ИПМ им.М.В.Келдыша РАН

Будущее-2021 • Труды конференции


Ю.Н. Орлов

# Методы распознавания языка текста на примере манускрипта Войнича

***Рекомендуемая форма библиографической ссылки***

Орлов Ю.Н. Методы распознавания языка текста на примере манускрипта Войнича // Проектирование будущего. Проблемы цифровой реальности: труды 4-й Международной конференции (4-5 февраля 2021 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2021. — С. 220-235. — https://keldysh.ru/future/2021/20.pdf https://doi.org/10.20948/future-2021-20

***Размещено также видео выступления***

# Методы распознавания языка текста на примере манускрипта Войнича

**Ю.Н. Орлов**

*Институт прикладной математики им. М.В. Келдыша РАН*

**Аннотация.** 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. 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

# Language recognition methods and Voynich Manuscript analysis

**Yu.N. Orlov**

*RAS Keldysh Institute of Applied Mathematics*

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

**Keywords:** letters frequency distribution, European languages groups, Voynich Manuscript, spectral portrait

## 1. Введение

Манускрипт Войнича (далее МВ) [1] – это рукопись, датируемая исследователями XVI в. 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 [2–4], недостаточно убедительно подкреплены полноценными статистическими исследованиями. Целью работы, однако, не является расшифровка рукописи. Не анализируется словарный состав, поэтому смысловая составляющая текста в работе не обсуждается. Вопрос 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. Исследования, проведенные в [5], 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 МВ. 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.

Представляется, что идея расшифровки МВ через анализ «слов» без идентификации языка не продуктивна. 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 [5], 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 модель-

ного распределения можно будет судить о долевом участии разных языков в написании таких двуязычных текстов.

Настоящая работа посвящена исследованию инвариантных свойств европейских языков. Для нахождения 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 Хёрста 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.

## 2. Статистика символов МВ и аппроксимация частот

Последующий анализ будет направлен на то, чтобы построить распределение символов МВ 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.

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 [5] 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. 1-2 лучше всего моделируются зависимостью (1), в которой $o=0$. Для французского, испанского и итальянского языков $o=-3$, для датского и шведского языков $o=-1$. Для финского языка $o=-6$, для эстонского $o=-4$.

