ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 52 за 2016 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Арутюнов А.А., Борисов Л.А., Зенюк Д.А., Ивченко А.Ю., Кирина-Лилинская Е.П., Орлов Ю.Н., Осминин К.П., Федоров С.Л., Шилин С.А.

Статистические закономерности европейских языков и анализ рукописи Войнича

Рекомендуемая форма библиографической ссылки: Статистические закономерности европейских языков и анализ рукописи Войнича / А.А.Арутюнов [и др.] // Препринты ИПМ им. М.В.Келдыша. 2016. № 52. 36 с. doi:10.20948/prepr-2016-52
URL: http://library.keldysh.ru/preprint.asp?id=2016-52

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.А. Арутюнов, Л.А. Борисов, Д.А. Зенюк,
А.Ю. Ивченко, Е.П. Кирина-Лилинская,
Ю.Н. Орлов, К.П. Осминин,
С.Л. Федоров, С.А. Шилин

Статистические закономерности европейских языков и анализ рукописи Войнича

Москва — 2016

Арутюнов А.А., Борисов Л.А., Зенюк Д.А., Ивченко А.Ю., Кирина-Лилинская Е.П., Орлов Ю.Н., Осминин К.П., Федоров С.Л., Шилин С.А.

Статистические закономерности европейских языков и анализ рукописи Войнича

Исследованы статистические закономерности распределения частот букв в текстах на европейских языках. Проанализирован уровень достоверности логарифмической аппроксимации упорядоченного распределения частот для текстов без огласовки, написанных одним алфавитом на одном и на двух языках. Предложены варианты языков, на которых мог быть написан Манускрипт Войнича, и рассмотрена его внутренняя структура. Построены спектральные портреты матриц условных вероятностей двухбуквенных сочетаний для текстов без огласовки и Манускрипта Войнича.

Ключевые слова: распределение частот буквенных сочетаний, группы европейских языков, Манускрипт Войнича, спектральный портрет

Arutyunov A.A., Borisov L.A., Zeniuk D.A., Ivchenko A.Yu., Kirina-Lilinskaya E.P., Orlov Yu.N., Osminin K.P., Fedorov S.L., Shilin S.A.

Statistical regularity of European languages and Voynich Manuscript analysis

The statistical properties of letters frequencies in European literature texts are investigated. The determination of logarithmic dependence of letters sequence for one-language and two-language texts are examined. The pare of languages is suggested for Voynich Manuscript. The internal structure of Manuscript is considered. The spectral portraits of two-letters distribution are constructed.

Key words: letters frequency distribution, European languages groups, Voynich Manuscript, spectral portrait

Работа выполнена при поддержке гранта РФФИ, проект № 16-01-00342

Содержание

Введение и постановка задачи

В настоящей работе проводится статистический анализ литературных текстов на европейских языках, среди которых рассматриваются представители индоевропейской и уральской языковых семей. Строятся распределения букв по частоте встречаемости в текстах достаточно больших объемов (более 100 тыс. знаков) с целью выявления закономерностей, присущих соответствующим лексиконам.

В монографии [1] были приведены некоторые результаты анализа частот употребления букв в европейских языках. Оказалось, что для большинства языков зависимость упорядоченных частот с высокой точностью является логарифмической, ее детерминация превышает 0,98. Параметры логарифмической зависимости определяются числом знаков в алфавите и допускают трактовку избыточности или недостаточности используемых знаков по отношению к выражаемому ими звуковому ряду.

Интерес представляет отклонение статистки букв текста от отмеченной логарифмической зависимости для двуязычных произведений, записанных как в полном алфавите, так и без учета гласных и смягчающих знаков, а также для текстов, написанных на так называемых искусственных языках. Можно ли по статистике частот употребления символов сделать достаточно достоверные предположения относительно языка текста? Этот вопрос возник у авторов в ходе обсуждения так называемого Манускрипта Войнича.

Манускрипт Войнича (далее МВ) [2] – это рукопись, датируемая исследователями XVI веком. Она состоит из последовательности знаков, трактуемых как буквы, из которых транскрипторы выделяют 22 различных символа. Эти знаки не являются элементами каких-либо известных алфавитов. Объем рукописи составляет порядка 170 тыс. знаков. В настоящее время рукопись хранится в библиотеке Йельского университета и имеет статус криптографической загадки.

Многочисленные исследования с целью расшифровки этого текста проводятся более ста лет, но безуспешно. Существующие версии об авторстве, содержании и языке рукописи, обзор которых можно найти в работах [3-5], на наш взгляд, недостаточно убедительно подкреплены полноценными статистическими исследованиями. Сразу подчеркнем, что нашей целью не является расшифровка рукописи. Мы не анализируем словарный состав, поэтому смысловая составляющая текста в работе не обсуждается. Вопрос, ответ на который мы попытаемся получить, состоит в следующем: является ли МВ осмысленным, но зашифрованным текстом, и на каком языке в таком случае он написан, или он представляет собой мистификацию, т.е. бессмысленный набор знаков? Может показаться, что для ответа как раз и требуется расшифровка текста, однако это вовсе не обязательно. Сначала следует выяснить, есть ли у осмысленных текстов некие общие статистические свойства, без знания которых невозможно провести нужную имитацию. Наши исследования показывают, что такие свойства имеются.

По поводу того, сколько и каких знаков в МВ, также нет единого мнения. Существует так называемая «европейская транскрипция» (EVA [6]) отображения знаков рукописи в латиницу. Кроме того, есть транскрипция Takahashi [7] – тоже в латиницу, но с другими частотами выделенных символов.

Различными исследователями предлагались многочисленные гипотезы о структуре Манускрипта. Считалось, что Манускрипт:

- написан с перестановкой букв;
- двум символам некоторого известного алфавита отвечает один символ рукописи;
- существует рукопись-ключ, без которой нельзя прочитать текст, ибо одинаковые символы в разных частях рукописи отвечают разным буквам;
- рукопись представляет зашифрованный двуязычный текст;
- первоначально из осмысленного текста были удалены гласные;
- текст содержит ложные пробелы между словами.

В контексте имеющихся гипотез несколько странно выглядят заявления отдельных исследователей о количестве якобы «слов» рукописи, о соответствии их распределения закону Ципфа (на самом деле не выполняющемуся для слов естественного языка), об информационной энтропии, якобы характерной для текстов определенной тематики. По поводу выдвигаемых идей такого рода, не имеющих достаточных статистических обоснований, выскажем ряд критических замечаний. Следует иметь в виду, что информационная энтропия, вычисленная по однобуквенным символам, вообще меняется в очень малых пределах для всех языков, а для вычисления энтропии, например, пентаграмм длина текста явно недостаточна, ибо вероятности буквосочетаний тогда определены недостаточно достоверно. Также отметим, что без расшифровки текста нелепо говорить о словах только исходя из наличия пробелов, зная, что пробелы могут быть ложными. При этом в различных бездоказательных в математическом плане концепциях на роль оригинального языка текста предлагаются иврит, испанский язык, русский, маньчжурский, вьетнамский и многие другие, включая «что-то арабское или индейское». В то же время наличие ложных пробелов следует рассматривать как вполне реальную составляющую структуры Манускрипта. Тогда расшифровка может представляться весьма проблематичной. Так, например, нижеследующий текст

«неле поговори тьословахтоль коисходяиз наличияпр обелов»

может быть весьма просто прочитан, если знать, на каком языке читать. Однако трудности возникают даже после самой примитивной шифровки. Для демонстрации потенциальных трудностей удалим сначала все гласные из вышеприведенной фразы. Тогда получим

«нлпгвртслвхтлксхдзнлчпрблв».

Далее допишем вполне понятную фразу на английском языке «let us talk about this text» и удалим гласные из нее: «ltstlkbtthstxt». Запишем остатки русской фразы квази-латиницей «nlpgvrtslvhtlkshdznlxprblv», обозначив «ч» через «х», и объединим с остатком английской фразы. Получим

«nlpgvrtslvhtlkshdznlxprblvltstlkbtthstxt».

А теперь разобъем эту «фразу» на «слова» и предложим криптологам прочитать, что тут написано:

«nlpgv rtsl vhtl kshd znlxprbl vltstlk btth stxt».

Не зная «точки сборки» фразы внутри одного и того же языка, правильно сделать это весьма затруднительно. Что уж говорить о расшифровке двухсот страниц подобных «фраз», особенно если учесть, что символы такого смешанного алфавита заменены на неизвестные значки! Кроме того, следует иметь в виду, что восстановление огласовки не однозначно.

Еще одно замечание, препятствующее осмысленному прочтению МВ, состоит в том, что листы могли быть пронумерованы после того, как рукопись попала к постороннему владельцу, так что порядок «слов» может оказаться искаженным. Более того, нет четкого указания на то, что МВ – это одна рукопись, а не две или три независимых, написанных некой тайнописью. Во всяком случае, судя «по почерку», исполнителей было несколько.

Следовательно, надо отойти от идеи увидеть в МВ слова. Содержательно можно обсуждать исключительно статистику отдельных символов, предполагая, что символы – это буквы, либо, если статистика их будет «не буквенная», что это – слоги или, по крайней мере, некоторые из них – слоги.

Исследования, проведенные в [1], показали, что распределения символов текстов по частоте встречаемости являются устойчивой характеристикой не автора или тематики текста, но языка. Предполагается, что столь же устойчивыми окажутся и распределения смеси текстов на разных языках, по уровню детерминации которых относительно некоторого модельного распределения можно будет судить о долевом участии разных языков в написании таких двуязычных текстов.

Возможно, что устойчивыми являются распределения текстов по символам внутри одной языковой группы, а для текстов, написанных на смешанном языке из двух разных групп (например, часть на английском, а часть – на французском), распределение окажется неустойчивым. В таком случае можно будет говорить о естественной кластеризации родственных языков по принципу близости распределений используемых ими частотно-упорядоченных символов при произвольной смеси пары языков в одном тексте. Интерес представляет также сравнение текстов, имеющих на уровне согласных общий алфавит, но написанных на языках из разных языковых семей, например на венгерском (уральская семья, финно-угорская ветвь) и английском (индоевропейская семья, западногерманская подгруппа). Кроме того, имеет смысл рассмотреть возможность того, что Манускрипт написан на искусственном языке.

Анализу выдвинутых предположений и исследованию инвариантных свойств европейских языков и посвящена настоящая работа. Для нахождения языковых инвариантов используются следующие статистики: расстояние между распределениями упорядоченных эмпирических частот буквосочетаний в норме L1; уровень детерминации логарифмической аппроксимации однобуквенных распределений для текстов без огласовки; показатель Херста

для ряда из количества букв, заключенных между двумя наиболее часто встречающимися одинаковыми буквами; спектральный портрет матрицы двухбуквенных сочетаний. Перечисленные индикаторы позволили провести формальную кластеризацию языков индоевропейской семьи по языковым группам, совпавшим с группами, которые были сформированы на основе историко-лингвистических исследований.

1. Статистика символов транскрипций Манускрипта Войнича

В табл. 1 приведены частоты символов МВ, полученные в результате транскрипции знаков рукописи в латиницу по двум версиям: EVA и Takahashi. При подсчете строчные и прописные буквы не различались (они, впрочем, не различаются и в оригинальной рукописи). Пробелы не считались при этом специфической «буквой».

Табл. 1. Частоты употребления символов Манускрипта Войнича

Символ	EVA	Takahashi
a	0.07456	0.07641
b		0.00051
c	0.06951	0.00254
d	0.06773	0.00269
e	0.10478	0.12940
f	0.00264	0.00598
g	0.00050	0.00019
h	0.09322	0.11559
i	0.06125	0.01472
k	0.05708	0.02901
l	0.05491	0.05624
m	0.00583	0.03713
n	0.03206	0.03988
o	0.13296	0.13616
p	0.00851	
q	0.02831	0.00870
r	0.03893	0.02402
s	0.03857	0.01541
t	0.03625	0.12789
u		0.00011
v	0.00005	
w		0.08296
x	0.00018	
y	0.09217	0.09445
z	0.00001	0.00001

Последующий анализ будет направлен на то, чтобы построить распределение символов MB по частоте встречаемости, сравнить его с аналогичными распределениями в европейских языках, выявить отклонения от уровня детерминации аппроксимирующей зависимости и определить, насколько велико расстояние между фактическим частотным распределением и его аппроксимацией в гистограммной норме L1.

Упорядоченность по убыванию частот встречаемости знаков в рукописи приведена на графиках рис. 1. Хотя эти распределения и близки по уровню детерминации аппроксимирующей зависимости (0,93), в деталях существенно различаются. Согласно исследованиям [1], график EVA (красная ломаная линия) характерен для германской группы языков, точнее для западногерманской подгруппы, а график Takahashi (зеленая ломаная линия) – для славянских и романских языков (рис. 2), а также и для германских, но для северогерманской подгруппы. Расстояние между этими двумя транскрипциями, частоты которых упорядочены по убыванию, в норме L1 равно 0,26, что примерно в три раза больше, чем между распределениями текстов без огласовки из одной языковой группы, и в 10 раз больше, чем между текстами с полным алфавитом. Это означает, что каждая из данных транскрипций отвечает принципиально разному прочтению MB, поэтому нельзя использовать их обе одновременно с целью уточнения статистики. Различия в транскрипциях связаны, по-видимому, с проблемой распознавания знаков Манускрипта, ибо не все они могут быть интерпретированы однозначно. Мы не будем обсуждать, правильно или нет распознаны знаки рукописи, а лишь исследуем статистические свойства представленных транскрипций.

