

Расстояния между пространствами языков и текстов

Введение

- Тексты можно рассматривать как метрические пространства
- Задача: сравнить структуры языков и текстов
- Цель — найти количественную метрику различий

Проблема

- Тексты ботов всё сложнее отличить от человеческих
- Поверхностные признаки (стиль, частоты) уже недостаточны
- Нужен метод, анализирующий глубинную семантическую структуру

Изначальный план

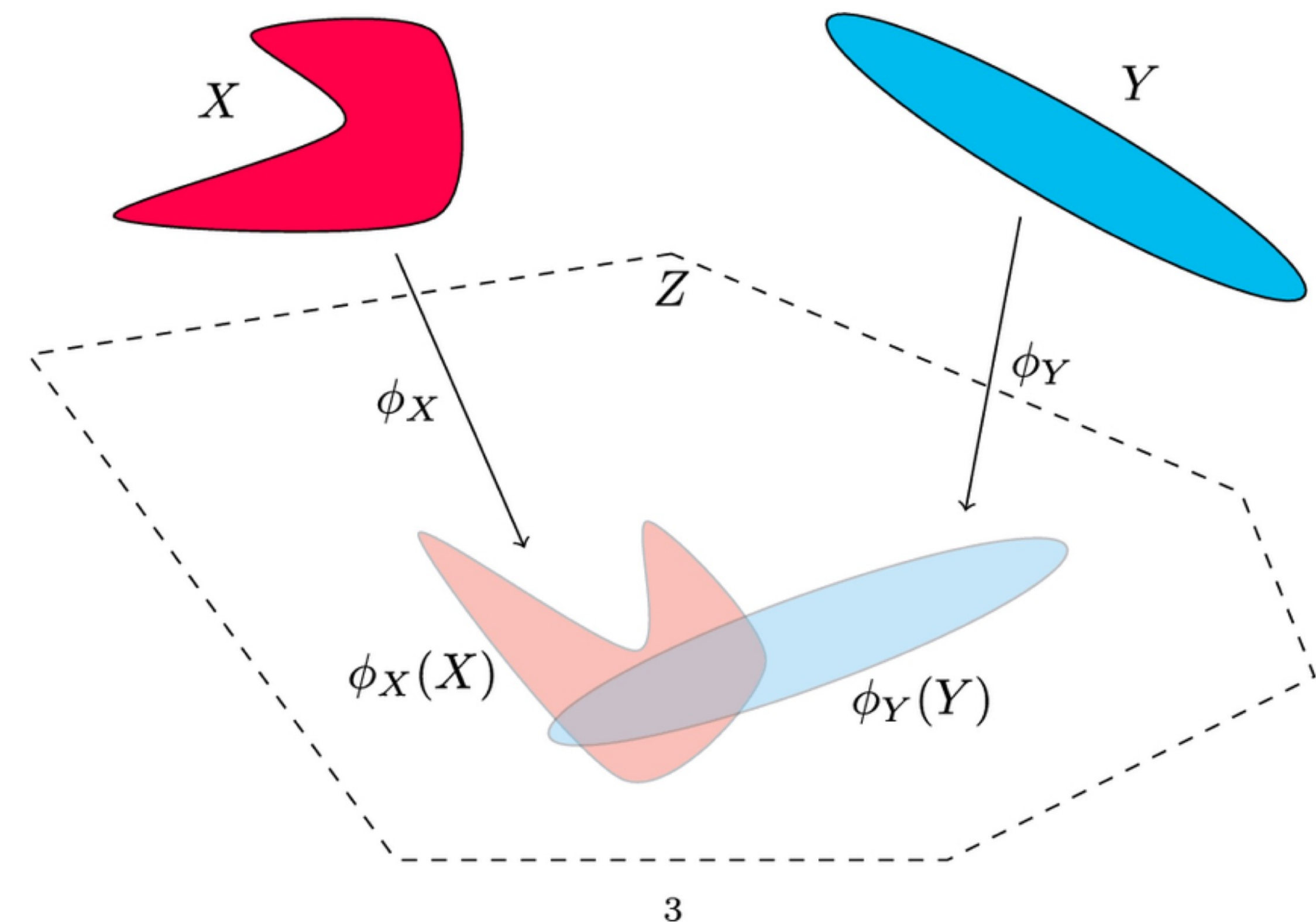
- Использовать расстояние Громова–Хаусдорфа (GH)

