ИПМ им.М.В.Келдыша РАН

Будущее-2024 • Труды конференции

В.А.Громов, Н.С.Бородин, А.С.Коган,
К.Н.Данг, А.С.Ерболова, Х.Баян

# Поймай бота: крупномасштабная структура естественного языка

***Рекомендуемая форма библиографической ссылки***

Громов В.А., Бородин Н.С., Коган А.С., Данг К.Н., Ерболова А.С., Баян Х. Поймай бота: крупномасштабная структура естественного языка // Проектирование будущего. Проблемы цифровой реальности: труды 7-й Международной конференции (15-17 февраля 2024 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2024. — С. 281-312. — https://keldysh.ru/future/2024/6-3.pdf https://doi.org/10.20948/future-2024-6-3

***Размещено также видео выступления***

# Поймай бота: крупномасштабная структура естественного языка

**В.А. Громов, Н.С. Бородин, А.С. Коган, К.Н. Данг, А.С. Ерболова, Х. Баян**

*НИУ ВШЭ, Покровский бульвар*

**Аннотация.** В современном мире специальные программы (боты) пишут комментарии, новости, отзывы, которые могут содержать ложную информацию. Вследствие этого крайне важно знать, написан ли тот или иной текст реальным человеком или ботом. Данная работа направлена на исследование семантических траекторий текстов на естественных языках для анализа вышеупомянутой проблемы. В исследовании используются понятия векторных вложений и их *n*-грамм, а также методы (1) кластеризации семантического пространства, (2) анализа позиции текстов на плоскости «энтропия-сложность», (3) оценки внутренних размерностей векторных представлений языка, (4) топологического анализа данных.

**Ключевые слова:** семантические траектории, обработка естественного языка, боты, кластеризация, плоскость «энтропия-сложность», внутренняя размерность, топологический анализ данных

# Spot the bot: large-scale natural language structure

**V.A. Gromov, N.S. Borodin, A.S. Kogan, Q.-N. Dang, A.S. Yerbolova, H. Bayan**

*National Research University Higher School of Economics*

**Abstract.** In the modern world, specialized programs (bots) write comments, news, reviews, which may contain false information. As a result, it is extremely important to know whether a given text was written by a real person or a bot. This work aims to study the semantic trajectories of texts in natural languages to analyse the aforementioned problem. The study utilizes the concepts of vector embeddings and their *n*-grams, as well as methods for (1) clustering the semantic space, (2) analysing the position of texts on the 'entropy-complexity' plane, (3) estimating the intrinsic dimensionalities of vector language representations, and (4) topological data analysis.

**Keywords:** semantic trajectories, natural language processing, bots, clustering; entropy-complexity plane, intrinsic dimensionality, topological data analysis

## 1. Введение

Стремительная эволюция искусственного интеллекта толкает человечество в мир постправды. Генеративный ИИ (чат-боты, LLM и т.д.), – важный движущий фактор процесса, имеет тенденцию искажать *языковую личность человека* (языковой аспект человеческой идентичности). В настоящее время языковая личность развивается под влиянием текстов, написанных человеком (в основном, с благими намерениями) – от детских сказок до университетских учебников по математическому анализу. Благодаря недавним достижениям в области генерации текстов мы вскоре с самого рождения окажемся погруженными в гигантский океан текстов («с привкусом, идентичным натуральному»), генерируемых ботами, искажающих языковую личность ребёнка. Это делает необходимым разработать системы искусственного интеллекта, способные различать тексты, созданные ботами, и тексты, написанные человеком.

Часто методы обнаружения стремятся идентифицировать конкретного бота, что представляется не совсем разумным: новые архитектуры ботов появляются с завидной регулярностью, что неизбежно приводит к устареванию ранее разработанных способов борьбы с ними. Вместо этого мы предлагаем акцентировать внимание не столько на ботах, сколько на том, что именно они стремятся имитировать – естественном языке.

Подход, предложенный Громовым и Мигриной [1], позволил взглянуть на естественный язык как на единую сложную систему, и, что более важно, самоорганизованно-критическую систему, литературные произведения в которой сравнимы со снежными лавинами. В настоящей работе предпринимаются попытки исследовать геометрические свойства этой системы и использовать их для обнаружения текстов, сгенерированных ботами.

Мы предполагаем, что множество всех векторных вложений (наблюдаемых слов) языка представляет собой репрезентативную выборку всех точек $d$-мерной поверхности и, следовательно, анализ такого множества позволит выявить структурные особенности этой поверхности. То же справедливо и для поверхностей, связанных с множествами всех наблюдаемых $n$-грамм. Забегая вперёд, следует отметить, что эти поверхности обладают нецелочисленными (фрактальными) размерностями, поэтому нам приходится отказаться от напрашивающегося термина «языковые многообразия» в пользу «языковых фрактальных структур». Мы предлагаем использовать термин *Hailonakea*[1] («знаковая необъятность» с гавайского) для

[1] Читатель может перейти по ссылке на статью в Википедии, https://en.wikipedia.org/wiki/Laniakea_Supercluster, чтобы ознакомиться с иллюстрацией сверхскопления Ланиакея, крупнейшего объекта наблюдаемой Вселенной. Каждый пиксель фотографии соответствует даже не галактике, а скоплению галактик; фотография даёт представление о крупномасштабной структуре нашей Вселенной. Название Ланиакея (*laniākea*) переводится с гавайского как «огромный рай», от *lani* «рай» и *ākea*

обозначения совокупности всех фрактальных языковых структур, для всех $n$-грамм, $n = 1, \ldots$, крупномасштабной структуры языка как сложной системы [2;3;4;5].

В настоящей работе рассматриваются несколько подходов исследования свойств фрактальных языковых структур для $n = 1$ (слов); $n = 2$ (биграмм) русского и белорусского языков: кластеризация[2], плоскость «энтропия-сложность»[3], внутренние размерности[4], топологический анализ данных[5]. Также стоит отметить, что мы ограничиваемся только литературными текстами, полагая, что литература составляет ядро соответствующего языка.

Дальнейший текст организован следующим образом. Во втором разделе рассматриваются работы, послужившие отправной точкой для настоящего исследования. В третьем формально ставится задача. Четвёртый раздел описывает методологию формирования векторных вложений, кластеризации, позиционирования на плоскости «энтропия-сложность», оценки внутренней размерности и оконтуривания «дырок»; пятый – результаты экспериментов. Наконец, в последнем разделе представлены выводы.

## 2. Обзор источников

Представляется полезным подойти к доступной литературе с двух точек зрения. Сначала мы затронем статьи, в которых естественный язык рассматривается как цельный естественнонаучный объект; затем коснёмся работ, посвящённых исследованию задачи детекции ботов; после затронем статьи, позволяющие изучать характеристики языка как заданного множества точек. Важно отметить, что настоящее исследование исходит из убеждения, что естественный язык представляет собой массив значений (понятий), а не массив слов (символов), и к нему следует относиться как к семантическому пространству (пространству значений).

---

«необъятный, неизмеримый». Мы признательны Н.Ю. Ютанову (†), главной астрономической обсерватории Российской академии наук в Пулково, а также его коллегам из обсерватории Мауна-Кеа (США) за предложенный термин.

[2] Мы предполагаем, что количество, положение и внутренняя структура кластеров для людей и ботов качественно различны.

[3] Каждый текст может рассматриваться как многомерный временной ряд, который можно расположить на такой плоскости [6].

[4] Мы определим понятие внутренней размерности формально в следующих разделах. Интуитивное представление можно получить, обратившись к обычной сфере – её внутренняя размерность равна двум, поскольку требуется не менее двух компонент (к примеру, широта и долгота), чтобы однозначно определить каждую её точку, вне зависимости от размерности пространства, в которое помещается сфера ([3;5] и т.д.).

[5] Слова и $n$-граммы распределены по семантическому пространству естественного языка неравномерно, они образуют «хабы» (скопления) и «дырки» (персистентные гомологии; области сущностей, употребляемых крайне редко).

К сожалению, большинство работ, изучающих естественный язык как целое, посвящены «словам», а не «значениям». Тем не менее, можно указать на несколько монографий [1;7;8], в которых «слова» исследуются с целью выявления свойств языковых процессов в качестве «значений». Так, Танака-Исии и её коллеги ([9] и ссылки в ней) анализируют длинные корреляции для английского и японского языков. Благодаря природе японского языка между иероглифами кандзи и «значениями» можно установить почти однозначное соответствие. Выводы, сделанные для иероглифов, можно осторожно распространить на элементы семантического пространства («значения»). В языке наблюдаются длинные корреляции, состоящие из слов, расположенных на расстоянии 10-15 позиций друг от друга[6].

Дебовски [8] исследует степенные законы естественного языка: он показывает, что именно этот класс распределений управляет большинством фундаментальных языковых процессов. Громов и Мигрина [1] рассматривают естественный язык как единое целое; авторы доказывают, что естественный язык представляет собой самоорганизующуюся критическую систему, при этом литературное произведение (или любой другой текст) представляет собой степенную «лавину» в семантическом пространстве. Кроме того, важно выделить исследования, посвященные моделированию языка как сложной сети. Например, Гарг и др. [2;3;4] исследуют динамику процессов естественного языка посредством построения и анализа графовых моделей языка. Как редкое исключение из работ, посвящённых «словам», можно сослаться на работу [6]: авторы преобразуют слова в векторные вложения, тем самым превращая литературные тексты в многомерные хаотические ряды.

Переходя к задаче детекции ботов, стоит отметить, что её решению посвящено довольно много статей. Значительная их доля используют нейросетевые подходы.

Первые работы использовали RNN. Например, нейросеть, основанная на архитектуре LSTM была предложена в [10]. Данная модель помимо текстов использовала метаданные об авторе. Насколько нам известно, первая модель, использующая только текст, была предложена в [11]. Эта модель основывалась на архитектуре BiLSTM. LSTM и GAN использовались для детекции ботов в социальной сети Twitter в [12]. В [13] приведено эмпирическое исследование предобученных и контекстуальных вложений в задаче детекции ботов в социальных сетях. Детальный обзор набора данных, содержащий фейковые и настоящие новости, был приведен в [14], где для классификации применялась модель deep diffusive network.

Также многие работы используют архитектуры типа «трансформер». Тонкая настройка моделей GPT, GPT-2 для детекции ботов показала хоро-

[6] Это значительно превышает длину типичной $n$-граммы в процедурах обработки естественного языка.

шие результаты [15]. Графовые методы широко используются для данной задачи [16;17]. Например, GCN и BERT были объединены в [18]. Естественно, существуют методы, не основанные на нейронных сетях. Например, лингвистическое выравнивание может помочь определить бота во взаимодействиях человек-бот [19].

Кластеризация, и пусть это будет плавным переходом в последнюю секцию нашего обзора, также применяется для выявления ботов. Например, в [20] использовался алгоритм DBSCAN для выявления комментариев на испанском языке, написанных ботами. Алгоритм кластеризации Уишарта применялся в [21] для определения фальшивых отзывов на сайте Yelp.com. Каждый вектор набора данных – взвешенное среднее векторов слов из этого предложения, где у разных частей речи был разный вес. В схожей работе [22], в которой вектором текста считалась сумма векторов входящих в него слов, был использован алгоритм кластеризации Зана.

Однако стоит снова отметить, что, в отличие от остальных работ, наше исследование направлено не на детекцию фиксированного бота. Мы пытаемся выявить структуру человеческого языка в целом, позволяя в теории выявлять различных ботов.

Существуют также и решения, использующие концепции из теории информации. Чу и др. [23] характеризуют различия между действиями ботов и людей в Твиттере, вычисляя энтропию статистики активности аккаунтов. Они обнаружили, что у людей энтропия выше, чем у ботов, что подчёркивает их более сложное поведение во времени. В нашей работе мы применяем аналогичные идеи к текстовым данным вместо метаданных.

Переходя к алгоритмам оценки внутренней размерности, в первую очередь, следует обратиться к работе [24]: в статье вводятся необходимые условия для того, чтобы некоторая функция являлась функцией внутренней размерности; соответствующие аксиомы опираются на понятие расстояния М. Громова между метрическими пространствами [25]. Далее можно в общих чертах разделить известные методы оценки внутренней размерности геометрического объекта (с использованием репрезентативной выборки его точек) на три класса, использующие, соответственно, траектории странных аттракторов, содержащих рассматриваемое многообразие [26;27]; графы данных (*proximity graphs*) [28]; персистентные гомологии [29;30].

