Закон аббревиации Ципфа и поэтический язык

Борис Орехов

Терминологическая метафора

Поэтический язык

Поэтический текст подчиняется всем правилам естественного языка, но на него накладываются дополнительные ограничения, такие как метро-ритмические нормы и организованность на различных уровнях (фонологическом, лексическом и др.).

Ю. М. Лотман

Динамика стиха, нарушение традиционных правил просодии, роль метра и его взаимодействие со словом для восстановления выразительности.

Ю. Н. Тынянов

Язык арабской поэзии

1. Последний гласный произносится долго, танвин опускается. Финальная фатха записывается как алиф: إِيَّاهُ يَعْنُونا «они имеют это в виду»
2. Показатель аккузатива (танвин фатх) –an > ā или опускается: أَصْبَحَ كَئِيبًا или أَصْبَحَ كَئِيبْ «стал грустным»
3. В усиленном наклонении также –an > ā, но –un > ūn: تَفْعَلَا < تَفْعَلَنْ (2 Sg m), но تَفْعَلُونْ < تَفْعَلُنْ (2 Pl m).
4. Та марбута пишется как та мамдуда: وَفَكَّاكُ العُناتِ вм. وَفَكَّاكُ العُناةِ «и освободитель пленников»
5. Слова, заканчивающиеся на ىً и اً просто теряют танвин: عَصا < عَصاً , فَتى < فَتىً. Слова, оканчивающие на -ٍ, теряют танвин, а также могут терять последний слабый: < قاضٍ قاضي или قاضْ .
6. Слитное местоимение 1 лица ـني или ـي может принимать форму ـنيا، ـيا или же последний ي может опускаться: أَكْرَمَني < أَكْرَمَنْ .
7. Долгие ـو и ـي могут заменяться на кясру и даму: صَنَعُ < صَنَعوا , يَدِ < يَدي .

Язык персидской поэзии

Желая уложиться в метрическую парадигму размера, Фирдоуси использует все узаконенные в классической просодии приемы: чередование полных и «облегченных» форм слова (мах, шах, пай, руй — мах, шах, па, ру), удвоение согласных (зар — зарр, типе — таппе), стяжение слогов (бегоза-р — богзар) и т. п.

Фирдоуси А. Шахнаме : в 6 т. / А. Фирдоуси ; Академия наук СССР, Отделение литературы и языка ; пер. [с фарси] Ц. Б. Бану-Лахути, В. Г. Берзнева ; редкол.: В. П. Волгин (пред.) и др. — Москва : Издательство АН СССР, 1957—1989. — (Литературные памятники). — С. 552.

Специфичное для тохарских метрических текстов

1. Девокализация конечного слога
2. Стяжение гласных
3. Опущение сильного ä
4. Опущение небеглого u:
    a. абстрактные существительные с суффиксом -une: āpāṣtuneyäṣ (YQ-32 b3) MV к āpāṣtuneyäṣ* Abl от āpāṣtune* абстр. сущ. от āpāṣät* прив. от pās- ʻохранять, соблюдатьʼ

Язык древнеирландской поэзии

1. Сохранение монофтонгов ō и ē (кл. др.-ирл. úa, ía): ōit ‘от тебя', ōim 'от меня'; переходная стадия ōa в tōatha; rōdaib (в тексте VII в. — ZGP, 1932, Bd 19, S. 198); fēda.
2. 2. Различение превербов to- и di-, которые сливаются в кл. др.-ирл. do-; t- в предударном слоге (кл. др.-ирл. d-).
3. -nd (кл. др.-ирл. -nn): артикль inda > inna.
4. Конечные слабые -l и -n в безударном слоге > сильные -ll, -nn (середина VII в.).
5. Сохранение -th, -ch в безударном слоге: hingnath.
6. Безударное -e- (кл. др.-ирл. -a-): ate-midiur, toceth.
7. Завершение апокопы на рубеже V—VI вв.: Gathbotha с сохранением конечного -a в Ульстерских анналах под 456 г.
8. Завершение синкопы к началу VII в.: Bresual, кл, др.-ирл. Bresal.
9. Союз с ‘и': sceo, -ch.
10. Беспредложный датив: dligther cech flathemoin.
11. Конечное положение глагола (тмесис и закон Бергина): ro- Columb o Chille -cúalammar; macne Ailb airmi.
12. Частое употребление суффигированных местоимений: tá-thum, glinnsius.