$$d_{GH}(X, Y) = \inf_{\varphi: X \rightarrow Z, \psi: Y \rightarrow Z} d_H(\varphi(X), \psi(Y)).$$

где

$$d_H(A, B) = \inf \{ \varepsilon > 0 : A \subseteq B_\varepsilon, B \subseteq A_\varepsilon \}.$$

GH сравнивает *формы* пространств и является эталонной геометрической метрикой.



Получить:

- расстояние между русским и английским
- расстояние между текстами человека и бота

Почему GH не подошёл

- NP-трудность поиска соответствий
- Корпуса: ~100 000 слов → неподъёмно

Вывод: GH не масштабируется к реальным языкам

Рассмотренные альтернативы

Gromov–Wasserstein

$$GW_2^2(\mu, \nu) = \min_{\pi \in \Pi(\mu_X, \mu_Y)} \sum_{i,j,k,l} \left(d_X(i, j) - d_Y(k, l) \right)^2 \pi_{ik} \pi_{jl}, .$$

- также NP-труднен в подсчете, но существуют эффективные аппроксимации

Энтропийный Gromov–Wasserstein

$$EGW_\varepsilon(X, Y) = \min_{\pi \in \Pi(\mu_X, \mu_Y)} \sum_{i,j,k,l} \left(d_X(i, j) - d_Y(k, l) \right)^2 \pi_{ik} \pi_{jl} + \varepsilon H(\pi),$$

- Теоретически привлекательный

Но:

- чувствительный гиперпараметр
- высокая стоимость вычислений
- сложности интерпретации

Итог: не выбран как основной метод

ИТОГОВЫЙ ВЫБОР

Sliced Gromov–Wasserstein (SGW)

$$\text{SGW}_2^2(X, Y) = \int_{S^{d-1}} \text{GW}_2^2((\pi_\theta)_\# \mu_X, (\pi_\theta)_\# \mu_Y) d\theta, \approx \frac{1}{L} \sum_{\ell=1}^L \text{GW}_2^2((\pi_{\theta_\ell})_\# \mu_X, (\pi_{\theta_\ell})_\# \mu_Y),$$

- Проекции в одномерные направления
- Быстро, масштабируемо
- Улавливает структуру расстояний
- Подходит для больших корпусов

Что сравнивали

Языки

- 100 000 наиболее частотных слов
- English, German, Hindi, Japanese, Korean, Russian

Корпусы текстов (русский)