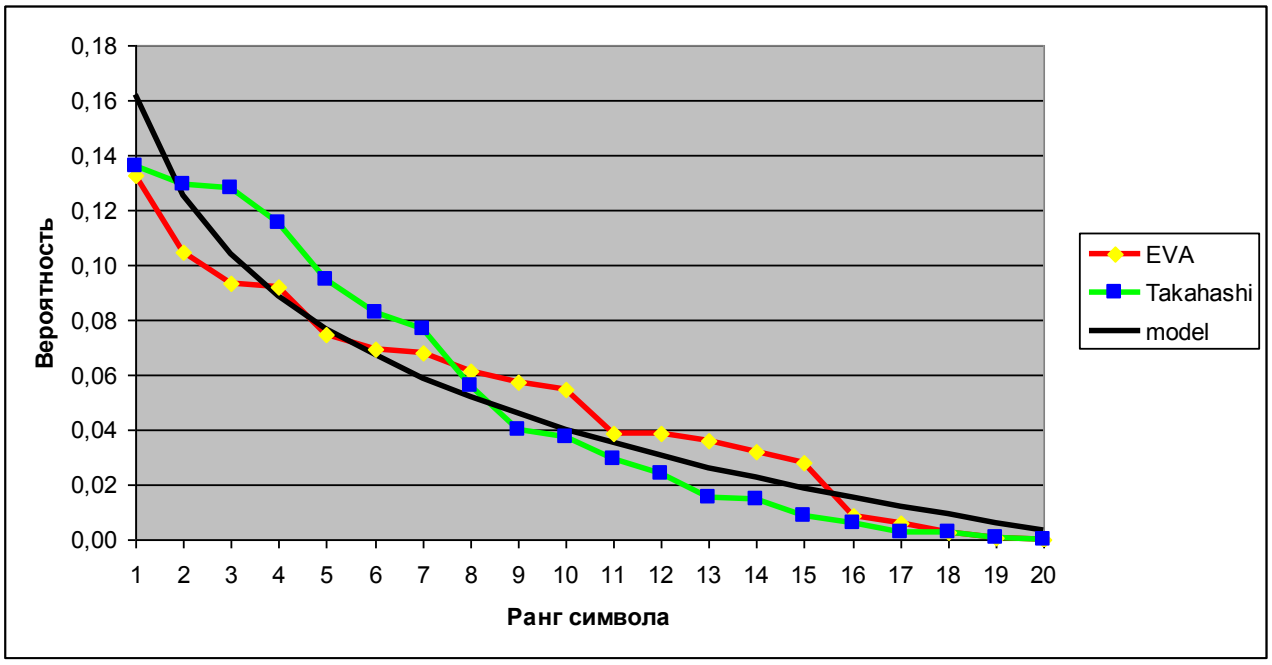

Рис. 1 – Упорядоченные частоты двух транскрипций Манускрипта и логарифмическая аппроксимация

Для большинства современных языков индоевропейской семьи характерна логарифмическая зависимость частоты буквы от ее ранга с достоверностью более 0,98. Уровень детерминации текстов без огласовки несколько ниже, но тоже достаточно высок – на уровне 0,96 (рис. 2, табл. 2).

Фактическое распределение упорядоченных частот символов для большинства текстов отличается от логарифмической аппроксимации в норме L1 в пределах 0,08-0,13, в том же промежутке лежат и расстояния между реальными распределениями на одном и том же языке безотносительно к тому, какой именно это язык (табл. 3). При этом 90%-ый доверительный интервал составляет [0,085; 0,115].

Отметим, что отклонения в норме L1 соответствующих аппроксимаций для обеих транскрипций Манускрипта примерно одинаковы и равны 0,17, что свидетельствует о недостаточной адекватности логарифмической модели применительно к рассматриваемому тексту.

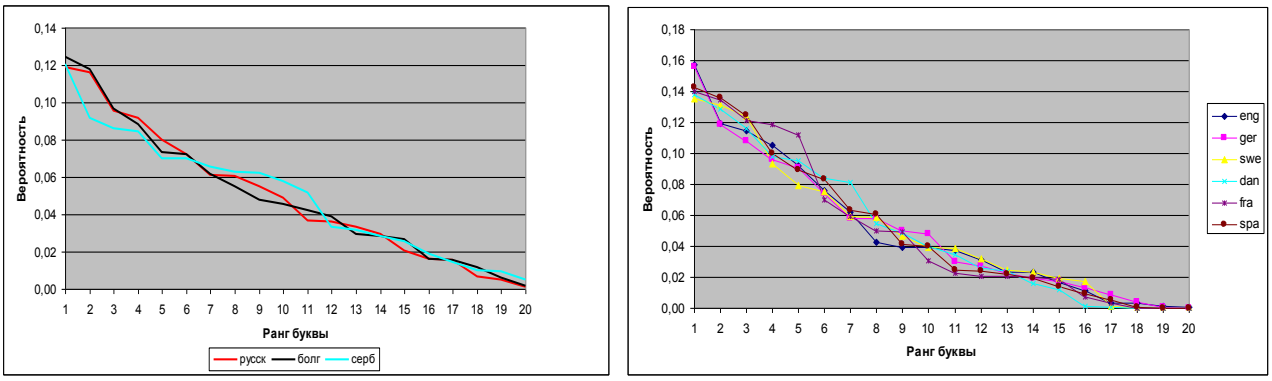

Рис. 2 – Упорядоченные частоты текстов без огласовки

В работе рассматриваются тексты в латинской [lat] и кириллической [кир] транскрипциях на следующих языках:

1 – *индоевропейская семья*

1.1 – *славянская группа*, подгруппы: восточная (русский [рус]), западная (польский [pol], чешский [che]), южная (сербский [serb/серб], хорватский [hr], болгарский [бол]);

1.2 – *германская группа*, подгруппы: северная (датский [dan], шведский [swe], норвежский букмол [nor]), западная (немецкий [ger], английский [eng], голландский [hol]);

1.3 – *романская группа*: итальянский [it], испанский [spa], французский [fra], румынский [rom];

1.4 – *греческий язык* [gre/гре];

1.5 – *баскский язык* [bask];

1.6 – *латинский язык* [lat];

2 – *уральская семья, финно-угорская ветвь*

2.1 – *угорская группа* (венгерский [hung]);

2.2 – *прибалтийско-финская группа* (финский [fin], эстонский [est]);

3 – *искусственные языки*

3.1 – эсперанто [esp], волапюк [vol];

3.2 – интерлингва [int];

3.3 – клингон [kl] (язык «обитателей» планеты Qo’noS);

3.4 – квенья (quenya) [qu] («эльфийский» язык).

Приблизительно 90 % рассмотренных языков имеют детерминацию модельной логарифмической зависимости упорядоченных по частоте символов выше, чем 0,96. Лишь распределения по буквам сербских (на латинице и кириллице), хорватских и румынских текстов без огласовки имеют гораздо меньшую точность аппроксимации.

Табл. 2. Детерминация логарифмической аппроксимации текстов без огласовки для некоторых европейских языков

[рус]	0,97	[ger]	0,98	[hr]	0,91
[бол]	0,97	[eng]	0,98	[pol]	0,96
[серб]	0,92	[hol]	0,98	[che]	0,96
[гре]	0,96	[dan]	0,96	[lat]	0,96
[fin]	0,96	[swe]	0,96	[it]	0,96
[est]	0,98	[nor]	0,96	[fra]	0,96
[hung]	0,96	[gre]	0,96	[spa]	0,96
[bask]	0,96	[serb]	0,84	[rom]	0,90

Расстояния между распределениями текстов на кириллице для славянской группы показывают, что русский, болгарский и сербский языки родственные: ближе всего русский и болгарский языки (расстояние 0,06), русский и сербский, как и болгарский и сербский, отстоят один от другого на 0,12. Отметим, что греческий язык в кириллической транскрипции отстоит от них более чем на 0,20 и в этом смысле не похож ни на один из славянских языков.

Для текстов на латинице расстояния между распределениями упорядоченных частот образуют кластеры (см. табл. 3) в смысле близости между собой в норме L1 в соответствии с языковыми группами. Так, например, статистика датского и шведского языков довольно близка, но она отличается от французского и итальянского языков, также близких один другому. Чешский и хорватский языки имеют близкую статистику, отличающуюся от статистик других упомянутых групп. Это показывает, что языки индоевропейской семьи, объединенные в группы или подгруппы, имеют близкие статистические свойства. Расстояния в норме L1 между распределениями из одной языковой

группы варьируются в довольно узких пределах 0,08-0,13, а между разными подгруппами они составляют 0,14-0,22.

Иначе обстоит дело с языками уральской семьи. Статистики согласных для финского и эстонского языков заметно различаются (расстояние между ними 0,22), хотя находятся они в одной прибалтийско-финской группе; линия, отвечающая венгерскому языку, удалена от обоих этих языков соответственно на 0,38 и 0,18, и находится ближе к языкам германской группы, расстояние до которых составило 0,16. Возможно, полученный результат обусловлен тем, что гласные играют более заметную роль в структуре этих языков.

Тексты на искусственных языках с целью выяснения их близости транскрипциям MB будут рассмотрены в отдельном разделе.

Табл. 3. Расстояния между распределениями частот текстов на латинице в норме L1 без огласовки, %

	ge	en	ho	da	sw	no	la	it	sp	fr	ro	bs	gr	fi	es	hu	po	ch	hr	se
ge		8	11	13	11	12	12	13	11	15	19	18	27	29	14	12	28	26	23	24
en			12	13	12	13	12	13	11	15	19	18	26	26	12	16	32	29	28	27
ho				10	11	11	19	21	19	22	27	25	23	27	20	18	31	28	27	32
da					11	10	13	13	9	14	13	18	27	27	12	16	35	31	30	25
sw						11	15	15	10	14	18	19	26	30	14	13	28	24	23	22
no							13	13	10	15	14	18	27	27	13	17	34	32	31	26
la								5	10	7	12	13	23	22	14	21	39	35	34	32
it									10	7	12	15	23	23	14	20	37	35	34	33
sp										11	13	16	22	25	14	16	36	32	30	28
fr											13	16	25	25	18	22	38	35	34	33
ro												20	30	31	21	18	41	38	33	25
bs													19	15	15	27	42	39	39	37
gr														14	21	31	48	45	44	43
fi															22	38	55	52	52	49
es																18	36	32	31	28
hu																	25	22	17	15
po																		5	12	22
ch																			9	20
hr																				12
se																				

Одинаковым цветом в табл. 3 выделены группы языков, попарно близких в смысле нормы L1 распределений упорядоченных частот в текстах без огласовки. Красным цветом отмечены некоторые неожиданно близкие пары языков. Объединение в кластеры проводилось по принципу парной близости

между всеми элементами кластера, а в случае неоднозначной принадлежности элемент считался отнесенным к кластеру с наибольшим числом элементов.

Прокомментируем полученные результаты.

1. По большей части в один кластер с расстояниями, меньшими 0,13, попадают языки, лингвистически относимые к одной группе. Это относится к германской, романской и славянской группам.

2. Греческий язык в латинской транскрипции обособлен от остальных рассматриваемых языков, как, впрочем, и финский, но один к другому они оказались ближе, чем каждый из них к другим языкам, расстояние между ними составило всего 0,14, что отмечено красным цветом в табл. 3.

3. Эстонский язык имеет явно другие статистические свойства, чем финский, в группу с которым он объединен лингвистами. Статистически эстонский и венгерский языки близки к германской группе, расстояние до которой у них составило 0,12-0,14.

4. Латинская транскрипция сербского языка близка только к хорватскому языку, расстояние между ними 0,12.

5. Славянская группа языков, использующих латиницу, плотно кластеризуется: это чешский, польский и хорватский языки.

6. Средневековая латынь хотя и вышла из употребления, тем не менее близка всем языкам романской группы. Имеющий неопределенный статус баскский язык удален от всех рассматриваемых языков, кроме латыни, до которой расстояние составило всего 0,13. Оба этих языка имеют одинаковую детерминацию логарифмической модели (0,96) с одинаковым числом эффективно используемых согласных (16).

Таким образом, метод кластеризации по принципу попарной близости распределений позволил получить вполне осмысленные результаты с точки зрения лингвистической классификации языков. Как уже говорилось выше, сами эти распределения с высокой точностью имеют логарифмический профиль.

Логарифмическая модель распределения символов была выведена С.М. Гусейном-Заде в [8] в предположении постоянства плотности распределения случайной точки $P(p_1,...,p_n)$ на n-мерном симплексе $\sum_{i=1}^{n} p_i = 1$, где p_i есть i-я по порядку частота употребления буквы в тексте. В [1] эта модель была модифицирована и применена для оценки полноты используемого алфавита в текстах на различных языках. Она имеет вид

$$f(k) = \frac{1}{n}\left(1 + \frac{1}{n+o}\ln\frac{n!}{k^n}\right). \qquad (1)$$

В этой формуле n – количество букв в алфавите, а параметр o есть ближайшее целое к подбираемому значению, отвечающему наименьшей ошибке аппроксимации фактического распределения по формуле (1). Смысл этого параметра состоит в том, что для рассматриваемого текста наиболее адекватен алфавит, число символов в котором есть $n+o$. Для русского,

немецкого, английского и венгерского языков эмпирические зависимости на рис. 2 лучше всего моделируются зависимостью (1), в которой $o=0$; этот вариант и показан на рис. 1 легендой «model». Для французского, испанского и итальянского языков $o=-3$, для датского и шведского языков $o=-1$. Для финского языка $o=-6$, для эстонского $o=-4$.