Закон аббревиации Ципфа

Что такое закон аббревиации Ципфа

— Закон аббревиации Ципфа: чем чаще слово — тем оно короче

— Не путать с классическим законом Ципфа о рангах (частота $\propto$ 1/ранг)

— Формулировка через корреляцию: ρ (частота × длина) < 0

— Чем ближе ρ к −1 — тем строже закон

Слог и буква как меры длины слова

— Две меры длины: буквы (орфографическая) и слоги (просодическая)

— Априорная гипотеза: для поэзии слоги «лучше» — просодия

— Буквы: число алфавитных символов

— Слоги: число гласных

— Проверяем обе меры — какая сильнее «ловит» закон в каждом подкорпусе

Поэтический язык

Данные: поэзия

— Поэтический подкорпус НКРЯ

— 15 469 127 словоупотреблений

— 25 295 групп «автор × год»

— 1 295 авторов; 569 091 словоформа (типов)

— Диапазон: 1705–2021 (297 лет с данными)

Поэзия: закон выполняется устойчиво

— 87% групп: $\rho < 0$ и $p < 0.05$

— Медиана ρ_letters по группам: −0.36

— Медиана ρ_sylls по группам: −0.34

— Общий пул: ρ_letters = −0.292, ρ_sylls = −0.273 ($p \approx 0$)

— 0 авторов из 1 035 с $\rho > 0$ — нарушителей нет

— Вывод: закон аббревиации — универсальное свойство русских языков

Поэзия: буквы vs слоги

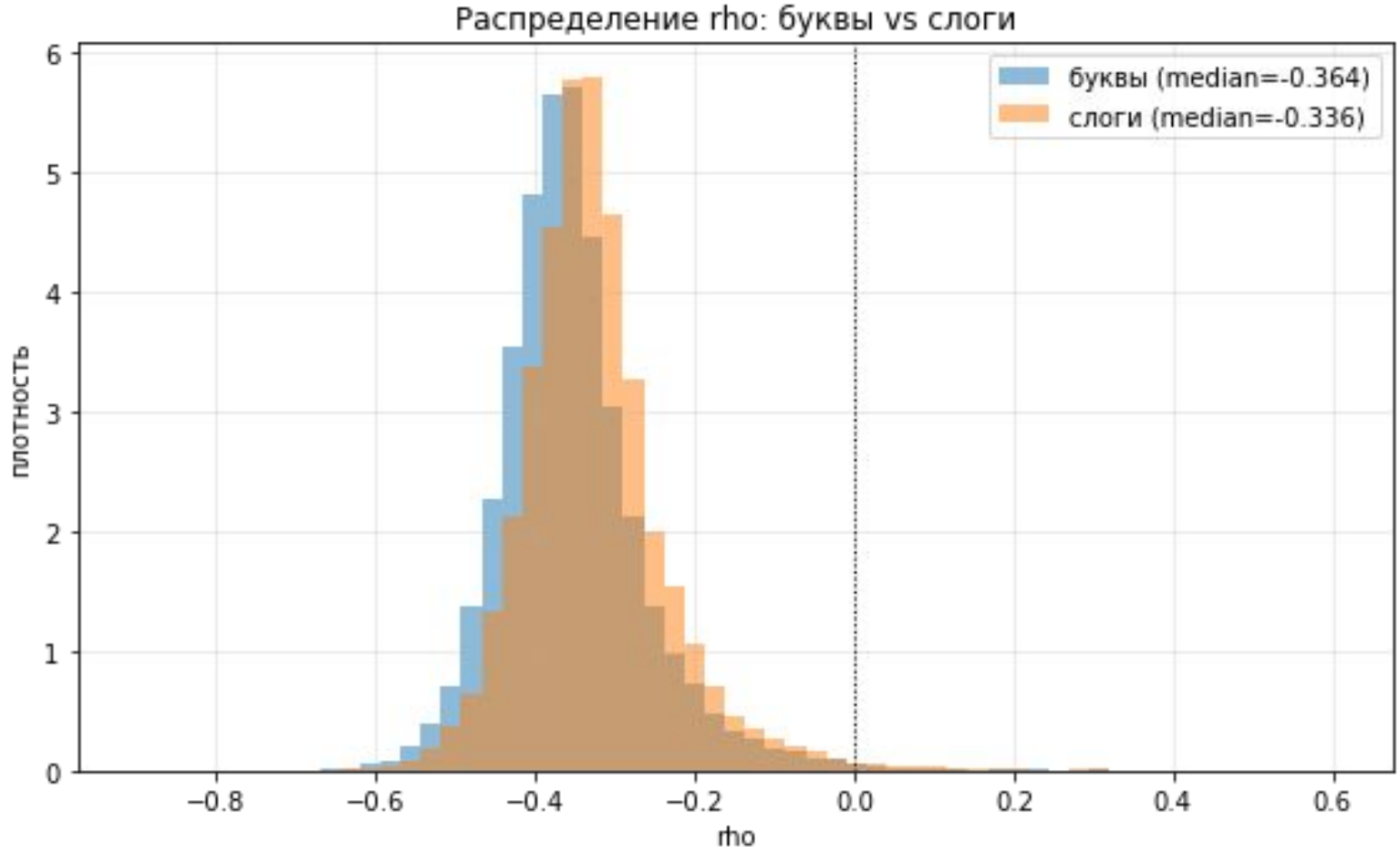
— Буквенная мера сильнее слоговой в 79% групп