В работах первого класса рассматриваются семантические и тональные временные ряды (см., например, [6] и ссылки в ней). Предполагается, что рассматриваемый временной ряд удовлетворяет теореме Такенса; это означает, что он движется в окрестности исследуемого странного аттрактора, что позволяет оценить размерность последнего. Канц и Шрейбер [26], а также Малинецкий с Потаповым [27] описывают основные способы определения размерностей странных аттракторов (топологической, хаусдорфовой, энтропийной), а также приводят методы их оценки. Работы по

ссылкам [31;32] посвящены способам оценки фрактальных размерностей для сложных систем, встречающихся в объективной реальности. К сожалению, эти методы (1) требуют гораздо более длинных временных рядов, чем любые доступные «языковые» временные ряды [27;33]; (2) кроме того, они зачастую не робастны в терминах изменяющихся данных.

Коста и др. [34] и Фараманд и др. [35] используют в своих работах граф $k$-ближайших соседей для оценки топологической размерности. Брито и др. [28] распространили метод и на другие графы данных.

Адамс и др. [29] используют понятие персистентной гомологии для вычисления внутренней размерности; Швайнхарт [30] произвёл строгий анализ этой оценки, связав её с энтропийной размерностью, а после распространил метод на фрактальные и мультифрактальные множества, т.е. на множества с нецелочисленными размерностями.

Наконец, методы топологического анализа данных и поиска персистентных гомологий используется и для изучения структуры естественного языка. В [36] автор рассматривает фильтрации над последовательностями слов в детских стишках; текстах, написанных детьми и текстах, написанных взрослыми. По баркодам (графикам, иллюстрирующих продолжительность существования той или иной гомологии) исследуется структура разных текстов, выявляются персистентные гомологии, иными словами, «дырки» в текстах.

### 3. Постановка задачи

Более формально необходимо:

- Для данного множества текстов $\mathfrak{I} = (\Omega_1, \cdots, \Omega_L)$ заданного естественного языка сформировать наборы $\aleph_n(d)$ векторных вложений всех уникальных $n$–грамм, $n = 1..N, d = 1..D$ ($d$ здесь и далее является размерностью пространства вложения; $n$ – количество последовательных слов в рассматриваемой $n$–грамме). Предполагается, что множество текстов $\mathfrak{I}$ формирует репрезентативную выборку текстов соответствующего языка;
- Для фиксированных $n$, используя наборы $\aleph_n(d), d = 1..D$ кластеризовать семантические пространства рассматриваемых естественных языков, проанализировать свойства кластеров как в статической парадигме, так и в динамической (семантические траектории);
- Отобразить множества траекторий текстов $\{\Omega\}$ на плоскость «энтропия-сложность». Исследовать зависимость режима траекторий (хаос, детерминированность) от значений параметров $n$ и $d$;
- Оценить внутренние размерности $n$–грамм $\hat{d}_i = f_i(\{\aleph_n(d)\})$ соответствующих языков, используя реализации различных,

разработанных нами, методов $f_i$. Проверить гипотезу о фрактальности языка для множеств $n$–грамм, $\aleph_n(d)$, используя те методы $f_i$, которые допускают нецелые значения размерности $\hat{d}$;

– Для множеств $\aleph_n(d)$ выявить персистентные гомологии[7] первого порядка $H_1$ и исследовать их характеристики.

## 4. Методология

4.1.1. Векторные представления естественного языка

*SVD.* Чтобы преобразовать исходный набор текстов (контекст) $\Im$ в множество точек $\aleph_n(d)$, мы применяем сингулярное разложение (SVD) к матрице совместной встречаемости (*co-occurrence matrix*) [37]:

Для контекста $\Im = (\Omega_1, \cdots, \Omega_L)$ и множества уникальных (лемматизированных) слов $\aleph = (\lambda_1, \cdots, \lambda_M)$ мы формируем $M \times L$ матрицу $S = (s_{ij})$, элементы которой рассчитываются как:

$$s_{i,j} = (1 - \varepsilon_i) \frac{k_{i,j}}{\sum_{i' \in \Omega_j} k_{i',j}}. \quad (4.1.1)$$

Здесь $k_{i,j}$ – число вхождений леммы $\lambda_i$ в текст $\Omega_j$; $\varepsilon_i$ – нормализованная энтропия слова $\lambda_i$ в $\aleph$:

$$\varepsilon_i = -\frac{1}{\ln L} \sum_{j=1}^{L} \frac{k_{i,j}}{\tau_i} \ln \frac{k_{i,j}}{\tau_i}, \quad (4.1.2)$$

$$\tau_i = \sum_{j=1}^{L} k_{i,j}. \quad (4.1.3)$$

Таким образом, наибольшие элементы $S$ соответствуют словам, часто встречающимся в конкретном тексте и редко встречающимся во всех остальных текстах этого контекста.

Далее к матрице $S$ применяется сингулярное разложение [38]:

$$S = U \Lambda V^T. \quad (4.1.4)$$

$U$ и $V^T$, соответственно, являются $M \times d$ и $d \times L$ прямоугольными матрицами; $\Lambda$ – $d \times d$ квадратная диагональная матрица, на главной диагонали которой расположены сингулярные значения в порядке убывания. Для расчёта SVD мы используем алгоритм Голуба–Кахана–Ланцоша [39]. Строки матрицы $U$ представляют собой векторные вложения слов $\aleph$.

[7] Число гомологий соответствует характеристике, широко исследуемой в топологическом анализе данных – числам Бетти.

Стоит отметить следующую особенность данного подхода: чтобы получить $d_2$-мерное вложение из $d_1$-мерного, $d_2 < d_1$, его следует усечь до первых $d_2$ компонент. В результате такого усечения будет потеряно наименьшее количество информации, так как сингулярные значения предварительно отсортированы в порядке убывания. Этот факт чрезвычайно важен для настоящего исследования: чтобы получить надёжные результаты, нам необходимо провести множество экспериментов для различных размерностей пространства вложения. С помощью этого метода достаточно построить набор вложений для некоторого большого $d$, а затем при необходимости обрезать их до меньших значений $d$. Другие доступные методы (например, глубокие нейронные сети) потребовали бы проведения вычислительно непомерно сложного моделирования для построения вложений для каждого $d$ отдельно. Здесь и далее векторное вложение $n$-граммы получается путём конкатенации эмбеддингов $n$ входящих в неё слов в том порядке, в котором они встречаются в тексте.

*Word2Vec (CBOW).* Однако, чтобы совсем не обойти стороной методы глубокого обучения, лидирующие на текущий момент в задачах NLP, мы также используем эффективный и в то же время вычислительно простой Word2Vec-подход к моделированию векторных вложений (CBOW) [40].

Векторные представления извлекаются из распределения вероятностей, полученного в результате обучения полносвязной нейронной сети. Основную идею CBOW можно описать в несколько шагов: (1) проход по языковому корпусу скользящим окном, (2) максимизация логарифмического правдоподобия центрального слова в зависимости от контекста: модель представляет собой перцептрон с одним скрытым слоем, на вход которого передаются необучаемые векторные представления слов; (3) на выходе возвращается вектор распределения вероятностей принадлежности некоторого элемента, скрытого в окне, каждому слову словаря. Таким образом, слово можно предсказать на основе окружающего его контекста.

Результирующая матрица $W$ (матрица весов на скрытом слое) представляет собой словарь обучаемых вложений (каждому слову присваивается вектор-столбец). Преимущество Word2Vec заключается в том, что алгебраические операции над векторами отражают семантические операции [40], а косинусная близость векторов означает семантическое сходство. При этом обучение матрицы Word2Vec существенно быстрее по сравнению со многими другими вариантами построения языковых моделей, поэтому в данной работе оно выбрано в качестве вспомогательного метода.

### 4.2 Кластеризация семантического пространства

Для кластеризации векторного пространства рассматриваемого[8] языка мы используем несколько алгоритмов кластеризации.

*Алгоритм Уишарта.*

В качестве основного метода выбран плотностный алгоритм Уишарта [41] в силу его робастности, умения отделять шум и отсутствия предположения о форме кластеров. Основная идея алгоритма заключается в итеративной оценке значимости кластеров, с помощью которой каждая точка либо помечается меткой кластера, либо относится к шуму.

***Обозначения*:**

$k$ – количество соседей (гиперпараметр)

$h$ – пороговый уровень значимости

$l$ – количество точек

$d(x, y)$ – расстояние между точками $x, y$

$d_k(x)$ – расстояние до $k$-ого ближайшего соседа точки $x$

$V_k(x)$ – объем гиперсферы радиусом $d_k(x)$ (объем минимальной гиперсферы с центром в точке $x$ и содержащей $k$ других точек)

$p(x) = \frac{k}{V_k(x)l}$ – оценка значимости точки $x$

$Z_n = \{x_j, j = 1 \dots n\}$ – множество вершин

$U_n = \{(x_i, x_j): d(x_i, x_j) \leq d_k(x_i), i \neq j\}$

$G(Z_n, U_n)$ – граф отношений (similarity relation graph)

$G(Z_q, U_q)$ – подграф графа $G(Z_n, U_n)$ на $Z_q = \{x_j, j = 1 \dots q\}$, $U_q = \{(x_i, x_j): x_i \in Z_q, x_j \in Z_q, (x_i, x_j) \in U_n\}$

Алгоритм Уишарта состоит из шагов:

1. Подсчёт $d_k(x)$ для всех точек набора данных
2. Сортировка набора данных в порядке возрастания $d_k(x)$.
3. Начало прохода по всем точкам: $q = 1$.
4. Рассмотрение подграфа $G(Z_q, U_q)$:

– Если $x_q$ оказывается изолированной вершиной, то создаётся новый кластер;

– Если $x_q$ соединена только с точками, которые принадлежат кластеру $l$, если этот кластер «закончен», то $x_q$ помечается шумом, иначе $x_q$ относится к кластеру $l$;

[8] Набор векторных вложений, построенный по описанной выше процедуре.

– Если $x_q$ соединена с точками, которые принадлежат кластерам $l_1, l_2, \dots, l_p, p > 1$ и все кластера «закончены», то точка $x_q$ помечается шумом;
– Если $x_q$ соединена с точками, которые принадлежат кластерам $l_1, l_2, \dots, l_p, p > 1$, но не все кластеры «закончены», то подсчитывается $n(h)$ – количество значимых кластеров при уровне значимости $h$. (Кластер $c$ будет значим, если $\max\limits_{x_i, x_j \in c} |p(x_i) - p(x_j)| \geq h$), если $n(h) > 1$ или среди $l_1, l_2, \dots, l_p$ есть шумовой кластер, то $x_q$ помечается шумом, все значимые кластеры помечаются «законченными», а все незначимые кластеры становятся шумом, а иначе кластеры $l_1, l_2, \dots, l_p$ объединяются в один и $x_q$ тоже к нему относится;

5. $q = q + 1$ и переход к шагу 4.
6. После обработки всех точек проверяются не «законченные» кластеры – если они значимы, то помечаются «законченными», иначе становятся шумом.

*Алгоритм Зана.*

Вторым задействованным подходом являются алгоритмы графовой кластеризации. Основным недостатком такого подхода обычно называют [42] низкую скорость работы, связанную с необходимостью построения специфических графов данных. В настоящем исследовании неизбежно конструирование остовного дерева наименьшего веса на этапе подсчёта внутренних размерностей, что позволяет обратиться к иерархическим алгоритмам, эксплуатирующим эту графовую структуру.

В сводке Ágnes и Fogarassy [42] авторы предоставляют две вариации алгоритма кластеризации остовным деревом: чёткую и нечёткую. Алгоритм Зана, использованный в данной работе, представляет класс чётких алгоритмов и по совместительству служит начальным приближением для нечёткого алгоритма Гато–Гева.