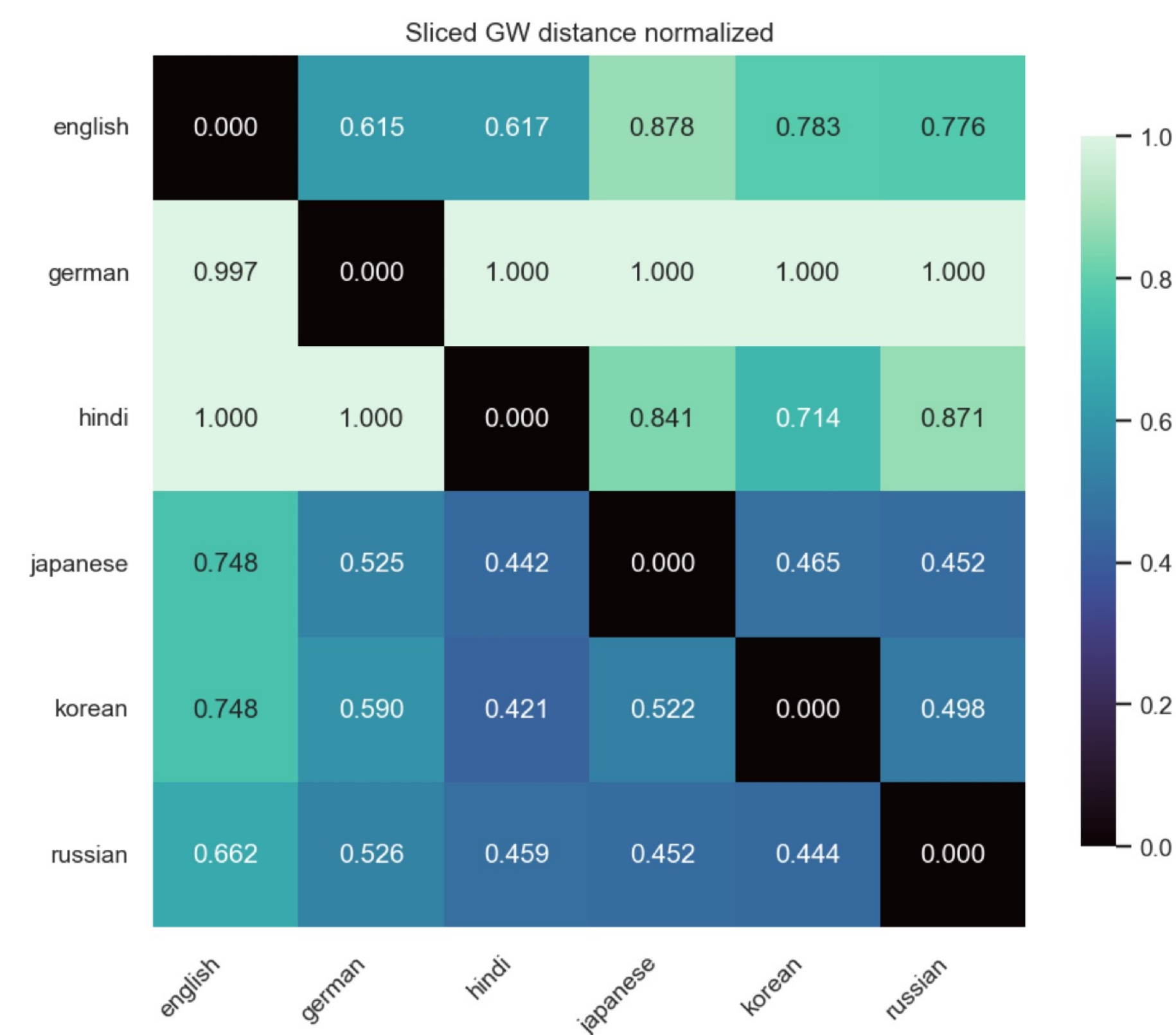
- human_russian
- balaboba
- gpt2
- mGPT

Методика

1. Формирование эмбеддингов (SVD, CBOW)
2. Представление языка как метрического пространства
3. Вычисление SGW для всех пар
4. Визуализация через MDS и тепловые карты