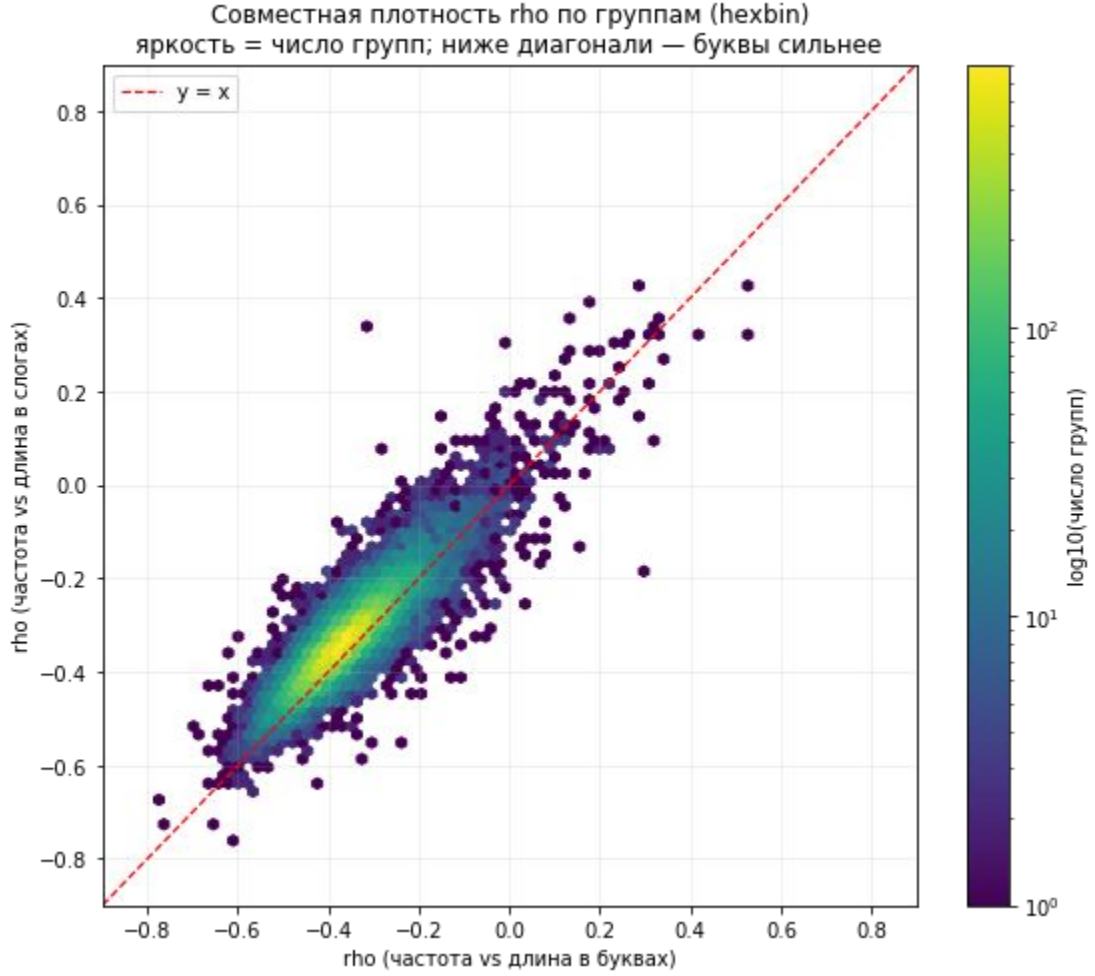
— Медиана разницы $|\rho_letters| - |\rho_sylls| = +0.028$

