Северо-Казахстанский университет им. М. Козыбаева

УДК 004.94

На правах рукописи

ЕРБОЛОВА АСЕЛЬ СЕРИКАНОВНА

Исследование структур естественного языка в задаче идентификации ботов

8D06101 – Информатика, вычислительная техника и управление

Диссертация на соискание степени
доктора философии (PhD)

Научный консультант
кандидат технических наук,
ассоц. профессор,
И.Г. Курмашев

Зарубежный консультант
доктор физико-математических наук,
профессор
В.А. Громов

Республика Казахстан
Петропавловск, 2026

СОДЕРЖАНИЕ

НОРМАТИВНЫЕ ССЫЛКИ..... 4
ОПРЕДЕЛЕНИЯ..... 5
ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ..... 6
ВВЕДЕНИЕ..... 7
1 СОВРЕМЕННЫЕ ПОДХОДЫ К АНАЛИЗУ СТРУКТУРЫ ЕСТЕСТВЕННОГО ЯЗЫКА И ЗАДАЧА ИДЕНТИФИКАЦИИ БОТОВ..... 12
1.1 Естественный язык как единая система..... 12
1.2 Хаотичность и сложность семантических траекторий..... 14
1.3 Фрактальные структуры и внутренние размерности текстов..... 16
1.4 Топологический анализ данных и поиск персистентных гомологий..... 18
1.5 Идентификация ботов на основе языковых особенностей..... 19
Выводы по первому разделу..... 22
2 МЕТОДЫ АНАЛИЗА КРУПНОМАСШТАБНОЙ СТРУКТУРЫ ЕСТЕСТВЕННОГО ЯЗЫКА..... 23
2.1 Статистические методы. Критерии согласия для степенных распределений..... 23
2.1.1 Подход Прюсснера..... 23
2.1.2 Подход Клаузета-Чализи-Ньюмана..... 24
2.2 Методы анализа хаотичности естественных языков в сематической траектории..... 26
2.2.1 Построение семантических траекторий текстов..... 27
2.2.2 Плоскость энтропии-сложности..... 28
2.3 Численные методы оценки внутренней размерности..... 30
2.3.1 Геометрическое представление естественного языка..... 31
2.3.2 Оценка внутренней размерности. Оценка Швайнхарта..... 32
2.3.3 Оценка внутренней размерности. Байесовская оценка..... 34
2.4 Методы топологического анализа данных и поиска персистентных гомологий..... 35
2.4.1 Семантическое пространство..... 35
2.4.2 Персистентная гомология..... 36
2.4.3 Контуры “дырок” естественного языка..... 36
2.5 Методы идентификации ботов..... 39
2.5.1 Сбор и предварительная обработка данных..... 41
2.5.2 Оценка размерности аттракторов семантических траекторий..... 44
2.5.3 Кластеризация n-грамм и показателей связности кластеров..... 44
Выводы по второму разделу..... 46
3 СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ КРУПНОМАСШТАБНОЙ СТРУКТУРЫ ЕСТЕСТВЕННОГО ЯЗЫКА..... 48
3.1 Языковые характеристики..... 48
3.1.1 Статистический анализ степенного распределения в языках..... 49

3.1.2 Кластеризация на основе статистических особенностей аналитических методов..... 51
3.2 Хаотичность естественных языков..... 53
3.2.1 Кластеры пространства признаков..... 56
3.3 Внутренние размерности естественных языков..... 60
3.3.1 Внутренние размерности для стандартных многообразий и фрактальных множеств..... 60
3.3.2 Внутренние размерности естественных языков..... 64
3.3.3 Кластеризация языковых представлений..... 67
3.4 Топологическая структура естественного языка..... 71
3.4.1 Выделение гомологии на участке векторного пространства..... 71
3.4.2 Выделение гомологий первого порядка..... 73
3.4.3 Классификация: люди и боты..... 82
Выводы по третьему разделу..... 85
4 ИДЕНТИФИКАЦИЯ БОТОВ..... 87
4.1 Результаты классификации текстов на естественном языке..... 87
4.1.1 Классификация на основе положения точки на плоскости “энтропия - сложность”..... 88
4.1.2 Классификация по характеристикам семантических траекторий..... 89
4.1.3 Подход, основанный на кластеризации n-грамм..... 90
Выводы по четвертому разделу..... 93
ЗАКЛЮЧЕНИЕ..... 94
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ..... 95
ПРИЛОЖЕНИЕ А – Акты внедрения..... 104
ПРИЛОЖЕНИЕ Б – Свидетельство об авторском праве..... 110
ПРИЛОЖЕНИЕ В – Корпус языковых данных и результаты статистического анализа..... 111
ПРИЛОЖЕНИЕ Г – Результаты оценки внутренней размерности и тестирования алгоритмов..... 117
ПРИЛОЖЕНИЕ Д – Экспериментальные результаты генерации текстов и их классификации..... 130

НОРМАТИВНЫЕ ССЫЛКИ

В настоящей диссертации использованы ссылки на следующие стандарты:

Закон Республики Казахстан. Об образовании: принят 27 июля 2007 года, №319-III (с изменениями и дополнениями по состоянию на 23.02.2023 г.).

Закон Республики Казахстан. О науке: принят 18 февраля 2011 года, №407-IV (с изменениями и дополнениями).

Инструкция по оформлению диссертации и автореферата. Высшая аттестационная комиссия МОН РК, 28 сентября 2004 года №377-3ж.

ГОСТ 7.32–2017. Отчет о научно-исследовательской работе. Структура и правила оформления.

ГОСТ 7.1–2003. Библиографическая запись. Библиографическое описание. Общие требования и правила составления.

ГОСТ 34.201–89. Информационная технология. Комплекс стандартов на автоматизированные системы. Виды, комплектность и обозначение документов при создании автоматизированных систем.

СТ РК 34.007–2002. Информационная технология. Телекоммуникационные сети. Основные термины и определения.

СТ РК ГОСТ Р ИСО/МЭК 12119–2006. Информационная технология. Пакеты программ. Требования к качеству и тестированию.

СТ РК ГОСТ Р 52292–2007. Информационные технологии. Электронный обмен информацией. Термины и определения.

ОПРЕДЕЛЕНИЯ

В настоящей диссертации применяют следующие термины с соответствующими определениями:

Бот – программная система, функционирующая автономно и осуществляющая генерацию либо обработку текстовой информации без непосредственного участия человека, имитируя поведение пользователя в цифровой среде.

