ТЕОРЕТИКО-ИНФОРМАЦИОННОЕ ИССЛЕДОВАНИЕ БЕЛОРУССКОГО ЯЗЫКА

Т. А. Хаткевич, Д. А. Коновалов, С. К. Яскевич

В работе проведен статистический анализ текстов прозы на белорусском языке. Для подробного исследования были выбраны произведения 12 белорусских писателей XX-XXI вв., написанные по правилам орфографии 1933 г.: А. Боровский, И. Мележ, И. Пташников, Я. Колас, Я. Брыль, К. Тарасов, К. Черный, М. Зарецкий, В. Орлов, В. Короткевич, В. Быков, З. Бедуля. А также проза 2 писателей, использовавших правила орфографии 1918 г.: В. Быков, К. Акула.

Для исследования в общем мы рассмотрели 650 текстов прозы различных писателей (31 Мб), написанных по правилам орфографии 1933 г., а также 120 текстов (6 Мб), написанных по правилам 1918 г.

Интерес вызвали показатели, связанные с криптоанализом различных шифров, в частности: частоты встречаемости k-грамм, использующиеся для криптоанализа шифров замены и гаммирования; индекс совпадения Фридмана, использующийся при криптоанализе шифра Вижинера; запрещенные биграммы и сочетаемость букв – для уменьшения сложности переборов при криптоанализе. Также подсчитаны значения энтропии Шеннона, что является полезным результатом в области кодирования информации, и построены модели открытых текстов.

Все результаты, приводимые в статье, (кроме частот встречаемости букв и средней длины слова) получены при анализе текстов с удалёнными разделителями между словами.

Хотелось бы заметить, что результаты аналогичных исследований для белорусского языка прежде не публиковались.

1. ЧАСТОТА ВСТРЕЧАЕМОСТИ БУКВ АЛФАВИТА

В табл. 1 приведены показатели, полученные при исследовании всех текстов прозы, как для правил правописания 1918 г., так и 1933 г. (буквы в таблицах записаны по убыванию частоты слева-направо, сверху-вниз).

Из таблицы видно, что, действительно, для правописания 1918 г. в большей степени характерно употребление мягкого знака (его частота здесь почти в 2 раза превышает частоту встречаемости в текстах, написанных по правилам 1933 г.), а также «яконье».

Для сравнения, в русском языке самые частые буквы – это О, Е, А, И, Т, Н [1].

Таблица 1

Орфография 1933 г.			
_ 0,1666	А 0,1346	Н 0,0495	І 0,0436
Л 0,0351	С 0,0347	Ы 0,0345	Р 0,0339
Я 0,0330	Е 0,033	К 0,0326	О 0,0318
Т 0,0284	Д 0,0277	У 0,0275	М 0,0258
В 0,0247	З 0,0246	П 0,0241	Ц 0,0208
Ў 0,0206	Г 0,0163	Б 0,0159	Ч 0,0133
Ь 0,0117	Ш 0,0114	Х 0,0099	Й 0,0082
Э 0,0071	Ж 0,0071	Ю 0,0055	Ё 0,0054
Ф 0,0008	‘ 0,0004		

Орфография 1918 г.			
_ 0,1605	А 0,1342	Н 0,0506	І 0,0416
Ы 0,0355	Я 0,0346	С 0,0344	Р 0,0339
Л 0,0335	К 0,0332	О 0,0308	Е 0,0301
Т 0,0284	Д 0,0281	У 0,0276	М 0,0259
В 0,0252	П 0,0241	З 0,0237	Ц 0,0212
Ь 0,0202	Ў 0,0201	Г 0,0166	Б 0,0158
Ч 0,0123	Ш 0,0114	Й 0,0105	Х 0,0096
Э 0,0085	Ж 0,007	Ю 0,0056	Ё 0,0043
Ф 0,0010	‘ 0,0003		

На основании полученных данных, был вычислен индекс совпадения Фридмана. Для вычисления использовалась следующая формула:

$I_c(x) = \sum_{i=1}^{33} p_i^2$, где p_i – вероятность буквы i.

Получены следующие результаты: для текстов правил орфографии 1918 г. I_c=0,0545, для текстов правил 1933 г. I_c=0,0554.

Сравним с данными для других языков. Например, для русского языка: I_c=0,053, для английского: I_c=0,066.[1]

2. СООТНОШЕНИЕ ГЛАСНЫХ И СОГЛАСНЫХ В БЕЛОРУССКОМ ЯЗЫКЕ

Получены следующие данные. В текстах правил 1933 г. из 31256768 знаков 13362318 (42,75%) гласных. Согласные, «ь» и «’» составляют 57,25%. Соотношения в текстах орфографии 1918 г. (6200330 знаков): 42,01% и 57,99% соотв. Среди писателей наибольшее содержание гласных у И. Пташникова: 43,20 %. Наименьшее у К. Тарасова: 41,99 % и у К. Акулы: 41,72 %.