Применительно к транскрипции EVA, для которой $n=22$, наилучшая аппроксимация достигается при $o=-2$. Это означает, что фактически в MB используется лишь 20 символов. Исключив два самых редких символа, получаем логарифмическую аппроксимацию с детерминацией 0,93 и отклонением в норме L1 от фактического распределения на уровне 0,167. То же наблюдение верно и в отношении транскрипции Takahashi. Соответствующие зависимости и были представлены выше на рис. 1.

Заметим теперь, что в большинстве европейских языков число согласных букв равно 20. Можно предположить, что исследуемая рукопись написана на одном из них, но без огласовки. Необходимым в статистическом смысле, но, разумеется, не достаточным условием для этого является, во-первых, близость одного из распределений транскрипций распределению выбранного языка (отклонение в норме L1 не превышает 0,10) и, во-вторых, примерно равное расстояние как от транскрипции, так и от выбранного языка до аппроксимирующей модельной зависимости (примерно 0,17). Анализируя данные на рис. 2, приходим к выводу, что из рассмотренных вариантов имеется лишь один подходящий – а именно, датский. Он отклоняется от модельной логарифмической зависимости на 0,172, транскрипция Takahashi отклоняется от нее же на 0,167, а сами эмпирические распределения языка Манускрипта и датского языка отклоняются одно от другого на 0,083 (см. рис. 3).

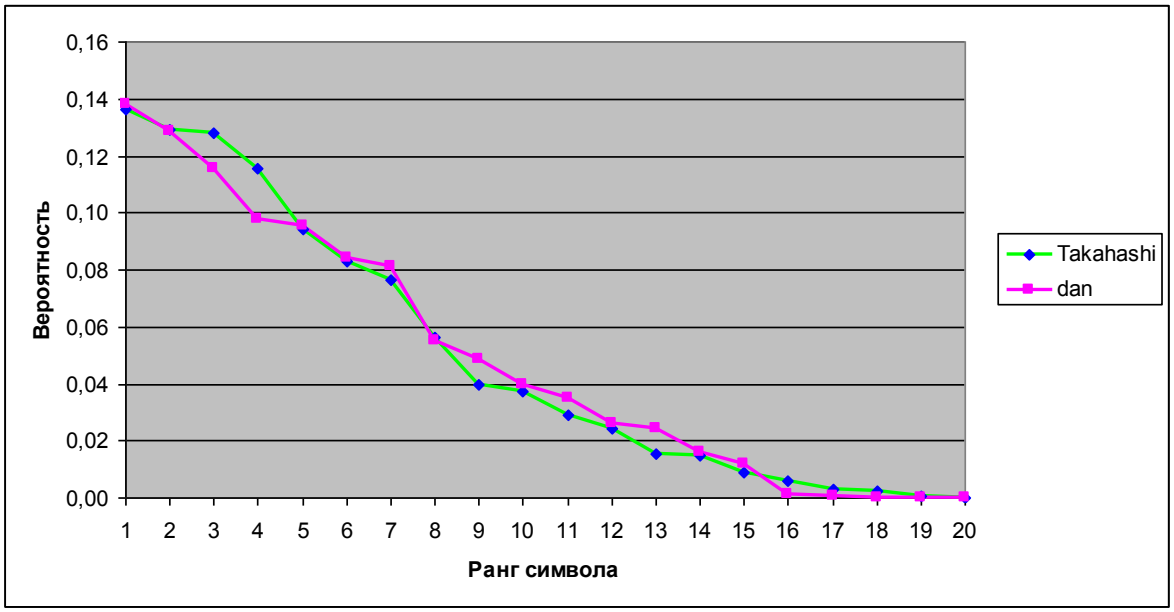

Рис. 3 – Распределения частот символов в транскрипции Takahashi и в текстах на датском языке без огласовки

При этом и детерминация логарифмической аппроксимации датского языка без огласовки, как и транскрипции Takahashi, равна 0,93. Близкие датскому шведский и норвежский (букмол) языки гораздо менее подходят на роль оригинального языка МВ, поскольку расстояния между всеми языками северогерманской группы одинаково и равно 0,11 (различия проявляются только в третьем знаке), а отличие шведского и норвежского от указанной транскрипции составляет 0,14, а не 0,08, как для датского языка.

Для транскрипции EVA подходящего языка среди рассмотренных европейских не нашлось.

Сделанные выводы основываются на статистическом анализе современных текстов. Перенос отмеченных свойств на рукописи XVI-XVII веков может быть сделан лишь в предположении, что в части использования согласных букв лексикон за это время не претерпел существенных изменений. К сожалению, в распоряжении авторов не имелось оригинальных текстов в достаточном для анализа количестве, поэтому высказанная гипотеза всего лишь иллюстрирует метод, применяемый для анализа текстов, и носит весьма предварительный характер. Можно лишь отметить, что анализ средневековых текстов на латыни привел к аналогичной зависимости: детерминация логарифмической модели без гласных составила 0,966. Этот факт позволяет предположить, что логарифмическая модель в определенном смысле инвариантна и для сравнения распределений символов по упорядоченности в текстах без гласных можно использовать как старые, так и современные тексты.

Подчеркнем, что предположительная идентификация языка еще не означает прочтения рукописи, поскольку инвариантом языка является кривая, а не положения на ней конкретных букв. В разных текстах на одном и том же языке упорядоченная последовательность букв «плавающая», хотя, разумеется, наиболее часто употребляемые символы в одном тексте не становятся редко используемыми в другом. Тем не менее однозначного соответствия между рангом и буквой ни в одном языке нет, ширина «окна миграции» ранга по текстам длиной порядка 200 тыс. знаков равна 5. Поэтому даже если определен оригинальный язык рукописи, гарантрировать расшифровку один этот факт еще не может. Требуется проводить дополнительный анализ по большому набору текстов, чтобы установить возможные сочетания букв в их упорядоченном распределении.

Итак, мы привели один аргумент в пользу того, что Манускрипт написан на некотором языке без использования гласных. Косвенным подтверждением выдвинутого тезиса является наличие в рукописи цепочек из трех одинаковых символов подряд, которые не встретились в текстах на латинице с полным алфавитом, но оказались присутствующими в них после удаления гласных. Так, например, в английском тексте без огласовки частота появления «bbb» составляет $3 \cdot 10^{-6}$, «lll» соответственно $4 \cdot 10^{-4}$ и «ttt» $8 \cdot 10^{-4}$. В транскрипции Takahashi также имеются цепочки «троек» с похожими частотами: «ttt» $5 \cdot 10^{-6}$, «lll» $2 \cdot 10^{-5}$, «ooo» $5 \cdot 10^{-5}$ и «eee» $8 \cdot 10^{-4}$.

Однако применительно к тому же датскому языку следует учесть и такую возможность: текст MB написан на языке с полным алфавитом, но без учета диакритических знаков, которыми весьма богаты языки северогерманской (и не только) группы. В то же время число реально используемых согласных могло быть меньше 20, поскольку буквы Q, X, W, Z используются в датском языке только в заимствованных словах, которых просто могло не быть в рукописи. Тогда число разных символов тоже оказалось бы равным 22, но уже по другим причинам. При этом могло быть искусственное добавление (повтор) гласной в том случае, когда требовалось исключить появление иного смысла слова при замене, например, апострофа или умляута на «чистую» букву. Однако анализ других статистик, проводимый далее, все-таки свидетельствует в пользу «безгласного» прочтения рукописи.

2. Распределение расстояний между одинаковыми символами

Рассмотрим текст как временной ряд случайной величины (буквы) со значениями $x(t)$, где t есть порядковый номер символа x от начала текста, а сам символ принимает значения из множества, называемого «алфавит». Длина цепочки символов текста, не содержащих внутри себя пару определенных символов, является весьма важной характеристикой языка, поскольку ее распределение тоже обладает устойчивостью.

В частности, анализ текстов на русском языке, проведенный в [1], показал, что последовательность расстояний между одинаковыми буквами (т.е. количество других букв, находящихся между двумя данными буквами) обладает следующими статистическими свойствами:

- автокорреляция с любым лагом близка к нулю, но приращения зависимы;
- распределение расстояний устойчиво по текстам разных авторов и жанров и после нормировки на частоту встречаемости не зависит от символа.

Естественно, что чем реже встречается символ, т.е. чем меньше его эмпирическая вероятность f, тем в среднем больше расстояние между двумя такими последовательно идущими символами. Если обозначить через L максимально наблюдаемое расстояние между двумя данными одинаковыми буквами, то с детерминацией 0,98 справедлива аппроксимация

$$\ln L \approx \exp\left(\frac{2}{f^{0,2}}\right). \qquad (2)$$

Вероятность того, что расстояние между двумя такими последовательными символами равно l, имеет максимум, с достоверностью 0,9 находящийся в точке $l^* = -3{,}8 - 2{,}6\ln f$. Для $l > l^*$ наблюдается следующая эмпирическая зависимость вероятности от частоты: $prob \propto \exp(-1{,}3lf)$.

При этом весьма важно, что для текста с полным алфавитом и для того же текста, но состоящего только из согласных, распределения расстояний между

одинаковыми символами заметно различаются. В полном тексте распределение расстояний имеет более длинный носитель и менее высокий максимум, чем в тексте без огласовки. На рис. 4 показаны распределения расстояний между часто встречающимися буквами «Т» для обычного текста на русском языке и для того же текста без учета гласных, а также аналогичные распределения для буквы «T» в английских текстах. Такие же взаимные расположения графиков распределений наблюдаются и для других букв во всех европейских языках.

Для MB распределения расстояний между парой одинаковых символов в обеих транскрипциях похожи между собой (рис. 5, 6), но отличаются от распределения для текстов с полным алфавитом и близки к распределениям для текстов без огласовки. Они характеризуются тем, что максимум расположен четко выше (на уровне 0,11-0,12), чем для текста с полным алфавитом (примерно 0,06-0,08). Это – второй аргумент в пользу того, что MB написан на некотором языке без огласовки. Однако это косвенный аргумент, поскольку, вообще говоря, значение максимума линейно зависит от числа используемых в тексте символов, и, следовательно, это значение может отвечать и специально разработанному для данной рукописи языку.

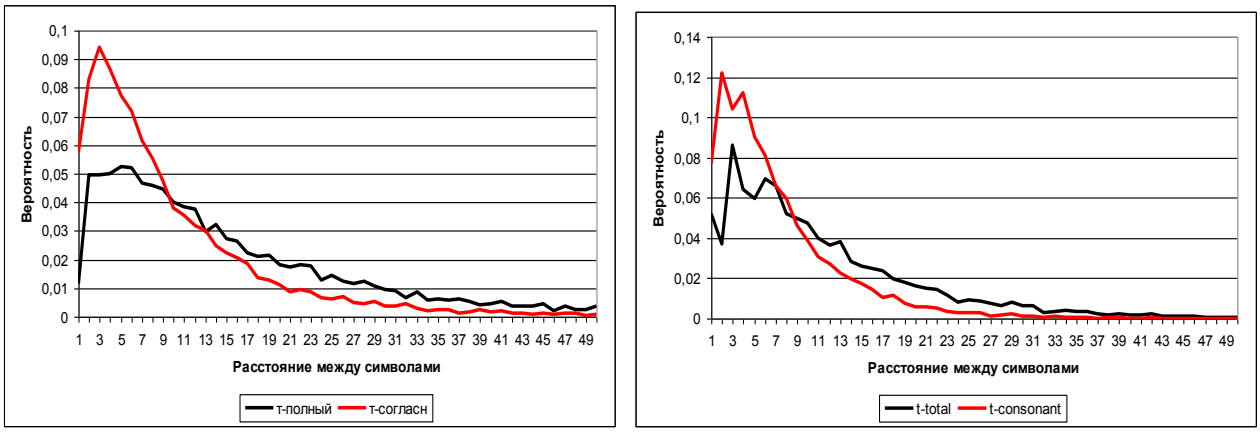

Рис. 4 – Распределение расстояний «Т-Т» в русских и английских текстах

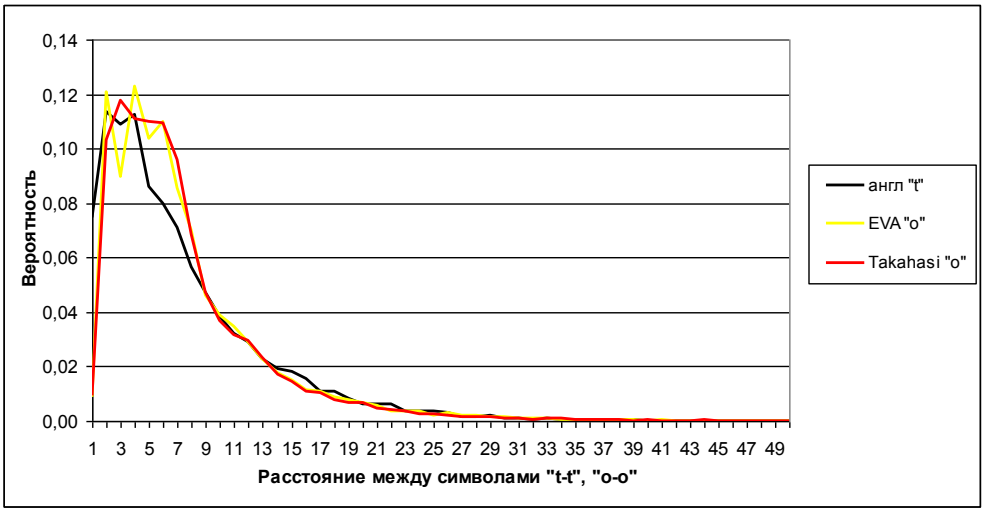

Рис. 5 – Распределение межсимвольных расстояний «О-О» для двух транскрипций и «Т-Т» для английского текста без гласных