Вектор вложения (*embedding*) – результат некоторого отображения слова или любой другой атомарной языковой сущности в конечномерное векторное пространство.

n-грамма – это последовательность из n смежных слов в тексте, где n является гиперпараметром.

Внутренняя размерность – численная характеристика множества данных, отражающая минимальное число независимых параметров, достаточных для аппроксимации его геометрической структуры в выбранном пространстве признаков.

Остовное дерево – подграф исходного графа, включающий все его вершины и являющийся деревом, то есть связным графом без циклов.

Минимальное остовное дерево – остовное дерево, для которого сумма весов рёбер минимальна среди всех возможных остовных деревьев данного графа.

Семантическое пространство языка – концептуальная структура, в рамках которой языковые единицы организованы таким образом, что их сочетания соотносятся с системой знаний и представлений о мире, закреплённых в коллективном языковом опыте.

Персистентная гомология – метод топологического анализа данных, позволяющий выявлять устойчивые топологические инварианты (компоненты связности, циклы, полости) в зависимости от параметра масштаба.

Кластеризация – задача разбиения множества объектов на непересекающиеся подмножества (кластеры) таким образом, чтобы объекты внутри кластера были более сходны между собой, чем с объектами из других кластеров.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ИИ	– Искусственный интеллект
ТАД	– Топологический анализ данных
МНК	– Метод наименьших квадратов
LLM	– Large Language Model
MST	– Minimum Spanning Tree
NLP	– Natural Language Processing
SVD	– Singular Value Decomposition
CBOW	– Continuous Bag of Words
BERT	– Bidirectional Encoder Representations from Transformers
k-NN	– k-Nearest Neighbors
APE/MAPE	– (Mean) Absolute Percentage Error
SVC	– Support Vector Classifier
DT	– Decision Tree
RF	– Random Forest

ВВЕДЕНИЕ

Представление о естественном языке как о единой системе, относящейся к классу самоорганизованно-критичных систем [1], вместе с современными методами построения эмбеддингов (векторных представлений) слов и n-грамм языка позволило [2] поставить вопрос об исследовании геометрических свойств этой системы: предлагается рассмотреть в качестве единого геометрического объекта множество всех слов языка, данных своими векторами представления (эмбеддингами). Предполагаем, что множество всех наблюдаемых в языке слов представляет собой репрезентативную выборку точек, лежащих на некоторой d-мерной поверхности, и потому исследование этого множества точек может дать информацию о геометрических свойствах данной поверхности, а точнее, фрактальной структуры – языковой фрактальной структуры.

В работе [2, p. 863360] для совокупности языковых фрактальных структур естественного языка для *n=1…*, предложен термин *Hailonakea* (с гавайского знаковая беспредельность). (Ср. по ссылке https://en.wikipedia.org/wiki/Laniakea_Supercluster Википедии Laniakea Supercluster можно найти фотографию самого крупного объекта в известной части Вселенной – суперскопления галактик Laniakea. Каждый пиксель на этой фотографии представляет собой даже не галактику, но скопление галактик; в целом фотография даёт нам представление о крупномасштабной структуре нашей Вселенной.)

В настоящей работе проводится исследование внутренних размерностей языковых фрактальных структур для *n=1* (слова); *n=2* (биграммы); *n=3* (триграммы) на материале русского и английского языков. Все вычислительные эксперименты, направленные на изучение поставленного вопроса, проводились над корпусами национальной литературы. Использование именно литературных текстов обусловлено тем, что национальная литература является ядром соответствующего языка. Следовательно, считаем, что вполне разумно использовать n-граммы, извлеченные из корпусов литературных шедевров, для исследования соответствующего языка.

В качестве второй цели настоящего исследования рассматривали решение задачи идентификации ботов, т.е. различения текстов, написанных людьми, и текстов, сгенерированных ботами. Более того, как в [3] предлагаем отличать не тексты одного (или нескольких) конкретных ботов от текстов людей, но тексты всех ботов от текстов всех людей. В качестве базовой гипотезы здесь проверяется гипотеза о том, что том, что тексты, написанные людьми, будут статистически значимо отличаться по указанным выше фундаментальным характеристикам от текстов, сгенерированных ботами.

В рамках решения задачи установления крупномасштабной структуры естественного языка необходимо:

– для заданного множества текстов естественного языка $\Im = (\Omega_1, \cdots, \Omega_N)$ построить множество всех d-мерных эмбеддингов n-грамм рассматриваемого естественного языка $\aleph_n(d), n = 1..N, d = 1..D$. При этом предполагается, что

множество текстов $\Im$ представляет собой репрезентативную выборку текстов, написанных на соответствующем естественном языке;

– для данного n по множествам $\aleph_n(d), d = 1..D$ построить множество характеристик, отвечающих элементам крупномасштабной, топологической структуры языка, а не локальным неоднородностям выборки.

Для решения задачи обнаружения ботов (чтобы отличить тексты, написанные людьми, от текстов, сгенерированных ботами), используем следующее постановку задачи [3, p. e2550]: для данного естественного языка рассматривается пространство всех текстов Ω как написанных людьми, так и сгенерированных всевозможными ботами. Пространство делится на подпространство $A = \{\alpha_1, \dots, \alpha_a\}$ текстов, написанных людьми, и подпространство $M = \bigcup_{j=1}^{L} M_j$, текстов, сгенерированных ботами, где $M_j = \left\{\mu_1, \dots, \mu_{m_j}\right\}$ – множество текстов, сгенерированных j-м ботом (ἄνθρωπος – человек, и μηχάνημα – машина). Также рассматривается пространство ботов M: какие-то из ботов (не все) тексты которых участвовали в формировании пространства Ω. Требуется построить признаки $\Lambda = \{\lambda_1, \dots, \lambda_k\}$ и построить на их основе классификатор $R = R(\Lambda)$ с F1-оценкой классификации выше порога r^*.

Выборка текстов, написанных людьми, случайным образом делится на обучающую и тестовую выборки, также строятся обучающая и тестовая выборки для текстов, сгенерированных ботами. Важной является процедура формирования обучающего и тестового множеств для текстов, сгенерированных ботами: не делим случайным образом на две части множество текстов, сгенерированных ботами, но делим случайным образом на две части множество ботов $\{M_j, j = 1..l\}$: тексты ботов, попавших в первое множество, используются для формирования классификатора, тексты ботов, попавших во вторую часть, – для его тестирования. При этом предполагается, что число текстов и распределение размеров текстов приблизительно одинаковы для частей обучающего множества, отвечающих людям и ботам; аналогичное предположение делается для тестового множества.

