

«География кончается вчера»

частотность топонимов в поэтическом корпусе

Борис Орехов (НИУ ВШЭ, ИРЛИ РАН, МГУ)

Корпус: объём и структура

- Материал: поэтический корпус русской поэзии XVIII–XXI вв.
- Текстов (XML-файлов) $\approx 140\,000$
- Общий объём $\approx 12,7$ млн словоформ
- Уникальных словоформ $\approx 1,1$ млн
- Хронологический охват 1720-е – 2020-е (31 десятилетие)
- Авторы $\approx 1\,000+$

Метод: пайплайн обработки

[1] Обход XML-корпуса



[2] Очистка текста от разметки и грависов



[3] NER-извлечение LOC-сущностей (Natasha / Slovnnet)



[4] Канонизация форм (pymorphy2 + унификация)



[5] Фильтрация через Wikipedia API

(оставляем только статьи с координатами)



[6] Агрегация по десятилетиям + визуализация

Технологический стек

- Парсинг XML
- Лемматизация
- Фильтрация
- Хранение
- Визуализация
- Статистика
- re, pathlib
- NERnatasha (NewsNERTagger), slovnet
pymorphy2, morph_vocab Natasha
- MediaWiki API (ru.wikipedia.org/w/api.php),
requests
- JSON, CSV, pandas
- matplotlib, contextily, folium, networkx
- scipy.stats, statsmodels

Ключевые методологические решения

- Одна сущность — один документ. Внутри одного текста сущность считается один раз, даже если упоминается многократно. Это устраняет эффект «длинных текстов».
- Канонизация форм. «Москвы», «Москве», «Москвой» сводятся к лемме москва через `ru morphology2` и унификацию дефисов.
- Фильтр через Wikipedia. Сущность оставляется только если у статьи есть координаты (P625 в Wikidata или `coordinates` в Wikipedia API). Это отсеивает ложные срабатывания NER (от, из, невинность).
- IPv4-only патч. В локальном окружении IPv6 до Wikipedia не работает — `urllib3.cn.allowed_gai_family = lambda: socket.AF_INET` снижает время запроса с 10 с до 0.3 с.
- Порог совместной встречаемости. В графе связей ребро проводится, если две сущности встретились вместе как минимум в 5 текстах — это отсекает случайные пары.

Результаты извлечения: что нашли

- Результаты: 23 929 LOC-сущностей → 4 127 с координатами
- Всего LOC-спанов от Natasha $\approx 24\,000$ Включая предлоги, абстракции, ошибки
- Уникальных после канонизации $\approx 23\,929$ Слитые формы
- После фильтра Wikipedia 4 127 У статьи есть координаты
- После объединения по координатам $\approx 3\,900$ «Петербург» = «Санкт-Петербург»