Для сравнения, в других языках процент гласных: русский – 43,20%, английский – 39,21%, немецкий – 39,27%, испанский – 47,95%. [1].

3. СРЕДНЯЯ ДЛИНА СЛОВА

Средняя длина слова в текстах правил 1918 г. составляет 5,2327, что превосходит соответствующий показатель у текстов орфографии 1933 г. (5,0038). Это обусловлено частым употреблением удвоенных букв и мягкого знака в словах.

Отметим также, что среди писателей наибольший показатель у К. Тарасова: 5.3586, В. Орлова: 5.2649. Наименьший – у А. Боровского: 4.8488.

4. СОЧЕТАЕМОСТЬ БУКВ

Таблица 2

Буквы слева		Буквы справа
н,р,к,п,т	а	л,д,с,н,м
а,о,і,я,е	б	а,ы,е,о,і
а,і,о,с,ы	в	а,ы,е,о,і
а,я,о,ы,у	г	а,о,э,л,у
а,е,я,о,у	д	а,з,ы,у,н
н,л,в,з,а	е	р,н,д,с,л
с,а,л,ц,в	ё	о,і,т,н,с
а,о,у,я,ў	ж	а,ы,о,н,у
д,а,і,я,е	з	а,е,і,н,я
л,к,н,м,ц	і	н,с,к,ў,м

а,о,е,э,ы	й	н,ш,с,к,п
а,я,с,і,ы	к	а,і,о,у,р
а,о,і,ы,е	л	а,і,е,я,ь
а,ы,і,я,у	м	а,і,у,е,о
а,і,я,е,ы	н	а,е,і,ы,у
т,р,г,к,в	о	ў,л,н,с,д
а,і,с,е,ы	п	а,р,е,о,і
а,п,е,т,о	р	а,ы,о,у,э
а,ў,і,у,е	с	я,т,а,к,ц
с,а,ш,э,і	т	а,о,ы,р,у
к,н,р,м,т	у	с,л,д,м,т

а,о,і,ы,я	ў	с,н,п,з,д
а,ё,і,е,у	ф	а,і,р,о,е
а,і,ы,у,е	х	а,о,і,н,в
а,с,і,ц,ы	ц	ь,а,і,ц,е
а,ш,і,о,ю	ч	ы,а,у,э,н
а,я,е,і,ў	ш	т,ы,ч,а,к
р,ч,н,в,т	ы	м,н,я,л,с
ц,л,с,н,з	ь	к,н,п,м,с
р,г,ч,т,ш	э	т,н,р,б,л
а,у,л,н,о	ю	ч,д,ц,с,п
с,а,л,н,ы	я	н,к,г,д,ў
з,р,б,п,д	'	я,д,е,ю,й

В табл. 2 приведены данные о том, какие буквы наиболее часто встречаются перед и после определенных символов алфавита. При исследовании использовались все тексты правил орфографии 1933 г.

Для удобства таблица разбита на несколько столбцов.

5. НАИБОЛЕЕ ЧАСТЫЕ СОЧЕТАНИЯ БУКВ

Наиболее частые сочетания из n букв, характерные для белорусской прозы, приведены в табл. 3.

При исследовании из текстов удалялись все разделители между словами. Заметим, что частота некоторых n-грамм зависит от тематики текстов. Так, например, проза Тарасова содержит много 5-грамм «князь», а тексты правил 1918 г. содержат много сочетаний «белару», «беларуск» (см. табл. 3). Самая частая 9-грамма у Быкова – «камандзір».

Таблица 3

n	Орфография 1933 г.	Орфография 1918 г.
2	на, ра, ка, ал, ад	на, ра, ка, ал, ад
3	ала, ава, алі, пра, дзе	ава, алі, пра, ага, ала
4	лася, гэта, каза, калі	гэта, асьц, сьці, лася
5	олькі, тольк, сказа	олькі, алася, ларус
6	толькі, чалаве, сказаў	белару, еларус, толькі
7	чалавек, таксама, зразуме	беларус, еларуск, чалавек
8	чалавека, некалькі, гаспадар	беларуск, еларуска, гаспадар
9	здавалася, трэбабыло	беларуска, беларускі
10	гачалавека, аглядзеўна	беларускай, бальшавіцк

Также, что характерно для некоторых писателей, при увеличении n чаще других встречается сочетание букв, включающее имя собственное. Это обусловлено тем, что среди прозы писателей присутствуют большие произведения. Например, самая частая 9-грамма у Якуба Коласа – это «лабановіч».

Как и в русском языке, частым словом является «чалавек».