Актуальность исследования обусловлена стремительным развитием технологий обработки естественного языка (NLP) и искусственного интеллекта (ИИ). В условиях увеличения объёма текстов, создаваемых как людьми, так и автоматизированными системами, возникает необходимость разработки эффективных методов, позволяющих различать эти типы контента для обеспечения информационной безопасности, мониторинга цифрового контента и устойчивости ИИ-систем. Применение современных подходов к анализу и классификации текстов подтверждает значимость данного исследования.

Предложенные в работе методы и подходы являются актуальными по нескольким причинам. Во-первых, исследование внутренних размерностей языковых фрактальных структур в сочетании с анализом семантических траекторий позволяет создать целостное представление о языковых закономерностях и способствует более глубокому пониманию динамики и структуры языковых систем.

Во-вторых, использование топологического анализа данных открывает новые возможности для изучения языковой системы и может повысить качество идентификации ботов. Это крайне важно в условиях нарастания угроз от автоматизированных систем манипуляции мнением и дезинформации. Таким образом, данная работа подчеркивает значимость и сложность естественного языка, предлагая эффективные методы и подходы к его анализу, что может существенно улучшить функционирование информационных систем в современном цифровом обществе.

Целью исследования является разработка методов для исследований структур языка, а также адаптация их для идентификации текстов, созданных людьми и сгенерированных ботами.

Для достижения данной цели решались следующие **основные задачи**:

1. Разработка характеристик языковых объектов и анализ их статистических свойств для выявления отличий между текстами, написанными людьми, и текстами, сгенерированными ботами.
2. Адаптация методов анализа хаотичности временных рядов для оценки сложности семантических траекторий текстов на естественном языке.
3. Адаптация методов оценки внутренних размерностей сложных геометрических объектов к исследованию языковых фрактальных структур.
4. Адаптация методов топологического анализа данных для изучения семантического пространства естественного языка, для выявления "дыр" для различных n-грамм (слова, биграммы, триграммы). Проведения сравнительного анализа характеристик крупномасштабных структур естественного языка в максимально широком круге языков.
5. Разработка классификационных моделей для идентификации текстов, созданных ботами, на основе выявленных языковых признаков и экспериментальная проверка их эффективности.

Объект исследования – языковые характеристики и структура естественного языка, эффективные при различении текстов, созданных людьми, и текстов, генерируемых автоматизированными системами (ботами).

Предмет исследования – методы крупномасштабной структуры естественного языка, её геометрических свойств, а также разработка методов идентификации текстов, созданных людьми, и текстов, сгенерированных автоматизированными системами (ботами).

Методы исследования, использованные в работе: статистические методы (тесты на различия, анализ распределений); методы анализа хаотичности сематической траектории для естественных языков; численные методы оценки внутренней размерности (геометрические и фрактальные методы n-грамм); топологический анализ данных (персистентная гомология); методы идентификации ботов (классификационные модели, разделение выборки).

Научная новизна полученных результатов состоит в следующем:

1. Проведён анализ естественных языков как самоорганизованно-критичных систем на множестве языков; выявлена уникальность эсперанто, демонстрирующего гауссово распределение в отличие от естественных языков, а

также установлено, что статистические характеристики степенных распределений, включая параметры степенного закона, могут служить надёжными метриками для понимания и классификации языков.

2. Представлены результаты анализа семантических траекторий на естественных языках, которые показывают хаотическую природу указанной стуктуры. Кластерный анализ выявил типологические кластеры, коррелирующие с энтропией и сложностью, подчеркивая значимость хаоса для языковой динамики.

3. Разработаны методы оценки внутренней размерности естественных языков, выявляющие их мультифрактальную природу и указывающие на инвариантность размерностей, устойчивую к изменениям методов извлечения векторных представлений.

4. Разработан метод выявления “дыр” в языке на основе методов топологического анализа данных. Применение методов расчёта персистентных гомологий первого порядка выявило крупномасштабные особенности естественного языка и позволило исключить артефакты выборки.

5. Предложены высокоэффективные алгоритмы классификации текстов, разделяющих тексты, созданные людьми, и сгенерированные ботами, достигающие свыше 96% для большинства языков.

Практическое значение полученных результатов состоит в том, что разработаны методы идентификации ботов для эффективного различения текстов, написанных людьми, и сгенерированных автоматизированными системами. Разработанные методы анализа языковых характеристик, таких как энтропия и сложность семантических траекторий, позволяют значительно повысить точность и эффективность систем автоматической идентификации ботов, что особенно важно в условиях растущей автоматизации контент-генерации и распространения дезинформации.

Предложенные методы топологического анализа данных и персистентной гомологии могут быть использованы для изучения структурных особенностей текстов, открывая возможности для анализа и классификации в различных областях. Проведенный сравнительный анализ естественных языков может быть полезен, способствуя переосмыслению классификаций языков с учетом их крупномасштабных структур.

Ключевые результаты, **выносимые на защиту:**

1. Модель естественных языков как самоорганизованно-критичных систем, применённая для сравнительного анализа широкого круга языков.

2. Модель анализа семантических траекторий для оценки хаотической природы языка, сравнительный анализ языков.

3. Модель анализа языковых фрактальных структур (внутренние размерности и “дырки”) для выявления мультифрактальности и подтверждения лингвистических родственных гипотез.

4. Классификационная модель для идентификации текстов, написанных людьми и сгенерированных ботами, подтвердившая практическую значимость методов.

Апробация работы. Апробация результатов диссертационного исследования проводилась в Национальном университете обороны Республики Казахстан, Карагандинском университете Казпотребсоюза, а также в Лаборатории анализа семантики Департамента анализа данных и искусственного интеллекта Факультета компьютерных наук Национального исследовательского университета «Высшая школа экономики» (Приложение В).

По результатам диссертационного исследования получено авторское свидетельство (Приложение А).

Публикации. Основные результаты диссертационной работы опубликованы в 7 печатных работах, в том числе: 1 статья в изданиях, рекомендованных КОКСН МОН РК; 1 публикация в международных конференциях; 5 статей в журналах, индексируемых в базе данных Scopus, с процентилем от 39 до 80.

Личный вклад в положения, выносимые на защиту. В работах, написанных в соавторстве, докторанту принадлежит выбор и построение метода решения, разработка и реализация алгоритмов и программ, получение и анализ результатов.