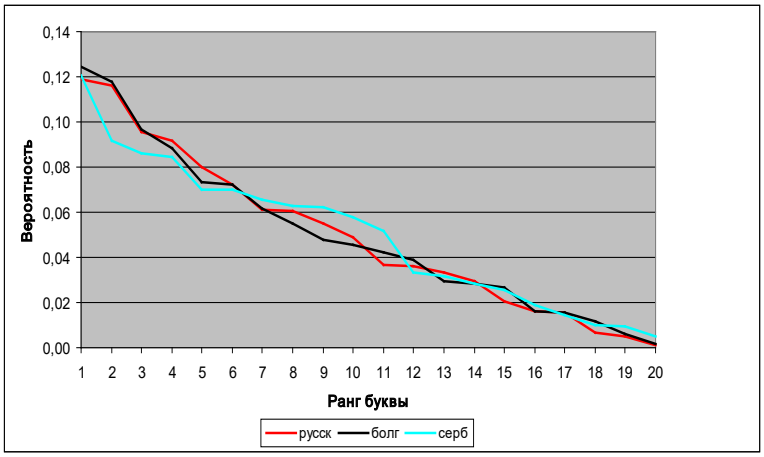


Рис. 1. Упорядоченные частоты символов в текстах на кириллице без огласовки

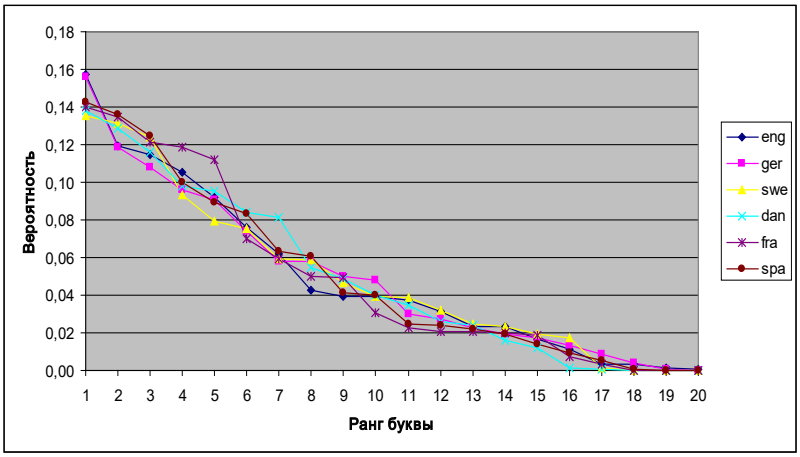


Рис. 2. Упорядоченные частоты символов в текстах на латинице без огласовки

Расстояния между распределениями символов в текстах 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 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 к 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.

Хотя эти распределения и близки по уровню детерминации аппроксимирующей зависимости (0,93), в деталях существенно различаются. 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, 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 МВ, 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. Насколько правильно распознаны знаки рукописи, здесь не обсуждается, исследуются только статистические свойства представленных транскрипций.

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).

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. При этом 90%-ый доверительный интервал составляет [0,085; 0,115].

Применительно к транскрипции EVA, для которой $n = 22$, наилучшая аппроксимация достигается при $o = -2$. Это означает, что фактически в МВ 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 (см. рис. 3). То же наблюдение верно и в отношении транскрипции Takahashi.

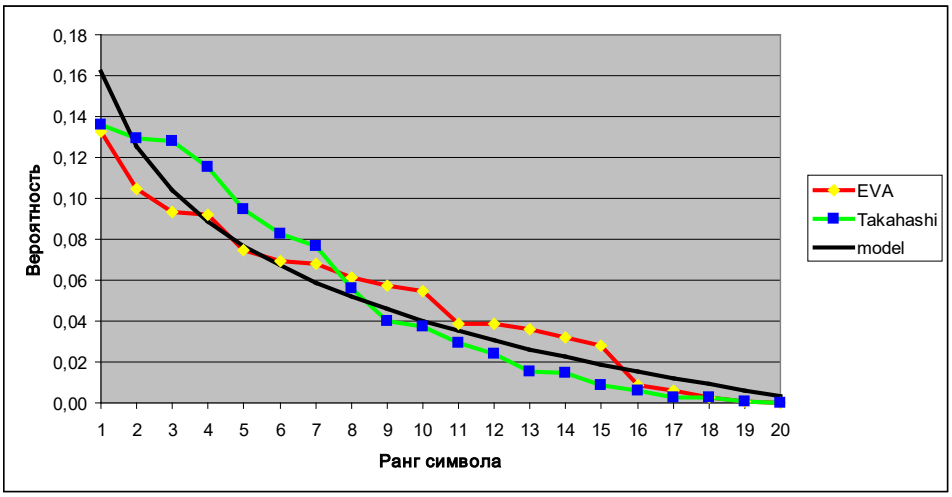


Рис. 3. Упорядоченные частоты двух транскрипций MB и логарифмическая аппроксимация

Однако, 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.

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). Из языков индо-европейской семьи в этом плане 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. 4).