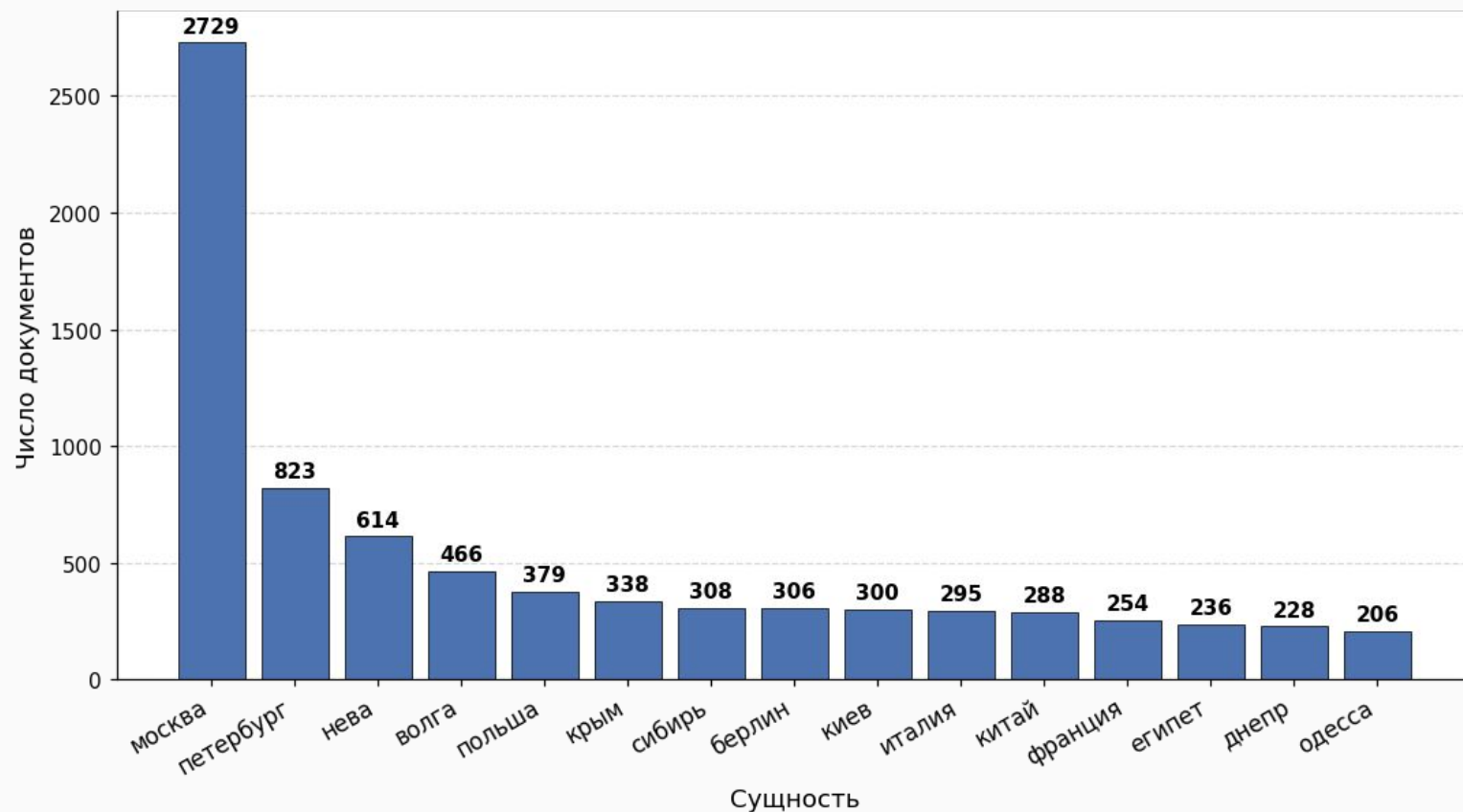
Что отсеялось и почему

not_found (нет статьи) ~25% от, из, невинность, любовь
— ложные срабатывания NER

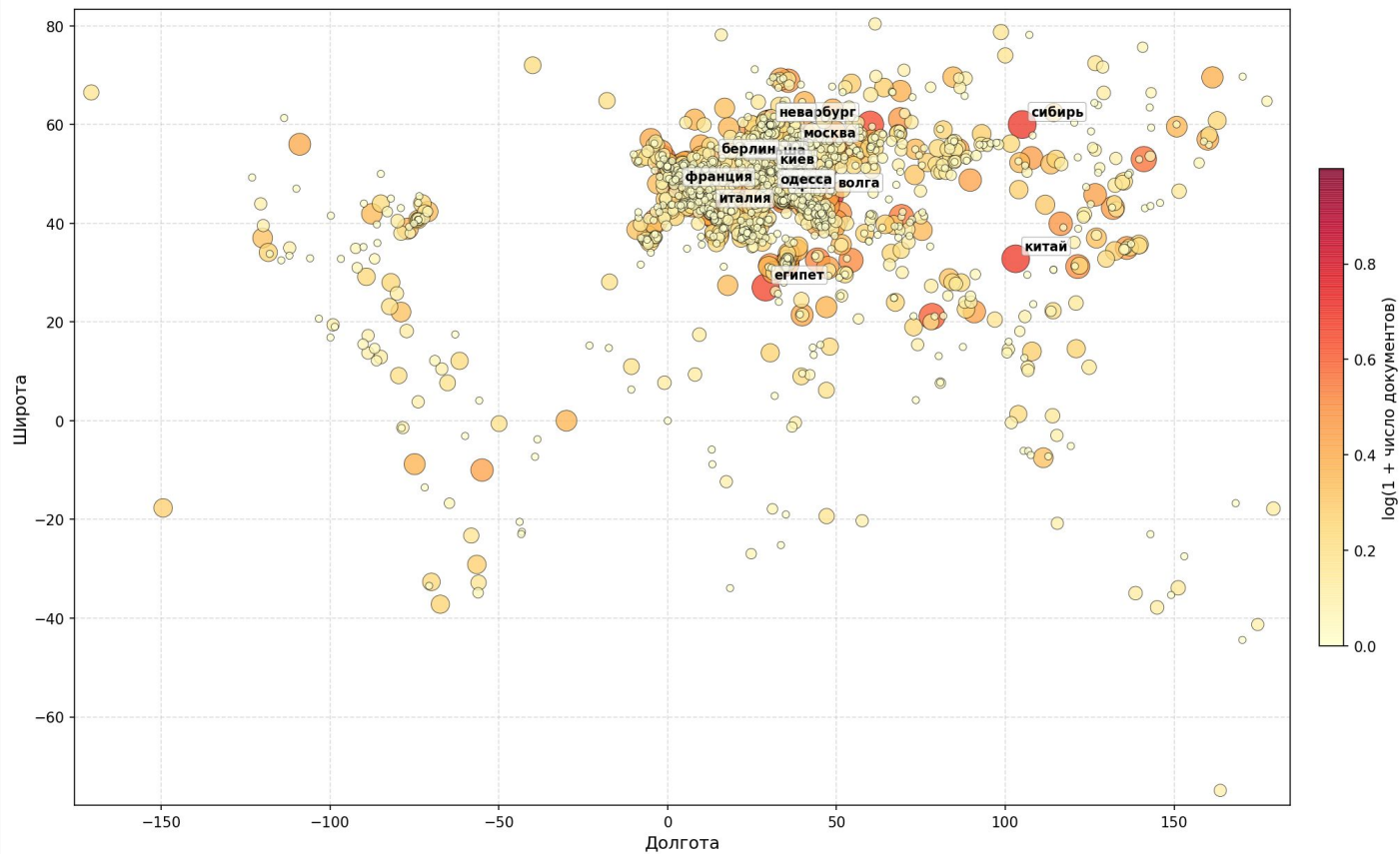
no_coords (статья без координат) ~5% абстрактные
понятия, персоны

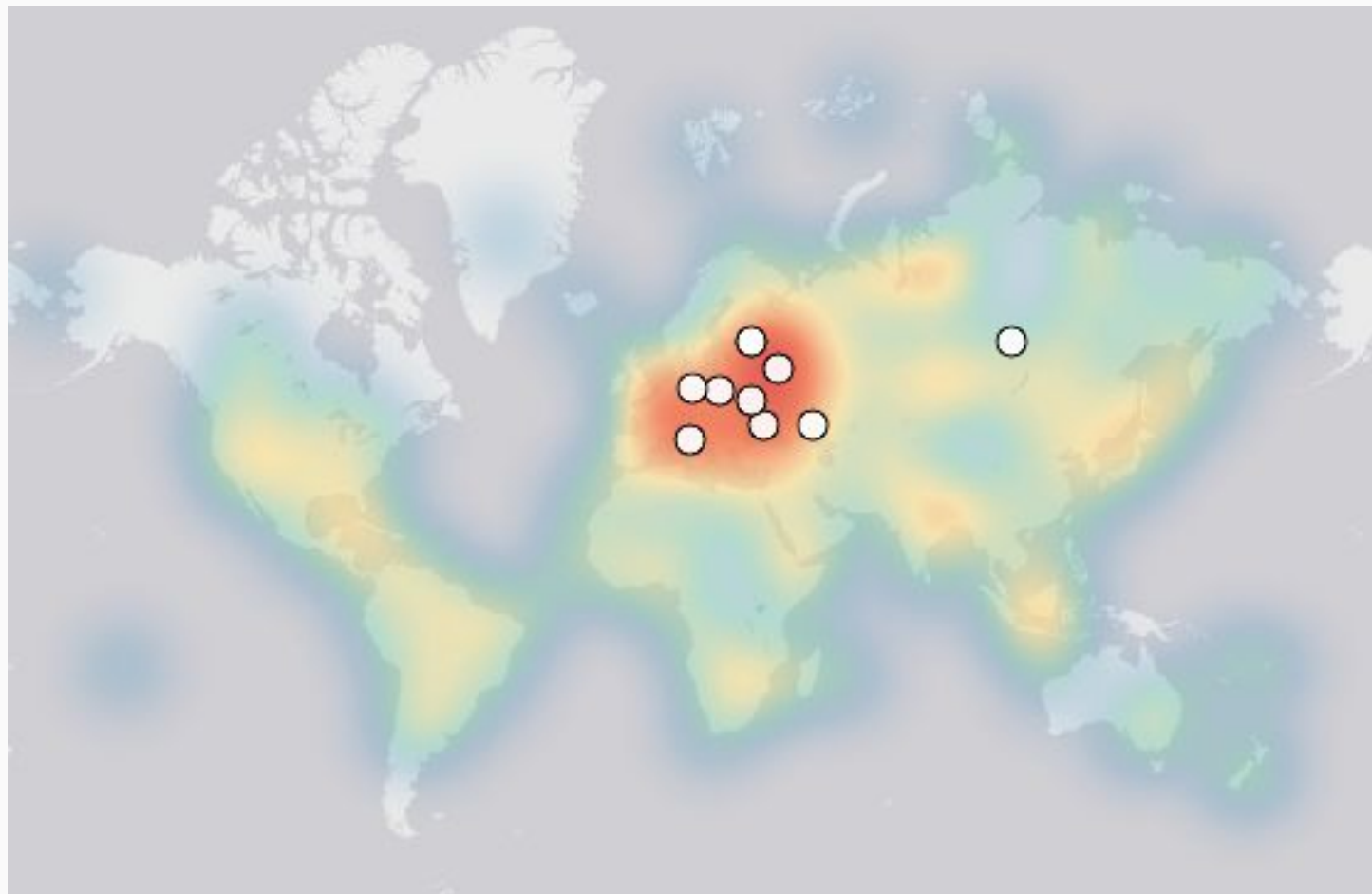
ok (с координатами) ~17% реальные топонимы