Структура работы. Диссертационная работа состоит из введения, четырёх разделов, выводов и списка использованных источников. Полный объём диссертации составляет 103 страницы, 40 рисунков, 16 таблиц, список использованных источников из 156 наименований.

1 СОВРЕМЕННЫЕ ПОДХОДЫ К АНАЛИЗУ СТРУКТУРЫ ЕСТЕСТВЕННОГО ЯЗЫКА И ЗАДАЧА ИДЕНТИФИКАЦИИ БОТОВ

Настоящий раздел структурирована следующим образом. В разделе 1.1, рассматриваются работы, анализирующие естественный язык как единую систему. Раздел 1.2 посвящён анализу литературы, исследующей меры хаотичности и сложности семантических траекторий текстов на естественных языках. В разделе 1.3 рассматриваются методы анализа фрактальной структуры естественного языка и получении оценок внутренних размерностей указанных фрактальных структур. Наконец, в последнем разделе (1.4) обсуждаются исследования, посвященные идентификация ботов.

1.1 Естественный язык как единая система

В данном разделе представлен анализ (к сожалению, не столь многочисленных исследований) исследований, в которых естественный язык рассматривается как единый объект, рассматриваемый в рамках естественнонаучной парадигмы.

Такого рода подход к анализу естественного языка послужил основой для создания нескольких теорий, описывающих универсальную грамматику [4, 5] и когнитивные механизмы, которые стоят за преобразованием универсальной грамматики в грамматику конкретного языка, т.е. набор синтаксических правил, которыми пользуется носитель данного языка. Вместе с тем, сторонники данного подхода исходят из убеждения, что естественный язык представляет собой набор слов (символов), подлежащих преобразованиям по формальным грамматическим правилам.

Исследование Эрнандес-Ферн Андес и др. [6] на примере каталанского языка обобщают ключевые законы, характерные для текстов на естественных языках на уровне морфема-слово, подчеркивая, что большинство из них имеют степенной характер. Важно отметить, что в подавляющем большинстве случаев соответствующие законы носят степенной характер; Торре и др. [7] приводят психофизиологические причины появления этих законов. Эти закономерности объясняются через взаимодействие психофизиологических процессов и физических принципов, определяющих речевую деятельность как сложную динамическую систему. Работа Башери и др. [8] расширяет данное объяснение, связывая формирование указанных закономерностей с когнитивными ограничениями человеческого мозга, такими как объем рабочей памяти и особенности обработки информации, что подчеркивает универсальный характер этих законов. Ван и Лю [9] указывают, что степенной показатель в законе Ципфа является показателем лексического разнообразия.

Полагаем, что гораздо более продуктивным является понимание языка как множество значений (концепций), объединённых в семантическое пространство данного языка (например, представленное пространством эмбеддингов или семантическим графом концепций, построенным по корпусу текстов рассматриваемого языка). К сожалению, как представляется, структура

пространства значений естественного языка (семантического пространства) методами естественных и компьютерных наук исследована в существенно меньшей степени, чем пространство слов языка. Вместе с тем, можем указать несколько работ, которые будучи посвящены в основном исследованию пространства слов, позволяют, в силу тех или иных причин, сделать некоторые умозаключения и о пространстве значений [4, p. 1-16; 10]. В частности, К. Танака-Исии в своей монографии [11] и серии статей [12-14] исследует долгие корреляции между словами в тексте на материале английского и японского языков. Особенность японского языка заключается в том, что иероглифы (кандзи) имеют четкое соответствие с их значениями, что позволяет устанавливать практически однозначные связи между символами и семантикой, что позволяет делать выводы и о структуре семантического пространства. В частности, в японском языке зафиксированы долгие корреляции между словами, расположенными на расстоянии 10-15 позиций друг от друга[1], что свидетельствует о сложных семантических связях в текстах языка. Л. Дебовский в своём исследовании [15] указывает, что базовые процессы в языке описывается степенными распределениями.

В работах исследование Громова и Мигриной [1, p. 1-6] естественный язык рассматривается как единая система в пространстве значений, причём авторы показывают, что язык представляет собой самоорганизовано-критичную систему. При этом тексты данного языка, как записанные, так и произнесённые, интерпретируются как «лавины» [16][2] в семантическом пространстве; здесь установлено, что размеры лавин подчиняются степенным законам распределения.

Здесь также следует отметить исследования, посвященные моделированию языка как сложной сети. Например, Гарг и др. [17, 18] исследуют динамику процессов естественного языка посредством построения и анализа графовых моделей языка. В монографии [18, p. 4-255] Гарг и его коллеги исследуют графовые модели языка и его применение в обработке естественного языка. Аналогичные подходы используются для изучения социальных структур и динамики дискурса в онлайн-средах [19].

Язык можно рассматривать как динамическую и самоорганизующую систему, развивающуюся через взаимодействия между индивидами. Исследования в области семиотической динамики демонстрируют, как сообщества создают и поддерживают общие семиотические системы для эффективного общения. В этом контексте язык предстает как постоянно изменяющаяся система, чьи элементы формируются и адаптируются для достижения максимальной успешности коммуникации при минимальных усилиях.

---

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

[2]Размеры этих лавин подчиняются степенным законам, что естественно для сложных систем: языковых (ср., например, работы (Antoni Hernández-Fernández, Juan María Garrido, Bartolo Luque, Iván González Torre Linguistic laws in Catalan; J. Milička, Václav Cvrček, David Lukeš Unpacking lexical intertextuality: Vocabulary shared among texts)), биологических и др.

В обзоре работ, связанных с лингвистическими последовательностями, следует отметить работу Садди и Уриагереки [20], которые выделяют два ключевых аспекта сложности: вычислительную сложность, связанную с грамматикой, и процедурную сложность, связанную с использованием когнитивных ресурсов для интерпретации и производства речи. Осознание необходимости рассмотрения естественного языка как объекта реального мира, восходящее к ранним работам Н. Хомского [5, p. 1-13], привело к формулировке исследовательской программы «Physics of Language» (PoL) (Важно также отметить работу Кривочена [21], где подчеркивается, что существующие модели синтаксиса естественного языка часто не отражают его сложности. Исследование Кривочена [22] акцентирует внимание на важности разработки динамической модели для построения фразовой структуры, принимающей во внимание вычислительные особенности синтаксических объектов и предлагающей использование смешанных фразовых маркеров, основываясь на сложности входных данных. В другой работе Х. Уриагерека [23] рассматривается вопрос сложности ряда с точки зрения минималистской программы Хомского, подчеркивающей экономичность и эффективность в объяснении языковых явлений. В упоминавшейся выше статье Кривочена [24] обосновывается хаотическая природа рассматриваемых последовательностей в рамках синтаксических механизмов языка. Также следует отметить исследование Миркина Б.Г. [25], которое рассматривает различные аспекты кластеризации данных и восстановления информации, что может способствовать дальнейшему пониманию структуры и динамики языка как системы.