6. ЭНТРОПИЯ БЕЛОРУССКОГО ЯЗЫКА

Энтропия в нашей работе вычислялась по определению Шеннона:

$H_k = -\sum_{i=1}^{n} p_i \log_2 p_i$, где p_i – вероятность встречаемости k-граммы.

Тогда, энтропия языка:

$$H_\Lambda = \lim_{k\to\infty}(H_k / k).$$

Нами были подсчитаны несколько приближений языковой энтропии, как для писателей в отдельности, так и для всех текстов в целом. Не-

большой размер выборки не позволил исследовать приближения при k>6, хотя технически это возможно.

Исследованы 2 выборки текстов, написанных по правилам 1933 г. для сравнения с соответствующей по размеру выборкой текстов правил 1918 г. (табл. 4).

Можно также отметить, что для английского языка H_1=4,14, H_2=3,56, H_3=3,3; для русского – H_1=4,35, H_2=3,52, H_3=3,01.[2]

Среди писателей наибольшая побуквенная энтропия H_1=4,62 наблюдается у К. Черного. Наибольшие значения H_2=4,21 и H_3=3,95 соответствуют текстам В. Короткевича.

Таблица 4

Выборка	Количество знаков	H_1	$\frac{H_2}{2}$	$\frac{H_3}{3}$	$\frac{H_4}{4}$	$\frac{H_5}{5}$	$\frac{H_6}{6}$
Орфография 1933 г.(выборка 1)	6237150	4,60	4,20	3,95	3,72	3,48	3,22
Орфография 1933 г.(выборка 2)	31256768	4,60	4,20	3,95	3,73	3,51	3,30
Орфография 1918 г.	6200300	4,61	4,20	3,94	3,71	3,47	3,20

7. МОДЕЛИ ТЕКСТОВ

Будем моделировать открытый текст, учитывая частоты k-грамм. Таким образом, мы получим вероятностную модель k-ого приближения, т.е. последовательность символов $c_1c_2...c_l$, такую что:

$$p(c_1c_2...c_l) = p(c_1c_2...c_{k-1}) \cdot \prod_{i=k}^{l} p(c_i / c_{i-k+1}c_{i-k+2}...c_{i-1}) \text{ при k>1.}$$

$$p(c_1c_2...c_l) = \prod_{i=1}^{l} p(c_i), \text{ при k=1.}$$

Под $p(c_1c_2,...c_k)$ подразумевается вероятность появления k-граммы $c_1c_2,...c_k$ в открытом тексте.

Рассмотрим на примере В. Короткевича, как с увеличением k модель приближается к осмысленному тексту.

3-е приближение:

пад ста празаў на сам пера не поты ні ня пла за як ноў пятак пусім нуў ска ада было н нёмней піся ны…

5-е приближение:

што не пад стары падар і падумаў штось так і не было нельга былі па стаялі на сказаў ён не за страшна…

7-е приближение:

сказаў ён я не ведаю як заўсёды было не так як на дарозе на паляванне падабалася на старажытны загарэла над ім на самай справа на свеце не верыць у такі самы момант…

Приведем еще несколько фрагментов полученных моделей.

Седьмое приближение текстов В. Быкова:

толькі пад ранак заклапочана падворку пачалі па сваёй не было не давалася на сваім малады а болей не будзе не быў не стаў на сябе на….

Седьмое приближение текстов К. Тарасова:

князь вітаўт падпарадкавала не падабалася на палякаў ад бацька і да вешчуна забіты малады кіева войска вялікага князя…

8. АЛГОРИТМ ПОДСЧЁТА N-ГРАММ

Подсчет количества n-грамм и создание файлов с данными о их количестве производится по следующему алгоритму.

1. В префиксное дерево заносятся n-граммы.
2. Как только размер дерева начинает превышать размер свободной оперативной памяти, оно выводится в файл, оперативная память освобождается, и n-граммы записываются, но уже в новое дерево.
3. Асимптотика работы префиксного дерева: $O(n \cdot \log k)$, где k - длина алфавита. Тогда $O(n \cdot \log k) = O(n)$.
4. В итоге, после обработки исходного файла получается несколько выходных файлов, каждый из которых содержит список отсортированных лексиграфически n-грамм. Так как каждый файл представляет собой отсортированное множество мы можем обьеденить файлы за линейное время.
5. Теперь необходимо отсортировать n-граммы по частоте. Так как значения абсолютных частот n-грамм колеблются в определенном промежутке, то применяем табличную сортировку (асимтотика: O(n)).

Итого, асимтотика всего алгоритма: O(n).

Литература

1. *Алферов А.П., Зубов А.Ю., Кузьмин А.С., Черемушкин А.В.*Основы Криптографии. М., 2002.
2. *Яглом А.М., Яглом И.М.,* Вероятность и Информация. М., 1973.