ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 24 МАТЕМАТИКИ Выпуск 3
2017

В. О. Миронкин, **Т. Д. Воронцова** (Москва, МИЭМ НИУ ВШЭ)**. Об энтропии летописных текстов.**

Летописное письмо достаточно трудно для однозначной оценки, так как создавалось оно параллельно с разработкой правил русской письменности, а значит, каждый летописец был волен использовать свой подход к написанию. Существуют примеры летописей, в которых правила написания меняются от начала произведения к его концу. При этом необходимо упомянуть и о существовании нескольких вариантов кириллицы, используемых отдельными авторами.

Оценка энтропии источника, порождающего сообщения из букв старославянского алфавита, и сопоставление его информационных характеристик с соответствующими характеристиками современного русского языка [5] представляют особый интерес при изучении русской словесности.

В табл. 1 представлен общий 49-ти буквенный старославянский алфавит.

Таблица 1. Алфавит источника мощности 49

Напомним ряд определений [6].

О п р е д е л е н и е 1. Энтропией источника на знак называется величина

$$H_l = \frac{1}{l} \sum_{c_l \in C_l} p(c_l) \log p(c_l),$$

где C_l — множество всех последовательностей, порожденных источником.

О п р е д е л е н и е 2. Энтропией источника сообщений называется величина

$$H_\infty = \lim_{l \to \infty} H_l.$$

З а м е ч а н и е 1. Точность вычисления данного предела существенно зависит от объема анализируемых данных, что может заметно сказаться на трудоемкости вычислений. При этом в некоторых случаях его оценка вовсе невозможна. Для оценки энтропии был выбран один из вариантов кириллицы, охватывающий летописи [1–4], представленные в табл. 2.

Таблица 2. Анализируемые летописные тексты

№ п/п	Название летописи	Век	Количество учтенных символов	Краткое содержание
1.	Повесть временных лет (по Ипатьевскому списку)	до XII	246 834	Образование государств. Распад Русской земли (сильное церковное и княжеское влияние на составителя)
2.	Киевская летопись	XII	316 015	Киевская земля. Борьба между Олеговичами и Мономаховичами, призывы к единению (церковные темы редки)
3.	Галицко-Волынская летопись	XIII	213 463	История Галиции и Волыни (используется двойная хронологическая сетка, отсутствует церковная тема)
4.	Суздальская летопись	XII	227 198	Библейские события. Основание Киева. Описание Владимиро-Суздальской Руси

Алфавит, используемый в указанных текстах, претерпел эволюционные изменения и был редуцирован до 45 символов: 44 буквенных символа (без символов Гервь, Yервль, Ижа, Эдо, Арь) и специальный символ ѯ с числовым значением 1000.

В ходе статистических исследований текстов, как правило, осуществляется их предварительная обработка, которая может заключаться, например, в удалении ряда служебных символов. В данном случае подобное форматирование текстов представляет собой нетривиальную задачу по ряду причин:

1. Отличительной особенностью старославянской письменности является отсутствие пробелов (рис. 1). Каноническое летописное письмо не учитывало разбиение текста на слова, однако летописцы могли по желанию пропускать расстояние между словами.

2. В основном дошедшие до нас летописи списаны с оригиналов. При этом постепенно происходила унификация служебных символов. Так, например, в оригинальных версиях текстов можно встретить привычный нам символ «точка» в разнообразных сочетаниях, не используемых в переписанных текстах: двоеточие (:), троеточие (⁖), четвероточие (⁘) и т. п., а также служебные знаки в виде креста (†), переноса (∼) и др.

З а м е ч а н и е 2. Согласно [7] самым часто употребляемым знаком являлась точка (в разных вариантах написания: вверху строки, внизу, а чаще — посередине) — знак, используемый после выговариваемых вместе слов.

3. Наличие знаков придыханий, не имевших смыслового звукового значения и служивших подсказкой читателю, где стоит сделать паузу, выделить букву ударением, иногда — сменить интонацию (летописи часто содержат в себе самые разнообразные их версии, не поддающиеся общему описанию).

4. Наличие служебных скобок, расставленных наборщиками в процессе редактирования оригинальных текстов. В них содержались комментарии к правкам: расста-

новка пробелов, подстановка опущенных гласных букв и т. д.).

5. Особое обозначение цифр. Числительные на письме обозначались комбинациями определенных букв, каждая из которых имела свое числовое значение, а написанное число вычислялось как сумма таких значений. Чтобы отличать слова от чисел, такие последовательности выделялись с обеих сторон точками посередине строки.

6. Использование цифрового символа ϟ для формирования хронологии.