1.2 Хаотичность и сложность семантических траекторий

Среди основополагающих работ в области количественной лингвистики и сложных систем особенно выделяются два основных направления исследований. Научные работы первого направления посвящены изучению количественных закономерностей в языке с применением методов статистического анализа и физического моделирования. Исследования [26, 27] демонстрируют, что лингвистические процессы подчиняются универсальным законам масштабирования, что подтверждается как статистическими данными, так и физическими моделями.

Другая работа посвящена сложным сетям в когнитивной науке, которая рассматривает язык как самоорганизующуюся систему, управляемую принципами сложных физических сетей. Эта структура особенно важна для понимания коллективной динамики и эмерджентных свойств эволюции и использования языка, включая фракталы [28].

Ma Qinghua and Xinxin W. [29] выделили ряд характеристик сложности языка, таких как неопределенность, нелинейность, адаптивность и хаотичность, которые указывают на неравенство его статуса. Еще одно важное утверждение, требующее рассмотрения, связано с тремя наиболее важными производными чертами "нелинейности" языка: его дисбалансом, эмерджентностью и интерактивностью. Наиболее важным наблюдением для целей настоящего

исследования является то, что в языке правила и хаотичность не являются взаимоисключающими. Правила при некоторых новых обстоятельствах и другие возникающие правила могут быть подвержены хаотическому изменению в геометрической прогрессии или экспоненциально, как предполагает изменение отношения формы и значения звукоподражания в генетической лингвистике [30]. Похожие идеи можно найти у [31], который рассматривает язык как сложную нелинейную систему, динамичную по своей природе из-за динамичности его пользователей, которые отличаются от других биологических видов своей креативностью. Динамичная природа языка и его носителей, следовательно, приводит к созданию несколько хаотичной системы.

Концепциями, лежащими в основе энтропии и стохастичности, являются хаотичность и случайность, которые подразумевают непредсказуемость и новизну поведения элементов информации. Однако в математической статистике и информатике хаос и стохастичность имеют разные характеристики. Хаотический процесс может быть непредсказуемым, даже зная начальное состояние, из-за влияния различных условий на поведение системы. Он также проявляет чувствительность к начальным условиям, в отличие от случайного процесса, где начальные условия не имеют значения. Хаотичность элементов не подчиняется правилам и определяется множеством условий, хотя энтропия как мера хаотичности может быть вычислена [32].

Признавая сложность естественного человеческого языка как системы и применяя стандартный теоретико-графовый метод, [33] предполагает, что вычислительные процедуры естественного человеческого языка решают динамически нарушаемое уравнение в человеческом мозге. Основываясь на уравнении закона Кирхгофа (закона электрического тока), утверждается, что рассчитано равновесие в любой сети, и делается вывод, что данное уравнение относится к уравнению с динамическим расстройством, таким образом доказывая, что нарушенная нефазность катализирует создание фаз. Сторонники генеративной лингвистики [34], в значительной степени опирающиеся на минималистскую модель в лингвистике, описывают операцию синтаксического слияния математически в терминах алгебр Хопфа. Они утверждают, что этот подход позволяет связать основной вычислительный механизм слияния с синтаксическими ограничениями конкретных языков. [35] выделили три особых класса синтаксических шаблонов, определенных в терминах спектра собственных значений матрицы, и обнаружили, что у них есть один общий шаблон - схема X-bar, которая, по-видимому, характеризует синтаксис человеческого языка. Следуя логике Хомского, [36] объясняют математические основы синтаксических отношений в языке, выдвигая идею матричного синтаксиса. Они утверждают, что их математическая структура разделяет некоторые аспекты квантовой механики, и вводят концепцию лингвистических цепочек как более экономичную теорию языка. Применяя алгоритмы и концепции теории квантовой механики и квантовой теории поля в области лингвистики, [37] показывают, что коллективные режимы в квантовой теории поля приводят к «многообразию концепций», соответствующих логическим

формам Н. Хомского, которая, по их мнению, даёт убедительные доказательства того, что язык является неотъемлемой частью мира природы.

Семантические траектории [38-40][3] литературных шедевров представляют собой динамические пути текста в семантическом пространстве, отражающие изменения значений слов. Такие траектории характеризуются хаотическим поведением, что делает их объектом изучения в контексте теории сложных систем. В работе [41] Громов и Данг исследуют хаотичность семантических и эмоциональных траекторий текстов из корпусов национальных литератур (вычисляя их энтропию и сложность) для нескольких языков (принадлежащих, в основном, индоевропейской языковой семье). Степени их хаотичности (энтропия и сложность) собирает в себе значительное число признаков, характеризующих как синтаксис языка (например, тип алгоритма построения минимального поиска в дереве синтаксического разбора предложения в рамках минималистской теории Н. Хомского [42], так и его семантику.

Задача сравнения (многомерных) временных рядов по степени их хаотичности предполагает использование робастных методов оценки хаотичности, что практически исключает (особенно, для многомерных временных рядов) классические подходы, связанные с оценкой старшими показателями Ляпунова [43, 44]. Среди более робастных методов можно отметить методы, основанные на классической и модифицированной информации по Фишеру [45], взаимной информации [46], рекуррентных графиках [47] а также вейвлет-многоуровневой комплексной сети для анализа многомерных нелинейных временных рядов [48]. Кроме того, существует методика оценки количества мотивов, найденных в рассматриваемой серии временных рядов [49-56], предсказуемость [57] и другие. Однако, с точки зрения рассматриваемой задачи наиболее эффективными и робастнымии здесь оказались методы, основанные на вычислении пермутационной энтропии и сложности рассматриваемого временного ряда [58, 59]. Также можно отметить обзорные работы Амиго, Келлера и Унакофовой [60] по энтропии и энтропийным величинам, а также исследование графовой пермутационной энтропии [61].