Основная идея обоих алгоритмов заключается в построении на полном графе множества точек $\{x_\ell\}_{\ell \in \mathbb{N}}$ с взвешенными метрикой близости рёбрами остовного дерева минимального веса $T(\{x_\ell\}_{\ell \in \mathbb{N}})$ с последующим удалением наиболее тяжёлых рёбер до выполнения критерия останова. Критерий определяется набором условий:

- алгоритм работает до тех пор, пока количество компонент связности (кластеров) в графе не достигнет значения заданного пользователем гиперпараметра и до тех пор, пока выполняется хотя бы одно из выбранных пользователем на этапе инициализации вспомогательных условий {criterion$_1$, criterion$_2$, criterion$_3$};

- наитяжелейшее ребро $e_i$ удаляется, если его вес больше в $a_1$ раз среднего веса ребра внутри кластера $E_j$, $|E_j| = K'$:

$$\text{criterion}_1 = |e_i| > \frac{a_1}{K-1} \sum_{e \neq e_i \in E_j} w(e); \quad (4.2.1)$$

- наитяжелейшее ребро $e_i$ удаляется, если его вес больше в $a_2$ раз среднего веса рёбер, соединяющих вершины $\{x_j\}_{j\in\mathbb{N}}$, лежащие в некоторой $\varepsilon$-окрестности вершин ребра $e_i$:

$$\text{criterion}_2 = |e_i| > \frac{a_2}{j} \sum_{e \in G(\{x_j\}_{j\in\mathbb{N}}) \subseteq T} w(e); \quad (4.2.2)$$

- для каждого сформированного кластера рассчитывается гиперобъём $V$, отбирается кластер $E_j$, $|E_j| = K$ с наибольшим объёмом $V_j$, удаляется наитяжелейшее ребро $e_i$, устранение которого понизит объём до порогового значения $a_3$:

$$\text{criterion}_3 = V_{j;1\ldots i-1}^{left} + V_{j;i+1\ldots K-1}^{right} \leq a_3; \quad (4.2.3)$$

Алгоритм естественным образом распространяется на пространство вложений языка и позволяет кластеризовать множество $n$-грамм. Метрикой близости в данном случае выступает Евклидово расстояние между $n$-граммами.

4.3. Энтропия-сложность

Текст может быть представлен временным рядом. Таким образом, тексты можно также исследовать методами из теории информации. В работе [43] предложен метод, позволяющий отличать хаотические ряды, с одной стороны, от простых детерминированных процессов, с другой стороны, от стохастических процессов[9]. Методы базируется на понятии порядковых паттернов (*ordinal patterns*) и для каждой такой порядковой структуры оценивается вероятность её появления, как частота её появления.

Для полученного распределения вычисляются характеристики из теории информации – энтропия Шеннона и сложность:

$$C[P] = D_{JS}[P, P_U] \cdot S[P], \quad (4.3.1)$$

где $P_U$ – равномерное распределение, $D_{JS}$ – дивергенция Дженсона-Шеннона, $S[P]$ – энтропия.

Данные метрики соотносят текст с точкой на плоскости, при этом её расположение относительно теоретических границ определяет тип ряда.

[9] В оригинальной работе рассматривается метод для одномерного временного ряда. Здесь используется его очевидное обобщение на случай многомерных временных рядов.

Каждому тексту, данному своей семантической траекторией, сопоставляется точка на плоскости «энтропия-сложность». Расположение точки относительно теоретических границ на указанной плоскости определяет тип ряда. Точки, соответствующие стохастическим процессам, расположены ближе к нижнему правому углу (высокая сложность и низкая энтропия), детерминированные процессы расположены в левом нижнем углу (низкие сложность и энтропия), хаотические процессы дают точки, близкие к вершине кривой верхней теоретической границы [43].

4.4. Алгоритмы оценки внутренней размерности

Определить понятие внутренней размерности заданного множества можно различными способами, взятыми из совокупности математических дисциплин [27]. В дальнейшем мы будем использовать следующие определения:

1. Топологическая размерность (размерность Лебега) метрического пространства [44] $(X;\rho)$, с заданной на нём функцией расстояния $\rho$, определяется как: $d_T = \inf\{i \in \mathbb{N} | \forall \epsilon > 0\ \exists C(X;\epsilon, i+1)\}$; $C(X;\epsilon, i+1)$ – конечное открытое $(i+1)$-кратное $\epsilon$ −покрытие для $X$.
2. Хаусдорфова размерность [45] задана как: $\sup\{\beta : \sup m(\epsilon,\beta) > 0 \,|\epsilon > 0\}$; $m(\epsilon,\beta) = \inf \sum_{\{A_i\}\in C(X;\epsilon)} (\operatorname{diam} A_i)^{\beta}$, $\operatorname{diam} A_i < \epsilon$.

Первое определение, в отличие от второго, не принимает возможным нецелое значение размерности[10] геометрического объекта.

Одной из ключевых целей настоящей работы является оценка внутренней размерности языковых объектов. Согласно Пестову [24] внутренняя размерность $d^*$ – это отображение $d^*(X;\rho,V) \to \mathbb{R}$ пространства $X$ с заданными на нём метрикой $\rho$ и мерой $V$, удовлетворяющее аксиомам концентрации меры, «плавной» зависимости от наборов данных и нормализации[11]. Таким образом, для каждого изучаемого «языкового» множества мы пытаемся оценить его топологическую и хаусдорфову размерности.

Для оценки внутренней размерности набора слов или $n$-грамм исследуемого[12] естественного языка мы применяем два подхода. Оба метода основаны на построении для рассматриваемой выборки графа данных и по-

[10] При применении к объекту с нецелой размерностью алгоритмы оценки обычно округляют значение до ближайшего целого числа.

[11] Аксиомы: (1) концентрации: для семейства пространств $(X_d)$ вида $(X;\rho,V)$ $\partial(X_d) \uparrow \infty$ тогда и только тогда, когда $(X_d)$ образует семейство Леви; (2) «плавной» зависимости от наборов данных: если метрика М. Громова $d_{conc}(X_d, X) \to 0$, тогда $\partial(X_d) \to \partial(X)$; 3) нормализации: $\partial(\mathbb{S}^d) = \Theta(d)$, $\mathbb{S}^d$ – $d$-мерная единичная гиперсфера.

[12] Набор векторных вложений, построенный по описанной выше процедуре.

следующем извлечении ее остовного дерева наименьшего веса (MST) [28, 30].

Причина использования алгоритмов на основе графов связана, прежде всего, с их робастностью [28]: для поиска локальных и глобальных паттернов таким алгоритмам достаточно определения частичного порядка соседей для каждой вершины, что значительно снижает влияние шума на выход алгоритма. Этот факт наиболее важен с точки зрения области применения: данные, собранные из Интернета, могут быть существенно зашумлены даже после стандартных методов предварительной обработки текста, не говоря уже о том, что внутренняя структура языка как системы практически не изучена, что существенно усложняет отделение артефактов, порождённых шумом, от важных структурных паттернов. Также следует отметить, что рассматриваемые нами алгоритмы являются непараметрическими, что снимает с исследования одну из наиболее трудоёмких задач поиска оптимальных гиперпараметров.

*Алгоритм Швайнхарта.* Первый из методов, описанных в данном разделе, использует теорему Швайнхарта [30], являющуюся обобщением теоремы Стила [46]:

***Теорема (Швайнхарт).*** *Пусть $V$ – $d^*$-Альфорсова регулярная мера, заданная на метрическом пространстве, и пусть $\{x_\ell\}_{\ell\in\mathbb{N}}$ – независимые и одинаково распределённые примеры из $V$. Пусть также на этих примерах определена величина $E_\alpha^0$:*

$$E_\alpha^0(x_1,\dots,x_\ell)=\sum_{e\in T_\ell(\{x_\ell\}_{\ell\in\mathbb{N}})}|e|^\alpha. \tag{4.4.1}$$

*Здесь $|e|, e\in T$ – весовой коэффициент, рассчитанный как евклидово расстояние между двумя вершинами ребра $e$, принадлежащего остовному дереву наименьшего веса $T(x_1,\dots,x_\ell)$. Тогда, если $0<\alpha<d^*$, то:*

$$\mathrm{const}_1\le\frac{E_\alpha^0(x_1,\dots,x_\ell)}{\ell^{1-\alpha/d^*}}\le\mathrm{const}_2, \tag{4.4.2}$$

$$\frac{\ln\left(E_\alpha^0(x_1,\dots,x_\ell)\right)}{\ln(\ell)}\to\frac{d^*-\alpha}{d^*} \tag{4.4.3}$$

*при $\ell\to\infty$; $\mathrm{const}_1$ и $\mathrm{const}_2$ – положительные, не зависящие от $\ell$ константы.*

Теорема позволяет оценить (фрактальную) размерность $\hat{d}_{\mathrm{Schw}}$ геометрического объекта, используя выборку его точек $\{x_\ell\}_{\ell\in\mathbb{N}}, x_\ell\in\mathbb{R}^d$: для этого вычисляются значения $E_\alpha^0$ для различных $\ell$, формирующие выборку для решения следующей задачи линеаризованной регрессии:

$$\ln\left(E_\alpha^0(x_1,\dots,x_\ell)\right)=\ln(\psi(\alpha,d^*))+\ln(\ell)\frac{d^*-\alpha}{d^*}, \tag{4.4.4}$$

Для решения задачи используется нелинейный метод наименьших квадратов с целью оценки размерности Хаусдорфа как верхнего предела размерности Минковского ([30], стр. 3), используя приближение $\frac{d^*-\alpha}{d^*} \approx \frac{\hat{d}_{\mathrm{Schw}}-\alpha}{\hat{d}_{\mathrm{Schw}}}$.

*Алгоритм Брито.* Второй подход [28], применённый нами для оценки (топологической) внутренней размерности, опирается на свойства случайных величин, порождаемых графами данных, таких как «$k$-ближайших соседей» ($k$-*nearest neighbours*) и «сферы взаимовлияния» (*spheres of influence*). В нашем случае использовалась сумма квадратов степеней вершин остовного дерева наименьшего веса. Многочисленные эксперименты показали, что данная статистика устойчива[13]: добавление шумовых компонент и (или) увеличение размерности пространства вложений $d$ существенно не меняют результирующие оценки (см. приложение В).

Далее краткое описание алгоритма: для выборки точек $\{x_\ell\}_{\ell\in\mathbb{N}}$ $d$-мерного пространства $(X_d, \rho, V)$ с ассоциированными на нём метрикой $\rho$ и вероятностной мерой $V$:

1. Инициализируется полный граф евклидовых расстояний и его остовное дерево наименьшего веса $T(x_1, \ldots, x_\ell)$ с целью расчёта статистики[14]:
$$M_\ell(d) := \frac{1}{\ell} \sum_{x_i \in T(\{x_\ell\}_{\ell\in\mathbb{N}} \subset X_d)} \big(\deg(x_i)\big)^2, \qquad (4.4.5)$$
2. Оцениваются параметры[15] распределения $f_{prior}(M_\ell | d = i) \sim N(\hat{\mu}_i, \hat{\sigma}_i^2)$ методом Монте-Карло:

---

[13] Брито и др. также предложили две другие статистики для оценки внутренней размерности [28]. К сожалению, эти две случайные величины оказались гораздо менее надёжными для наших данных.

[14] Брито и др. [28] (а также Стил и др. [47]) показали, что математическое ожидание степени вершины для остовного дерева наименьшего веса остаётся постоянным независимо от количества узлов в графе данных. В частности, когда данные выбираются из непрерывного распределения на $\mathbb{R}^d$, доля узлов, степень которых равна $j$, почти наверное сходится к пределу, зависящему исключительно от $j$ и $d$. Кроме того, Брито и др. [28] продемонстрировали монотонные изменения, происходящие в остовном дереве при увеличении внутренней размерности: и число листьев (вершин степени = 1), и число вершин с большими степенями, демонстрируют устойчивый рост. Это говорит о том, что среднее значение степеней вершин остовного дерева наименьшего веса, возведённое в степень больше 1, может быть рассчитано для выявления различий между топологическими размерностями.

$$\hat{\mu}_i \coloneqq \frac{1}{K}\sum_{j=0}^{K} M_\ell\left(i;\{x_\ell\}_j{\sim}U(\mathbf{0_i},\mathbf{1_i})\right), \tag{4.4.6}$$

$$\hat{\sigma}_i^2 \coloneqq \frac{\ell}{K-1}\sum_{j=0}^{K}\left(M_\ell\left(i;\{x_\ell\}_j{\sim}U(\mathbf{0_i},\mathbf{1_i})\right)-\hat{\mu}_i\right)^2. \tag{4.4.7}$$

Здесь $K$ – количество выборок, по $\ell$ точек в каждой, взятых из $i$-мерного единичного гиперкуба.