— Устойчиво во времени: во всех 29 декадах 1730–2010

— $\rho_letters \times \rho_sylls$: 79% точек выше диагонали $y = x$: т.к. $\rho < 0$, «выше» = ближе к нулю = слоги слабее

— Гипотеза «слоги лучше для поэзии» отклонена


Распределение rho: буквы vs слоги
буквы (median=-0.364)
слоги (median=-0.336)
плотность
rho
0
1
2
3
4
5
6
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


Совместная плотность rho по группам (hexbin)
яркость = число групп; ниже диагонали — буквы сильнее
y = x
rho (частота vs длина в слогах)
rho (частота vs длина в буквах)
log10(число групп)
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
10^2
10^1
10^0

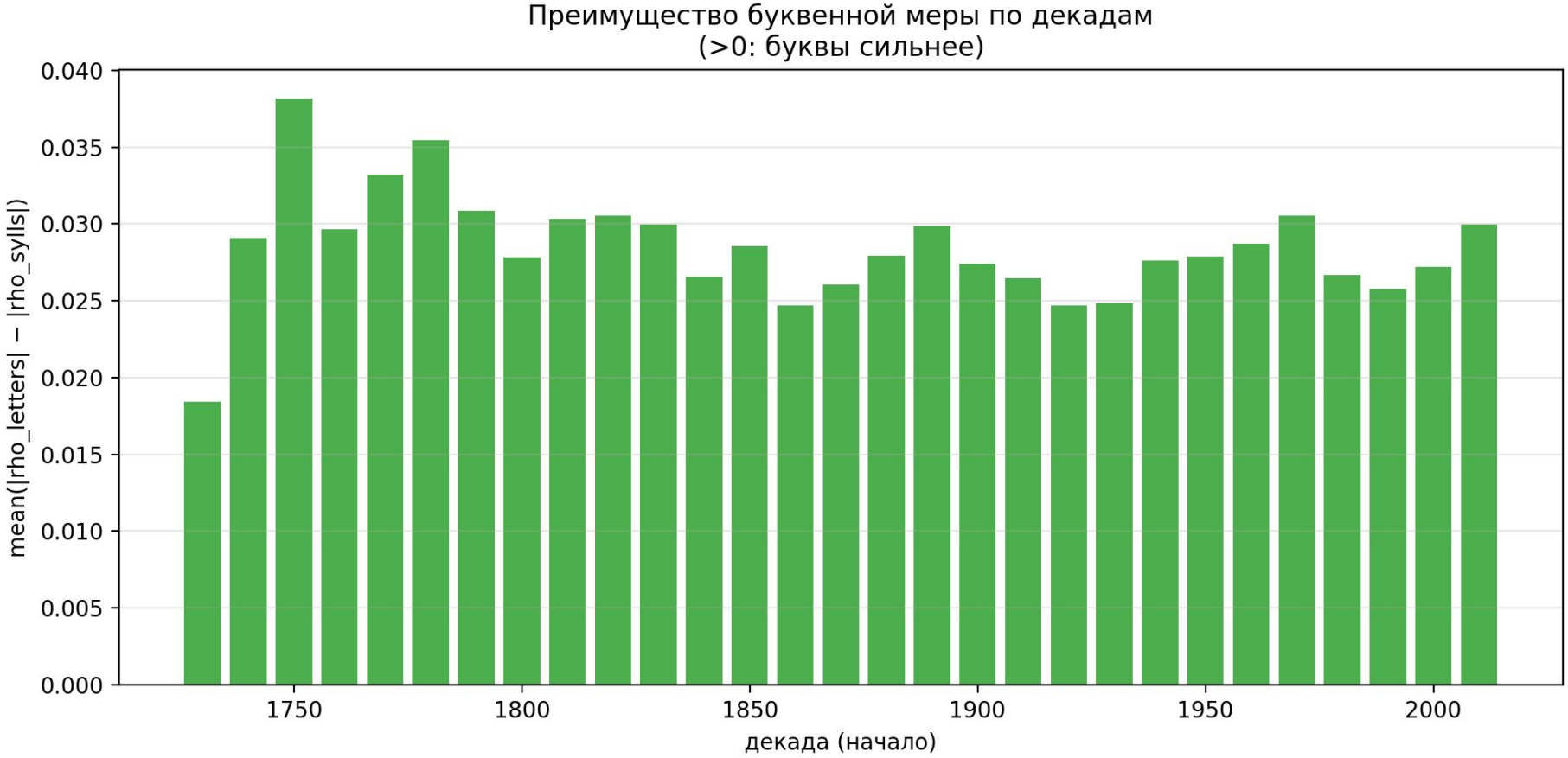
Поэзия: динамика по декадам

— Преимущество букв над слогами — константа ~0.02–0.04

— Все 29 декад 1730–2010: значение > 0, тренда нет

— Закон аббревиации — устойчивое свойство, а не мода эпохи

— Но сила закона по годам — варьирует


Преимущество буквенной меры по декадам
(>0: буквы сильнее)
mean(|rho_letters| − |rho_sylls|)
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
1750
1800
1850
1900
1950
2000
декада (начало)

«Прозаический» язык

Данные: проза

— Основной подкорпус НКРЯ

— 83 480 926 словоупотреблений (в 5,4 раза больше поэзии)

— 1 331 505 словоформа (типов)

— 138 531 XML-файл; все совпали с метаданными

— Диапазон: 1800–1899 (100 лет — XIX век)

Проза: закон выполняется

— 100 лет: $\rho < 0$ и $p < 0.05$

— Общий пул: $\rho_{letters} = -0.168$, $\rho_{sylls} = -0.114$

— 0 лет с $\rho > 0$ — нарушителей нет

— Закон аббревиации выполняется, но слабее, чем в поэзии

— Буквы сильнее слогов

Поэзия vs проза: общий пул

— Поэзия: ρ_letters = −0,292, ρ_sylls = −0,273

— Проза: ρ_letters = −0,168, ρ_sylls = −0,114

— Разрыв по буквам: 0,124; по слогам: 0,158

— В поэзии закон в 1,7–2,4 раза сильнее

— Но: разные эпохи (поэзия 1705–2021, проза 1800–1899)

