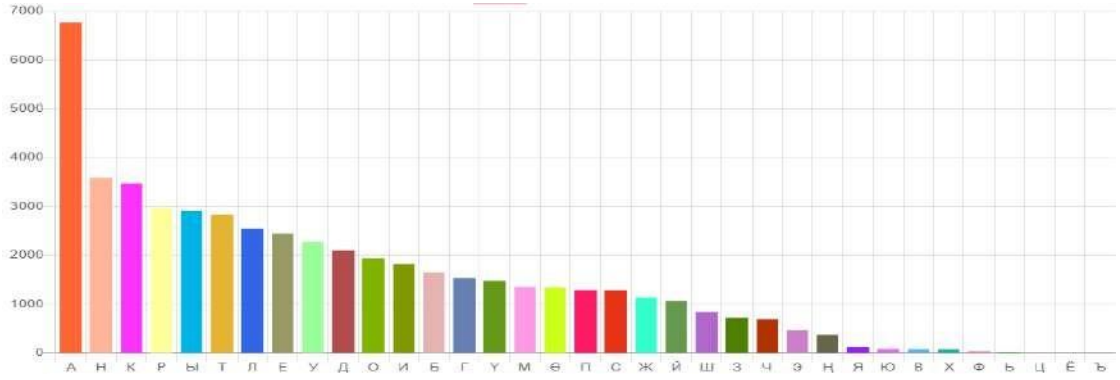


*Рис. 1. Гистограмма частотности букв в кыргызском тесте*

Как видно из гистограммы самой часто используемой буквой в кыргызском тексте является буква А, на втором месте Н, на третьем месте буква К а самые редко используемые буквы Ц, Ё, Ь, а буквы Щ, Ъ отсутствовали в рассмотренных текстах польностью.

Для сопоставления частотности букв в текстах на кыргызском и русском языках представим ниже в виде Табл 2. результаты осредненных данных частотности букв русского алфавита [2], в порядке убывания, представленные Национальным корпусом русского языка (НКРЯ ) - (официальное пополняемое собрание текстов).

*Таблица 2. Результаты обработки текстовых данных на русском языке [2].*

| **Буквы** | **О, о** | **Е, е** | **А, а** | **И, и** | **Н, н** | **Т, т** | **С, с** | **Р, р** |
|---|---|---|---|---|---|---|---|---|
| **Частотность в %** | 10.48 | 8.483 | 7.998 | 7.367 | 6.7 | 6.318 | 5.473 | 4.746 |

| **В, в** | **Л, л** | **К, к** | **М, м** | **Д, д** | **П, п** | **У, у** | **Я,я** | **Ы, ы** |
|---|---|---|---|---|---|---|---|---|
| 4.533 | 4.343 | 3.486 | 3.203 | 2.977 | 2.804 | 2.615 | 2.001 | 1.898 |
| **Ь, ь** | **Г, г** | **З, з** | **Б, б** | **Ч, ч** | **Й, й** | **Х, х** | **Ж, ж** | **Ш, ш** |
| 1.735 | 1.687 | 1.641 | 1.592 | 1.45 | 1.208 | 0.996 | 0.94 | 0.718 |
| **Ю, ю** | **Ц, ц** | **Щ, щ** | **Э, э** | **Ф, ф** | **Ъ, ъ** | **Е.., е..** | | |
| 0.638 | 0.486 | 0.361 | 0.331 | 0.267 | 0.037 | 0.013 | | |

Сопоставление данных из Табл. 1, 2 частотности букв в кыргызском и руссском текстах показывает, что, если в кыргызском тексте буква А имеет самый высокий процент частотности, то в русском тексте таким является буква О, а частотность буквы А находится на 3-ем месте. Интересно отметить, что буквы Ё, Ш, Щ, Ц, Ъ, Ф и в кыргызских и в русских текстах имеют самые низкие проценты частотности.

Отметим, что по разработанной нами компьютерной программе можно обрабатывать тексты сколь угодно большого объема с достаточно быстрой скоростью. Программа оригинальна и без особого труда может адаптироваться применительно к другим национальным языкам, тем самым является достаточно универсальной.

***Список цитируемых источников***


1. https://www.e-reading.club/chapter.php/1002975/39/Tomas_Skarlett_-_Korporaciya_Pops.html
2. https://dpva.ru/Guide/GuideUnitsAlphabets/Alphabets/FrequencyRuLetters/
3. https://lingvoforum.net/index.php?topic=90098.0
4. http://kitepkana.kg/kyzyk-eken/zhanylyktar/orozkuldun-kyyaly.html
5. **5**.http://www.literatura.kg/articles/?aid=798
6. http://bizdin.kg



***Рецензенты: Панков П.С.*** *– доктор физико-математических наук, профессор, член-корр. НАН КР*