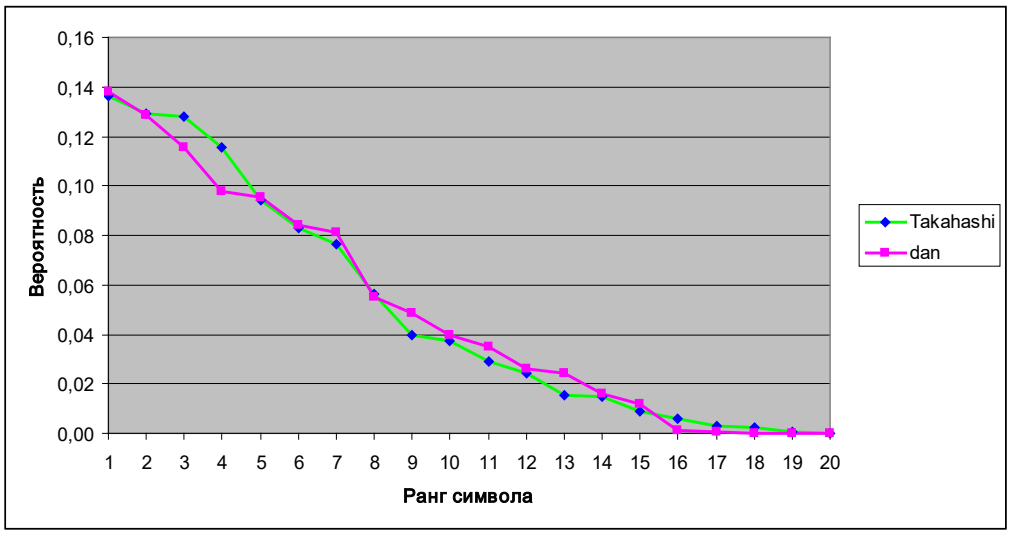


Рис. 4. Распределения частот символов в транскрипции Takahashi и в текстах на датском языке без огласовки

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 MB, 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.

Итак, приведен один аргумент в пользу того, что MB 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» $8 \cdot 10^{-4}$.

### 3. Использование индикатора Хёрста

Рассмотрим текст как временной ряд случайной величины (буквы), а буква 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.

Индикатор Херста вводится как показатель изменчивости временного ряда и 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:

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

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

$$H_N(t) = \frac{1}{N} \sum_{k=1}^{N} \left( \xi(t,k) - \bar{\xi}_N(t) \right) \left( 1 + \ln(k/N) \right). \qquad (2)$$

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

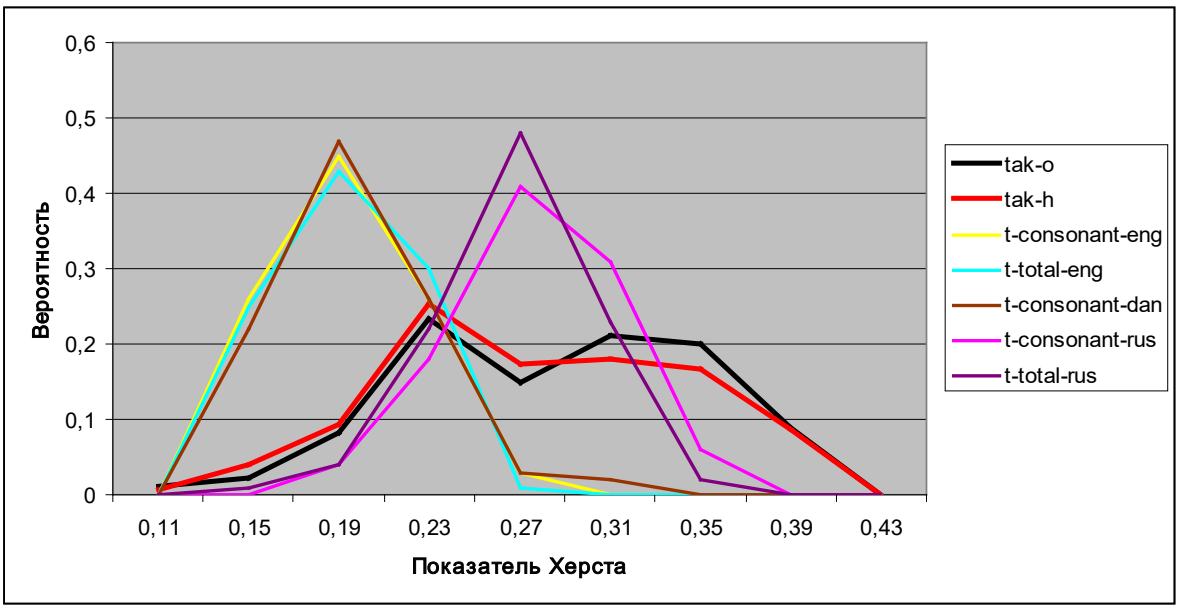


Рис. 5. Распределения показателей Хёрста для рядов расстояний между наиболее часто встречающимися буквами в текстах

Видно, что распределения на рис. 5 для русского и английского языков имеют максимумы в четко различающихся точках, а распределения для датского и английского языков практически совпадают. Распределение показателя Хёрста является индикатором языковой группы. Отсюда следует, что выдвинутый в разделе 2 вариант с датским языком в качестве оригинального языка MB следует исключить, поскольку распределения показателя Хёрста для рукописи имеют значительные отличия от аналогичных распределений для обычных текстов. Для MB это распределение более пологое и смещено вправо (черная и красная линии на рис. 5), 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. Следовательно, рукопись содержит записи на разных языках.