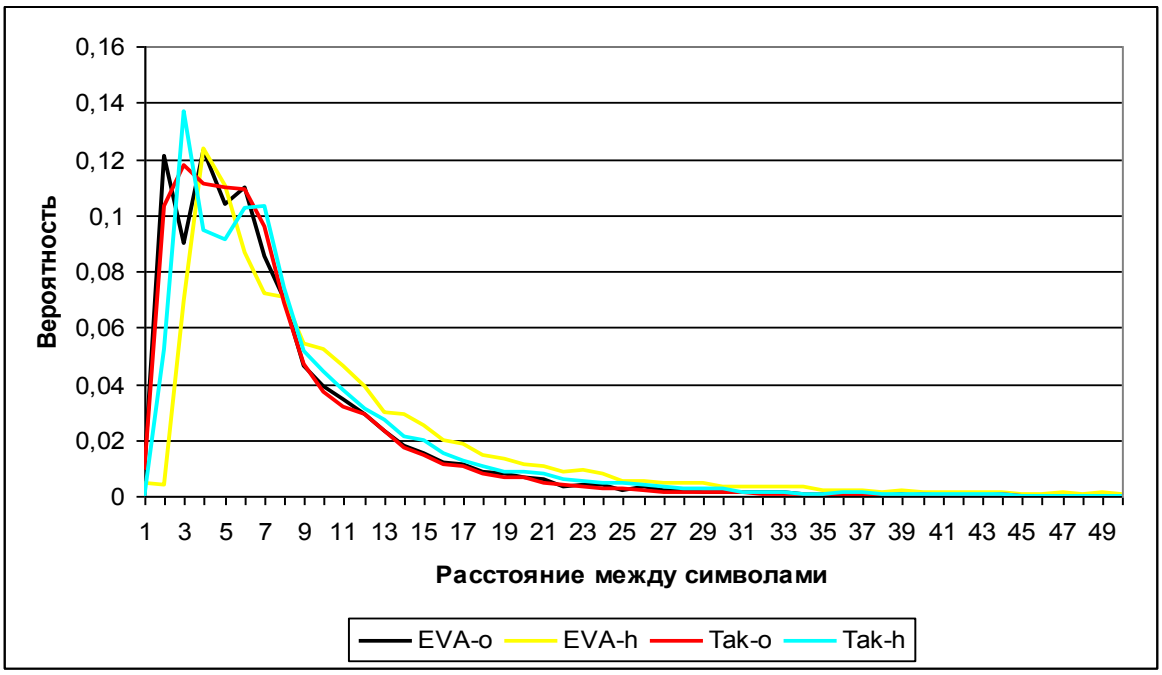

Рис. 6 – Распределение межсимвольных расстояний для двух транскрипций

Заметим теперь, что появление символов в тексте происходит не «просто так», а связано с использованием тех или иных слов в зависимости от смысловой направленности произведения. Поэтому расстояния между буквами не полностью случайны. Тогда может существовать параметр, отвечающий за долговременную память между последовательностью появления букв в тексте. В работе [9] на роль такого параметра был выбран показатель дробной производной в уравнении Фоккера-Планка относительно выборочной плотности функции распределения приращений временного ряда. Применительно к задаче, рассматриваемой в настоящей работе, временной ряд образован последовательностью расстояний между выбранными буквами. Мы рассмотрим здесь менее трудоемко вычисляемый индикатор, который соответствует вышеуказанному показателю дробной производной для так называемых самоподобных процессов – а именно, показатель Херста [10]. Этот показатель определяется следующим образом.

Для данного временного ряда $b(t)$ строится ряд $x(t)=b(t+1)-b(t)$ первых разностей и вводится скользящее среднее приростов по выборке длины k:

$$\bar{x}(t,k)=\frac{1}{k}\sum_{i=t-k+1}^{t}x(i).$$

Затем вычисляется накопленное отклонение от среднего (размах):

$$R(t,k)=\max_{j\le t}\left(\sum_{i=t-k+1}^{j}\left(x(i)-\bar{x}(t,k)\right)\right)-\min_{j\le t}\left(\sum_{i=t-k+1}^{j}\left(x(i)-\bar{x}(t,k)\right)\right).$$

Вычисляются также скользящая дисперсия рассматриваемого временного ряда по выборке длины k

$$\sigma_x^2(t,k)=\frac{1}{k}\sum_{i=t-k+1}^{t}\left(x(i)-\bar{x}(t,k)\right)^2,$$

логарифм отношения размаха к шуму и его выборочное среднее:

$$\xi(t,k)=\ln\left(\frac{R(t,k)}{\sigma_x(t,k)}\right),\quad \bar{\xi}_N(t)=\frac{1}{N}\sum_{k=1}^{N}\xi(t,k).$$

Показатель Херста $H_N(t)$ по выборке длины N на шаге t определяется как коэффициент регрессии величины $\xi(t,k)$ на логарифм длины выборки и вычисляется по формуле:

$$H_N(t)=\frac{1}{N}\sum_{k=1}^{N}\left(\xi(t,k)-\bar{\xi}_N(t)\right)\left(1+\ln(k/N)\right). \qquad (3)$$

Выяснилось, что расстояния между одинаковыми буквами независимо от огласовки для всех рассматриваемых языков образуют так называемый антиперсистентный ряд, поскольку показатель Херста для ряда из этих расстояний существенно меньше, чем критическое значение 0,5, отвечающее белому шуму. Распределения показателя Херста, построенного по выборке длины $N=5000$, показано для некоторых языков на рис. 7.

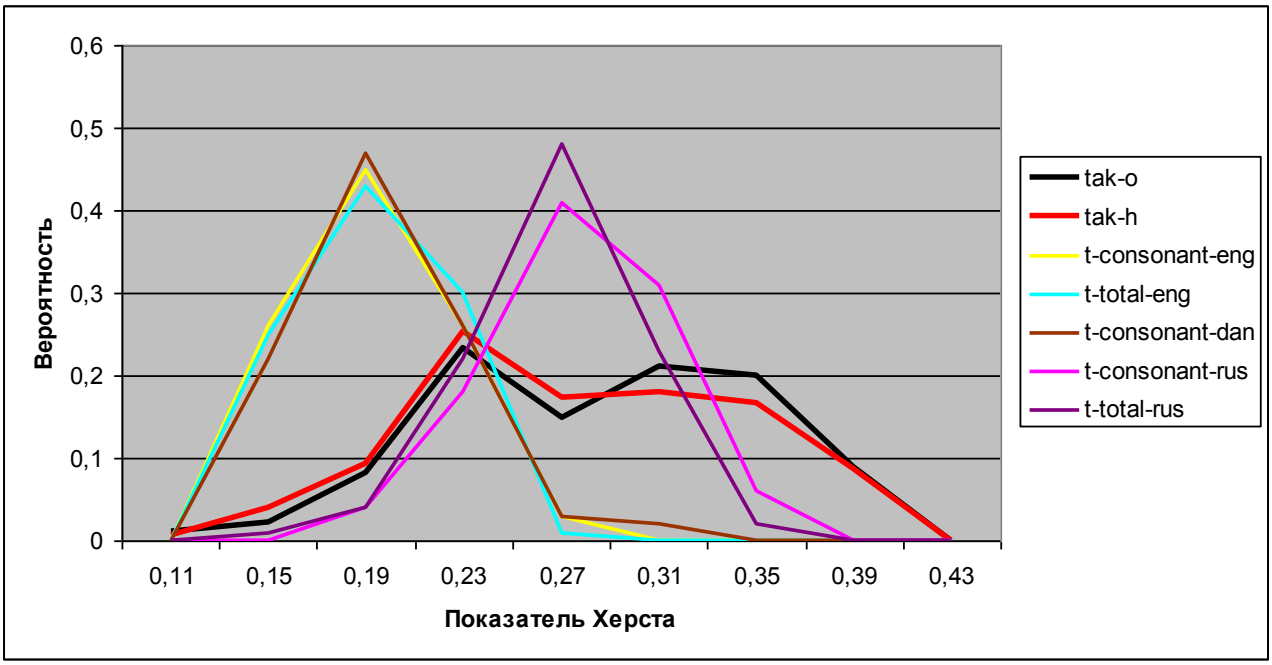

Рис. 7 – Распределения показателей Херста для рядов расстояний между наиболее часто встречающимися буквами в текстах

Видно, что распределения на рис. 7 для русского и английского языков имеют максимумы в четко различающихся точках, а распределения для датского и английского языков практически совпадают. Возможно, что распределение показателя Херста является индикатором языка, на котором написан текст, или соответствующей языковой группы, но это есть тема

отдельного исследования. Пока же можно сделать вывод о том, что выдвинутый в разделе 1 вариант с датским языком в качестве оригинального языка МВ следует исключить, поскольку распределения показателя Херста для рукописи имеют значительные отличия от аналогичных распределений для обычных текстов. Для МВ это распределение более пологое и смещено вправо (черная и красная линии на рис. 7), что свидетельствует о большей случайности в расположении символов, чем для текстов на одном из европейских языков. Это означает, что статистика языка рукописи отличается от текстов, написанных на одном языке. Анализ открывающихся в этом случае вариантов будет проведен в следующих разделах. Будут рассмотрены две возможности: рукопись написана на специально разработанном для нее языке, либо же она содержит записи на разных языках.

Отметим, что применительно к анализу МВ в случае смешения нескольких языков показатель Херста как индикатор бесполезен, поскольку заранее не известно, в каком месте текст написан на том или ином языке. Однако важно, что распределение этого показателя оказалось практически одинаково как для текста с полным алфавитом, так и для текста без гласных, так что хотя оно и не обладает в этом смысле индикативными свойствами, но, по-видимому, удостоверяет, что при выраженной унимодальности соответствующий текст написан на одном языке.

3. Статистика частот символов в искусственных языках

Одно из объяснений отклонения статистики символов Манускрипта от статистики «обычного» языка может состоять в том, что язык рукописи – искусственный. Существует порядка сотни таких языков, большая часть которых построена на различных идеях объединения естественных языков, но есть и языки, на которых разговаривают вымышленные герои фантастических литературных произведений в сказочных странах. Эти языки, впрочем, не сильно отличаются от естественных в смысле своих статистических свойств. На рис. 8 показаны распределения для некоторых из них.

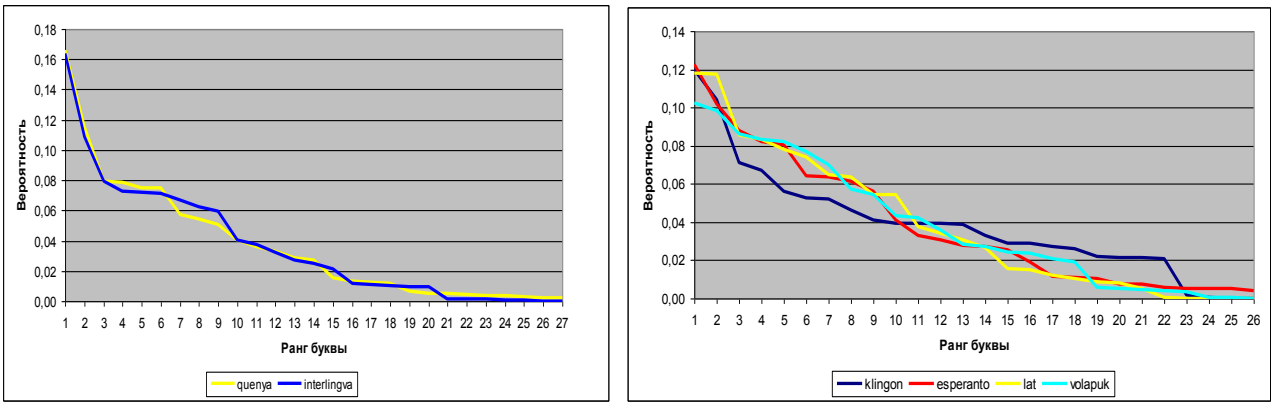

Рис. 8 – Распределение символов по упорядоченности
в некоторых искусственных языках

Оказалось, что «эльфийский» язык quenya Дж.Р.Р. Толкиена очень близок к языку interlingva, а эсперанто и волапюк близки латыни. Клингонский язык в этом плане стоит особняком. Расстояния между всеми этими распределениями сравнительно небольшие – от 0,09 до 0,13, но надо учесть, что в данном примере рассматривались полные алфавиты, для которых в случае близких естественных языков расстояния варьируются от 0,03 до 0,05.

Заметим теперь, что один и тот же фактор – наличие цепочек из трех одинаковых символов подряд – может быть интерпретирован как с точки зрения безгласного прочтения MB, так и в рамках гипотезы об искусственном языке рукописи, поскольку символы могут играть и синтаксическую роль. В этом смысле текст на эсперанто весьма близок к транскрипции EVA (рис. 9): расстояние между ними в норме L1 составило всего 0,11, причем основные различия – в малых частотах, а не в больших.

Достоверность логарифмической аппроксимации в искусственных языках несколько хуже, чем в текстах на обычных языках индоевропейской группы, и находится на уровне детерминации текстов без огласовки. Для клингонского языка детерминация составляет 0,95, для эльфийского 0,96 и для эсперанто 0,97. Разумеется, это не означает, что любой искусственный язык имеет такую же высокую детерминацию, но все-таки детерминация MB статистически значимо ниже этих величин.

Мы не претендуем здесь на подробный анализ искусственных языков. Наша цель – дать некоторую частную демонстрацию предположения о том, что если язык достаточен для написания осмысленного текста большого объема, то его статистические свойства ожидаемо близки свойствам естественных языков.