Топ-15 географических сущностей по числу документов

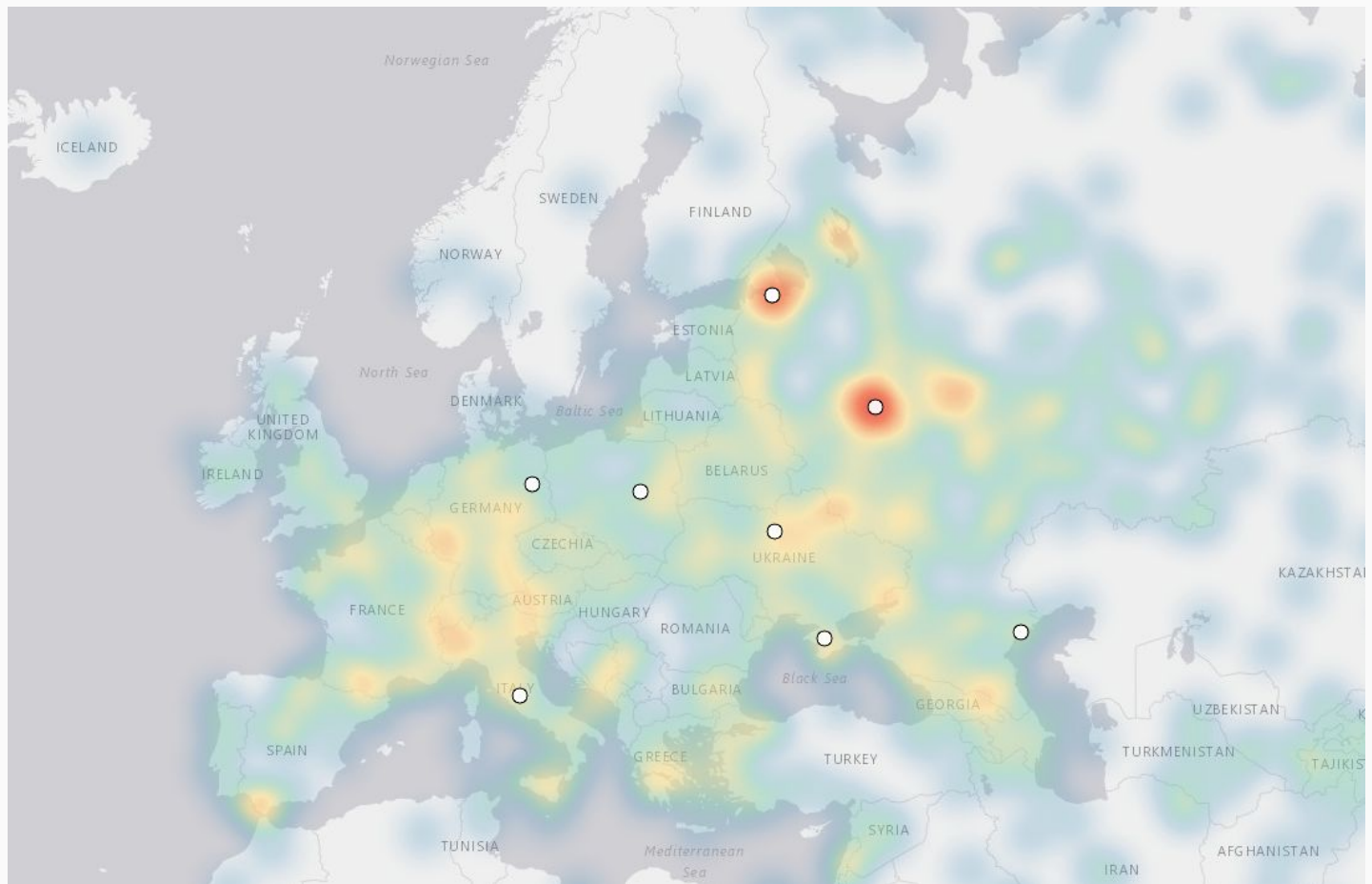


Плотность географического распределения ЛОС-сущностей
(размер и цвет точки $\propto$ числу документов, логарифмическая шкала)