## 4. Анализ двуязычных текстов

Наблюдаемую в большинстве европейских языков дерминацию текстов без огласовки на уровне 0,96 можно понизить до 0,93, как в MB, 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 в. 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 [9], где определенная зависимость была отмечена.

Приведем здесь результаты статистического анализа частот в современных текстах, написанных на двух языках, но в одном алфавите. Рассмотрим сначала тексты из одной языковой группы.

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.

Также и для распределений символов для других текстов – англо-немецких, франко-итальянских и вообще текстов на языках одной группы или подгруппы – детерминация логарифмической аппроксимации смеси приблизительно совпадает с детерминацией текстов на одном языке и составляет те же 0,96.

Рассмотрим теперь примеры смешения текстов из разных языковых групп индоевропейской семьи. На рис. 6 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.

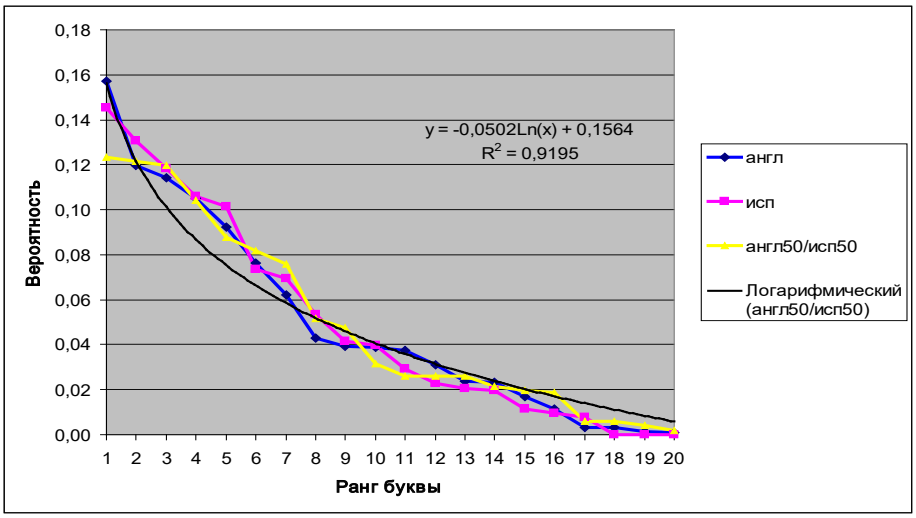


Рис. 6. Аппроксимация испано-английских текстов

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.

Построенные эталонные распределения символов в текстах на определенных языках 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. 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

$$j = \arg\min \left\| F_i(x) - G(x) \right\|. \qquad (3)$$

Результаты анализа выборок длины 1000 символов в транскрипции Takahashi привели к следующим результатам (рис. 7).