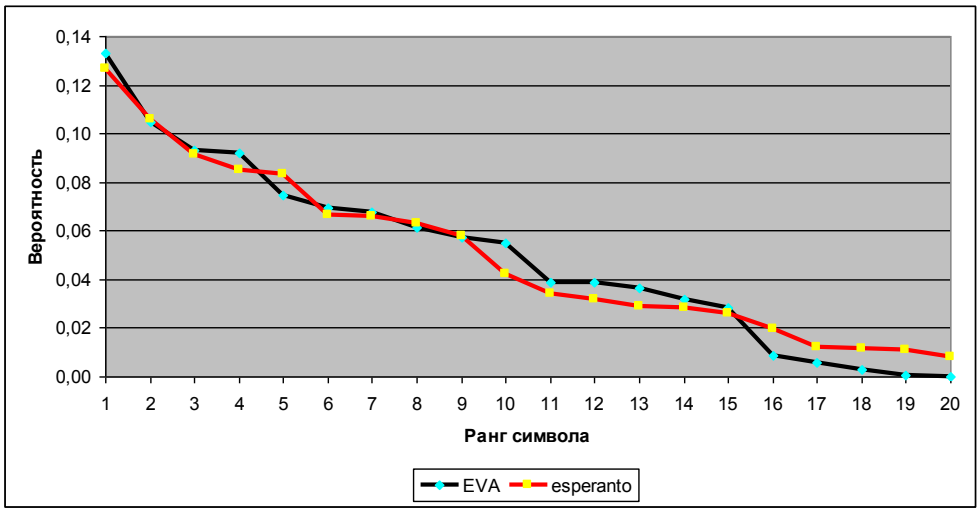

Рис. 9 – Распределение символов по упорядоченности для текста на эсперанто и для транскрипции EVA

На рис. 10 приведены распределения показателя Херста для ряда расстояний между одинаковыми символами для искусственных языков. Также для удобства сравнения приведен график распределения показателя Херста для MB из рис. 7.

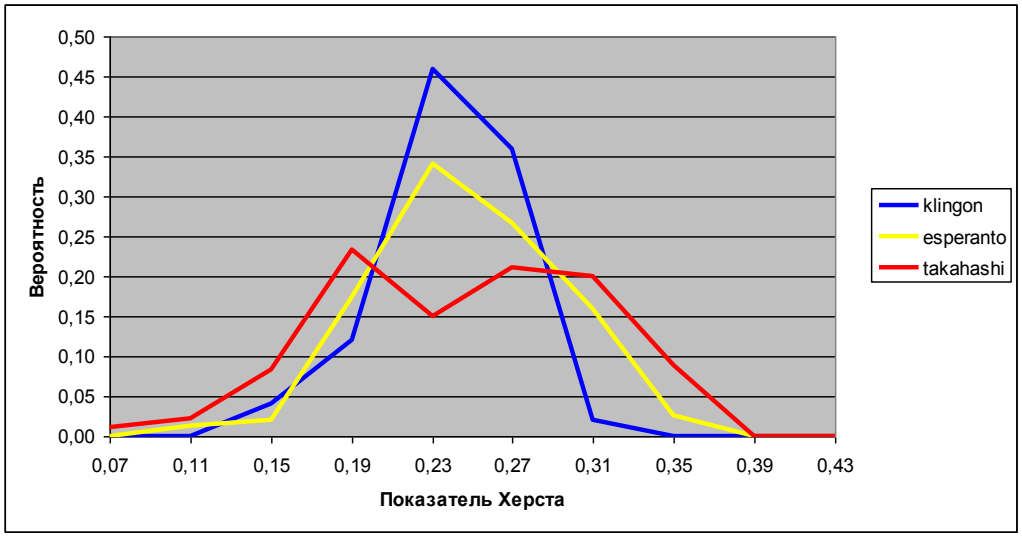

Рис. 10 – Распределения показателей Херста для рядов расстояний между одинаковыми буквами в текстах на искусственных языках

Сравнивая рис. 7 и рис. 10, видим, что показатели Херста для текстов на естественных и искусственных языках ведут себя примерно одинаково, и это поведение отличается от МВ. Видимо, и для искусственных языков в случае их достаточной проработки характерно антиперсистентное поведение ряда расстояний между одинаковыми буквами в текстах. Следовательно, вариант искусственного языка рукописи на данном этапе также надо исключить.

Как уже говорилось выше, более пологая форма и отсутствие унимодальности распределения показателя Херста для МВ могут быть объяснены тем, что в тексте перемешаны несколько языков, в частности, два.

4. Статистика частот символов в двуязычных текстах

Найденный в п.1 вариант языка, на котором мог бы быть написан МВ, не исчерпывает другие возможности построить текст, статистика символов которого близка к транскрипции Takahashi. Именно то, что отклонение логарифмической аппроксимации частот МВ от реального распределения, равное приблизительно 0,17, что существенно больше, чем для большинства текстов на европейских языках, свидетельствует о том, что МВ мог быть написан на двух языках, имеющих общий алфавит. Последнее условие не обязательно, но оно значительно упрощает исследование.

Наблюдаемую в большинстве европейских языков дерминацию текстов без огласовки на уровне 0,96 можно понизить до 0,93, как в МВ, если считать, что текст написан на двух языках, имеющих один алфавит (например, латиницу), и этот текст после удаления гласных и перекодировки превращается в то, что мы знаем как Манускрипт Войнича. При этом мы предполагаем, что одинаковые буквы в разных языках не обозначались в рукописи разными символами, что, конечно же, сильно сужает поле поиска. Заметим все же, что, поскольку детерминация транскрипции Takahashi выше, чем 0,9, возможность

использования разных алфавитов маловероятна. Для такого использования надо знать частоты употребления символов в каждом из алфавитов и сгруппировать переобозначенные символы надлежащим образом, что для XVI века представляется не очень реалистичным, особенно если учесть, что нужный для этого регрессионный анализ был изобретен значительно позднее. По этой же причине следует предположить, что текст рукописи осмысленный, иначе отклонение от статистики букв, специфической для естественного лексикона, было бы гораздо больше.

Таким образом, в этом разделе мы принимаем следующие рабочие гипотезы в отношении МВ:

1. Манускрипт является двуязычным текстом с общим алфавитом.
2. Перед перекодировкой из текста были удалены гласные.
3. Перекодировка состояла в однозначной замене буквы символом.
4. Пробелы в тексте не считаются символами.

Тогда следует выяснить, какие пары языков с общим алфавитом и в какой пропорции могли бы рассматриваться как языки Манускрипта, из одной ли они языковой группы или из разных и каких именно, а также как сильно влияет на статистические свойства текстов их тематическая направленность. Применительно к текстам на русском языке влияние жанра на алфавитное (но не на упорядоченное по частоте) распределение рассматривалось в работе [11], где определенная зависимость была отмечена.

Приведем здесь результаты статистического анализа частот в современных текстах, написанных на двух языках, но в одном алфавите. Это следует сделать во всяком случае для того, чтобы проверить гипотезу о снижении детерминации модели (1) при смешении языков текстов. Для тестирования этой гипотезы соединим два текста примерно равных объемов, каждый из которых написан на своем языке, но алфавиты обоих текстов одинаковые. Тексты будем анализировать без учета гласных.

Рассмотрим сначала тексты из одной языковой группы.

На рис. 11 показаны усредненные распределения текстов на русском и болгарском языках без огласовки. Кроме того, приведен график модельной зависимости (1) для 20-буквенного алфавита при значении параметра $o = 0$. Оказалось, что и «чистые», и 50/50 смешанные тексты на русском и болгарском языках имеют близкие распределения с одной и той же детерминацией, равной 0,96, а отклонение фактического распределения от модельного равно 0,10. Эта смесь, очевидно, имеет другие статистические свойства, чем МВ в любой из двух транскрипций.

На рис. 12 показаны аналогичные распределения для англо-немецких текстов. Оказалось, что и для них, как и для франко-итальянских текстов и вообще текстов на языках одной группы или подгруппы (в рамках трех групп: славянской, германской и романской) детерминация логарифмической аппроксимации смеси приблизительно совпадает с детерминацией текстов на одном языке и составляет те же 0,96.

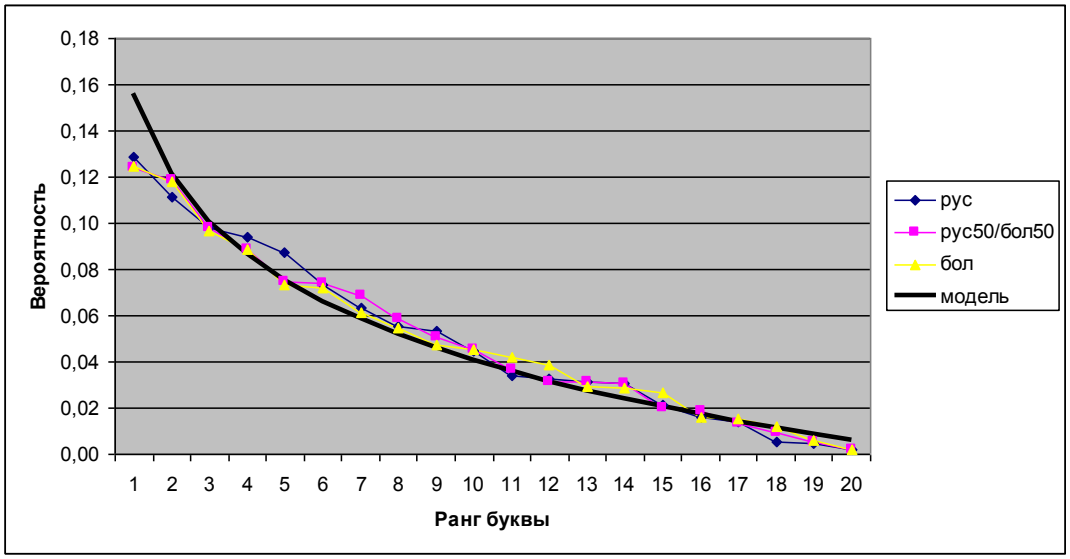

Рис. 11 – Аппроксимация русско-болгарских текстов

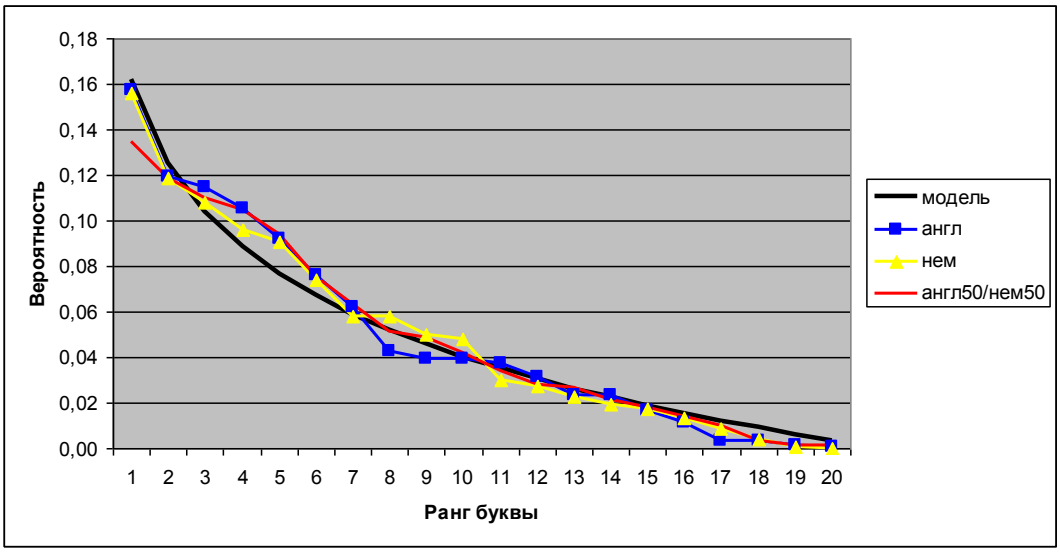

Рис. 12 – Аппроксимация англо-немецких текстов

Таким образом, выяснилось, что языки одной группы не только имеют близкие распределения упорядоченных частот в текстах без огласовки, но и смесь из них имеет ту же детерминацию логарифмической аппроксимации, что и составляющие. Было бы интересно проверить это наблюдение на текстах XVI века, но в данной работе мы ограничимся только анализом статистики современных текстов, понимая, что полученные выводы не являются строго доказательными применительно к МВ.

Рассмотрим теперь примеры смешения текстов из разных языковых групп индоевропейской семьи. На рис. 13 показаны распределения частот в испано-английских текстах без огласовки. Отметим, что смешение в равных долях английского и испанского текстов приводит к детерминации на уровне 0,92 с отклонением 0,17 аппроксимации в норме L1 от фактического распределения. По виду статистики это смешение похоже на транскрипцию Takahashi, расстояние между двумя распределениями составило 0,09.