1.3 Фрактальные структуры и внутренние размерности текстов

В настоящем параграфе рассматриваются подходы к оценке внутренней размерности геометрических объектов, в том числе фрактальных. Здесь, прежде всего, стоит обратиться к работе [62]: в статье вводятся необходимые условия для того, чтобы функция была функцией внутренней размерности; соответствующие аксиомы основаны на расстоянии М. Громова между метрическими пространствами [63].

В целом, известные методы оценки внутреннего размера геометрического объекта (используя репрезентативную выборку его точек) на три класса:

[3]Здесь следует подчеркнуть, что мы употребляем термин “семантическая траектория” в совсем ином смысле, чем это принято, например, в контексте теории экспертных систем: в ней данный термин описывает последовательность логических или семантических изменений, которые вовлекают информацию в процессы анализа и обработки и др.

1. Странные аттракторы. Основаны на исследовании траекторий движения на многообразиях с использованием теории странных аттракторов [64, 65].

2. Графы данных. Используют граф-теоретические методы для определения внутренней размерности [66].

3. Персистентные гомологии. Применяют персистентные гомологии и анализ топологических данных для вычисления фрактальной размерности [67, 68].

Канц и Шрейбер [65, p. 3-367], а также Малинецкий с Потаповым [64, с. 3-335] описывают основные способы определения размерностей странных аттракторов (топологической, хаусдорфовой, энтропийной). Авторы также приводят методы их оценки, такие как реконструкция аттрактора (которая позволяет визуализировать динамику системы), расчет корреляционной размерности и использование фрактальной геометрии, такие как расчет размерности по мере увеличения масштаба. К сожалению, эти методы имеют свои недостатки: 1) они требуют гораздо более длинных временных рядов, чем любые доступные «языковые» временные ряды [49, p. 8474-8477; 64, с. 3-335; 66, p. 263-276; 69]; 2) кроме того, зачастую не робастны в терминах изменяющихся данных. Отметим также работу [41, p. 113934], посвящённую анализу “языковых” временных о странных аттракторах рассматриваются семантические и эмоциональные временные ряды.

В работах Коста и др. [70] и Фараманд и др. [71] используется метод графов ближайших соседей для оценки топологической размерности многомерных данных. Метод Коста и др. основывается на построении графа, в котором для каждой точки определяются её ближайшие соседи (k-NN), что позволяет анализировать локальную топологическую структуру данных; локальная размерность оценивается через изучение расстояний между точкой и её соседями. Метод [71, p. 265-271] позволяет оценить локальную размерность в окрестности каждой точки, а затем комбинировать эти локальные оценки для получения глобальной оценки размерности. Этот алгоритм эффективен при работе с низкоразмерными подмногообразиями в высокоразмерных пространствах. Брито и др. [66, p. 263-276] распространили метод и на другие типы графов данных.

Другой подход к оценке фрактальной размерности основывается на персистентной гомологии с оценкой [67, p. 1-30]. Швайнхарт [68, p. 107291] предложил подход к оценке фрактальной размерности, с использованием энтропийной размерностью. Подход применим к исследованию фрактальных и мультифрактальные множества.

Недавние исследования, проведенные E.D. Santis, G. De Santis и A. Rizzi [72] и L.C. Ribeiro, A.T. Bernardes и H. Mello, сосредоточены на фрактальной структуре языка. В статье Santis и др. [72, p. 10143-10159] предложена методология анализа морфологической организации текстов для оценки размерностей мультифрактальные структур. Ribeiro и др. в своей работе [28, p. e0285630] применяют методы фрактальной геометрии для анализа языковых данных. Опираясь на алгоритм word2vec для создания векторных представления

слов, авторы выявляют семантические связи и различия между языками, а также исследуют их фрактальные характеристики.

1.4 Топологический анализ данных и поиск персистентных гомологий

Переходя к обзору исследований численных методов топологического анализа данных (ТАД), отметим, что ТАД является мощным инструментом для исследования структурных особенностей в разных областях. Одно из главных преимуществ ТАД заключается в его применимости к сложным системам [4, p. 1-16; 73-76]. В области космологии Бермехо и др. [77] авторы выделяют дыры и пустоты среди космической паутины. Исследование персистентных гомологий широко используется в биологии и медицине. Авторы статьи [78] предоставляют широкий обзор способов применения методов ТАД в таких областях, как клиническая помощь и прецизионная медицина, анализ медицинских изображений, точность медицинской диагностики, биологические исследования (включая "омические" науки), структурная биология, иммунология, эпидемиология и многое другое. В исследовании [79] разработан подход к анализу пространственной организации ДНК на основе локализованной взвешенной персистентной гомологии, который позволяет анализировать топологические свойства ДНК на уровне локальных доменов, что способствует выявлению функциональных характеристик и структурных вариаций. В продолжение обсуждения применения топологических методов в различных доменах, стоит обратить внимание на использование топологических признаков в задачах классификации белков. В исследовании авторов Dey и Mandal [80] рассматривается подход, в котором симплициальные комплексы используются для моделирования иерархической структуры белковых молекул. В рамках этого подхода вычисляются персистентные гомологии для фильтрации комплексной структуры белков. Авторы [81] также изучают дыры и пустоты в географических данных поиском персистентных гомологий. В [82, 83] описывается применение ТАД в нейронауках и изучении связи с мозгом.

Современные подходы к анализу языка всё чаще включают методы топологического анализа данных, в частности персистентную гомологию, для формализации структурных особенностей естественно языковых систем. В работе [84] рассматривается применение методов топологического анализа данных и персистентной гомологии для изучения структуры естественного языка. Автор анализирует последовательности слов в детских стихотворениях и сравнивает их с текстами, написанными взрослыми, чтобы выявить топологические характеристики, такие как "дыры" в текстах. Используя баркоды для визуализации персистентных гомологий, работа демонстрирует, как топологические методы могут обогатить анализ текстов и помочь понять их внутреннюю структуру и семантику. В работе [85] авторы исследуют применение персистентной гомологии для анализа текстов и классификации персидских стихотворений, написанных различными поэтами, используя методы топологического анализа данных для выявления структурных особенностей и семантических связей в произведениях. Авторы [86] исследуют применение

топологических признаков для классификации юридических текстов, демонстрируя, что информация, извлеченная из топологических структур, может быть использована для решения сложной задачи установления соотношения между юридическими документами, что позволяет улучшить точность классификации в контексте юридического вывода. В работе [87] авторы исследуют наличие логических и литературных "дыр" в заголовках и аннотациях научных статей, используя методы топологического анализа данных для выявления несоответствий, что позволяет более эффективно обнаруживать фальсифицированные публикации. Множество исследований по данной тематике акцентируют внимание на ключевом аспекте методов топологического анализа – их интерпретируемости [85; 88].