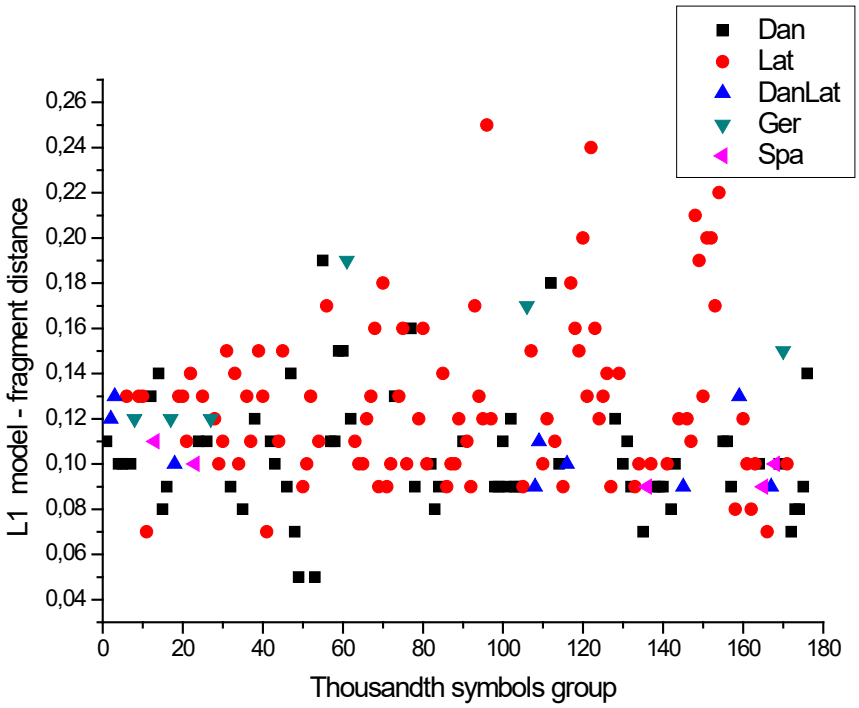


Рис. 7. Идентификация языка фрагментов МВ

Периодически язык текста оказывается близок тем или иным языкам рассматриваемых групп.

## 5. Использование спектральных портретов матриц биграмм

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:

$$P_{ij} = \frac{F(i,j)}{f(i)}, \quad f(i) = \sum_j F(i,j). \tag{4}$$

Из (4) 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 [10], число $\lambda$ принадлежит $\varepsilon$-спектру $\Lambda_\varepsilon(P)$ матрицы $P$, если существует такая возмущающая ее матрица $\Delta$, что $\|\Delta\| \le \varepsilon\|P\|$ и $\det(\lambda I - P - \Delta) = 0$.

При исследовании расположения точек спектра представляют интерес замкнутые гладкие кривые $\gamma_\varepsilon$, представляющие изолинии $\varepsilon$-спектра. Контур $\gamma_\varepsilon$ разбивает весь $\varepsilon$-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:

$$\kappa_\gamma(P) = \frac{\|P\|^2}{l_\gamma} \oint_\gamma \|R(\lambda)\|^2 \, d\lambda \,. \tag{5}$$

Здесь $l_\gamma$ есть длина контура $\gamma$. Величина $\kappa_\gamma(P)$ выбрана как индикатор точности разделения спектра потому, что если на некоторой кривой $\gamma$ 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 $\gamma_\varepsilon$, оказалось несколько собственных значений, то с указанной точностью $\varepsilon$ их естественно считать совпадающими.

Представляет интерес сравнить спектральные портреты матриц (4) для двух транскрипций МВ, а также для текстов германской и романской групп без огласовки. Результаты вычислений представлены на рис. 8-113. Одинаковым цветом закрашены области, в которых находятся собственные значения матриц, если элементы этих матриц известны с точностью, указанной в легенде.