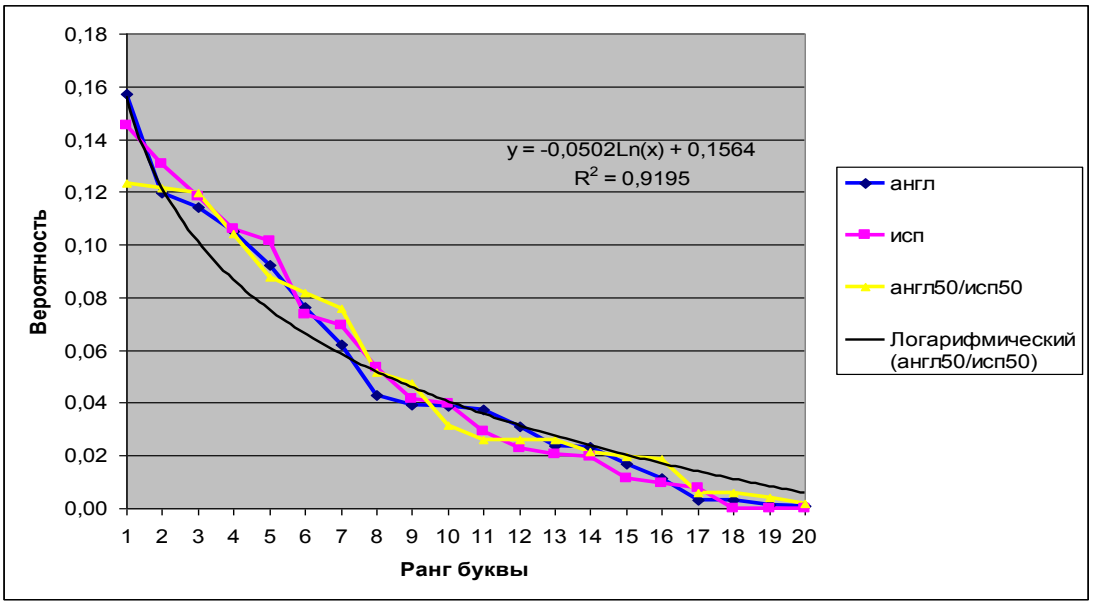

Рис. 13 – Аппроксимация испано-английских текстов

Характерно, что на расстоянии 0,08-0,10 находятся все распределения, отвечающие разным испано-английским текстам, взятым в пропорции 50/50. Следовательно, имеет смысл поискать пропорцию между объемами текстов на этих языках, при которой детерминация повысится до 0,93. Такая пропорция находится: ей отвечает примерно 60 % английского текста и 40 % испанского. Расстояние между распределениями этой смеси и транскрипции Takahashi в норме L1 составило 0,08. Тем самым статистическая гипотеза о таком языковом составе текста MB может считаться вполне допустимой.

Попутно выяснилось, что тексты, написанные на разных языках одной группы, не меняют своего распределения упорядоченных частот символов при смешивании текстов в любых пропорциях. Для текстов же на языках разных групп распределение смеси либо меняется по сравнению с распределениями оригиналов, либо расстояния между распределениями значительно превосходят наблюдаемый уровень кластеризации. Подчеркнем, что эти выводы относятся только к литературным текстам (т.е. написанным профессиональными писателями), из которых потом удалены гласные и смягчающие знаки.

Надо заметить, что Манускрипт по ряду признаков может считаться не столько литературным произведением, сколько представлять собой некоторый специализированный текст (например, рецептурную книгу астрологического характера). В связи с этим необходимо исследовать, как зависит статистика упорядоченных частот от жанровой и тематической специфики текста. В [1] было выяснено, что большая совокупность текстов разных жанров, объединенных в один фолиант (порядка 5 млн знаков без огласовки) имеет одно и то же логарифмическое распределение с нулевым значением параметра o в (1) независимо от языка. Однако чисто профессиональные тексты в [1] не рассматривались.

Подчеркнем, что если использовать полный алфавит, то каждая деятельность (физика, история, медицина и т.п.) имеет определенную

«жанровую» функцию распределения частот текстов. Эта функция хотя и находится в пределах точности аппроксимации (1) для данного языка, все же имеет тематическую специфику, позволяющую с ошибкой на уровне 0,2 классифицировать тексты по тематике, исходя только из однобуквенного распределения.

На рис. 14 приведены некоторые распределения тематических текстов на русском языке без огласовки. Оказалось, что нюансы между разными темами несущественны, уровень детерминации флуктуирует около 0,96, а отличие фактических распределений от модельной зависимости (1) изменяется в пределах 0,09-0,13, что примерно совпадает с аналогичными характеристиками для литературных текстов.

Следовательно, можно считать, что тематика текста не оказывает существенного влияния на распределение его упорядоченных согласных, и потому выводы, полученные при рассмотрении литературных текстов, имеют расширительное применение и для текстов по специальной тематике.

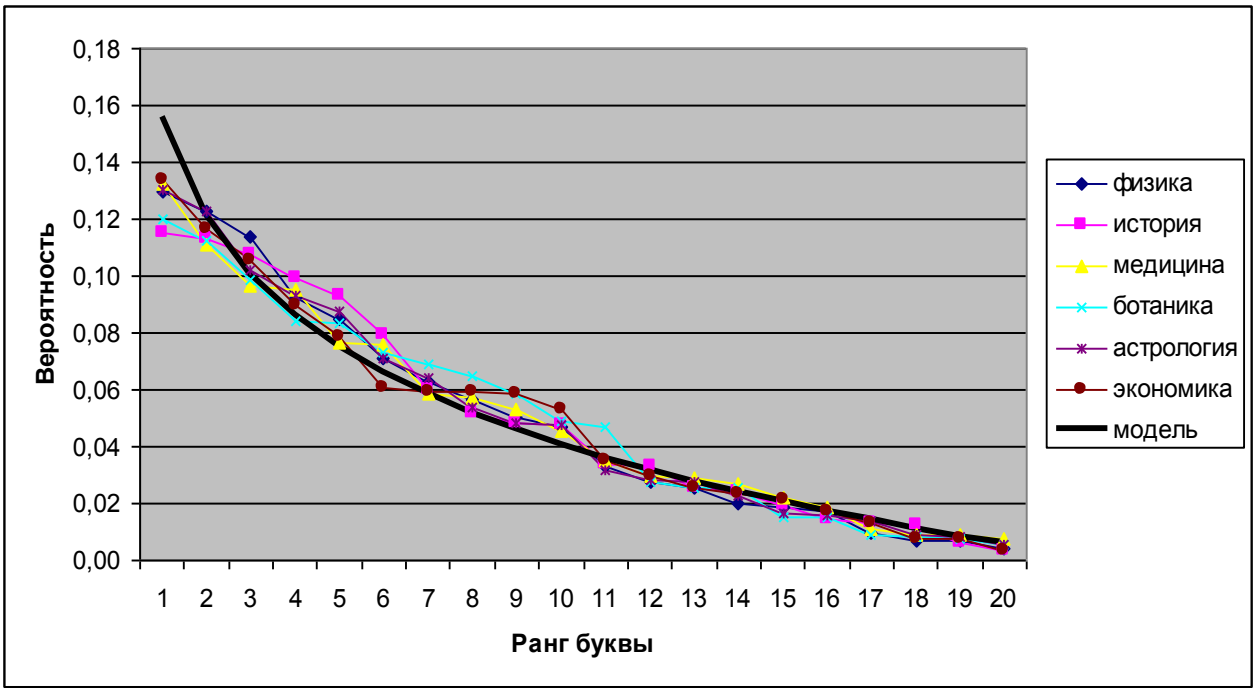

Рис. 14 – Аппроксимация распределений частот
в специализированных текстах на русском языке без огласовки

Такими же свойствами обладают распределения специальных текстов и на других языках. Например, юридические тексты и законодательные акты на английском, немецком, голландском, шведском и датском языках (как представителей германской группы), записанные без гласных, имеют логарифмическую аппроксимацию с высокой детерминацией, близкой к 0,99, но при значении $o = -3$ (см. рис. 15).

Следовательно, можно считать, что тексты без огласовки не имеют выраженной жанровой специфики, проявляющейся в распределении частот.

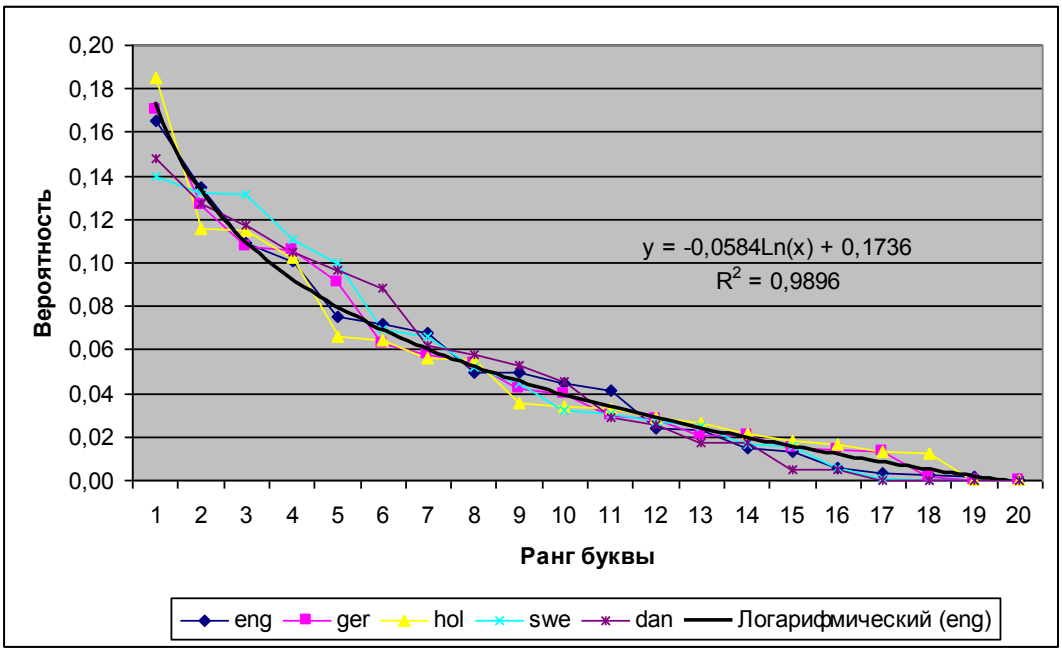

Рис. 15 – Аппроксимация распределений частот в юридических текстах на языках германской группы без огласовки

5. Идентификация языка фрагмента текста

Построенные в п.2 распределения можно использовать для ответа на вопрос, где в тексте рукописи используется преимущественно один язык (например, испанский), а где смешанный. Для этого надо применить метод идентификации выборочных функций распределения для малых выборок, предложенный в [12]. Суть метода состоит в следующем. Пусть имеются эталонные функции распределения (паттерны) $F_i(x)$ и некоторый фрагмент временного ряда, выборочная функция распределения которого есть $G(x)$. Тогда этот фрагмент считается выборкой из распределения $F_j(x)$ с номером

$$j = \arg\min \left\| F_i(x) - G(x) \right\|. \qquad (4)$$

Норма, в которой вычисляется расстояние между функциями в (4), выбирается из соображений минимальной ошибки идентификации на тестовом массиве данных. В [12] было выяснено, что для выборок малых длин порядка 50-200 значений и для плотностей распределений, интегрально уклоняющихся одна от другой в норме L1 на величину порядка 0,1-0,2, наилучшей нормой является норма L1 между функциями распределения, т.е.

$$\left\| F(x) - G(x) \right\| = \int \left| F(x) - G(x) \right| dx. \qquad (5)$$

В качестве эталонных плотностей распределений в примере с МВ выступают два: это эмпирическое распределение $f_1(k)$ текста на одном языке (например, на латыни) без огласовки, имеющее логарифмическую аппроксимацию с детерминацией 0,96, и эмпирическое распределение $f_2(k)$ смеси текстов на языках из двух разных групп (испанском и английском), имеющее логарифмическую аппроксимацию с детерминацией 0,93, близкое к

графику «смеси» на рис. 13. Для удобства сравнения с ранее вычисленными расстояниями между распределениями будем использовать норму L1 между плотностями вероятностей, т.е. вычисляем расстояния между эталонами и выборочной плотностью функции распределения $g_n(k,t)$, построенной по выборке длины n в скользящем окне с единичным шагом по времени:

$$\rho_i(n,t)=\sum_{k=1}^{20}\left|f_i(k)-g_n(k,t)\right|,\quad i=1,2. \tag{6}$$

Аргумент t отвечает номеру символа в тексте МВ, которым оканчивается выборка.

Сравнивая между собой одновременные расстояния $\rho_1(n,t)$ и $\rho_2(n,t)$, можно идентифицировать соответствующие фрагменты как написанные преимущественно на одном языке, если $\rho_1<\rho_2$, или на двух языках, если $\rho_2<\rho_1$. Если взять достаточно большую длину выборки n, то в каждый момент времени будет идентифицироваться ситуация $\rho_2<\rho_1$ (см. рис. 16), что может показаться недостаточно информативным с точки зрения внутренней структуры текста. Тем не менее, по расхождению расстояний между фрагментом текста МВ длиной в 30 тыс. и эталонами можно понять, что МВ явно состоит из трех частей (без учета первых 30 тыс.): 45 тыс., 70 тыс. и 40 тыс. знаков. Первая и последняя части, по-видимому, содержат фрагменты, написанные на латыни, а середина содержит текст на испанском и английском (немецком или датском) языках.

Рис. 16 – Ряды расстояний от фрагмента текста длиной 30 тыс. знаков до эталонов латыни и англо-испанской смеси

В то же время, как и ожидалось, на выборках малых длин фрагменты устойчиво идентифицируются как написанные на языке определенных групп: латынь, романская группа, германская группа. Временной ряд расстояний $\rho_i(n,t)$ для выборок длиной $n=200$ (5-7 строк МВ в транскрипции Takahashi) показан на рис. 17. Видно, что периодически язык текста оказывается близок тем или иным языкам рассматриваемых групп и стабильно удален от «смеси». Также наблюдаются фрагменты, когда расстояния до всех эталонов примерно

одинаковы, что может означать либо отсутствие нужного эталона в нашей библиотеке, либо переходное состояние, когда в скользящем окне мы захватываем часть текста на одном языке, а часть на другом. То, что таких мест сравнительно мало, свидетельствует об узости окна: по-видимому, на одном языке написаны одна-две страницы рукописи, что составляет 1500 знаков.