Результаты: Языки

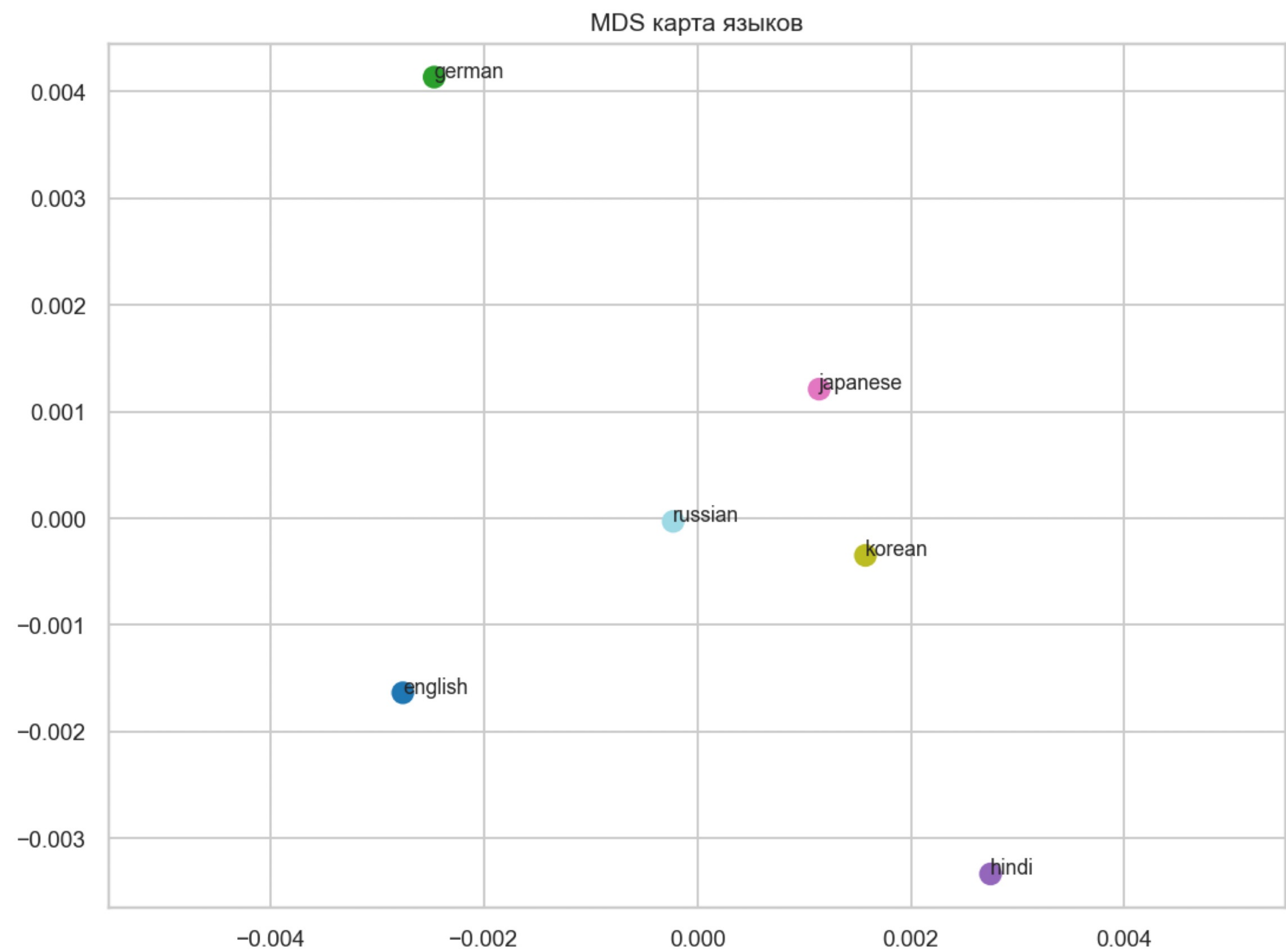
Матрица SGW (6 языков)