— Поэтому нужно честное сравнение на общем окне

Честное сравнение: окно 1800–1899

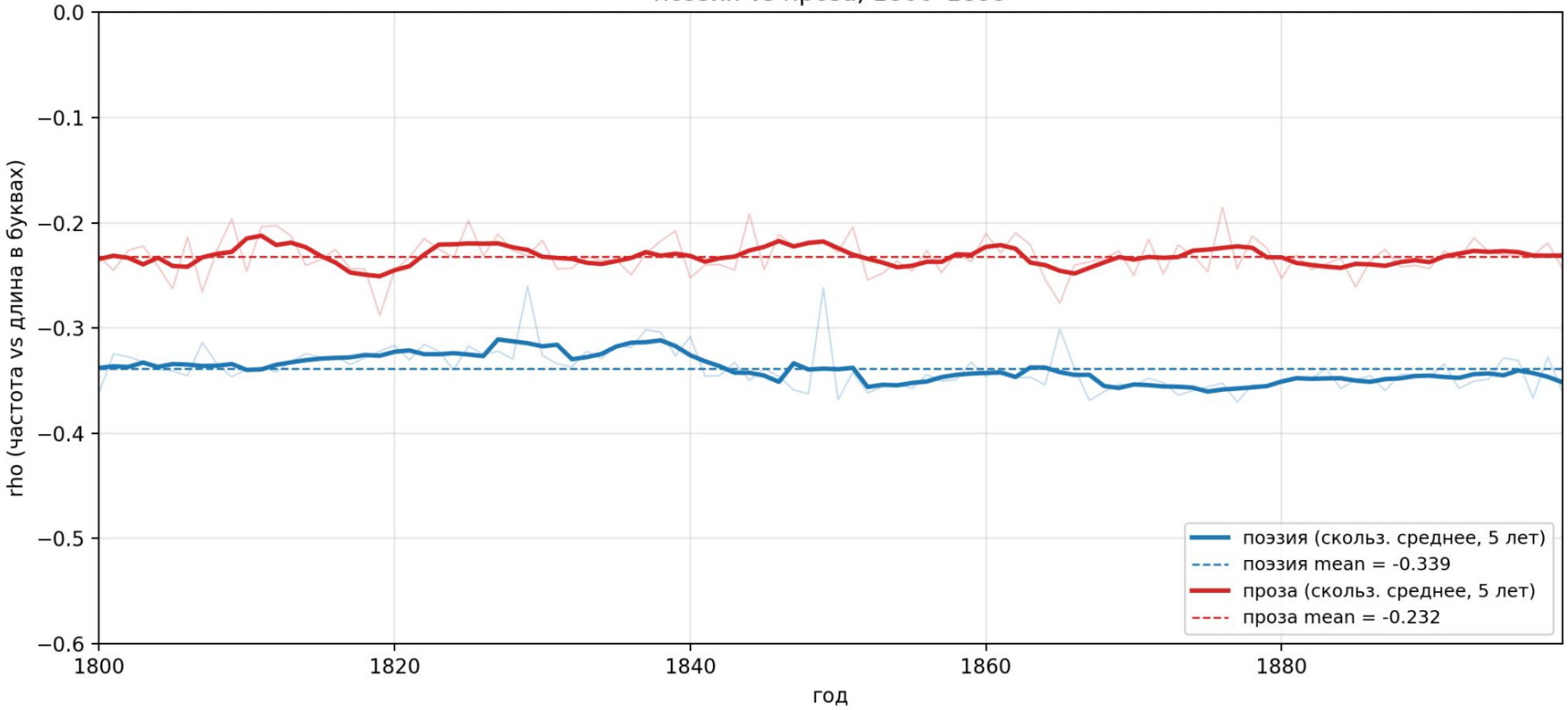