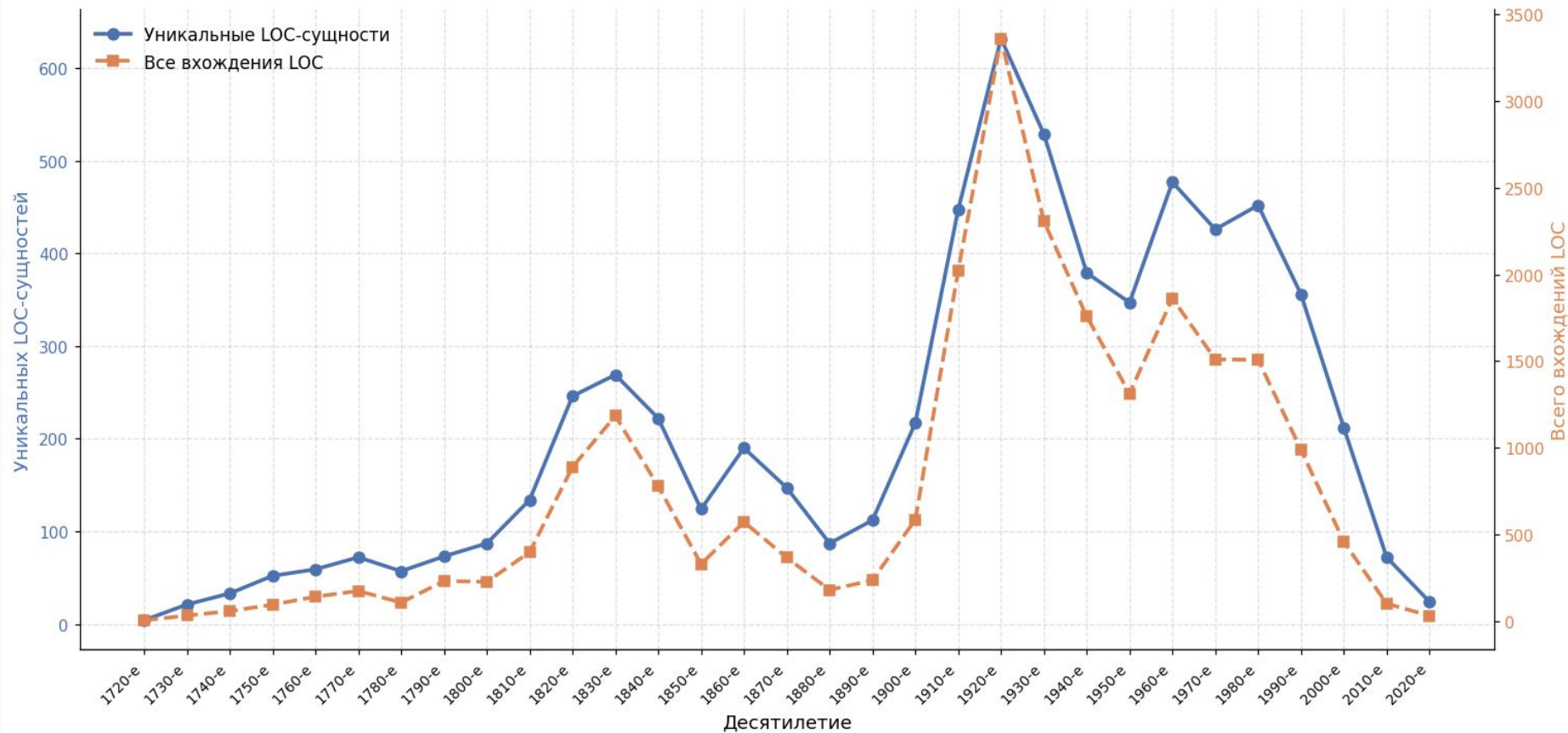
Данные по всему корпусу

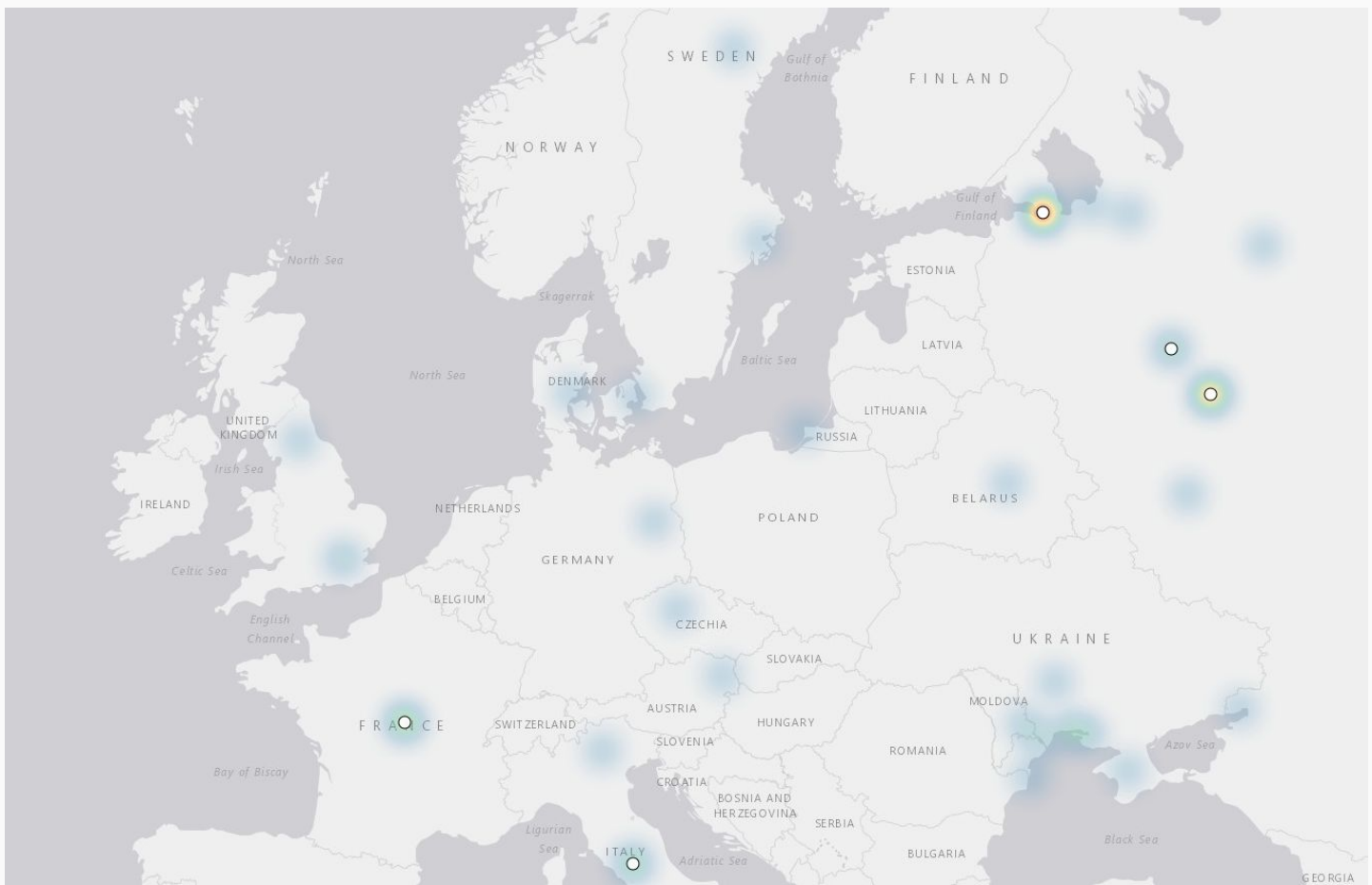


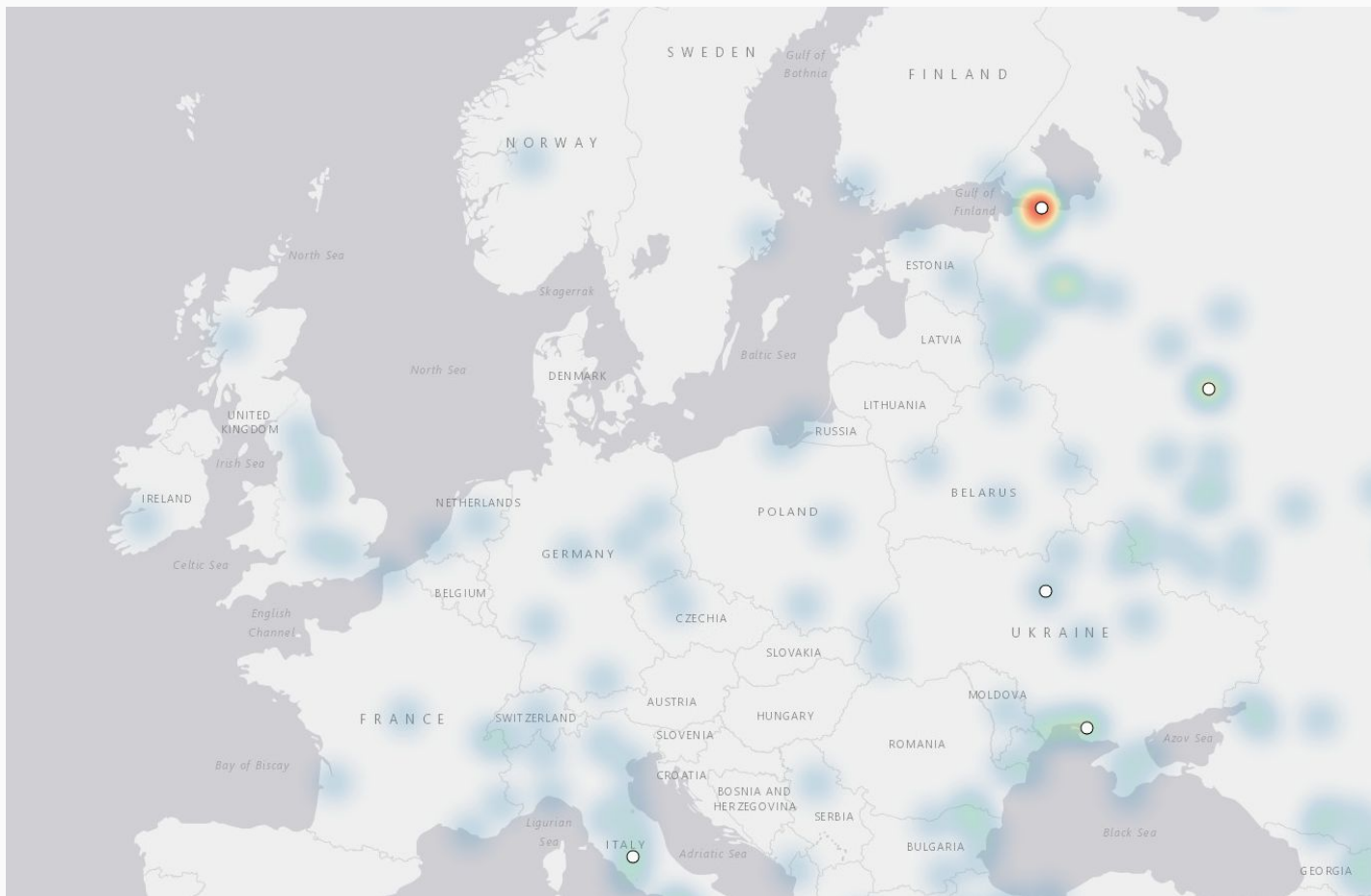
Данные по всему корпусу

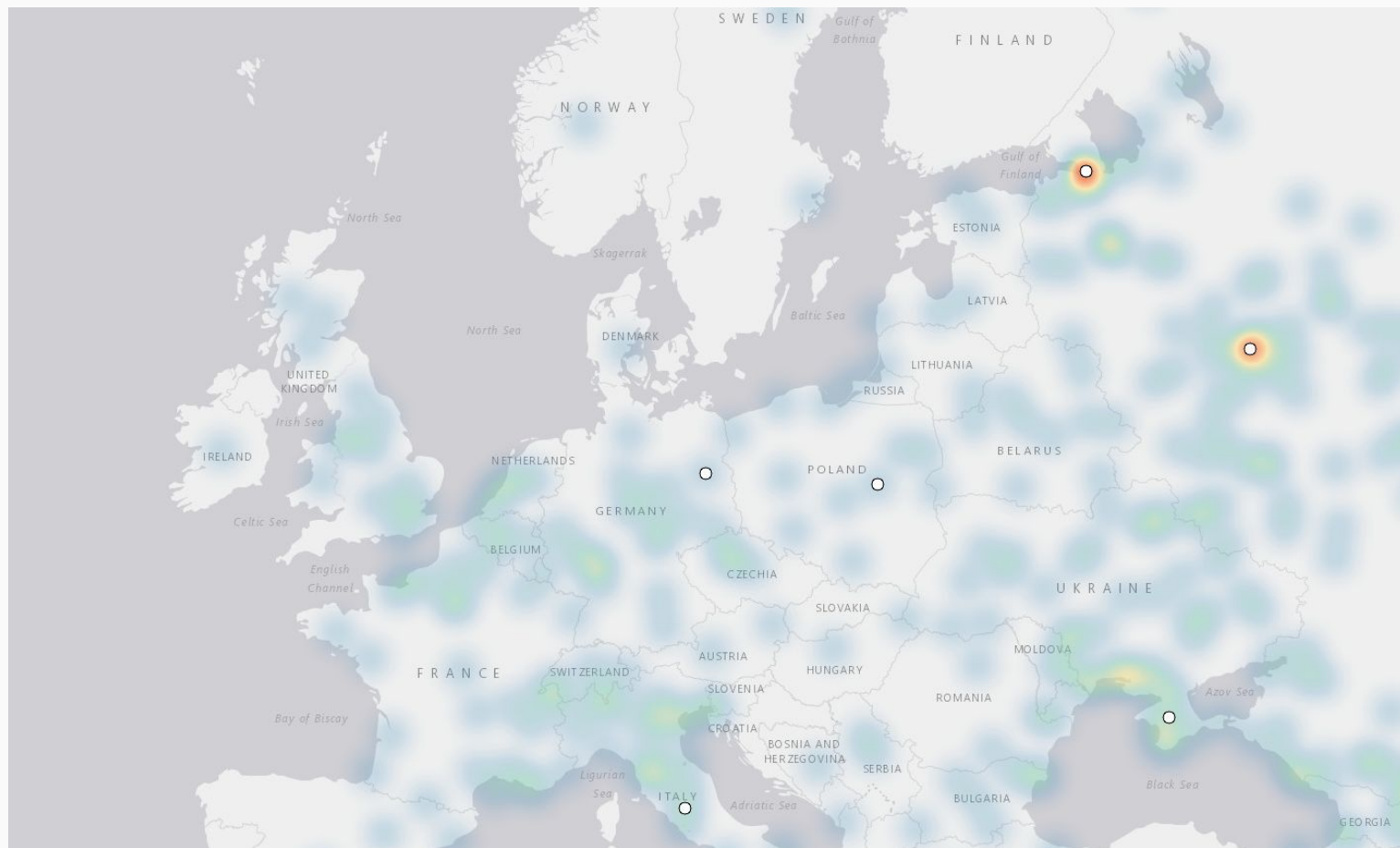
Временная динамика

Динамика LOC-сущностей по десятилетиям





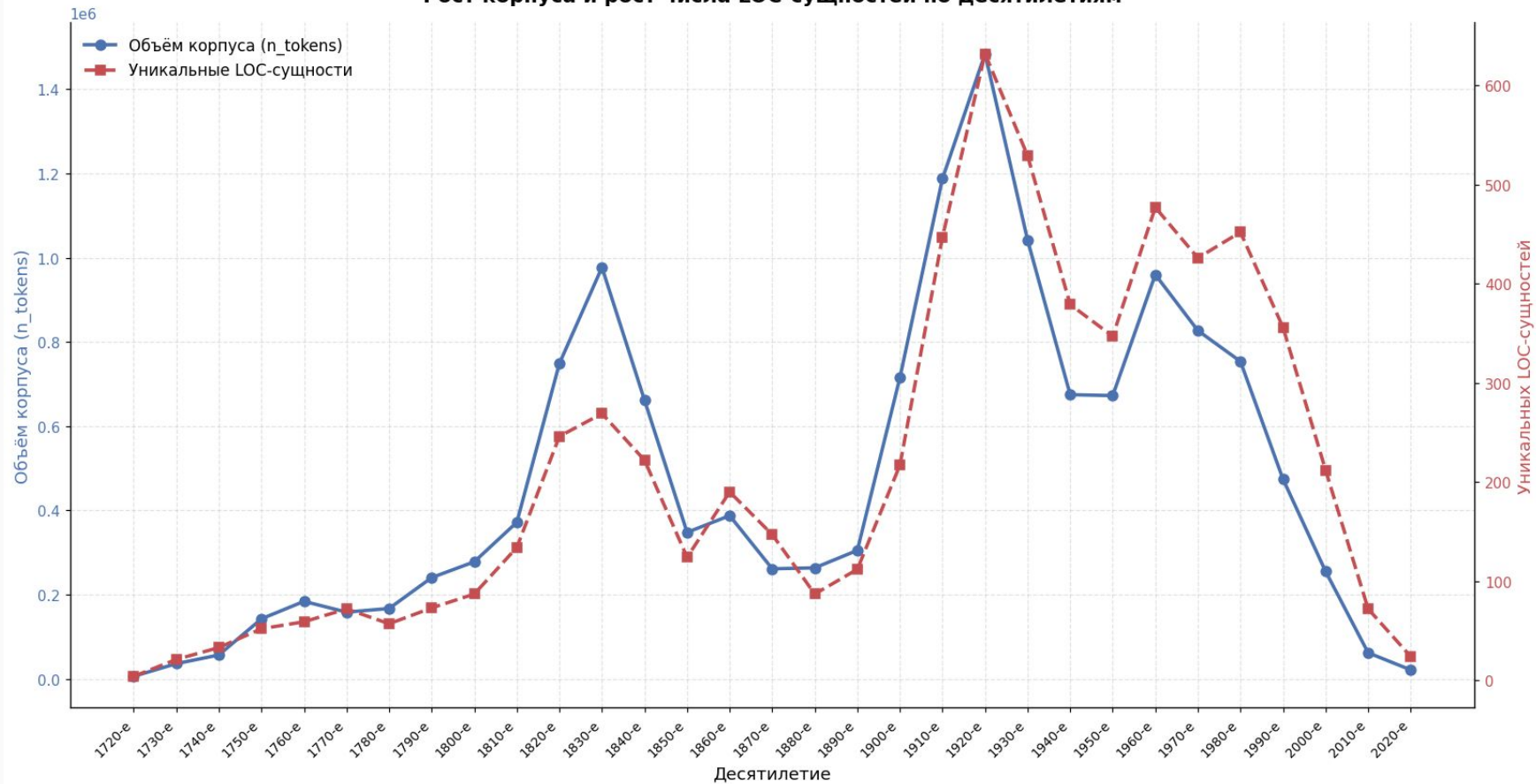




Число LOC растёт потому, что растёт корпус?

- Наивное наблюдение: чем позже десятилетие, тем больше уникальных LOC-сущностей.
- 1720-е: 4 сущности → 1920-е: 632 сущности. Рост в 158 раз.
- Но за тот же период корпус вырос с 6 060 до 1 484 117 словоформ — в 245 раз.
- Вопрос: это реальный рост географического разнообразия поэзии или просто механический эффект объёма?

Рост корпуса и рост числа LOC-сущностей по десятилетиям



Ищем зависимости

Данные и метрики

Данные:

- 31 десятилетие — от 1720-х до 2020-х.
- 12 метрик в каждой строке.
- ~12,7 млн словоформ в общей сложности.

Метрики:

- Объём: число документов, число словоформ, число уникальных словоформ.
- Лексическое разнообразие: TTR, MSTTR, Herdan's C, Guiraud's R, энтропия, hapax ratio.
- LOC: уникальные сущности, вхождения.
- Производные: LOC на 1000 токенов, LOC на документ.

ЛОС и объём корпуса растут синхронно

- Пик обеих кривых — 1920-е: 1 484 117 токенов, 632 сущности.