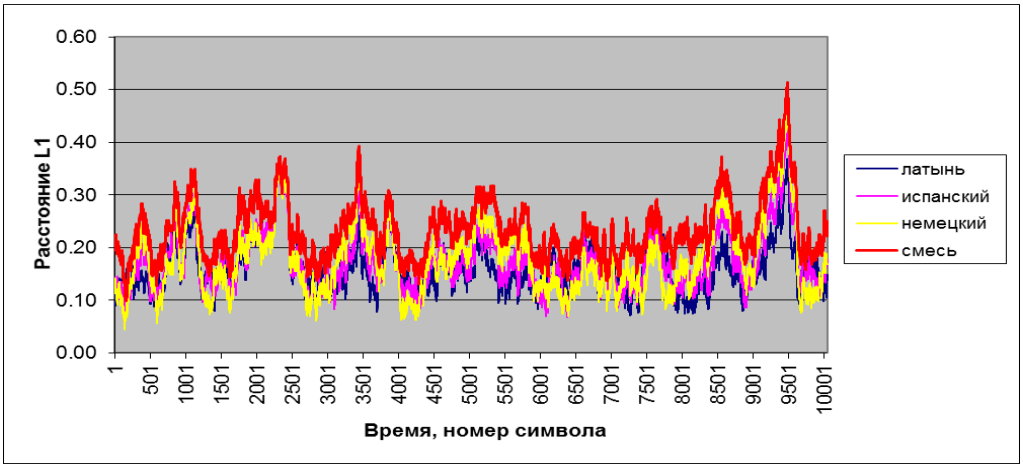

Рис. 17 – Ряды расстояний от фрагмента текста длиной 200 знаков до эталонов

Разумеется, сделанный вывод о структуре текста – лишь одна из возможностей; авторы допускают, что может существовать эталон некоего языка, находящегося ближе к фрагментам МВ, чем выбранные в данной работе. Но все же следует подчеркнуть, что совокупность аргументов в пользу того, что исследуемый текст написан на двух или более европейских языках, пока не встречает противоречий.

6. Анализ спектрального портрета Манускрипта Войнича

Рассмотрим матрицу P_{ij} эмпирических условных вероятностей того, что в некотором месте текста находится символ j при условии, что слева от него находится символ i. Эта матрица выражается через двухбуквенное $F(i,j)$ и однобуквенное $f(i)$ распределения вероятностей:

$$P_{ij} = \frac{F(i,j)}{f(i)}, \quad f(i) = \sum_j F(i,j). \tag{7}$$

Из (7) следует, что матрица P_{ij} имеет одно из собственных значений, равное 1, и этому значению отвечает собственный вектор $f(i)$. Другие собственные числа этой матрицы характеризуют устойчивость частот пар букв для фрагментов текста. Согласно С.К. Годунову [13], число λ принадлежит ε-спектру $\Lambda_\varepsilon(P)$ матрицы P, если существует такая возмущающая ее матрица Δ, что $\|\Delta\| \le \varepsilon \|P\|$ и $\det(\lambda I - P - \Delta) = 0$.

Резольвентой матрицы P называется матрица

$$R(\lambda) = (\lambda I - P)^{-1}. \tag{8}$$

В терминах резольвенты ε-спектр определяется следующим образом: число λ принадлежит ε-спектру $\Lambda_\varepsilon(P)$ матрицы P, если

$$\|R(\lambda)\| \geq \frac{1}{\varepsilon\|P\|}. \qquad (9)$$

На этой формуле основан практически применяемый численный алгоритм определения областей в комплексной плоскости параметра λ, отвечающих спектральным пятнам в зависимости от величины возмущения ε. При исследовании расположения точек спектра представляют интерес замкнутые гладкие кривые γ_ε, представляющие изолинии ε-спектра. Контур γ_ε разбивает весь ε-спектр $\Lambda_\varepsilon(P)$ на две части – лежащие внутри и вне его. Параметр дихотомии $\kappa_\gamma(P)$ оценивается нормой квадрата резольвенты (9) на данной кривой:

$$\kappa_\gamma(P) = \frac{\|P\|^2}{l_\gamma}\oint_\gamma \|R(\lambda)\|^2 d\lambda. \qquad (10)$$

Здесь l_γ есть длина контура γ. Величина $\kappa_\gamma(P)$ выбрана как индикатор точности разделения спектра потому, что если на некоторой кривой γ нет точек спектра $\lambda(P)$, то норма резольвенты на такой кривой конечна: $\|R(\lambda)\|_\gamma < \infty$, как и интеграл от нее по этой кривой.

Если внутри области, ограниченной кривой γ_ε, оказалось несколько собственных значений, то с указанной точностью ε их естественно считать совпадающими. Тогда подпространство с базисом из собственного и присоединенных векторов для такого кратного собственного значения будет инвариантным подпространством для оператора P. Проектор Π на это инвариантное подпространство определяется формулой:

$$\Pi_\gamma = \frac{1}{2\pi i}\oint_\gamma R(\lambda)d\lambda.$$

Удобно рассматривать радиальную дихотомию, т.е. дихотомию, задаваемую кривой $\lambda = re^{i\varphi}$ при фиксированном значении r. Тогда параметр дихотомии $\kappa_r(P)$ является нормой эрмитовой матрицы $H_r(P)$, имеющей интегральное представление

$$H_r(P) = \frac{1}{2\pi}\int_0^{2\pi}\left(P^+ - re^{-i\varphi}I\right)^{-1}\left(P - re^{i\varphi}I\right)^{-1}d\varphi, \quad \kappa_r(P) = \|P\|^2\cdot\|H_r(P)\|. \qquad (11)$$

Интеграл в (11) сходится только в том случае, если на окружности $\lambda = re^{i\varphi}$ нет собственных значений матрицы P. Эта формула используется при численном нахождении ε-спектра матрицы в виде линий уровня для L_2-нормы резольвенты, которые (линии) и приведены ниже.

Представляет интерес сравнить спектральные портреты матриц (7) для двух транскрипций МВ, а также для текстов германской и романской групп без

огласовки. Результаты вычислений представлены на рис. 18-19. Одинаковым цветом закрашены области, в которых находятся собственные значения матриц, если элементы этих матриц известны с точностью, указанной в легенде.

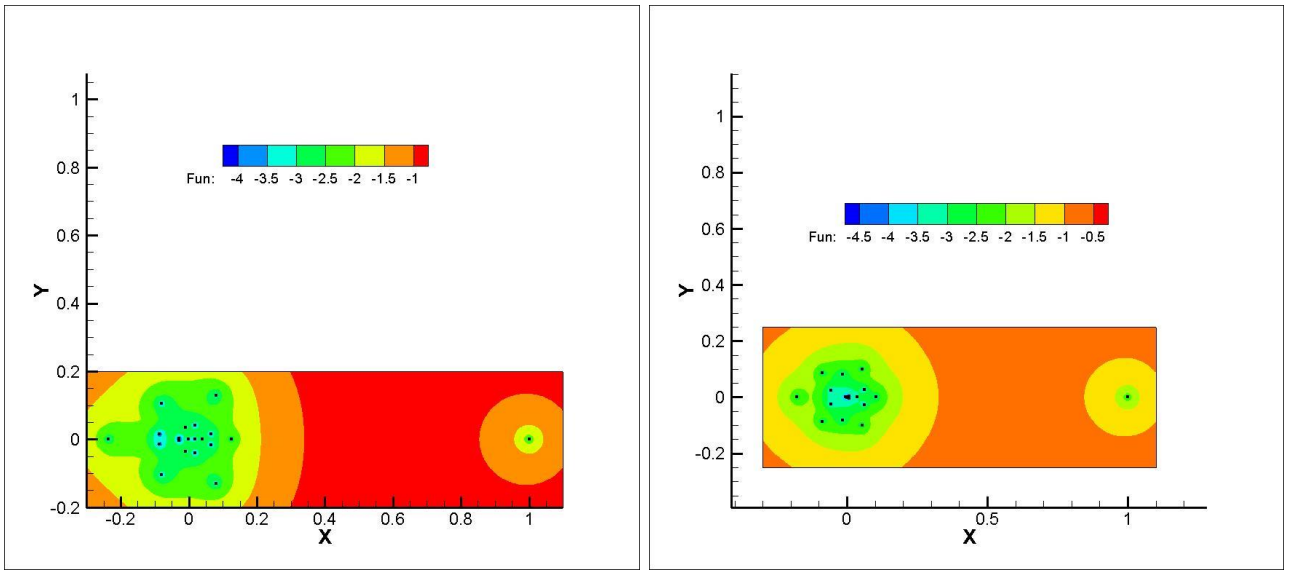

Рис. 18 – Спектральные портреты текстов без огласовки
на английском языке (слева) и на латыни (справа)

Все матрицы вида (7) имеют одно обособленное собственное значение, равное единице. Остальные собственные значения образуют структуру, характерную для того или иного языка. Интерес представляют действительные собственные значения, ядро вблизи нуля, а также большие по модулю комплексные собственные значения. Для всех европейских языков область расположения спектра приближенно ограничена кругом радиуса 0,2 (зеленая область на рис. 18). Как было выяснено в [1], для текстов в полном алфавите область расположения спектра имеет вид не круга, а эллипса, большая полуось которого равна приблизительно 0,5, а малая по-прежнему равна 0,2.

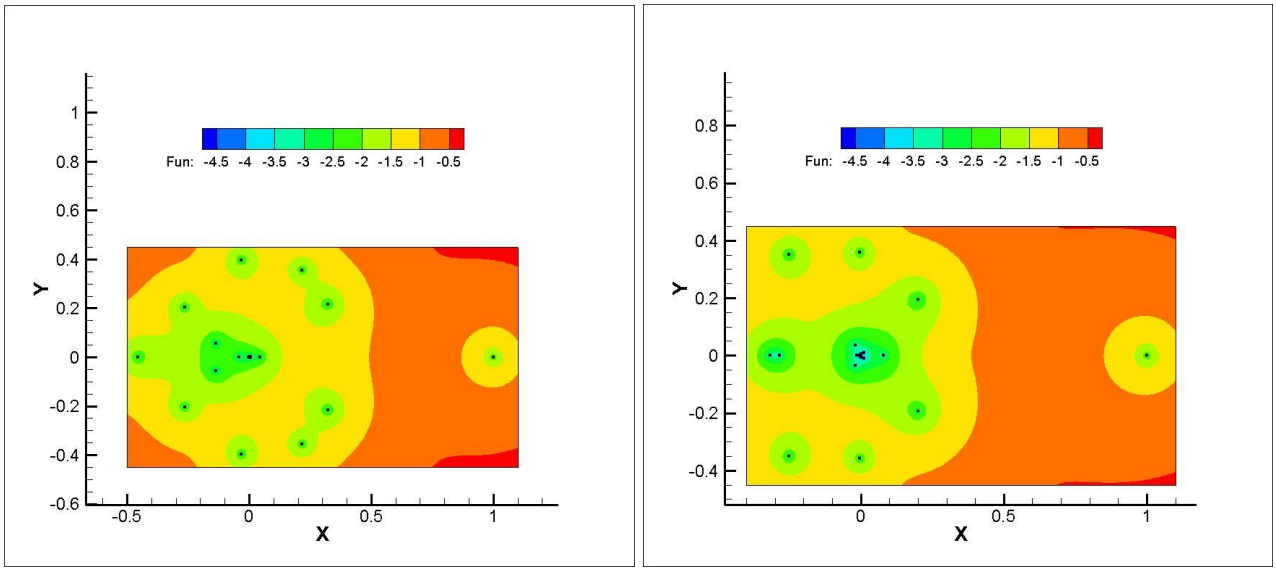

Рис. 19 – Спектральные портреты транскрипций МВ:
EVA (слева) и Takahashi (справа)

Сравнивая рис. 18 и рис. 19, видим, что области равной точности в нахождении собственных значений матриц (7) для МВ и обычных текстов (как в полном алфавите, так и без огласовки) заметно отличаются. Принципиальное значение имеет то, что для обеих транскрипций МВ круг (не эллипс!) расположения собственных значений имеет примерно в два раза больший радиус, чем для естественных языков. При этом спектр EVA сдвинут влево, а спектр Takahashi – вправо. Отличие спектральных портретов транскрипций отвечает различиям в распределениях упорядоченных частот (красная и зеленая кривые на рис. 1), для которых выпуклости указанных кривых меняются в противофазе. Характерно, что обе транскрипции имеют пять несвязных спектральных зон равной точности 10^{-2} (светло-зеленый цвет на рис. 19).

То, что собственные значения транскрипций МВ лежат в круге, а не в эллипсе, отличает именно тексты без гласных. В два раза же больший радиус этого круга свидетельствует о том, что возможные соседства пар символов более вариативны, чем для одного языка. Тем самым полученные в этом разделе результаты не противоречат выдвинутой концепции составного языка МВ и дополняют ее еще одним статистическим аргументом. Представляется важным подчеркнуть, что все эти аргументы принципиально различны, т.е. выражают особенности независимых статистик, указывающих на то, что трактовка МВ как составной рукописи вполне допустима.

В заключение этого раздела авторы выражают глубокую благодарность к.ф.-м.н. О.Б. Феодоритовой за помощь в проведении расчетов спектров матриц и интерпретации особенностей спектральных портретов.

7. Замечания о структуре Манускрипта

В заключительной части нашего исследования обсудим принципиальную возможность статистического выявления фрагментов рукописи, написанных, если этот термин применим к тексту на неизвестном языке, «в сходной манере».