— 100 лет у обоих подкорпусов

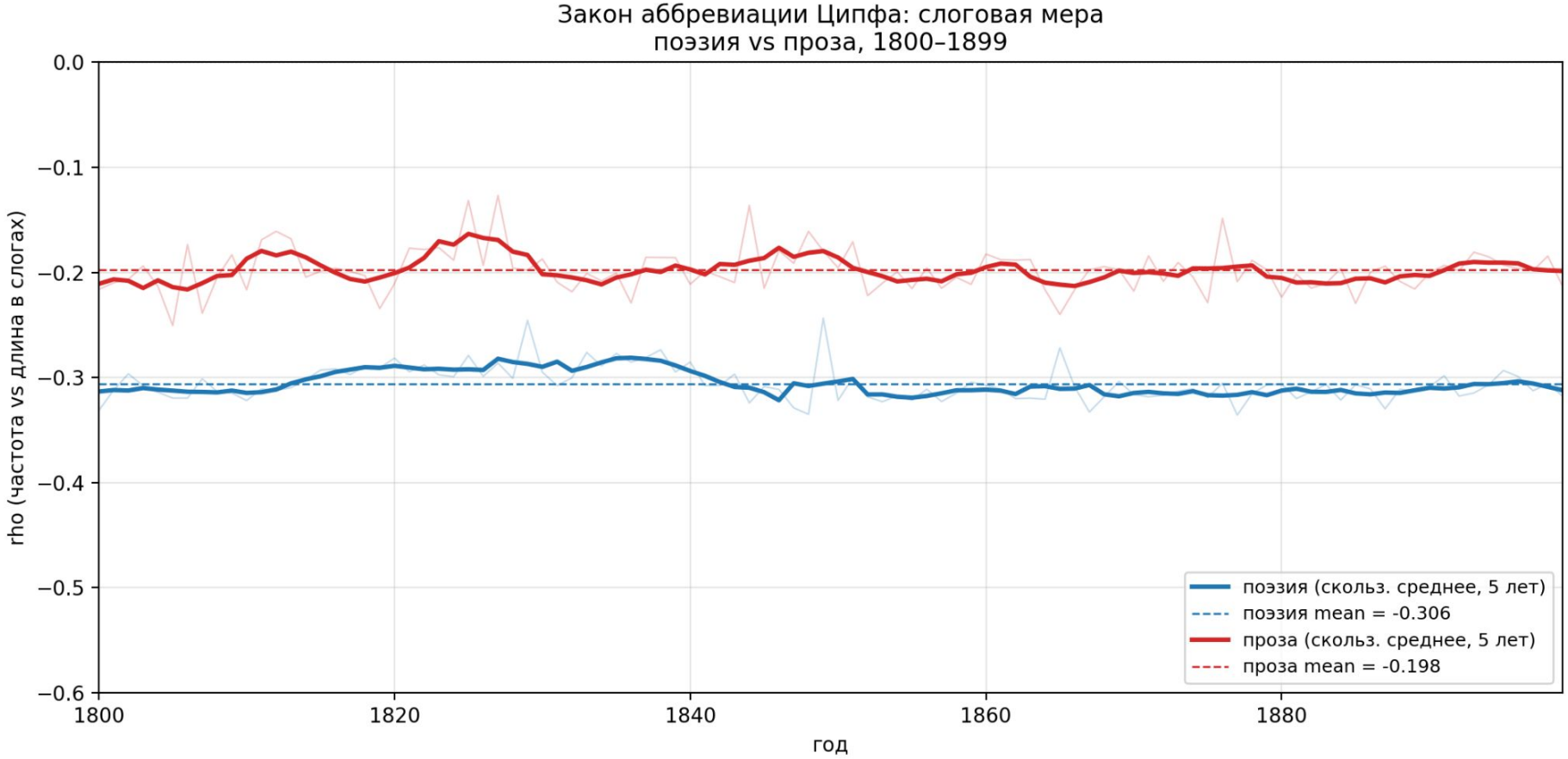
— Поэзия: mean ρ_letters = −0,339, ρ_sylls = −0,306