Инсайты

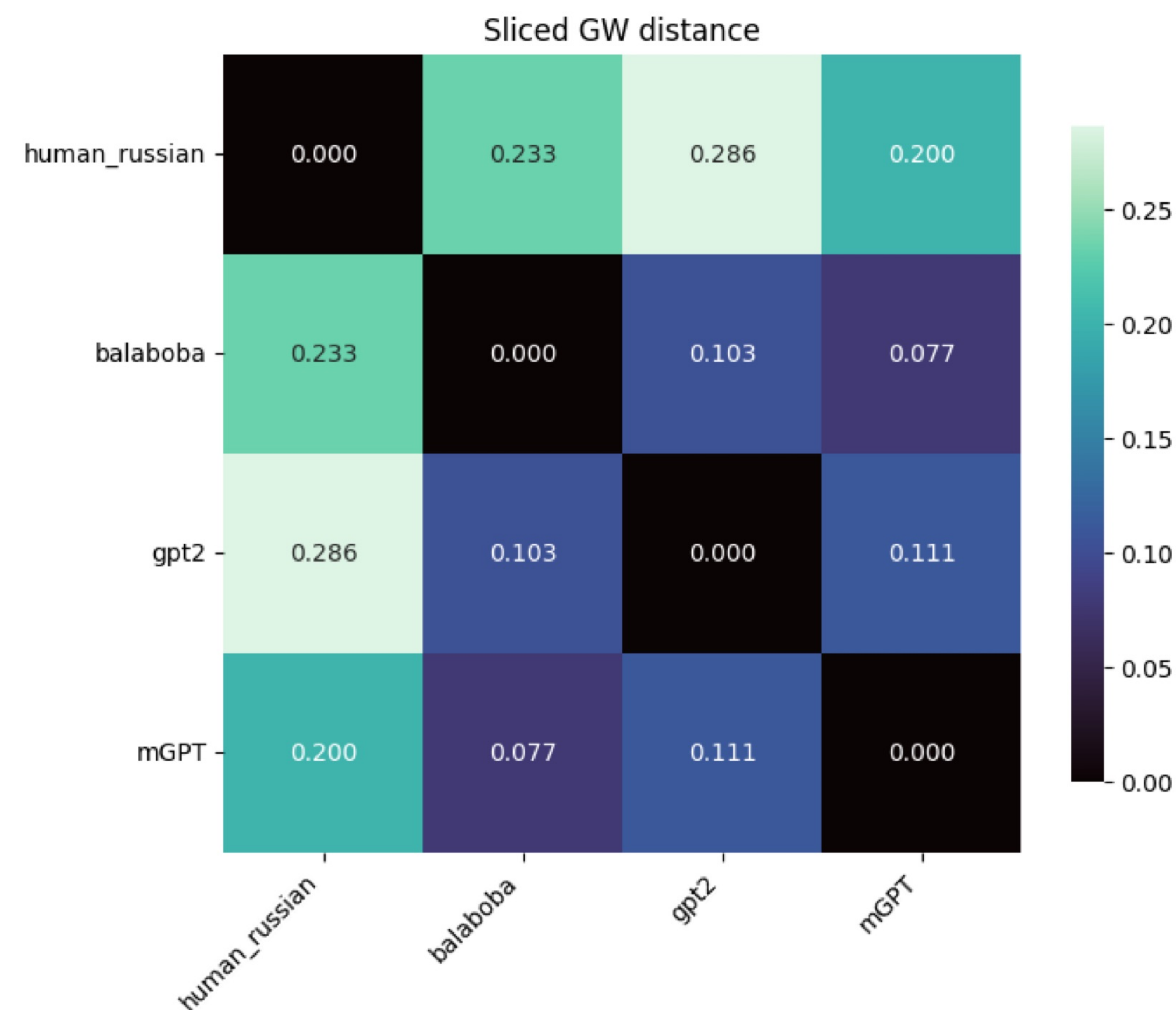
- Все расстояния — стабильные, отличимые
- Russian ↔ English: 0.0038
- Языки формируют структуры, отражающие различия эмбедингов