1.5 Идентификация ботов на основе языковых особенностей

Значительная часть работ, посвящённых идентификации ботов, посвящена работе с метаданными, т.е. с информацией об аккаунтах ботов, с информацией, содержащейся в их профилях, динамики генерации сообщений ботами и другие характеристики, не вытекающие непосредственно из текстов, порождённых ботами. Также можно использовать информацию о взаимодействиях между учетными записями, которая допускает представление в виде графа. Графовые методы широко применяются для решения задачи обнаружения ботов. Например, Дайя и др. [89] предложили систему обнаружения ботов на основе графов для коммуникационного графа, в котором пользователи представлены узлами, а их взаимодействия – ребрами. Такой подход позволяет анализировать паттерны взаимодействий и выявлять аномалии, характерные для ботов. Меснардс и др. [90] установили, что боты в социальных сетях, как правило, больше взаимодействуют с людьми, чем с другими ботами, и эта функция (heterophily) может быть использована для их обнаружения. Ли и др. [91], разработали метод BotFinder, который представляет собой метод для обнаружения социальных ботов в онлайн-сетях, основанный на использовании вложений графов Node2Vec [92] для представления пользователей и применении методов обнаружения сообществ, что позволяет эффективно идентифицировать ботов. Фам и др. [93] представили Bot2Vec – усовершенствованный алгоритм Node2Vec [93, p. 101771] для обнаружения ботов в социальных сетях, ориентированный на представление узлов внутри сообществ. Такой подход позволяет учитывать структуру взаимодействий и повышает точность выявления ботов, демонстрируя при этом гибкость в применении к различным типам социальных сетей и используя методы, которые учитывают, как локальные, так и глобальные структуры графа. Существуют различные модели, которые используют графовые сверточные сети: BotRGCN в Feng и др. [94], Squeeze GCN в Фу и др. [95], SEGCN в Лю и др. [96]. В своем обзоре Лата М. [97] предлагает подробную классификацию методов идентификации ботов. Опираясь на результаты соревнования по идентификации ботов, проведенного Агенством исследовательских проектов (DARPA), автор делает вывод, что "обнаружение

ботов должно быть полууправляемым", то есть алгоритмы этого класса должны проверяться и корректироваться людьми.

Соответственно, более перспективным нам представляется подход к решению задачи идентификации ботов, использующий только тексты, сгенерированные ботами. Подавляющее большинство исследований данного класса посвящено работе с одним конкретным ботом и посвящено построению своеобразного антибота, обычно, с помощью той или иной нейросетевой модели. Так, например, в работе [98] используется тонкая настройка предварительно обученные генеративных трансформеров (GPT, GPT-2) для идентификации твиттер-ботов, основываясь исключительно на текстах. В работе [99] приводится исследование предобученных эмбеддингов, таких как glove, word2vec, fastText, и контекстуальных (ELMo) применительно к задаче идентификации ботов.

Нам представляется, что здесь наиболее интересным теоретически и значимым практически является исследование задачи идентификации ботов в той постановке, в которой она рассматривается в настоящей работе, – как задачи анализа всех текстов того или иного естественного языка с целью проведения границы между текстами, написанными людьми, и текстами, сгенерированными ботами [3, p. e2550]. Авторам известно весьма мало работ, которые работают с задачей идентификации ботов в такой постановке. Вместе с тем мы можем отметить ряд работ, в которых авторы, непосредственно не касаясь задачи идентификации ботов, разрабатывают методы анализа структуры семантического пространства естественного языка, которые позволяют, во всяком случае, потенциально рассматривать указанную задачу.

В данном случае первый подход основан на анализе простых текстовых характеристик. Например, в статье Кан и др. [100] анализируются простейшие текстовые характеристики, такие как лексические и синтаксические параметры, для выявления различий в стилях общения между людьми и игровыми ботами. Исследуются частота символов, общая частота букв и средняя длина слов в сообщениях чата, что помогает оценить сложность и разнообразие слов, отражая характер человеческого общения. Также анализируется использование пунктуации для выявления стилистических особенностей, а синтаксические характеристики, включая частоту служебных слов, помогают дополнительно различить способы общения. Кардайоли и др. [101] моделируют пользователя Twitter, используя набор стилистических особенностей, и различают аккаунты ботов и пользователей-людей, оценивая согласованность стиля их сообщений. Чакраборти и др. [102] сочетают выделение текстовых признаков с графовым подходом для обнаружения фальшивых пользователей в Twitter, используя информацию как о текстах, так и о сетевых взаимодействиях между пользователями. Громов и Данг [103] преобразуют текст в многомерный временной ряд, чтобы рассчитать его энтропию и сложность, чтобы отличить временные ряды роботов от временных рядов людей.

Во втором подходе к тематике обнаружения бот-активности применяется анализ настроений текстов, который позволяет различать тексты, написанные людьми, и генерируемые ботами. Эмоциональные характеристики, присущие

этим текстам, проявляют значительные различия. Этот подход, по-видимому, в настоящее время вызывает ажиотаж в научном сообществе, о чем свидетельствует множество публикаций по этой теме, к примеру [104, 105], Heidari и соавторы [106] применяют сложные текстовые характеристики, включая лексические и синтаксические особенности, для анализа твитов на английском и голландском языках с целью извлечения признаков, которые помогают различить тексты ботов и реальных пользователей. Liao и соавторы [107] предлагают новую многоуровневую графическую нейронную сеть (MLGNN) для анализа настроений в тексте, которая улучшает результаты анализа, эффективно извлекая информацию с помощью графовых структур; Lin et al. [108] и Galgoczy et al. [109] объединяют BERT с анализом настроений текста для выявления вредных новостей. Кроме того, методы прогностической кластеризации [49, p. 8474-8477; 50, p. 1-20; 51, p. 1827-1837; 52, p. 3317-3321; 110], оказываются многообещающими для автоматического обнаружения ботов, выявляя характерные подпоследовательности во временных рядах. Во многих исследованиях используется размеченная информация для обучения нейронных сетей. Например, [111] разработали набор данных пользователей Twitter с ретвитами из ненадежных или надежных источников новостей с использованием лингвистической информации для их идентификации. Рен и Джи [112] рассматривают эффективную модель обнаружения спама с ложным мнением, отмечая при этом ограничения методов обучения под наблюдением и необходимость дальнейшего изучения методов без присмотра.