3. Рассчитываются значения функции правдоподобия относительно эквивалентности между желаемой оценкой, $\hat{d}$, и каждым предполагаемым кандидатом, значением истинной топологической размерности, $i$, посредством использования теоремы Байеса для вычисленной на реальных данных $M'_\ell$:

$$p\{\hat{d}=i\} = p(i|M'_\ell) = \frac{N\left(M'_\ell;\hat{\mu}_i,\frac{\hat{\sigma}_i^2}{\ell'}\right)}{\sum_j N\left(M'_\ell;\hat{\mu}_j,\frac{\hat{\sigma}_j^2}{\ell'}\right)}, \tag{4.4.8}$$

Таким образом, получается дискретное условное распределение вероятностей относительно эквивалентности размерности Лебега каждому кандидату.

4. Вычисляется оценка внутренней (топологической) размерности $\hat{d}_{BQY}$ как математическое ожидание $E[\hat{d}]$, округлённое до ближайшего целого числа:

$$\hat{d}_{BQY}(M'_n) = \text{round}\left(\sum_i p(i|M'_\ell) * i\right). \tag{4.4.9}$$

Брито и др. доказали ([28], стр. 276-277), что $\hat{d}_{BQY} \to d_T$, почти наверное, при $K, \ell, \ell' \to \infty$. За $\hat{d}_{BQY}$ и $d_T$ обозначены оценка алгоритма и подлинная топологическая размерность, соответственно.

Для валидации полученных оценок мы использовали метод, основанный на анализе формальных понятий (FCA) [48;49]. Авторы подхода отмечают, что его оценки для данных, порождаемых системами реального мира, довольно грубы; тем не менее, они могут быть использованы для контроля порядка оценок методов, описанных в данном разделе.

4.5. Топологический анализ данных

---

[15] Доказано, что последовательность случайных величин $M_\ell$ сходится при $\ell \to \infty$ к случайной величине с нормальным распределением (Брито и др. [28], стр. 268). Алгоритм использует это как априорное знание о характеристиках моделируемого распределения.

При изучении структуры векторного пространства важно изучить его топологию – геометрические отношения между объектами, составляющими пространство. Топологический анализ данных предоставляет большое множество инструментов для изучения формы и структуры сложных данных. В данной работе особое внимание уделяется изучению данных с различными размерностями, поэтому важно учитывать изменение размерности при исследовании топологических признаков. Таким образом, необходимо находить персистентные гомологии (*persistent homology*), то есть уникальные гомологии в структуре векторного пространства, которые при этом сохраняются при изменении масштабов анализа.

В построенных семантических пространствах для русского и белорусского языков производится поиск гомологий первого порядка $H_1$, так называемые «дырки» в пространстве (гомологии нулевого порядка $H_0$ – компоненты связности). Таким образом производится попытка нахождения «слепых зон» естественного языка – областей с малоупотребляемыми языковыми единицами (слова/биграммы и т.д.).

Для нахождения персистентных гомологий над векторным пространством строится фильтрация Вьеториса–Рипса.

***Определение***. *Комплекс Вьеториса–Рипса* диаметра $\epsilon$ – это симплициальный комплекс

$$VR(\epsilon) = \{\sigma | diam(\sigma) \leq \epsilon\}, \quad (4.5.1)$$

***Определение***. Возрастающая последовательность $\epsilon_1, \epsilon_2, \ldots$ порождает *фильтрацию Вьеториса–Рипса* – последовательность увеличивающихся симплициальных комплексов $VR(\epsilon_1) \subseteq VR(\epsilon_2) \ldots$ .

По фильтрации Вьеториса–Рипса, изменяя диаметры комплексов, мы отслеживаем моменты «рождения» и «смерти» разных классов гомологии и выделяем наиболее продолжительные.

При изучении структуры естественного языка также встаёт вопрос о границах найденных «дыр», т.е. какие слова/словосочетания их оконтуривают. Используется метод, описанный в статье [50], для поиска *homology representatives* – элементов, по которым можно восстановить гомологию. Данный метод позволяет выделить цепочки объектов, “обводящих «дырки»” в векторном пространстве.

## 5 Эксперименты

Во всех последующих экспериментах для получения векторных вложений юни- и биграмм русского и белорусского языков мы формируем корпусы национальной литературы с помощью текстов, доступных для скачивания из открытых источников.

К каждому тексту, входящему в любой из корпусов, мы применяем алгоритм предобработки, включающий:

– удаление стоп-слов (артикли, союзы, ругательства, вводные слова и т.п.);

– токенизацию (преобразование имён собственных, числительных и т.п. в соответствующие метки классов);

– лемматизацию (выделение начальных форм слов);

### 5.1 Кластеризация русского и белорусского языков

В результате экспериментов по сравнению алгоритмов кластеризации был выбран алгоритм Уишарта в силу необходимости работать с чрезвычайно большими выборками данных. Для подбора гиперпараметров алгоритма использовался индекс Калинского-Харабаза. После получения кластеризаций датасетов векторов $n$-грамм, для каждого кластера были подсчитаны по 8 метрик, их описания и формулы представлены в табл. 1.

***Обозначения:***

$NC$ – количество кластеров

$n_i$ – количество точек в кластере $i$

$\widetilde{n_i}$ – общее количество точек (с учётом повторов) в кластере $i$

$C_i$ – кластер $i$

$c_i$ – центр кластера $i$

Таблица 1. Метрики, считающиеся для каждого кластера.

| | |
|---|---|
| Количество уникальных векторов в кластере. Нормирование по размеру самого большого кластера (размер – количество уникальных векторов) | $\frac{n_i}{max_{j=1}^{NC} n_j}$ |
| Количество уникальных векторов в кластере. Нормирование по общему количеству уникальных векторов | $\frac{n_i}{\Sigma_{j=1}^{NC} n_j}$ |
| Количество векторов с повторениями в кластере. Нормирование по размеру самого большого кластера (размер – количество векторов с повторениями) | $\frac{\widetilde{n_i}}{\Sigma_{j=1}^{NC} \widetilde{n_j}}$ |
| Количество векторов с повторениями в кластере. Нормирование по общему количеству векторов с повторениями | $\frac{\widetilde{n_i}}{max_{j=1}^{NC} \widetilde{n_j}}$ |
| Максимальное расстояние до центроида | $max_{x \in C_i} d(x, c_i)$ |
| Среднее расстояние до центроида | $\frac{\Sigma_{x \in C_i} d(x, c_i)}{n_i}$ |
| Максимальное расстояние между точками в кластере | $max_{x,y \in C_i, x \neq y} d(x, y)$ |
| Среднее расстояние между точками в кластере | $\frac{\Sigma_{x,y \in C_i, x \neq y} d(x, y)}{n_i(n_i - 1)}$ |

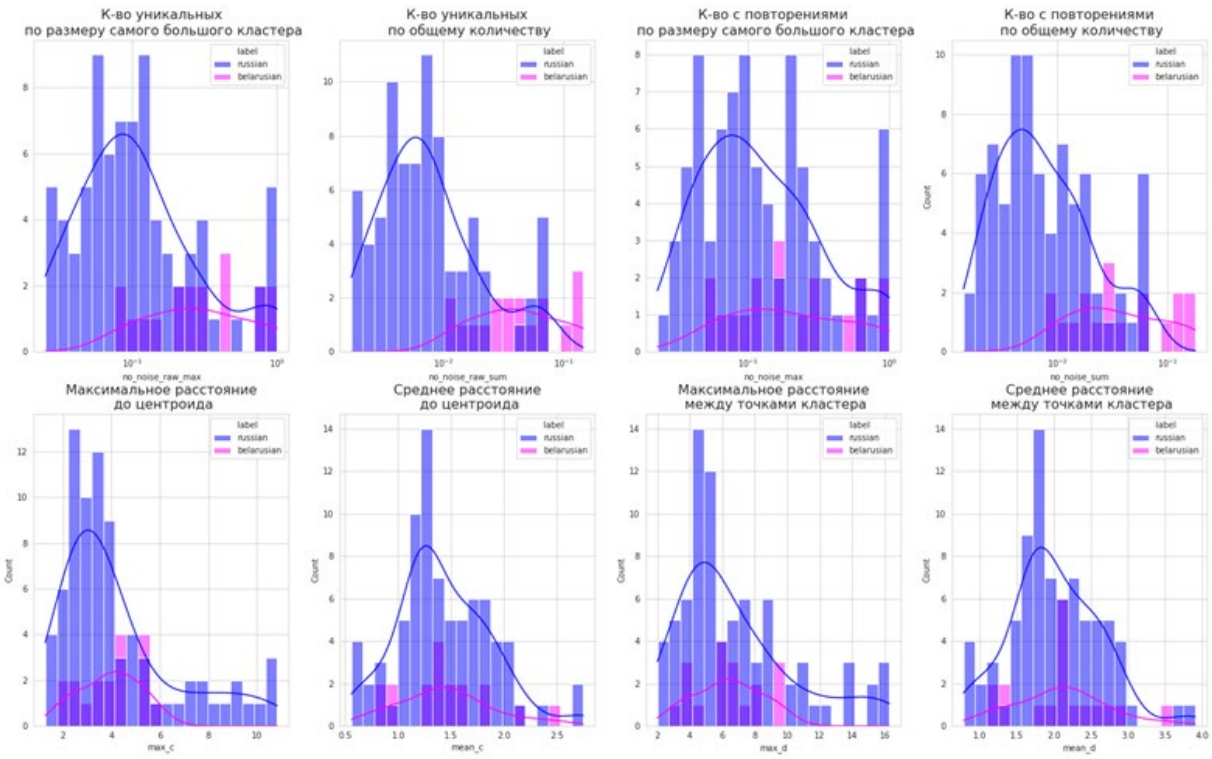


Рис. 1. Гистограммы метрик кластеров векторных вложений CBOW

В результате для кластеризаций обоих языков было получено по 8 массивов, содержащих метрики. Протестированы гипотезы о том, что выборки метрик происходят из одного распределения с использованием критерия Уилкоксона. В результате тестировалось 8 гипотез и была сделана поправка на множественное тестирование методом Хольма–Бонферрони (так как высока вероятность ошибки первого рода). Уровень значимости установлен 5%.

При использовании эмбеддингов CBOW ($d = 8, n = 2$) для русского получилось 80 кластеров, для белорусского – 18 кластеров. Гистограммы полученных метрик представлены на рис. 1.

После поправки на множественное тестирование было выявлено 3 (из 8) статистически значимых различий. Таким образом, можно сказать, что для кластерных структур русского и белорусского выделены статистически значимые различия.