- Провал — 1880–1890-е: 263 563 и 305 099 токенов, 87 и 112 сущностей.
- Минимум начала периода — 1720-е: 6 060 токенов, 4 сущности.
- синхронность настолько высока, что возникает подозрение — одно объясняет другое.

Прямая связь: LOC vs объём корпуса

$R^2 = 0.859$ — объём корпуса объясняет 85.9% вариации числа LOC.

Корреляции:

- Pearson: +0.927 ($p = 7.1e-14$).
- Spearman: +0.953 ($p = 1.5e-16$).

Связь линейная и
сверхсильная

Объём —
доминирующий фактор

А что с
лексическим
разнообразием?

Богатство языка

- Объём корпуса объясняет 85.9% вариации LOC.
- Остаётся 14.1% — что их объясняет?
- Гипотеза: лексическое разнообразие эпохи.
- Логика: богатый язык → больше слов → больше топонимов.

Guiraud's R

- Что считает: число уникальных слов, делённое на квадратный корень из общего числа слов.
- Сколько разных слов на единицу длины текста, с поправкой на то, что длинные тексты naturally дают больше повторов.
- Что любит: большие тексты, в которых много новых слов.
- Что не любит: повторы — если одно слово встречается часто, R падает.
- Аналогия: количество разных блюд на банкете, поделённое на корень из числа гостей. Чем больше гостей, тем сложнее удивить — но если удивляешь, то по-настоящему.