— Проза: mean ρ_letters = −0,232, ρ_sylls = −0,198

— Разрыв: 0,107 (буквы) и 0,108 (слоги) — почти одинаков

— Значит: различие поэзия/проза не зависит от меры длины


Закон аббревиации Ципфа: буквенная мера
поэзия vs проза, 1800–1899
rho (частота vs длина в буквах)
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
1800
1820
1840
1860
1880
год
поэзия (скольз. среднее, 5 лет)
поэзия mean = -0.339
проза (скольз. среднее, 5 лет)
проза mean = -0.232


Закон аббревиации Ципфа: слоговая мера
поэзия vs проза, 1800–1899
rho (частота vs длина в слогах)
год
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
1800
1820
1840
1860
1880
поэзия (скольз. среднее, 5 лет)
поэзия mean = -0.306
проза (скольз. среднее, 5 лет)
проза mean = -0.198

Почему в поэзии сильнее?

— Частые короткие слова (предлоги, союзы, местоимения) необходимы для метра

— Лексический состав: в поэзии выше доля служебных и местоименных слов

— Эллипсис: поэзия допускает пропуск глаголов → больше коротких частых слов

Буквы vs слоги: итог по двум подкорпусам

— Поэзия: |ρ_letters| > |ρ_sylls| в 98% лет

— Проза: |ρ_letters| > |ρ_sylls| в 100% лет

— Буквенная мера систематически сильнее — и в поэзии, и в прозе

— Значит: это свойство графики, а не поэзии

Отклонения

Как и где закон аббревиации выполняется хуже

— 1 035 авторов с n_types ≥ 200 — 0 нарушителей (ρ > 0)

— «Худшие» — в основном малые выборки (n_types 200–400): шум

— Но есть крупные с реальным ослаблением:

— Гнедич: 26 196 типов, ρ = −0.248

— Бахтерев: 5 197, ρ = −0.240

— Садофьев: 2 661, ρ = −0.247

— У 6 авторов ослабление устойчиво по обеим мерам (Рок, Логинов, Колоколов, Кудрейко, Бахтерев, Дон-Аминадо)

— Гипотеза: стилистическая особенность, а не шум

— 297 лет, 0 нарушителей

— «Худшие» годы — крупные (не шум):

— 1829 (34 707 типов, ρ = −0.260), 1849 (22 797, −0.262)

— 1837–1840 (ρ ~ −0.30), 1991 (13 529, −0.300), 2006 (−0.308)

— «Лучшие» годы — XVIII век:

— 1751, 1754, 1758, 1746, 1785 (ρ ~ −0.37…−0.40)

— Ослабление в «худшие» годы устойчиво по буквам и слогам

Поэзия: динамика по эпохам

— XVIII век (классицизм): закон сильный, |ρ| ~ 0.37–0.40

— 1820–1840-е (романтизм/реализм): ослабление, |ρ| ~ 0.26–0.31

— Конец XIX века: снова сильнее (1877: −0.371)

— 1990–2000-е: новое ослабление (−0.28…−0.31)

— Интерпретация: чем разнообразнее и «прозаичнее» лексика эпохи, тем слабее экономия

Проза XIX века: динамика плоская

— 100 лет, 0 нарушителей

— Средний ρ_letters по декадам: −0.225 … −0.241 (разброс 0.016)

— Средний ρ_sylls: −0.178 … −0.210 (разброс 0.032)

— Тренда нет — закон аббревиации стабилен весь XIX век

— «Худшие» годы — крупные: 1876 (92 392 типа, −0.185), 1844 (−0.192), 1809 (−0.196)

— Проза не реагирует на литературные течения, в отличие от поэзии

Итоги

— Закон аббревиации универсален: нет нарушителей ни в одном срезе

— В поэзии сильнее, чем в прозе: разрыв ~0.107 на честном окне 1800–1899

— Буквы сильнее слогов и в поэзии (98% лет), и в прозе (100% лет) — свойство русского языка

— Поэзия «дышит» по эпохам, проза — нет: U-образная vs плоская динамика