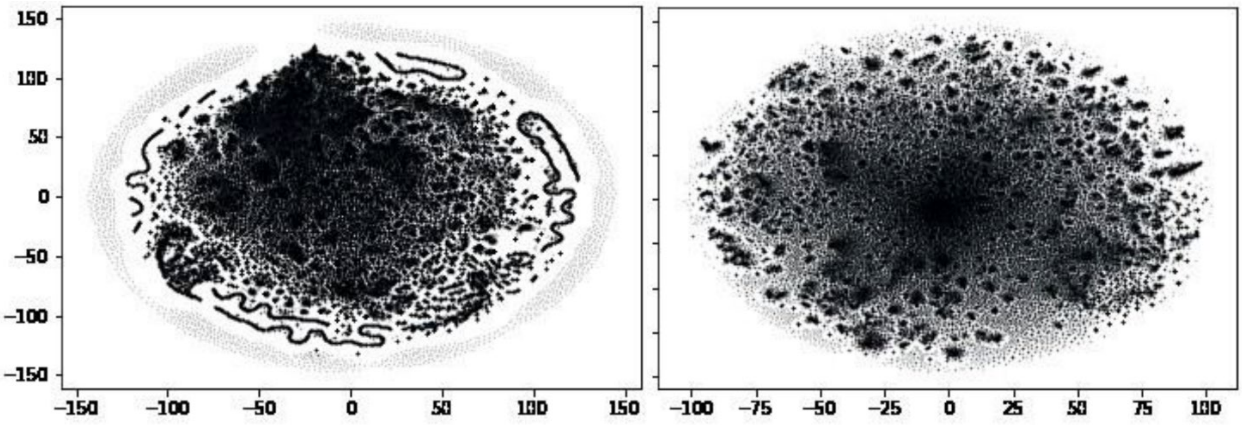


Рис. 2. t-SNE визуализация языковых пространств для русского (слева) и белорусского (справа) языков

5.2. Анализ позиций на плоскости «энтропия-сложность» для русского и белорусского языков

Мы полагаем, что меньшие $d$ и $n$, для которых большинство текстов, по-видимому, принадлежат к области чисто случайных процессов, больше не отражают природу текстов. Мы можем рассматривать эти значения $d$ и $n$ как нижнюю границу, так что семантические траектории для больших $d$ и $n$ действительно отражают соответствующий текст. С другой стороны, поскольку траектории литературных произведений ограничены по размеру, мы не можем адекватно оценить энтропию и сложность для очень больших значений $d$ и $n$ – это определяет верхнюю границу.

На рис. 3 приведены значения $d$, $n$, позволяющие анализировать семантические траектории русского и белорусского языков. Синяя область соответствует допустимым значениям $d$ и $n$; для значений над оранжевой линией тексты попадают в детерминированную область; для значений ниже синей линии – шум. Заметим, что при фиксированных значениях $d$ диапазоны допустимых значений $n$ для белорусского языка выше, чем для русского языка. Для $d = 1$ минимально допустимое число слов в белорусском языке – 8, тогда как диапазон допустимых $n$ для русского языка – от 5 до 8. Для $d = 2$ диапазон $n$ для белорусского языка – 5 до 7, для русского – от 4 до 5. Более подробно вопрос о допустимых значениях для разных языков рассмотрен в статье [6].

5.3. Внутренние размерности русского и белорусского языков

Для оценки внутренней размерности языковых фрактальных структур мы приняли значения размерности пространства вложения $d = \{5, 10, 15\}$. Важным ориентиром для нас стали исследования по применению метода анализа пар энтропия-сложность [6]: согласно результатам статьи, при $n = 2$ допустимые значения $d$ не превышают $15$.

При использовании алгоритма Швайнхарта, для каждого языка мы:

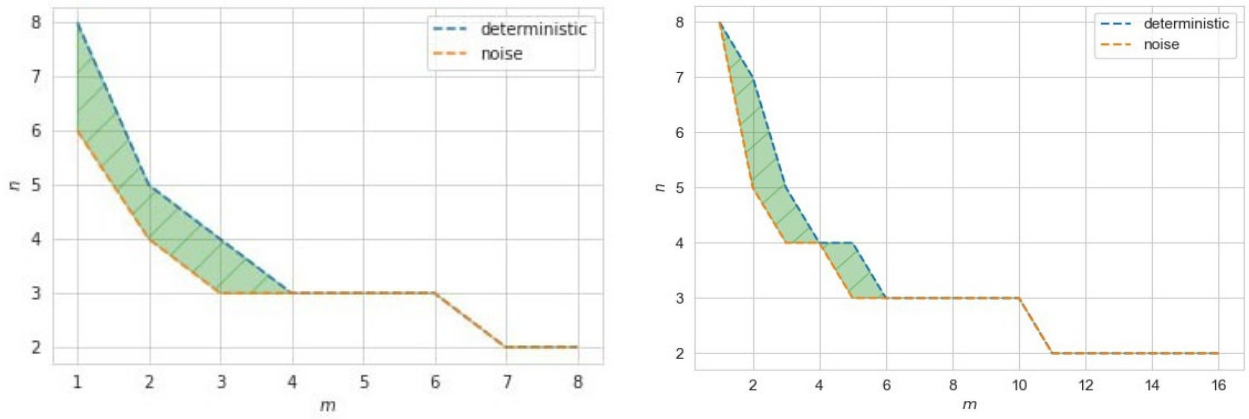


Рис. 3. Допустимые значения параметров $n$ и $d(m)$ для русского (слева) и белорусского (справа) языков

1) формируем последовательность выборок $n$-грамм, варьируя их объём $\ell$ от $100{,}000$ $n$-грамм до объёма полного датасета;
2) используя все выборки, для последовательности $\alpha$ от $1e-4$ до $10$ с шагом $s \approx 1e-1$, мы оцениваем параметры регрессии (4.2.4) и рассчитываем $\hat{d}_{Schw}$;
3) отбрасываем недопустимые $\alpha$ по следующим критериям:
   a. (1) доверительные интервалы для линии регрессии слишком велики (превышают $10\%$);
   b. (2) доверительные интервалы для параметра $\frac{\hat{d}-\alpha}{\hat{d}}$ слишком велики (превышают $10\%$);

При использовании алгоритма Брито, для каждого языка мы:

1) Для $d = i \in 2..15$ генерируем $\{x_1^j, \dots, x_{1e+6}^j\}_{j=1..100}, x_i^j \sim U(\mathbf{0}_i, \mathbf{1}_i)$ выборки, здесь $\mathbf{0}_i, \mathbf{1}_i$ – $i$-мерные нулевой и единичный векторы, соответственно, $i$ – потенциально возможное истинное значение (топологической) внутренней размерности;
2) оцениваем параметры распределений $\hat{\mu}_i$ и $\hat{\sigma}_i^2$ для каждого $i$;
3) формируем последовательность выборок $n$-грамм, варьируя их объём $l$ от $100{,}000$ $n$-грамм до объёма полного датасета;
4) для каждого значения $\ell$ вычисляем оценки $\hat{d}_{BQY}(\{x_1, \dots, x_\ell\})$.

Табл. 2 и 3 резюмируют результаты. На рис. 4 показаны типичные зависимости использованных оценок; для оценок алгоритмом Швайнхарта характерна типичная мультифрактальная зависимость (предлагаем читателю сравнить с рис. А2). Оценки внутренних размерностей представляются весьма надёжными: для различных методов, значений $n$ и $d$, способов по-

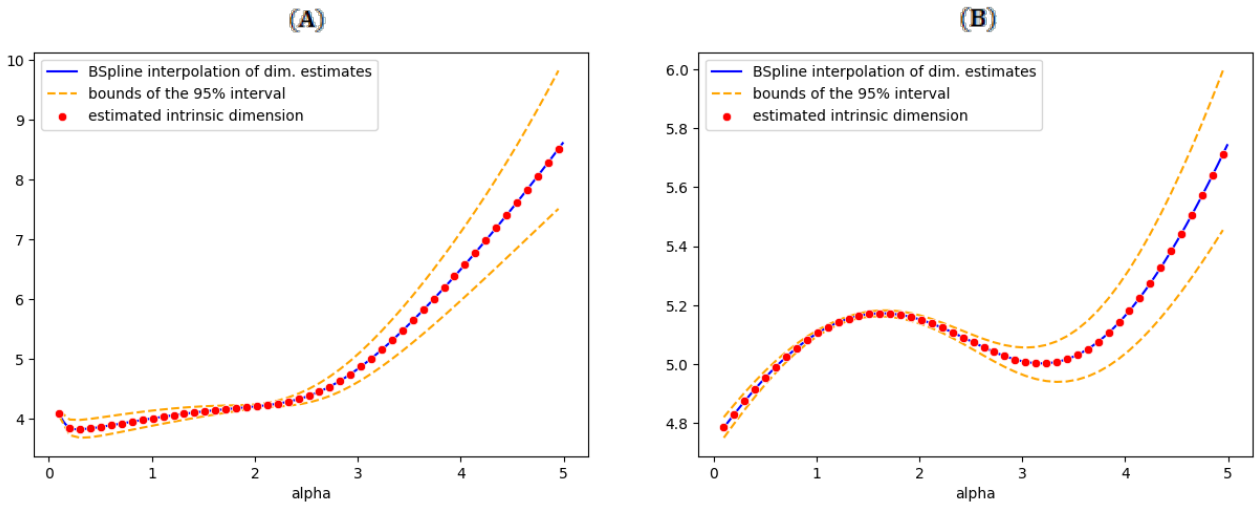


Рис. 4. Зависимость оценок алгоритма Швайнхарта от параметра $\alpha$ для: (A) SVD-представлений русской национальной литературы ($n = 2, d = 5$); (B) SVD-представлений белорусской национальной литературы ($n = 2, d = 5$). Красные точки являются оценками алгоритма, синяя линия представляет собой B-сплайн интерполяцию зависимости оценок от $\alpha$, оранжевый пунктир – верхнюю и нижнюю границы $95\%$ доверительного интервала. Графики для языков идентичны графикам для мультифракталов

лучения векторных представлений (SVD, CBOW) получаются схожие результаты[16].

Это позволяет сделать вывод, что внутренняя размерность языка является характеристикой самого языка и что геометрически язык является мультифракталом. Принадлежность фазового пространства естественного языка к классу мультифракталов позволяет выдвинуть следующие гипотезы:

- язык как динамическая система является странным аттрактором [1;27];
- тексты естественного языка – хаотические ряды [1;6].

Границы верхней оценки внутренней размерности составляют 9 для обоих языков (мы используем алгоритм Брито исключительно для проверки результатов на адекватность). Наши результаты хорошо согласуются с [48].

[16] Результаты для $d = 15$ можно объяснить тем фактом, что имеющиеся выборки недостаточно велики (читатель может сравнить с результатами [13]).

Таблица 2. Оценки внутренней размерности для SVD-представлений естественного языка

| language | $n$ | $d$ | $\hat{d}_{BQY}$ | $\min \hat{d}_{Schw}$ | $\max \hat{d}_{Schw}$ | $\alpha$ |
|---|---|---|---|---|---|---|
| RUSSIAN | 1 | 5 | 5 | 4.65 | 5.52 | 0-4.21 |
| | | 10 | 8 | 7.14 | 8.20 | 0-5.03 |
| | | 15 | 15 | 9.81 | 12.78 | 0-7.33 |
| | 2 | 5 | 5-6 | 5.11 | 7.79 | 0-10 |
| | | 10 | 8-9 | 6.48 | 8.52 | 0-8.88 |
| | | 15 | 13-14 | 8.26 | 10.70 | 0-10 |
| BELORUSIAN | 1 | 5 | 4 | 4.59 | 5.67 | 0.1-1.62 |
| | | 10 | 6-7 | 5.45 | 6.55 | 0.1-1.62 |
| | | 15 | 9-10 | 6.14 | 7.85 | 0.1-1.62 |
| | 2 | 5 | 5 | 4.78 | 5.77 | 0.1-4.79 |
| | | 10 | 7-8 | - | - | - |
| | | 15 | 10-11 | - | - | - |

Таблица 3. Оценки внутренней размерности для CBOW-представлений естественного языка

| language | $n$ | $d$ | $\hat{d}_{BQY}$ | $\min \hat{d}_{Schw}$ | $\max \hat{d}_{Schw}$ | $\alpha$ |
|---|---|---|---|---|---|---|
| RUSSIAN | 1 | 5 | 4-5 | 4.45 | 4.62 | 0-0.72 |
| | | 10 | 5-6 | 5.04 | 7.18 | 0-1.21 |
| | | 15 | 7-8 | 5.98 | 9.57 | 0-1.21 |
| | 2 | 5 | 5 | 3.82 | 6.79 | 0-4.16 |
| | | 10 | 5 | 4.85 | 9.43 | 0-8.28 |
| | | 15 | 9-10 | 5.73 | 10.51 | 0-8.79 |
| BELORUSIAN | 1 | 5 | 4 | 4.17 | 4.92 | 0-0.97 |
| | | 10 | 5 | 5.32 | 6.74 | 0-1.77 |
| | | 15 | 7 | 6.27 | 9.21 | 0-1.77 |
| | 2 | 5 | 5 | 3.21 | 4.76 | 0-2.65 |
| | | 10 | 7 | 4.47 | 8.63 | 0-8.89 |
| | | 15 | 9 | 5.11 | 10.09 | 0-10 |

5.4. Анализ топологии русского и белорусского языков

На рис. 5-6 изображены выделенные наиболее устойчивые гомологии первого порядка в семантическом пространстве русского и белорусского языков, соответственно (SVD, $d = 2$).