Энтропия (Shannon entropy)

- Насколько предсказуемо следующее слово? Если все слова встречаются примерно одинаково часто — энтропия высокая. Если одно слово занимает 90% текста, а остальные по разу — низкая.
- Что любит: равномерность. Текст, где 1000 слов встречаются по 1 разу, энтропийно богаче, чем текст, где 1 слово встречается 1000 раз.
- Что не любит: доминирование одного слова.
- Аналогия: оркестр. Если все инструменты играют примерно одинаково — звук богатый (высокая энтропия). Если один барабан заглушает всех — бедный (низкая).

MSTTR (Mean Segmental TTR)

- Что считает: среднее отношение уникальных слов к общим по сегментам фиксированной длины (обычно 1000 слов).
- Простыми словами: «Разрезаем текст на куски по 1000 слов и в каждом считаем долю новых слов. Потом усредняем».
- Что любит: стабильность. Если текст равномерно богат — MSTTR высокий. Если богатство сосредоточено в начале, а дальше повторы — низкий.
- Что не любит: резкие перепады.
- Аналогия: проверка качества воды в разных точках реки. Если везде чисто — средний балл высокий. Если чисто только у истока — низкий.

Аналогия

Четыре способа оценить богатство леса:

- Guiraud's R — сколько разных видов деревьев на квадратный корень из площади.
- Энтропия — насколько равномерно распределены виды (или один вид занимает 90% леса).
- MSTTR — сколько видов в среднем на каждом гектаре.
- Herdan's C — как быстро появляются новые виды по мере углубления в лес.

Лексические метрики vs LOC: сильные связи

Guiraud's R:

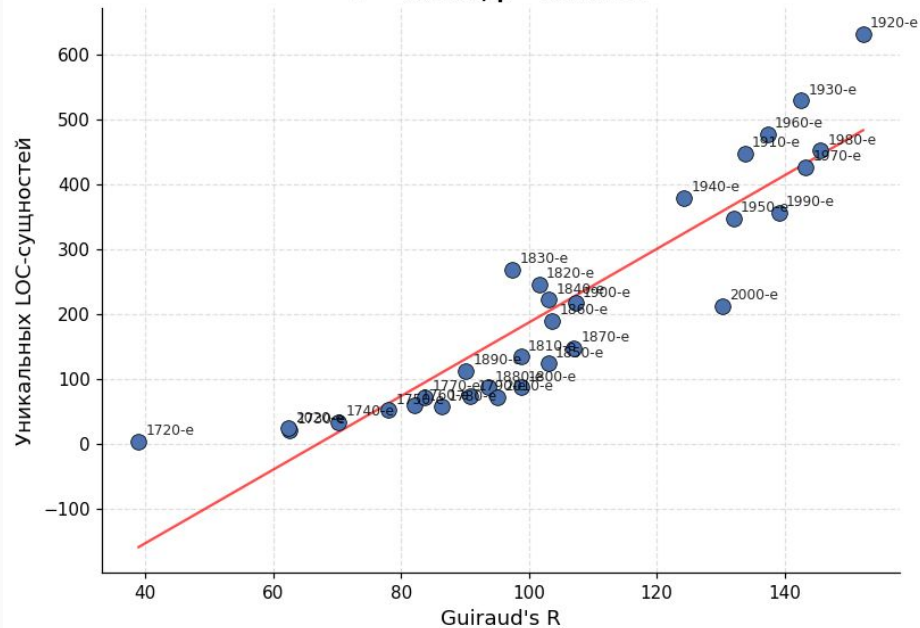
- Pearson: +0.910 ($p = 1.4e-12$).
- Spearman: +0.946 ($p = 1.0e-15$).
- Самая высокая линейная корреляция среди всех лексических метрик.

Энтропия:

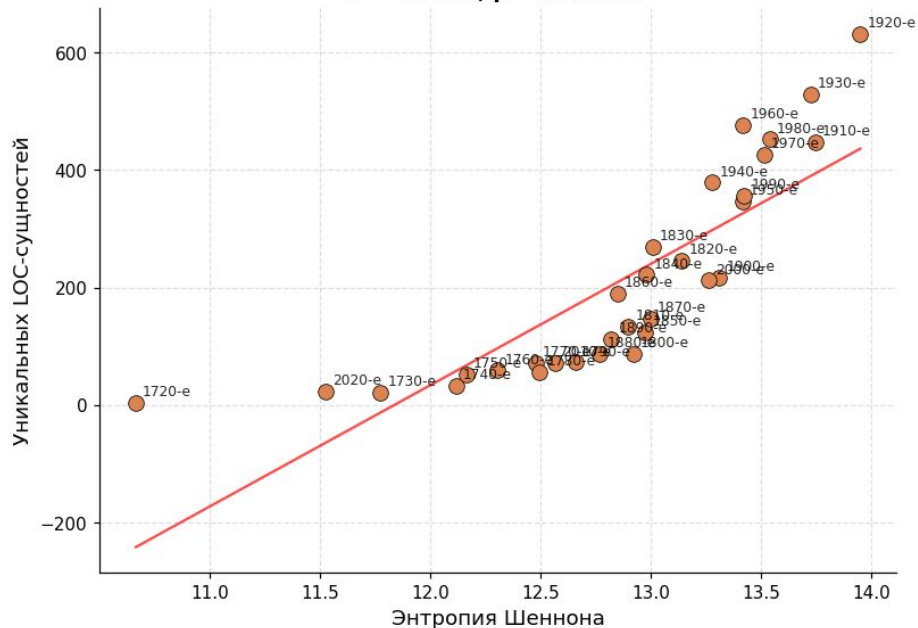
- Pearson: +0.829 ($p = 8.6e-09$).
- Spearman: +0.973 ($p = 4.7e-20$).
- Самая высокая ранговая корреляция — порядок десятилетий почти идеально совпадает.