MDS-карта языков



Результаты: Человек vs Бот

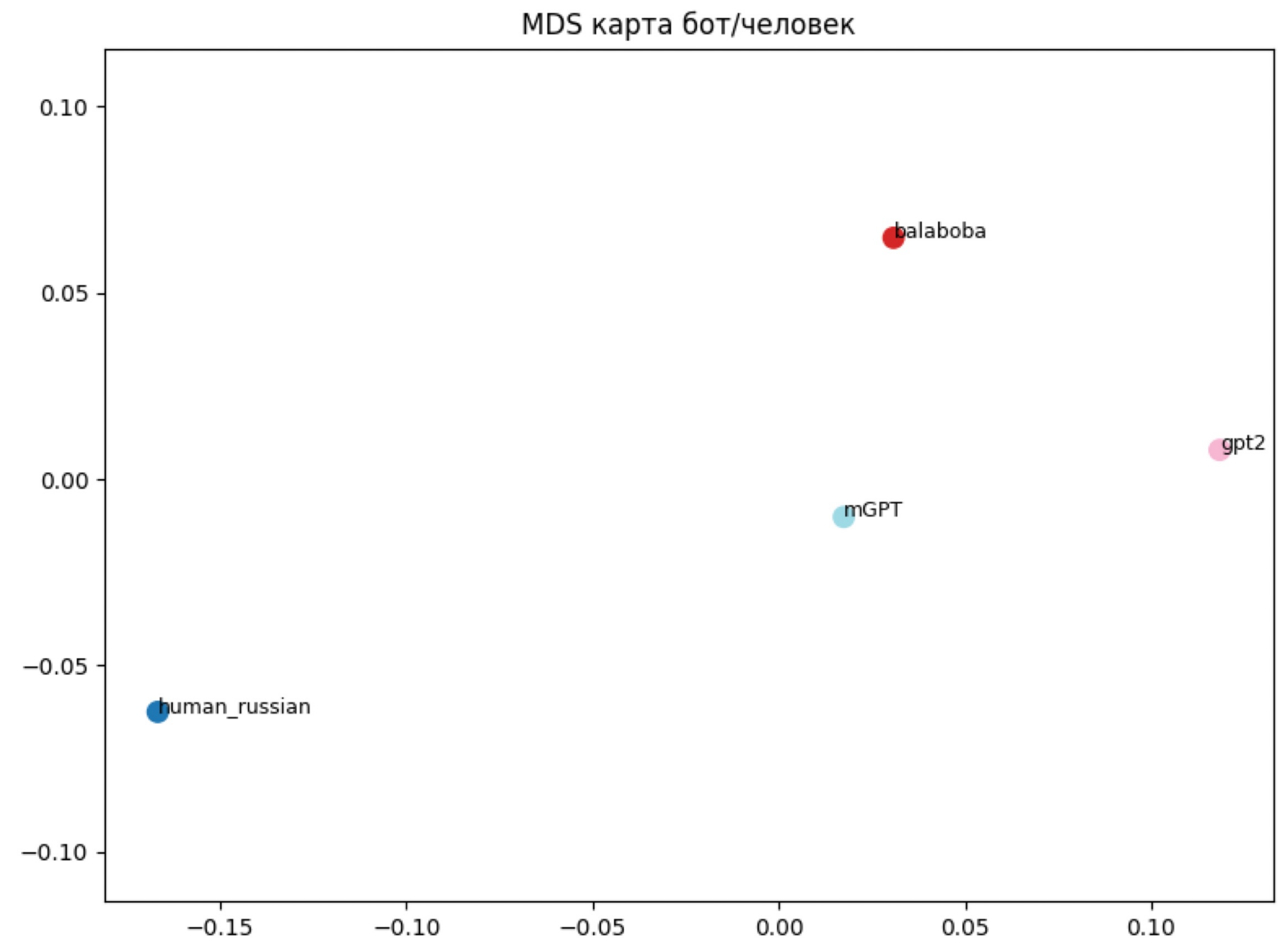
Матрица SGW (4 корпуса)



Что это значит?

- Бот-корпуса ближе друг к другу, чем к человеку
- human_russian — отдельный «кластер»
- Наибольшее отличие — human_russian ↔ gpt2

MDS корпусов



Выводы

- SGW оказался практичным и устойчивым методом
- Позволяет количественно сравнивать структуры языков
- Различает тексты человека и бота на уровне эмбеддингов

Но:

- не гарантирует 100% детекцию
- лучше использовать в составе гибридных методов

Перспективы

- Расширение числа языков
- Работа с n -граммами ($n > 1$)
- Комбинация SGW с ML-классификацией