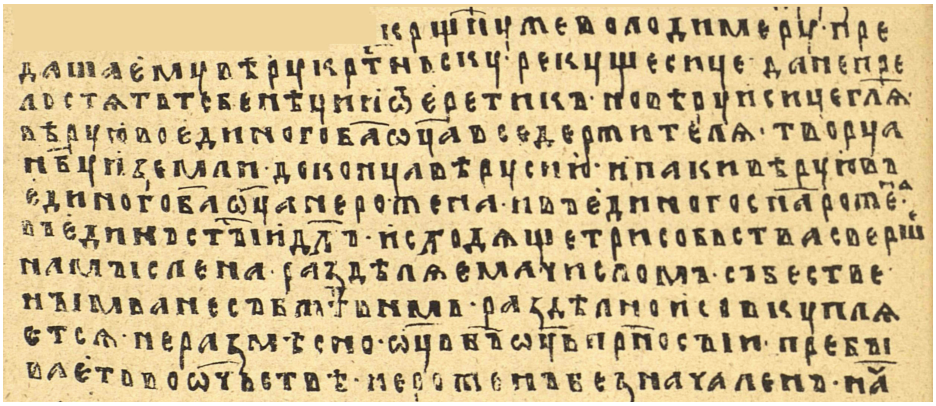

Рис. 1. Фрагмент старославянского текста без пробелов

З а м е ч а н и е 3. Для выбранных текстов в ходе исследования была проведена предварительная обработка, учитывающая указанные особенности (удаление пробелов, комментариев, спецсимволов и т. д.).

В табл. 3 представлены значения энтропий 45-ти буквенного источника сообщений на знак в зависимости от длины $k \in \overline{1,11}.$.

Таблица 3. Энтропия источника на знак

k	1	2	3	4	5	6	7	8	9	10	11
Hk	4,736	4,313	3,981	3,611	3,195	2,798	2,461	2,190	1,973	1,790	1,645

На рис. 2 изображено поведение энтропии на знак исследуемого источника (A^*) и источника, моделирующего тексты русской художественной литературы XXI века (A).

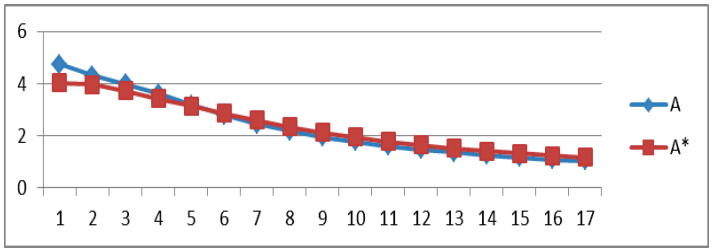

Рис. 2. Асимптотика энтропии двух источников

Из рис. 2 видно, что энтропии на знак для источников, порождающих последовательности русского и старославянского алфавитов, ведут себя приблизительно одинаково. При длине последовательностей меньшей 6 символов появление очередной кириллической буквы — событие более информативное, чем для современного языка, однако,

начиная с шестиграмм и далее, энтропия на знак источника текстов старославянского языка становится меньше, чем для современного языка.

При этом появление 7-й буквы в тексте снижает неопределенность в 2 раза, появление 11-й — в 3 раза, а 17-й — в 4 раза по сравнению с неопределенностью, соответствующей одной букве текста. Экспериментально полученное распределение букв старославянского и современного русского алфавитов приведено на рис. 3.

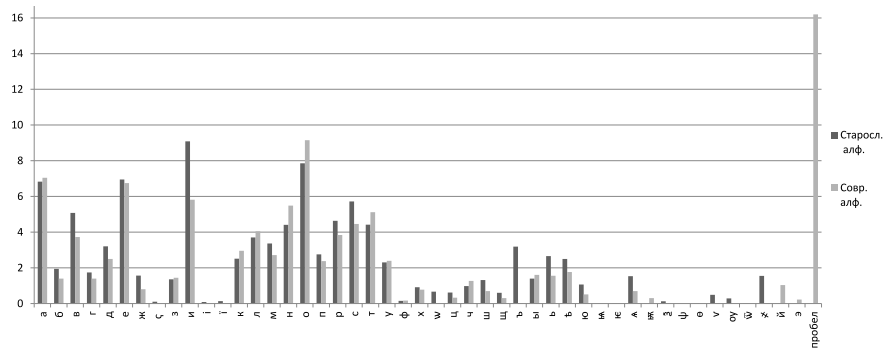

Рис. 3. Гистограмма частот букв старославянского и русского алфавитов

Как видно из рис. 3 главное отличие частотных характеристик алфавитов A^* и A заключается в использовании пробела. При этом заметим, что вышедшие из обращения символы старославянского алфавита имеют малые частоты встречаемости (скорее всего этим и объясняется их эволюционное исчезновение). Остальные символы имеют примерно схожее распределение для обеих моделей языков. Однако для некоторых кириллических символов произошли заметные изменения. Так, например, для гласных букв: использование буквы «и» уменьшилось на 33%, буквы «о» — увеличилось на 15%. Среди согласных букв сократилась частотность буквы «в» — на 33%, «с» — на 25а использование буквы «н», наоборот, увеличилось на 20%.

СПИСОК ЛИТЕРАТУРЫ

1. Киевская летопись. — Электронное издание: http://www.lrc-lib.ru/.
2. Галицко-Волынская летопись. — Электронное издание: http://www.lrc-lib.ru/.
3. Повесть временных лет по Ипатьевскому (Академическому) списку. — Электронное издание: http://www.lrc-lib.ru/.
4. Суздальская летопись по Лаврентьевскому списку. — Электронное издание: http://www.lrc-lib.ru/.
5. *Вильбоа Н. В.*, *Лось А. Б.*, *Миронкин В. О.* Об исследовании информационных характеристик естественных языков. — Обозрение прикл. и промышл. матем., 2016, т. 23, в. 1, с. 3–16.
6. *Духин А. А.* Теория информации. М.: Гелиос АРВ, 2007.
7. *Карский Е. Ф.* Славянская кирилловская палеография. М.: Наука, 1979.