Два сильнейших кандидата: Guiraud's R и энтропия

Guiraud's R vs LOC
 $r = +0.909$, $p = 1.42e-12$

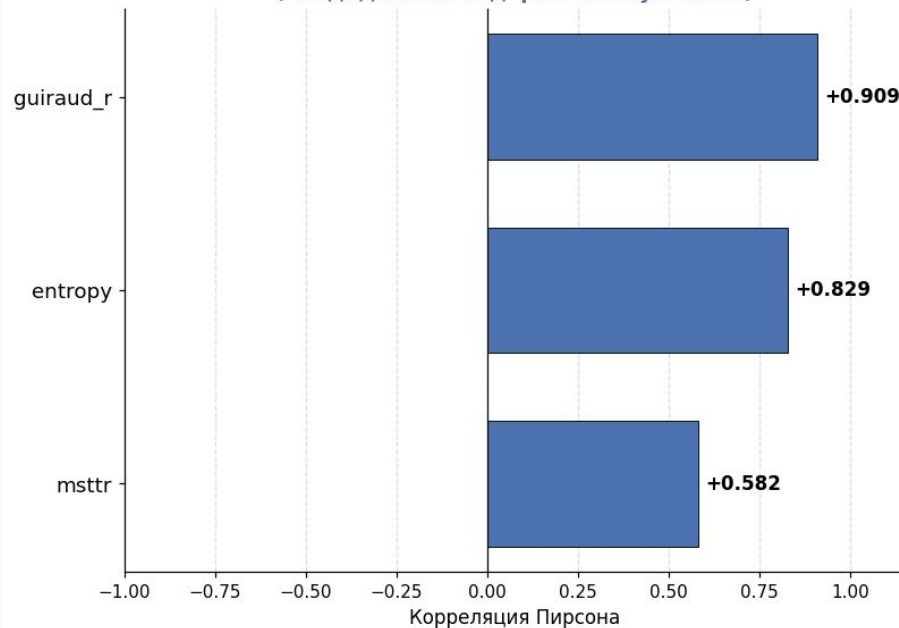


Энтропия vs LOC
 $r = +0.829$, $p = 8.61e-09$

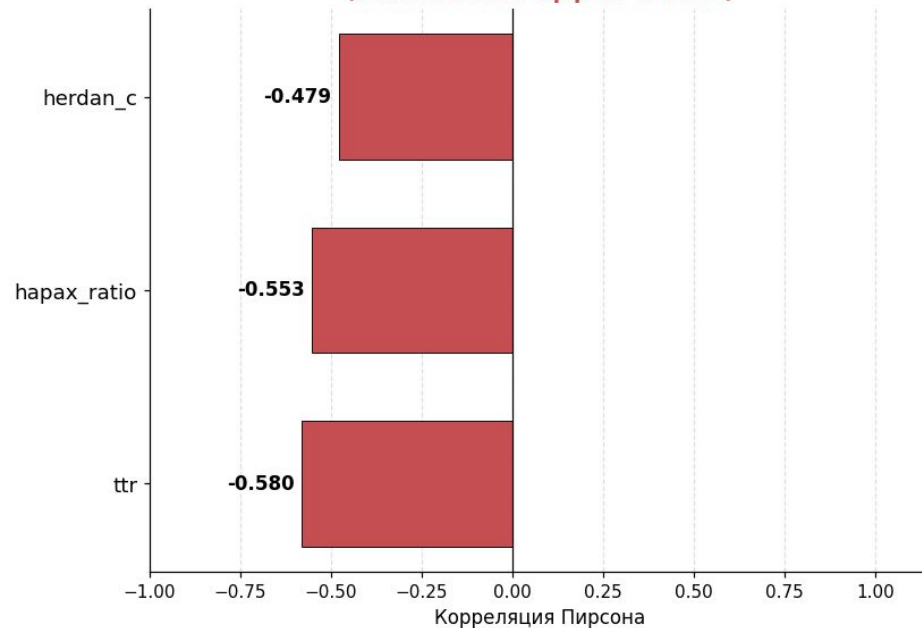


Два класса лексических метрик по связи с LOC

Класс 1: растут с LOC
(кандидаты на содержательную связь)



Класс 2: падают с ростом корпуса
(механический эффект объёма)



MSTTR

Pearson: +0.582 ($p = 6.0e-04$).

Spearman: +0.504 ($p = 3.8e-03$).

Что это значит:

- MSTTR даёт заметно более слабую корреляцию, чем Guiraud's R и энтропия.
- Разница: MSTTR измеряет богатство внутри сегментов фиксированной длины, без нормировки на весь текст.

Лексические метрики vs LOC

TTR (Type-Token Ratio):

- Pearson: -0.580 ($p = 6.2e-04$).
- Spearman: -0.796 ($p = 8.9e-08$).
- Чем больше корпус, тем ниже TTR — механический эффект.

Нарых ratio:

- Pearson: -0.553 ($p = 1.3e-03$).
- Spearman: -0.718 ($p = 5.4e-06$).
- Доля слов с частотой 1 падает с ростом корпуса

Лексические метрики vs LOC

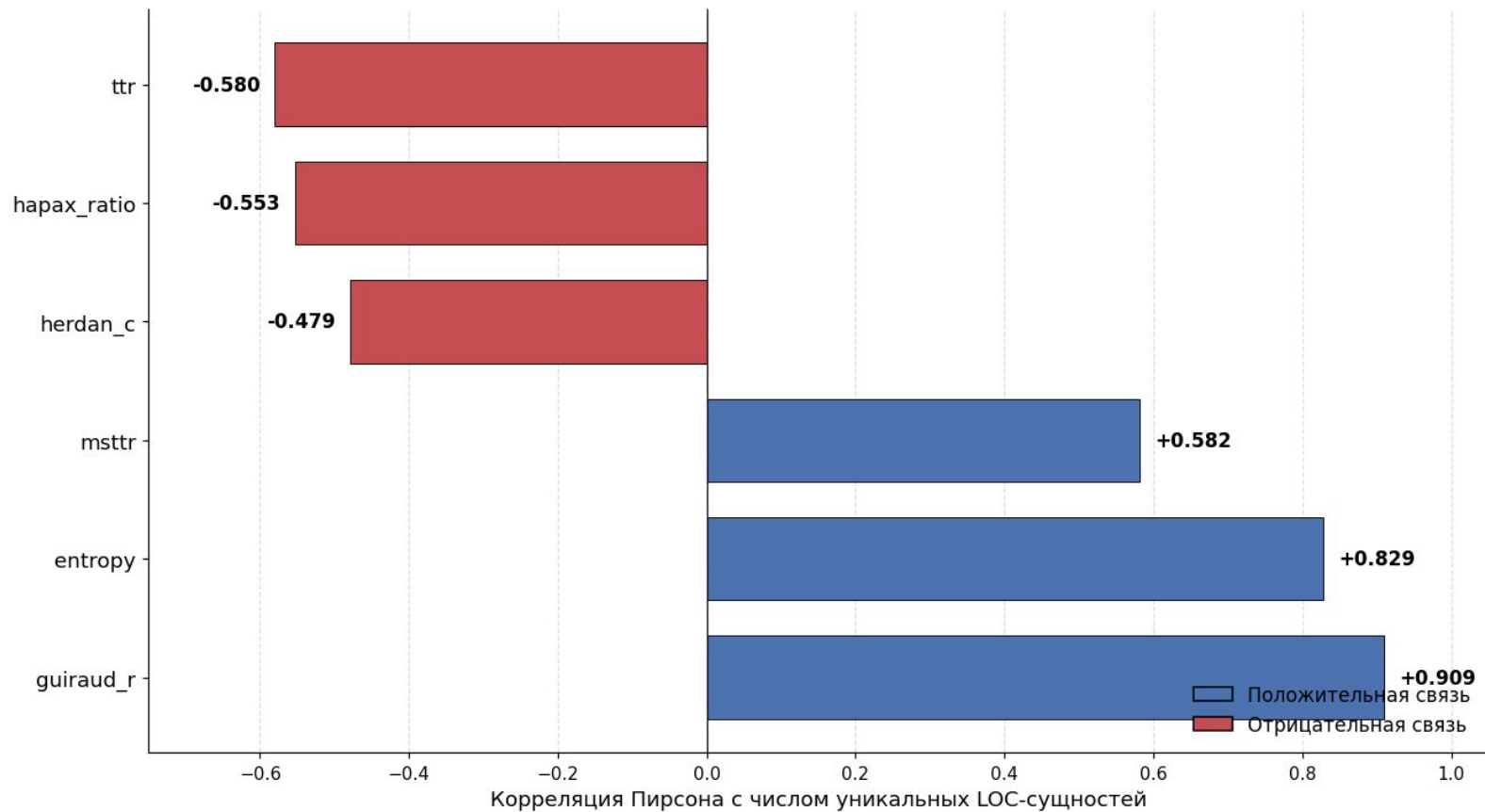
Herdan's C:

- Pearson: -0.479 ($p = 6.4e-03$).
- Spearman: -0.607 ($p = 2.9e-04$).
- Отрицательная связь — неожиданно для метрики, которая должна нормировать объём.

Что это значит:

- Все три метрики зависят от объёма — чем больше корпус, тем ниже их значение.
- Поэтому их отрицательная корреляция с LOC — артефакт объёма, а не содержательный сигнал.