В следующем списке мы предоставляем слова, находящиеся на границах выделенных «дыр» (по убыванию их персистентности):

**Русский:**

– любовь, писать, один, сад, убийство, над, где, дом;

– птица, мальчик, спросить, отказать, какой, нибудь;
– пора, Павлович, солнце, идти, мой, зеркало;
– нет, государь, страх, тот, карета, бывать, театр, сильный, охота, рассказ, отец, ни, барышня;
– сам, благодарить, советовать, танцевать, чай, сделать, Бог;
– то, квартальный, император, христианский, генерал, песня, гора;
– студент, медведь, интеллигенция, где, граф, брат;
– мама, товарищ, писать, университет, ну;
– квартальный, стол, гость, это, палец, христианский, император;
– гость, власть, мать, опять, божий, работать, июль, государство, с, деньга, стол;

**Белорусский**[17]:

– хацець, год, мець, добры, галава, бачыць, дзень;
– першы, іншы, жыццё, зрабіць, зусім;
– дзіця, павінны, дзверы, горад, пакуль, зразумець, зараз, ляжаць;
– няма, дарога, даць, таму, месца, вада, адразу, пайсьці, чакаць;
– браць, ніякі, пытанне, край;
– тады, раз, галава, добры, стаяць, больш, зямля, пачаць, зноў;
– сувязь, сустрэцца, дазволіць, пачатак, ходзіць, уздыхнуць, птушка, чуцца, задумацца;
– даць, дарога, няма, пайсьці;
– цяжкі, павінны, чорны, малады, далей, жанчына;
– недзе, нельга, падаць, патрэбны, бераг, галоўны, вецер, партызан, лепш, паспець, пасля, кніга, прыхадзіць, усміхнуцца, воля, старацца, спаць, забіць, выпадак, загадаць, салдат, мабыць, многа, моцна;

**Заключение**

В настоящей работе представлен новый объект исследования – языковая фрактальная структура; мы предполагаем, что множество всех $n$-грамм векторных вложений естественного языка представляет собой репрезентативную выборку этого фрактального множества.

Мы кластеризовали семантические пространства русского и белорусского языков и получили существенные качественные различия в структуре кластеров.

Мы рассмотрели тексты русского и белорусского языков как временные ряды и посчитали метрики энтропии и сложности. Были вычислены допустимые значения параметров $d$ (размерность векторного представления слова) и $n$ (число слов в $n$-грамме), при которых тексты попадают в область хаотических рядов. При фиксированных значениях $d$ диапазоны

[17] В приложении С представлен перевод на русский язык.

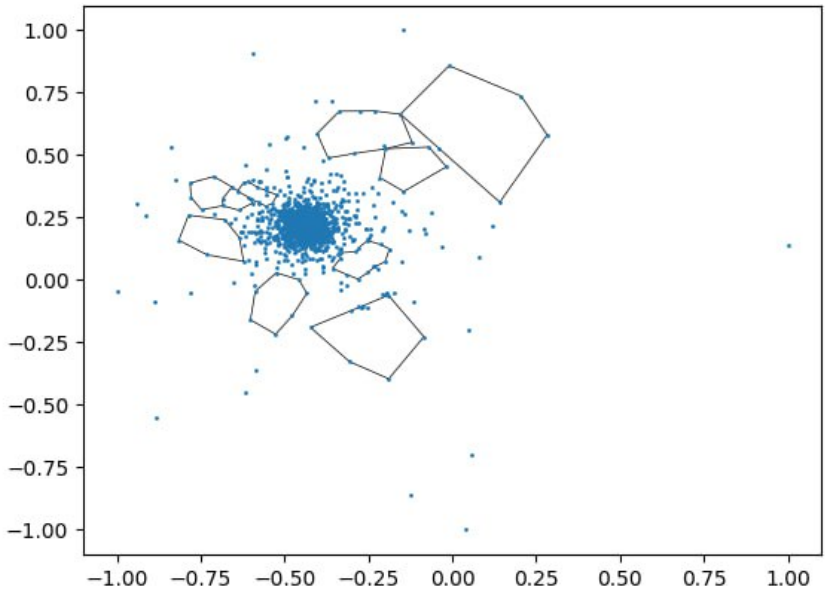


Рис. 5. Гомологии первого порядка в семантическом пространстве слов русского языка.

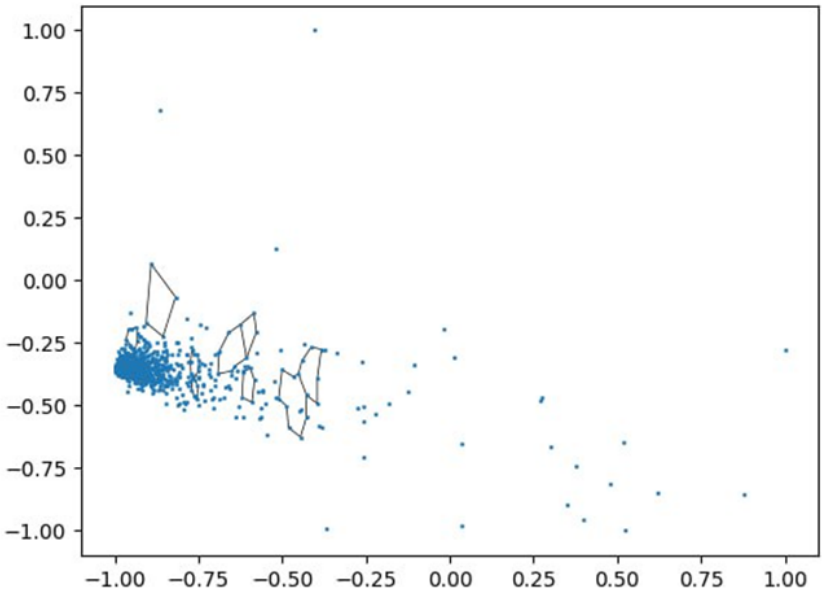


Рис. 6. Гомологии первого порядка в семантическом пространстве слов белорусского языка.

допустимых значений $n$ для белорусского языка выше, чем диапазоны для русского языка.

Мы установили, что для русского и белорусского языков языковые структуры обладают спектрами нецелочисленных внутренних размерностей и, следовательно, являются мультифракталами. Для белорусского языка верхние границы этих размерностей приближаются к $9$ для юнограмм и $10$ для биграмм; для русского языка, аналогично, к $9$ для юни-

грамм, и 10 для биграмм. Таким образом, для естественных языков внутренняя размерность оказывается совсем небольшой.

Методами топологического анализа данных мы исследовали структуру векторного пространства слов русского и белорусского языков и выделили в них гомологии первого порядка – «дырки»-области с малоупотребляемыми словами.

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 году.

**Литература**

1. *Gromov V.A., Migrina A.M.* A language as a self-organized critical system // Complexity 9212538 (2017).
2. *Garg M., Gupta A.K., Prasad R. (eds.)* Graph learning and network science for natural language processing. – CRC Press, 2022.
3. *Garg M., Kumar M., Samanta D.* Towards pattern recognition with network science and natural language processing for information retrieval // Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods ICPRAM. – 1, 495-500 (2023).
4. *Garg M., Kumar M.* The structure of word co-occurrence network for microblogs // Physica A, 512, 698-720 (2018).
5. *Markovič R., Gosak R. et al.* Applying network theory to fables: complexity in slovene belles-lettres for different age groups // Complex Networks 7, 114-127 (2018).
6. *Gromov V.A., Dang Q.N.* Semantic and sentiment trajectories of literary masterpieces // Chaos, Solitons and Fractals 175(1), 113934 (2023).
7. *Allott N., Lohndal T., Rey G.* Synoptic introduction // A companion to Chomsky / 1st ed. – New Jersey: Wiley Blackwell; 2021. P.1-17.
8. *Dębowski Ł.* A simplistic model of neural scaling laws: Multiperiodic Santa Fe processes. arxiv:2302.09049
9. *Tanaka-Ishii K.* Statistical universals of language / 1st ed. – Cham: Springer; 2021.
10. *Ferrara E., Kudugunta S.* Deep neural networks for bot detection. 2018. Information Sciences 467, 312-322 (2018).
11. *Wei F., Nguyen T.* Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. arxiv:2002.01336
12. Najari S., Salehi M., Farahbakhsh R. GANBOT: A GAN-based framework for social bot detection // Soc. Netw. Anal. Min. 12, 4 (2022).
13. *Garcia-Silva A., Berrio C., Gomez-Perez J.M.* An empirical study on pre-trained embeddings and language models for // Proceedings of the 4th Workshop on Representation Learning for NLP: 148–155 (2019).
14. *Zhang J., Dong B., Yu P.S.* FAKEDETECTOR: Effective fake news detection with deep diffusive neural network. arxiv:1805.08751

15. *Garcia-Silva A., Berrio C.* Understanding transformers for bot detection in twitter. arxiv:2104.06182
16. *Feng S., Wan H. et al.* BotRGCN: Twitter bot detection with relational graph convolutional networks. arxiv:2106.13092
17. *Zhao J., Liu X. et al.* Multi-attributed heterogeneous graph convolutional network for bot detection //Information Sciences 537, 380-393 (2020).
18. *Guo Q., Xie H. et al.* Social bots detection via fusing BERT and graph convolutional networks //Symmetry 14(1), 30 (2022).
19. *Bhatt P., Rios A.* Detecting bot-generated text by characterizing linguistic. arxiv:2106.01170
20. *Bonilla R.F.* Spot the bot: Motifs in semantic spaces. 2020.
21. *Бритков Р.А.* Выявления ботов в социальных сетях: Выделение характерных последовательностей в семантическом пространстве. 2019.
22. *Гайсин А.Г.* Кластеризация характерных последовательностей семантических траекторий. 2019.
23. *Chu Z., Gianvecchio S. et al.* Detecting automation of twitter accounts: Are you a human, bot, or cyborg? // IEEE Transactions on Dependable and Secure Computing 9(6), 811-824 (2012).
24. Pestov V. Intrinsic dimension of a dataset: What properties does one expect? // International Joint Conference on Neural Networks, Orlando, FL, USA, 2007. P.2959-2964.
25. *Gromov M.* Metric structures for Riemannian and non-Riemannian spaces / 1st ed. – Boston: Birkhäuser Boston, 2007.
26. *Kantz H, Schreiber T.* Nonlinear time series analysis / 2nd ed. – Cambridge: Cambridge University Press, 2003.
27. *Малинецкий Г.Г., Потапов А.Б.* Современные проблемы нелинейной динамики. – М.: Эдиториал УРСС, 2000.
28. *Brito M.R., Quiroz A.J., Yukich J.E.* Intrinsic dimension identification via graph-theoretic methods // Journal of Multivariate Analysis 116, 263-277 (2013).
29. *Adams H., Aminian M. et al.* A fractal dimension for measures via persistent homology // Abel Symposia 15, 1-31 (2020).
30. *Schweinhart B.* Fractal dimension and the persistent homology of random geometric complexes. arxiv:1808.02196
31. *Caruso F, Oguri V.* The cosmic microwave background spectrum and an upper limit for fractal space dimensionality // The Astrophysical Journal 694, 151-153 (2009).
32. *Manrubia S.C., Sole R.V.* Self-organized criticality in rainforest dynamics // Chaos, Solitons and Fractals 7, 523-541 (1996).
33. *Gromov V.A., Shulga A.N.* Chaotic time series prediction with employment of ant colony optimization // Expert Systems with Applications 39, 8474-8478 (2012).