Как было отмечено во Введении, принятая нумерация листов МВ не обязательно имеет правильное упорядочение. Статистическое исследование фрагментов МВ в скользящем окне, выполненное в разделе 5, показало (см. рис. 16), что МВ можно трактовать не как единый документ, а как два-три независимых произведения, написанных, допустим, неким «братством» на понятном только им языке. В этом разделе мы рассмотрим статистические аспекты анализа МВ на однородность отдельных листов. Следуя методике [1], для решения этой задачи будем использовать функционал расстояния между распределениями символов в их алфавитном упорядочении. Для определенности анализируем транскрипцию Takahashi.

Анализ литературных текстов на европейских языках, выполненный в [1], показал, что между двумя достаточно большими (от 10 тыс. знаков) текстами одного автора на одном и том же языке расстояние в смысле нормы в L1 для

алфавитно упорядоченных распределений символов составляет 0,03-0,07, для разных авторов это расстояние лежит в интервале 0,04-0,13, а для разных языков независимо от авторов расстояние между текстами составляет 0,20-0,50. Применительно к МВ [2] имеет смысл проверить, насколько близки между собой распределения условных частей рукописи в соответствии с имеющимися иллюстрациями. Не претендуя на оригинальность, мы выделим традиционную «ботаническую» часть МВ (это листы 1-57 – см. рис. 20), «женские тела» (листы 75-85) и «астрологию» (листы 103-116). На возможные нарушения последовательной нумерации листов указывает то, что листы 87, 90, 93-96 имеют явно «ботанический» вид, листы 58, 68, 69 похожи на последующую «астрологическую» часть, а листы 65, 66 было бы удобно отнести к «женским телам». Не ясна также тематическая принадлежность листов 67, 70-73, 86, 88, 89, 99-102, орнаментированных некими «ступками» и растениями.

В табл. 4 приведены расстояния между выделенными частями МВ, понимаемые как расстояния в норме L1 между соответствующими распределениями упорядоченных частот.

Табл. 4. Расстояния между распределениями символов в частях МВ

	ботаника	**тела**	**астрология**	**ступки**
ботаника		**0,30**	**0,20**	**0,25**
тела			**0,18**	**0,27**
астрология				**0,20**
ступки				

Приведенные в табл. 4 расстояния характерны для текстов, написанных на разных языках. При этом тексты могут быть одинаковыми (например, роман на английском и он же на французском) или разными. Важно, что тексты, написанные на одном и том же языке, отличаются в норме L1 не более чем на 0,13. Длина каждого из фрагментов МВ более 10 тыс. знаков, так что вывод о языковой неоднородности МВ представляется обоснованным. Отметим, что половина каждой части отличается от другой половины на 0,10, поэтому выделение частей МВ в соответствии с рисунками вполне логично.

Проанализируем, к каким частям МВ близки отдельные листы, которые не имеют однозначной трактовки. На одном листе МВ (две страницы) содержится от 500 до 2 тыс. символов в зависимости от размеров рисунков. Лист из «своей» части отстоит от нее на расстояние 0,10-0,30, а от «чужих» на 0,25-0,50. Попробуем идентифицировать принадлежность вышеперечисленных неоднозначных листов посредством близости их распределений к распределениям частей МВ.

Рис. 20 – Примеры листов MB [2] в соответствии с частями табл. 4

«Ботаническая» часть оказалась ближайшей для следующих листов: 69 (расстояние до эталона 0,36), 86 (расстояние 0,16), 87 (расстояние 0,30), 93-96 (расстояние 0,15).

К «телам» близки листы 65 (расстояние 0,32) и 66 (расстояние 0,17).

К «астрологии» близок лист 58 (расстояние 0,33).

Остальные неоднозначно интерпретируемые листы 67, 68, 70-73, 88-90 близки к «ступкам» (листы 99-102) с расстояниями 0,17-0,32.

Таким образом, 4 листа из 23 были идентифицированы не так, как следовало бы в соответствии с рисунками. Это листы: 68 (ступки вместо ожидаемой астрологии), 69 (ботаника вместо астрологии), 86 (ботаника вместо ступок) и 90 (ступки вместо ботаники). Важно, однако, не это, а то, что некоторые 4 листа (не все из них те, которые идентифицированы «не так») отстоят от ближайших к ним эталонов на весьма большие расстояния, превосходящие 0,32 (последний дециль расстояний для «достоверно своих» листов), что показывает недостаточную надежность вывода по ним. Доля таких фрагментов среди рассматриваемых 23 листов составила 0,16. Заметим, что точность аналогичной идентификации по литературным текстам [1] составила 0,15, что близко к сделанным оценкам для МВ и типично для этого метода.

Подчеркнем, что вывод о неоднородности (уже структурной, а не лингвистической) рассматриваемого текста подтверждается двумя различными статистическими исследованиями – «тонкой» процедурой анализа распределений в скользящем окне (раздел 5) и статической кластеризацией «тематически похожих» фрагментов. Результаты обеих процедур оказались вполне совместимыми: различающиеся фрагменты, найденные в разделе 5, могут претендовать на тематическое объединение с теми фрагментами, расстояния до которых значимо меньше, чем до других, причем эти расстояния примерно равны расстояниям между частями фрагментов, считающихся внутренне однородными. Это свидетельствует о корректности предложенного объединения. Тем самым возможна новая точка зрения на МВ как на рукопись, написанную не только несколькими языками, но и как вообще не на одну, а на две-три различных рукописи.

Укажем теперь точность, с которой получены представленные в работе результаты. При статистическом распознавании образов путем сравнения с эталоном критическим местом является точность, с которой известен сам эталон. Под эталоном в нашем случае понимается вероятностное распределение текста по символам. Если текст состоит из N знаков и записан алфавитом из n символов, то распределение этих символов в тексте определено с точностью ε, которая находится численно из уравнения [1]:

$$\frac{u_{1-\varepsilon/2}}{\varepsilon}=\frac{\sqrt{N}}{\Sigma_N(n)},\quad \Sigma_N(n)=\sum_{j=1}^{n}\sqrt{f_N(j)\cdot\left(1-f_N(j)\right)}. \qquad (12)$$

Здесь u_γ есть γ-квантиль нормального распределения, которым для простоты аппроксимируется соответствующий квантиль распределения Стьюдента порядка N при больших значениях N, а $f_N(j)$ есть эмпирическая

частота встречаемости j-го символа в данном тексте длины N. В частности, для логарифмической модели упорядочения (1) при $n=20$ и $o=0$ значение суммы в (12) равно 3,93; применительно к МВ с числом символов $N=1{,}7\cdot10^5$ его фактическое распределение таково, что $\Sigma_N(n)=3{,}65$. Правая часть уравнения относительно ε в (12) для теоретической модели равна 105, а для МВ – соответственно 113. Этим значениям отвечают близкие точности $\varepsilon=0{,}02$, различающиеся в третьем знаке после запятой. Аналогично выясняется, что точность распределения частот для одного листа (1500 знаков) составляет 0,1. Следовательно, отличия между распределениями фрагментов на уровне 0,08-0,13, а листов от фрагментов – на уровне 0,20-0,40 вызваны не статистическим шумом выборок, а отвечают существу дела.

Согласно полученным оценкам точности, достоверные области спектральных портретов МВ на рис. 19 соответствуют светло-зеленой легенде пиктограмм. Таким образом, с указанной точностью статистических оценок различие между выборками корректно определено.

Заключение

В результате проведенных статистических исследований было установлено следующее.

Во-первых, групповая классификация индоевропейских языков может быть осуществлена формальной математической операцией – попарной кластеризацией распределений упорядоченных частот текстов без огласовки. Во-вторых, внутри подгрупп родственные языки могут быть смешаны без существенного изменения таких распределений. В-третьих, для уральской семьи кластеризация языков по указанному выше правилу не проходит, т.е. это правило не является универсальным. В-четвертых, показатель Херста (или его распределение) представляется устойчивым инвариантом языка. В-пятых, спектральные портреты текстов на языках индоевропейской семьи имеют сходные черты в расположении групп собственных значений.

В дальнейшем предполагается продолжить исследования в направлении поиска языковых инвариантов с целью установления статистических связей между различными языковыми группами и семьями. Возможно, это позволит лучше понять процессы, лежащие в основе самоорганизации (происходящей через посредство людей, конечно) набора слов в лексикон.

Что касается Манускрипта Войнича, то наиболее вероятной гипотезой о структуре языка, на котором он написан, является такая: МВ написан на смешанном языке без огласовки, 60 % текста написано на одном из языков западногерманской группы (английский или немецкий), а 40 % текста – на языке романской группы (итальянский или испанский) и/или на латыни. Аргументами в пользу такого вывода являются следующие: статистика символов МВ похожа на статистику осмысленного текста, но поведение показателя Херста для расстояний между одинаковыми символами значительно

отличается от текстов, написанных на одном языке – естественном или искусственном; в то же время существует смесь, обладающая требуемыми статистическими свойствами; также и расстояния между алфавитными распределениями крупных частей МВ характерны для текстов, написанных на разных языках.

Кроме того, по-видимому, последовательность листов МВ может быть уточнена, если считать, что листы тематически должны быть собраны воедино. Считать ли части Манускрипта разными произведениями или одним, пока не ясно, поскольку большие расстояния между частями характерны для разных языков, а не разных произведений. Для последних они существенно меньше.

И все же на один из самых интригующих вопросов для многих о том, что на самом деле представляет собой Манускрипт Войнича, откуда он появился и кто, а главное, зачем его создал, авторы не могут пока дать однозначного ответа, ибо для этого требуется реальная и обоснованная расшифровка рукописи. Проведенное исследование позволяет предположить следующее (на правах исторической реконструкции).

Возможно, что некая небольшая группа (алхимиков?) – учитель и его немногочисленные ученики – разработали алфавит на основе современного им шрифта. На данном, весьма неплохо проработанном, надо сказать, шрифте они записали несколько текстов для внутреннего употребления, причем сами авторы, судя по легкости письма (символы не нарисованы каллиграфом, а написаны, причем многие из них слитно – см. рис. 20), хорошо понимали, что написано. Впрочем, сначала мог быть изготовлен черновик с переводом обычного текста на шифр, а уже затем этот шифр был записан в виде изучаемой нами рукописи. Однако эта, допустим, «алхимическая школа» достаточно быстро перестала существовать по неизвестным нам причинам, оставив после себя несколько (предположительно три) текста, которые и дошли до нас: «ботанический», «анатомический» и «астрологический». Хранились они, скорее всего, вместе, а после того как попали к другим алхимикам, уже никогда не использовались по причине того, что никто не смог их прочесть. Все последующие владельцы этих текстов не имели достоверного понятия о том, что попало им в руки: десяток страниц случайно переместился в неподходящие для них места, и лишь после этого страницы были пронумерованы одним из новых владельцев (видимо, для того, чтобы страницы не перепутались окончательно). Дальнейшее известно из стандартных описаний Манускрипта: тексты попали в папскую библиотеку, где были обнаружены Войничем и впервые им описаны.

Нам, конечно, неизвестно, о чем конкретно написано в этих текстах. Однако мы надеемся, что при помощи настоящей работы, опираясь на развитые методы и создавая новые, кто-либо из будущих исследователей сумеет это выяснить. Но, почти наверное, мы никогда не узнаем, что же в точности произошло с «алхимической школой», которая когда-то, пытаясь сохранить свои тайны для узкого круга посвященных, создала этот манускрипт.

В заключение авторы благодарят Н.Г. Звегинцеву за указание на ряд особенностей рукописного текста и иллюстраций, учет которых позволил более корректно провести кластеризацию отдельных частей Манускрипта Войнича и осуществить проверку текста на однородность.

Литература

1. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.
2. Shailor B.A. Voynich catalog record. Yale University Beinecke Rare Book & Manuscript Library.
3. Pelling N. J. The curse of the Voynich: the secret history of the world's most mysterious manuscript. – Surbiton, Surrey: Compelling Press, 2006. – 230 p.
4. Barabe J.G. Materials analysis of the Voynich Manuscript. Yale University Beinecke Rare Book & Manuscript Library.
5. Levitov L. Solution of the Voynich Manuscript: A liturgical Manual for the Endura Rite of the Cathari Heresy, the Cult of Isis. – Walnut Creek, California: Aegean Park Press, 1987. – 182 p.
6. Landini G., Zandbergen R. A well-kept secret of mediaeval science: The Voynich manuscript //Aesculapius. – 1998. – V. 18. – P. 77-82.
7. Транскрипция Takahashi. http://voynich.no-ip.com/folios/
8. Гусейн-Заде С.М. О распределении букв русского языка по частоте встречаемости // Проблемы передачи информации, 1988. Т. 24, вып. 4, с. 102.
9. Зенюк Д.А., Клочкова Л.В., Орлов Ю.Н. Моделирование нестационарных случайных процессов кинетическими уравнениями с дробными производными // Журнал Средневолжского математического общества, 2016. – Т.18. № 1.
10. Кириллов Д.С., Короб О.В., Митин Н.А., Орлов Ю.Н., Плешаков Р.В. Распределения показателя Херста нестационарного маркированного временного ряда / Препринты ИПМ им. М.В. Келдыша. № 11, 2013. – 16 с. http://library.keldysh.ru/preprint.asp?id=2013-11
11. Орлов Ю.Н., Осминин К.П. Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика, 2010. Т. 26, № 2, с. 95-108.
12. Власюк А.А., Орлов Ю.Н. Точность идентификации выборочных распределений временных рядов в зависимости от типа распределения, нормы и длины выборки // Препринты ИПМ им. М.В. Келдыша. № 17, 2015. – 25 с. http://library.keldysh.ru/preprint.asp?id=2015-17
13. С.К. Годунов. Современные аспекты линейной алгебры. – Новосибирск: Научная книга, 1997. – 388 с.