Сырые корреляции: лексические метрики vs LOC



Что остаётся после контроля на объём

TTR:

- Сырая корреляция: -0.580
- Частная корреляция:
 $+0.239$ ($p = 0.19$)
- Не значима — связь была артефактом

MSTTR:

- Сырая корреляция: $+0.582$
- Частная корреляция:
 $+0.082$ ($p = 0.66$)
- Полностью исчезла —
связь была артефактом
объёма

Что остаётся после контроля на объём

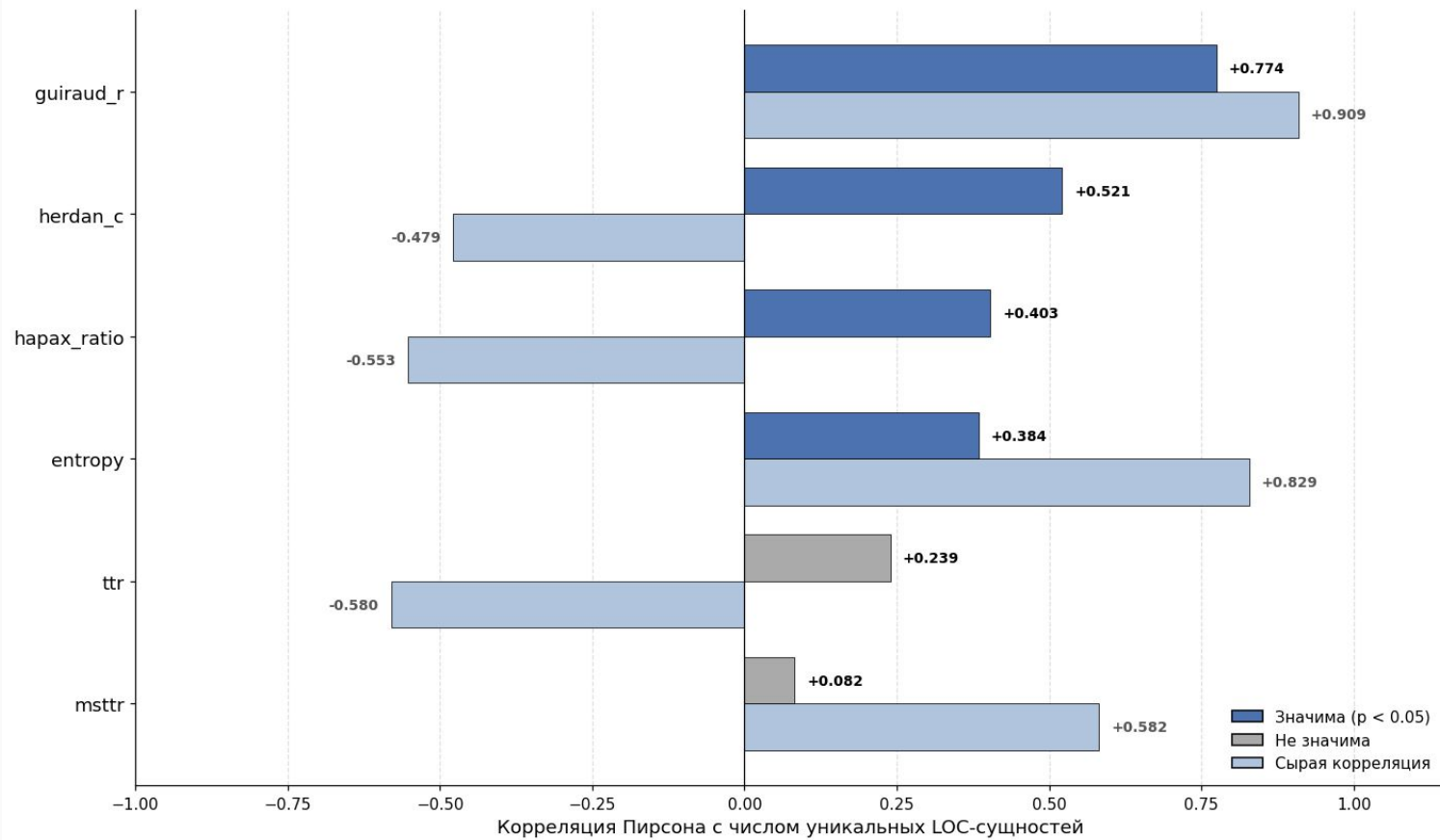
Guiraud's R:

- Сырая корреляция: +0.910
- Частная корреляция:
+0.774 ($p = 3.2e-07$)
- Потеря при контроле: всего
0.14 — связь устойчива

Энтропия:

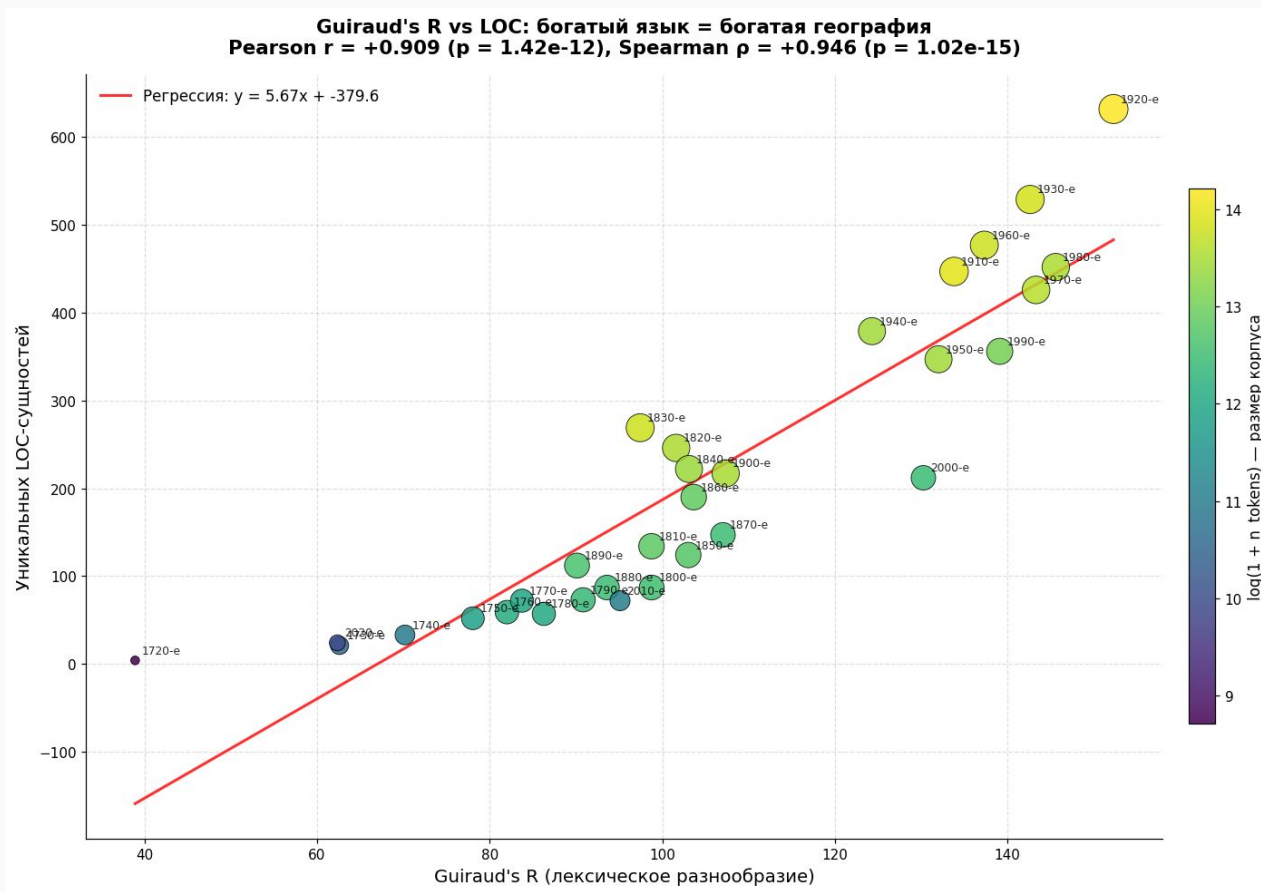
- Сырая корреляция: +0.829
- Частная корреляция:
+0.384 ($p = 0.033$)
- Заметное падение — связь
слабая, но значимая

Сырые vs частные корреляции: что остаётся после контроля на объём



Даже при одинаковом объёме корпуса десятилетия с более разнообразной лексикой имеют больше LOC. $R^2 \approx 0.6$ — 60% остаточной вариации LOC объясняется лексическим разнообразием.

Связь между лексическим разнообразием эпохи и числом LOC



Богатый язык = богатая география.

Эпохи лексического богатства
оказываются и эпохами
географического богатства

Географическое разнообразие
русской поэзии — не
механическое следствие роста
корпуса, а самостоятельный
процесс расширения
поэтического языка.