34. *Costa J.A., Girotra A., Hero A.O.* Estimating local intrinsic dimension with k-nearest neighbor graphs // IEEE/SP 13th Workshop on Statistical Signal Processing, Bordeaux, France, 2005. P.417-422.
35. *Farahmand M., Szepesvari C., Audibert J.-Y.* Manifold-adaptive dimension estimation // Proceedings of the 24th international conference on Machine learning, 2007. P.265–272.
36. *Zhu X.* Persistent homology: An introduction and a new text representation for natural language processing. https://www.ijcai.org/Proceedings/13/Papers/288.pdf
37. *Bellegarda J.R.* Latent semantic mapping: Principles and applications / 1st ed. – Cham: Springer, 2007.
38. *Kalman D.* A singularly valuable decomposition: The SVD of a matrix // The college mathematics journal 27, 2-23 (1996).
39. *Golub G., Kahan W.* Calculating the singular values and pseudo-inverse of a matrix // SIAM Journal on Numerical Analysis, 2(2), 205-224 (1965).
40. *Mikolov T., Chen K. et al.* Efficient estimation of word representations in vector space. arxiv:1301.3781
41. *Wishart D.* A numerical classification methods for deriving natural classes // Nature 221, 97-98 (1969).
42. *Vathy-Fogarassy Á., Abonyi J.* Graph-based clustering and data visualization algorithms. – Springer Publisher Company, 2013.
43. *Rosso O.A., Larrondo H.A. et al.* Distinguishing noise from chaos // Phys. Rev. Lett. 99, 154102 (2007).
44. *Čech E.* Topological spaces / 1st ed. – Prague: Pub. House of the Czechoslovak Academy of Sciences, 1966.
45. *Hausdorff F.* Dimension und äußeres Maß // Mathematische Annalen 79, 157-179 (1918).
46. *Steele J.M.* Growth rates of euclidean minimal spanning trees with power weighted edges // The Annals of Probability 16, 1767-1787 (1988).
47. *Steele J.M., Shepp L.A., Eddy W.F.* on the number of leaves of a euclidean minimal spanning tree // Journal of Applied Probability 24, 809-826 (1987).
48. *Kuznetsov S.O., Gromov V.A. et al.* Formal concept analysis for evaluating intrinsic dimension of a natural language // 10th International Conference on Pattern Recognition and Machine Intelligence, Kolkata: Springer, 2023. In press.
49. *Kuznetsov S.O.* Pattern structures for analyzing complex data // Lecture Notes in Computer Science 5908, 33-44 (2009).
50. *Čufar M., Virk Ž.* Fast computation of persistent homology representatives with involuted persistent homology. arxiv:2105.03629
51. *Arneodo A., Bacry E., Muzy J.-F.* Random cascades on wavelet dyadic trees // Journal of Mathematical Physics 39, 4142-4164 (1998).

# Приложение А. Тестирование алгоритмов оценки внутренних размерностей на геометрических объектах с известной внутренней размерностью.

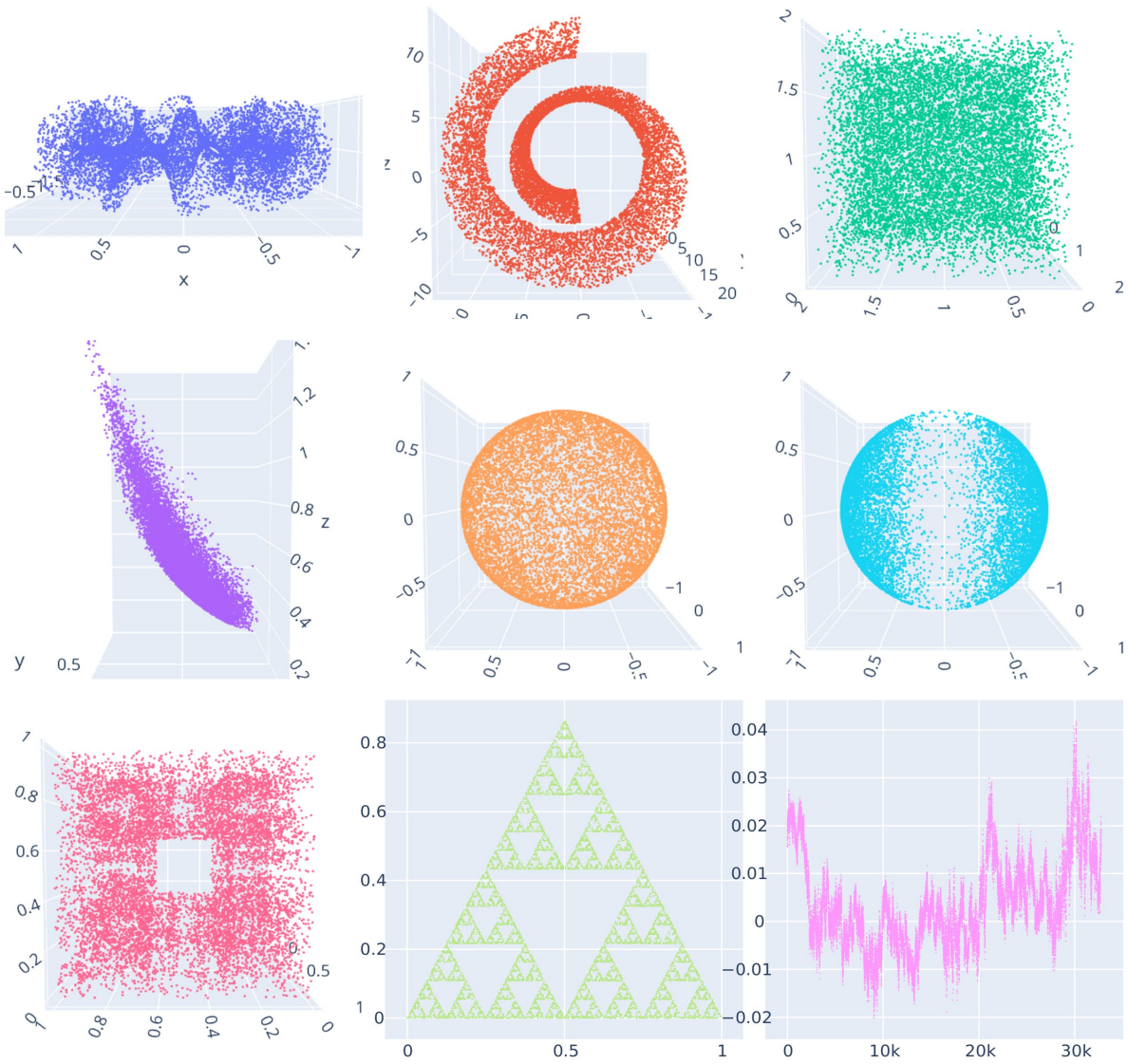


**Рис А1. Геометрические объекты, использованные для тестирования методов оценки внутренней размерности.** Точки распределены по поверхности равномерно, если иное не указано явно. Многообразия (геометрический объект целочисленной хаусдорфовой размерности): самый верхний ряд: лента Мёбиуса, швейцарский рулет, единичный куб; средний ряд: параболоид, единичная сфера, единичная сфера (точки распределены неравномерно – смесь двух гауссовых распределений с модами, соответствующими двум полюсам сферы). Фрактальные и мультифрактальные объекты: самый нижний ряд: губка Менгера, ковёр Серпинского, логнормальный вейвлет-каскад (мультифрактал; по оси абсцисс отложены отметки времени, по оси ординат — переменная состояния).

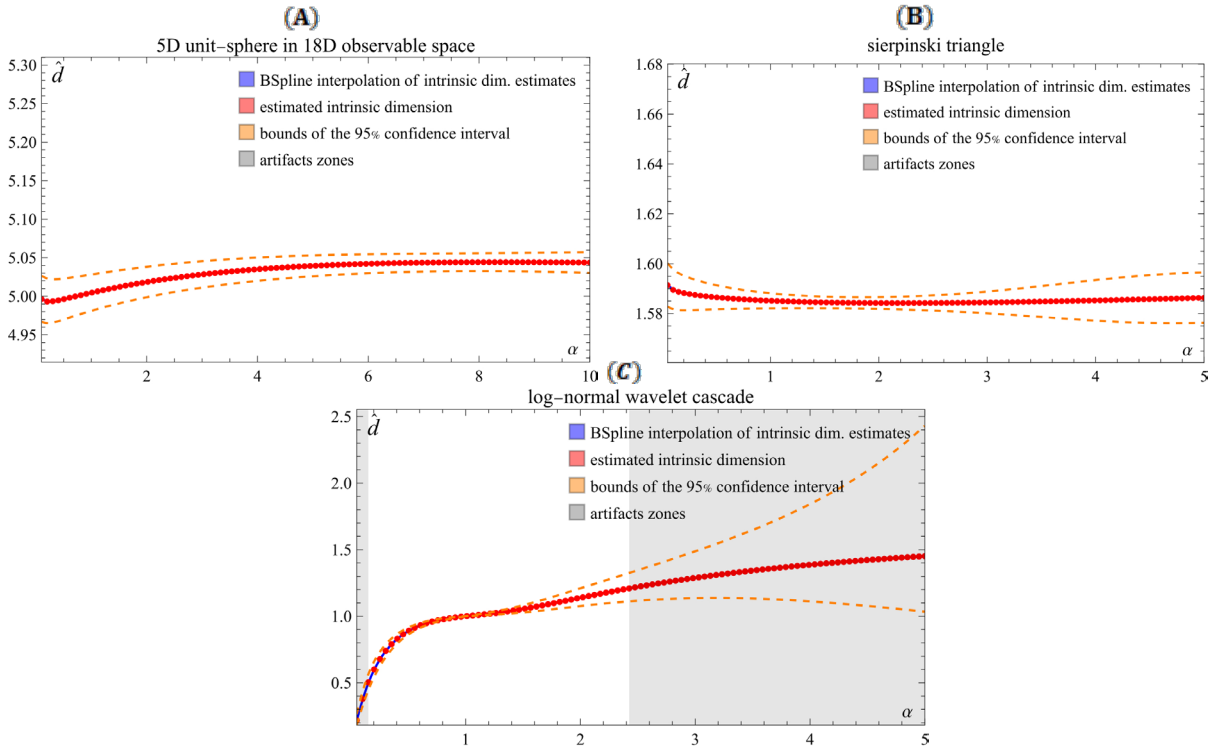


**Рис. А2. Зависимость оценок алгоритма Швайнхарта от параметра $\alpha$ для:** (A) единичной сферы (многообразие, $d_H = 5$, $d = 18$); (B) ковра Серпинского (фрактальное множество, $d_H = 1.58$, $d = 2$); (*C*) логнормального вейвлет-каскада (мультифрактал). Красные точки являются оценками алгоритма, синяя линия представляет собой B-сплайн интерполяцию зависимости оценок от $\alpha$, оранжевый пунктир – верхнюю и нижнюю границы $95\%$ доверительного интервала. На рисунке также представлены критерии допустимости параметра $\alpha$, такие как расхождение доверительных интервалов (серые зоны).

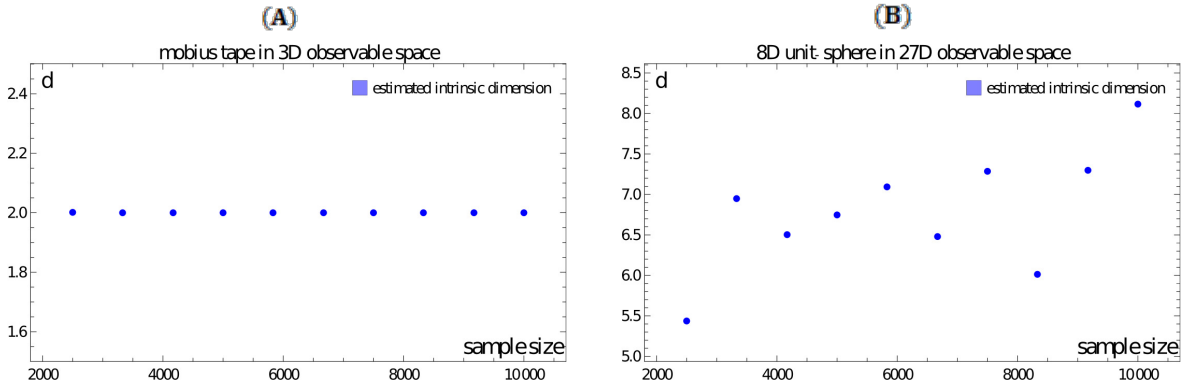


**Рис. А3. Зависимость оценок алгоритма Брито (не округлённых до целого) от объёма выборки для:** (A) ленты Мёбиуса (многообразие, $d_T = 2$, $d = 3$); (B) единичной сферы (многообразие, $d_T = 8$, $d = 27$). По мере уменьшения величины отношения внутренней топологической размерности к размерности пространства вложения, при прочих равных условиях, снижается и точность алгоритма.