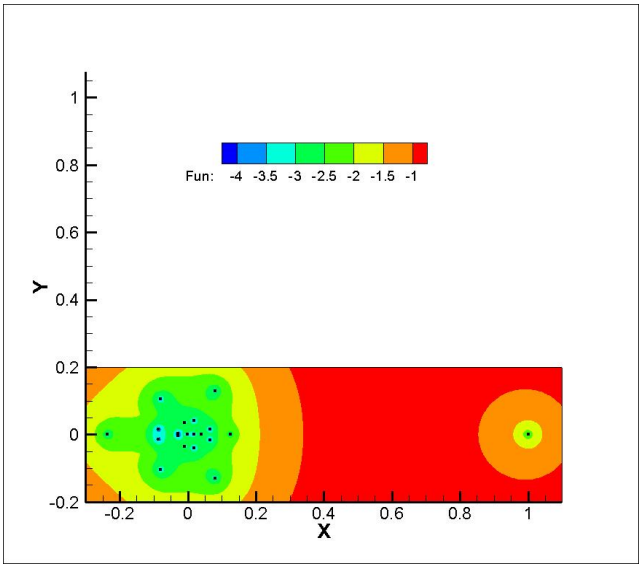


Рис. 8. Спектральный портрет текста без огласовки на английском языке

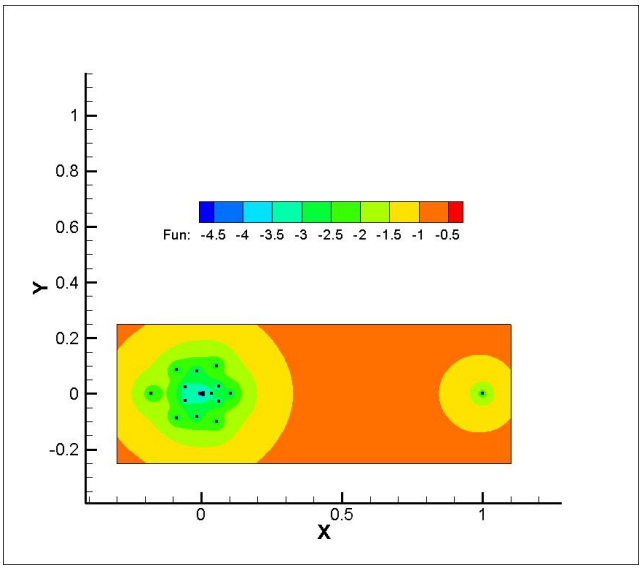


Рис. 9. Спектральный портрет текста без огласовки на латыни

Все матрицы вида (4) 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. 10-11). Как было выяснено в [5], 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.

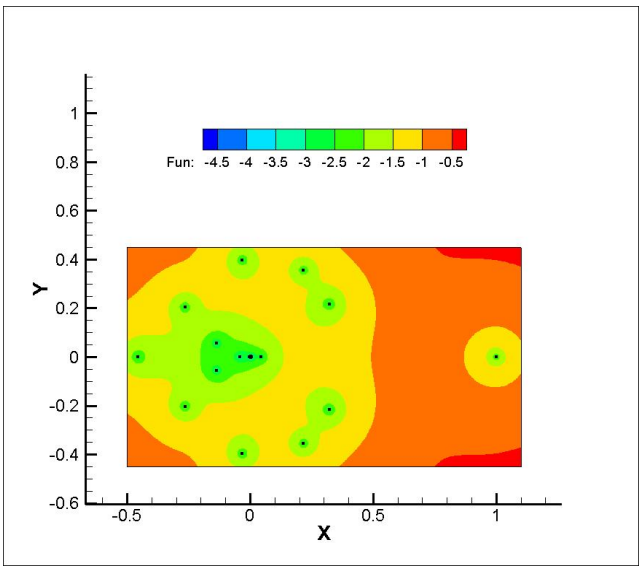


Рис. 10. Спектральный портрет транскрипции EVA

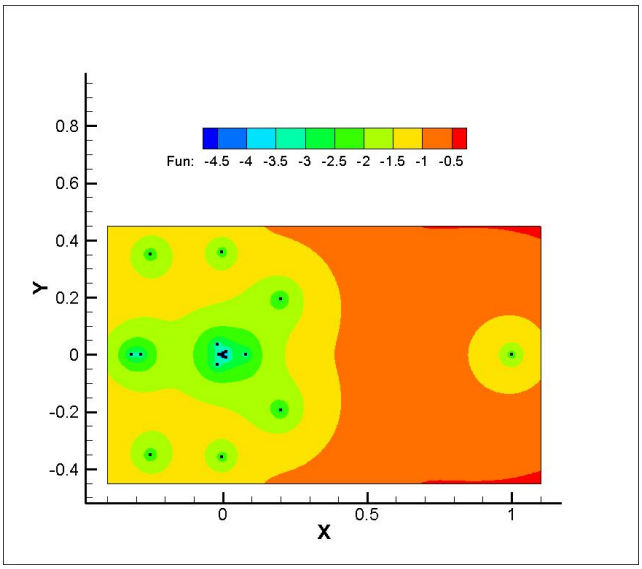


Рис. 11. Спектральный портрет транскрипции Takahashi

Сравнивая рис. 8-9 и рис. 10-11, видим, что области равной точности в нахождении собственных значений матриц (4) 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, 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. 11).

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.

Работа выполнена при поддержке РФФИ (проект 19-01-00602).

**Литература**

1. *Shailor B.A.* Voynich catalog record. Yale University Beinecke Rare Book & Manuscript Library.

2. *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.
3. *Barabe J.G.* Materials analysis of the Voynich Manuscript. Yale University Beinecke Rare Book & Manuscript Library.
4. *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.
5. *Орлов Ю.Н., Осминин К.П.* Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.
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. *Орлов Ю.Н., Осминин К.П.* Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика, 2010. Т. 26, № 2, с. 95-108.
10. *Годунов С.К.* Современные аспекты линейной алгебры. – Новосибирск: Научная книга, 1997. – 388 с.