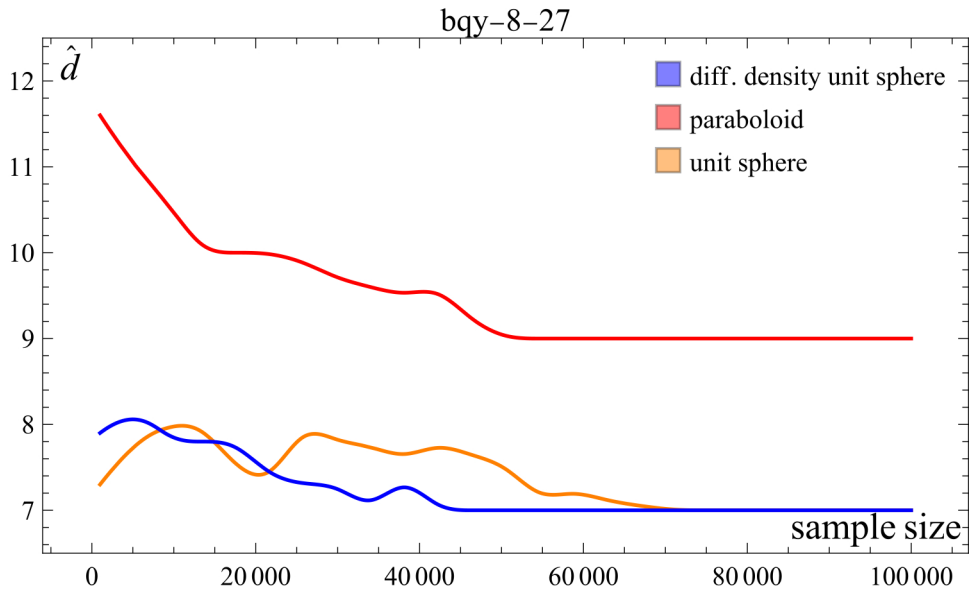


**Рис. B3. Зависимость оценок алгоритма Брито от объёма выборки для различных многообразий $d_T = 8$ и $d = 27$.** Ответы алгоритма не округляются до ближайшего целого, чтобы показать сходимость по мере увеличения выборки. Алгоритм сходится к постоянной внутренней размерности для всех рассматриваемых наборов данных.

Таблица A1. Абсолютная процентная ошибка (APE) оценок алгоритма Швайнхарта для различных геометрических объектов

| | Mobius tape APE | Swiss-roll APE | Unit-cube APE | Unit-sphere APE | Unit-sphere (Gauss.) APE | Menger sponge APE | Sierpinski carpet APE |
|---|---|---|---|---|---|---|---|
| 1.00 | 0.47% | 0.43% | 1.35% | 0.07% | 0.68% | 1.56% | 0.07% |
| 2.00 | 0.77% | 0.43% | 1.57% | 0.03% | 1.26% | 1.68% | 0.03% |
| 3.00 | 1.15% | 0.47% | 1.77% | 0.03% | 2.41% | 1.76% | 0.11% |
| 4.00 | 1.59% | 0.55% | 1.98% | 0.08% | 3.79% | 1.82% | 0.18% |
| 5.00 | 2.10% | 0.68% | 2.20% | 0.12% | 5.02% | 1.85% | 0.24% |
| 6.00 | 2.66% | 0.85% | 2.43% | 0.16% | 5.99% | 1.87% | 0.28% |
| 7.00 | 3.22% | 1.08% | 2.69% | 0.19% | 6.77% | 1.90% | 0.32% |
| 8.00 | 3.72% | 1.36% | 2.97% | 0.22% | 7.43% | 1.92% | 0.35% |
| 9.00 | 4.11% | 1.69% | 3.28% | 0.25% | 8.05% | 1.97% | 0.38% |
| 10.00 | 4.33% | 2.04% | 3.62% | 0.29% | 8.58% | 2.03% | 0.42% |
| mean | 2.41% | 0.96% | 2.38% | 0.14% | 5.00% | 1.84% | 0.24% |

Таблица A2. Абсолютная процентная ошибка (APE) оценок алгоритма Брито для различных геометрических объектов. В каждом столбце указаны размерность пространства вложения $d$ и топологическая размерность объекта $d_T$

| | Unit-sphere APE $d=9$ $d_T=2$ | Unit-sphere APE $d=18$ $d_T=5$ | Unit-sphere APE $d=27$ $d_T=8$ | Paraboloid APE $d=9$ $d_T=2$ | Paraboloid APE $d=18$ $d_T=5$ | Paraboloid APE $d=27$ $d_T=8$ | Mobius tape APE $d=3$ $d_T=2$ | Swiss-roll APE $d=3$ $d_T=2$ |
|---|---|---|---|---|---|---|---|---|
| 250 | 0.03% | 17.3% | 32.0% | 0.07% | 8.73% | 1.21% | 0.07% | 49.7% |
| 333 | 0.00% | 16.6% | 13.1% | 0.03% | 54.3% | 24.1% | 0.00% | 0.00% |
| 417 | 0.01% | 5.5% | 18.7% | 0.01% | 2.30% | 28.4% | 0.00% | 0.00% |
| 500 | 1.59% | 12.5% | 15.6% | 0.01% | 17.8% | 40.7% | 0.00% | 0.00% |
| 583 | 0.00% | 17.5% | 11.3% | 0.00% | 23.4% | 4.90% | 0.00% | 0.00% |
| 667 | 0.00% | 1.29% | 19.0% | 0.00% | 36.8% | 36.8% | 0.00% | 0.00% |
| 750 | 0.00% | 4.28% | 8.93% | 0.00% | 30.8% | 39.2% | 0.00% | 0.00% |
| 833 | 0.00% | 15.3% | 24.8% | 0.00% | 30.3% | 32.7% | 0.00% | 0.00% |
| 917 | 0.00% | 9.85% | 8.77% | 0.00% | 4.60% | 28.5% | 0.00% | 0.00% |
| 1000 | 0.00% | 12.1% | 1.45% | 0.00% | 17.0% | 25.6% | 0.00% | 0.00% |
| mean | 0.00% | 11.2% | 15.3% | 0.02% | 22.6% | 26.2% | 0.01% | 4.98% |

## Приложение B. Поведение алгоритма Брито при уменьшении отношения внутренней размерности $d_T$ к размерности пространства вложения $d$

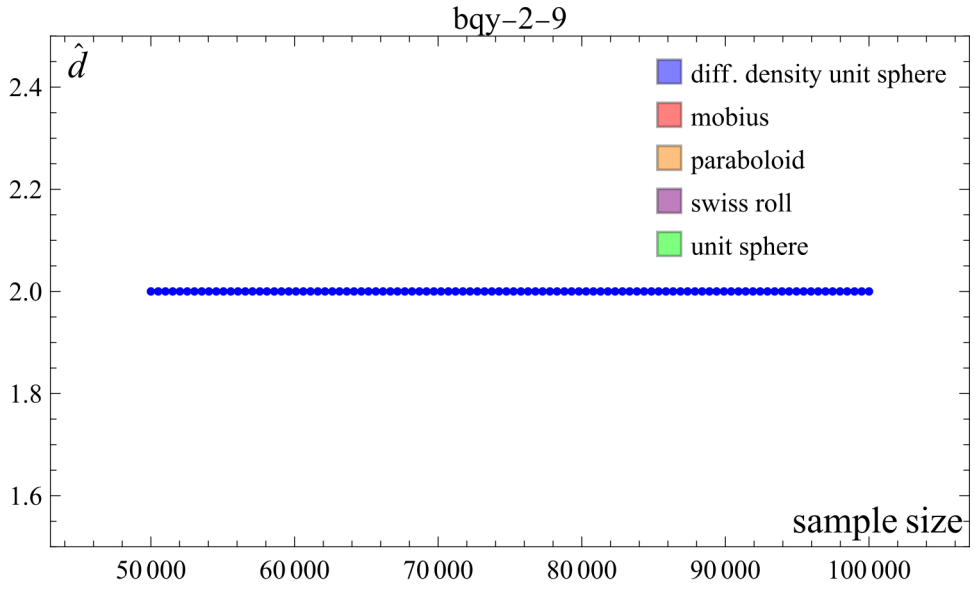


**Рис. B1. Зависимость оценок алгоритма Брито от объёма выборки для различных многообразий $d_T = 2$ и $d = 9$.** Ответы алгоритма не округляются до ближайшего целого, чтобы показать сходимость по мере увеличения выборки. Алгоритм изначально сходится к постоянной внутренней размерности для всех рассматриваемых наборов данных.

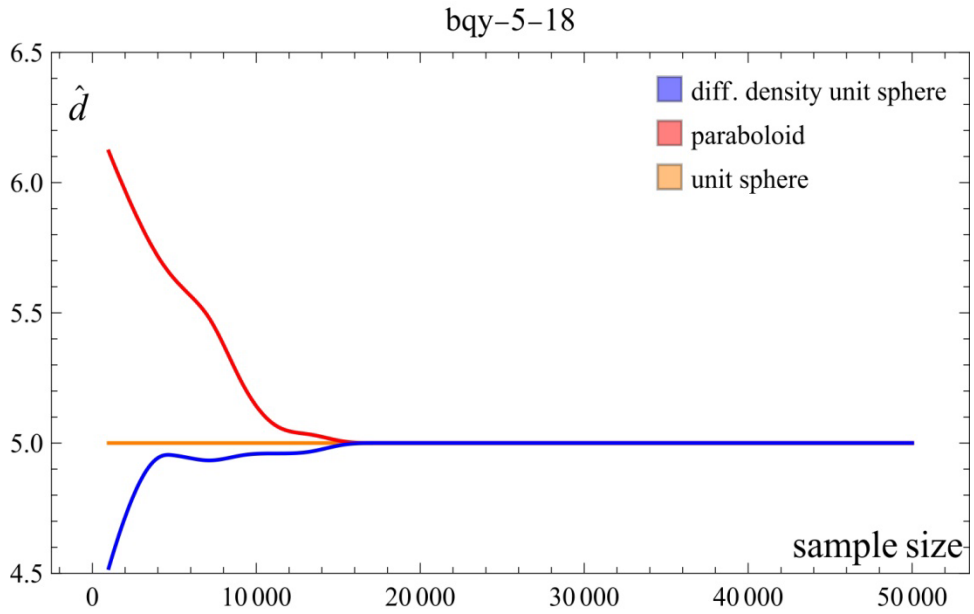


**Рис. B2. Зависимость оценок алгоритма Брито от объёма выборки для различных многообразий $d_T = 5$ и $d = 18$.** Ответы алгоритма не округляются до ближайшего целого, чтобы показать сходимость по мере увеличения выборки. Алгоритм сходится к постоянной внутренней размерности для всех рассматриваемых наборов данных.

## Приложение C. Перевод границ персистентных $H_1$-гомологий белорусского языка

– хочу, год, иметь, хорошо, голова, увидеть, день;
– первое, другое, жизнь, вообще;
– ребёнок, должен, дверь, город, пока, понимаю, сейчас, лечь;
– нет, дорога, дайте, следовательно, место, воду, немедленно, идите, подождите;
– взять, нет, вопрос, край;
– затем, время, голова, хорошо, стоять, ещё, приземлиться, начать, снова;
– связь, встретиться, разрешить, начать, идти, вздохнуть, птица, услышать, подумать;
– дай, дорога, нет, иди;
– тяжёлый, должен, черный, молодой, дальше, женщина;
– где-то, нельзя, упасть, нужно, берег, главный, ветер, партизан, лучше, успеть, после, книга, прийти, улыбнуться, будет, попробовать, спать, убить, дело, приказ, солдат, наверное, много, сильно.