В монографии [11, p. 3-243] и ряде других работ [12, p. 481-501; 13, p. e0164658; 14, p. 2150033] указывает в качестве одного из признаков, отличающих текст, написанный человеком, наличие длинных корреляций между словами в тексте (см. также [113, 114]. Здесь можно предположить, что ботам (даже обученным с помощью весьма сложных нейросетевых моделей) будет сложно симулировать) такого рода последовательности (поскольку в подавляющем своём большинстве они учатся на локальной информации и выучиваются отслеживать только локальные связи внутри одного n-грамма).

В работе [115] рассматриваются энтропийные характеристики естественного языка. [10, p. 3-380] исследует характеристики естественных языков; при этом устанавливается, что в большинстве случаев наблюдаемый тип распределения – это степенное распределение. Здесь, как нам представляется, есть перспективы для решения задачи идентификации ботов – боту, генерирующему тексты, будет сложно соблюдать все законы, рассмотренные в указанных работах.

Подводя итог, в естественных языках можно выделить две группы признаков: локальные и глобальные (целостные). Локальные характеристики основаны на свойствах отдельных слов или n-грамм; глобальные - на свойствах всего текста или даже самого языка. Часто локальные характеристики не позволяют отличить тексты, написанные человеком, от текстов, созданных ботами, в то время как глобальные характеристики могут выполнять эту работу.

Выводы по первому разделу

Опираясь на материал, изложенный в данном разделе, можно прийти к следующим выводам:

1. Проведён всесторонний анализ ключевых исследований, рассматривающих естественный язык как самоорганизованно-критичную систему. Изучены различные аспекты сложности и хаотичности языка, с особым вниманием на теории универсальной грамматики и когнитивные механизмы, связывающие формальные грамматические правила с реальным использованием языка. Эти исследования подтвердили, что естественный язык не является просто набором символов для формальных преобразований, а представляет собой сложное взаимодействие с психофизиологическими процессами человека.

2. Проведен анализ величин семантической сложности и хаотичности текстов, а также методов, позволяющих рассматривать языковую фрактальную структуру. Важно отметить, что многие работы подчеркивают степень хаотичности и сложности языковых процессов, что открывает новые возможности для применения методов статистического анализа и теории сложных систем.

3. В рамках дальнейшего анализа были применены разнообразные методы, включая графовые модели и топологический анализ данных, что представляет собой интересный подход к исследованию языковых структур. Особое внимание было уделено выявлению взаимосвязи между текстами, написанными людьми, и текстами, генерируемыми автоматическими системами, что позволяет установить границы между двумя типами содержания и оценить их структуру.

4. В этом контексте значительное внимание уделено различным методам для идентификации текстов, написанных людьми, и текстов, сгенерированных ботами. Исследованы уникальные характеристики, присущие ботам, которые позволяют отличать их тексты от человеческих. Рассмотрены методы анализа, основанные на текстовых характеристиках, эмоциях и семантическом пространстве. Эти исследования, проведенные в данной главе, закладывают основу для следующих анализов, направленных на дальнейшее изучение языковых фрактальных структур и их применения в практических задачах, таких как автоматизация обнаружения ботов в текстах.

2 МЕТОДЫ АНАЛИЗА КРУПНОМАСШТАБНОЙ СТРУКТУРЫ ЕСТЕСТВЕННОГО ЯЗЫКА

В работе представлены результаты анализа 52 естественных языков, принадлежащих к 18 языковым семьям (Приложение В).

Исследование охватывает 100% самых распространенных языковых семей (и 12% от общего числа семей) в мире [116], на которых говорит 74,3% населения планеты (Приложение В). Все тексты загружены из открытых источников и были проверены носителями языка; переводы и техническая литература были удалены. Предварительно обрабатываем тексты, удаляя числа, знаки препинания, иностранные слова и лишние символы.

2.1 Статистические методы. Критерии согласия для степенных распределений

В данном исследовании рассматриваем естественные языки как самоорганизованно-критичные системы [1, p. 1-6; 117]. Формально, для естественного языка $\aleph_i$, представленного своим набором текстов Ω_i, проверяется статистическая гипотеза H_0 о том, что случайная величина "число слов в тексте, написанном на этом языке", подчиняется степенному закону распределения:

$$p(x) = ax^{-\tau}, \;\; x > x^{min} \tag{2.1.1}$$

где $x = |\omega|$, $\omega \in \Omega_i$ представляет собой количество слов в случайно выбранном тексте ω, написанном на языке $\aleph_i$. Здесь τ и x_{min}обозначают показатель степени и параметры нижнего среза соответственно. Параметр a определяется из условий нормализации. Расчетные значения τ и x_{min}группируются для классификации соответствующих языков.

Для проверки этой статистической гипотезы H_0 использовали два подхода (гипотеза подразумевает, что размеры лавин подчиняются степенному закону распределения). Первый подход применяет понятие коллапса данных (см. монографию [118]. Второй подход использует синтетические наборы данных, сгенерированные восстановленным степенным законом [119]. Оба подхода исключают самые маленькие и самые большие элементы выборки (ниже и выше минимальных и максимальных пределов соответственно), которые обычно не соответствуют степенному закону распределения, чтобы оценить, соответствует ли данные степенному закону распределения в среднем диапазоне.

2.1.1 Подход Прюсснера Метод Прюсснера [118, p. 126-128] использует функцию плотности вероятности в виде:

$$p(x) = ax^{-\tau} g\left(\frac{x_{min}}{bL^D}\right), \quad x > x_{min} \tag{2.1.2}$$

Распределение подчиняется модифицированному степенному закону в интервале, ограниченном нижним и верхним отсечениями x_{min} и x_{max}.

Необработанные данные (с размером выборки L) группируются с использованием экспоненциального сглаживания (binning). Для проверки нулевой гипотезы о том, что выборка извлечена из распределения с степенным законом, строится зависимость $p(x)\ x^{-\tau}$ от $\xi = \frac{x}{x_c(L)}$, где $x_c(L) = bL^D$ для разных L; соответствующие графики накладываются друг на друга. Данное явление получило название коллапса данных. Оценка качества соответствия степенному распределению включает следующие этапы:

1. Сглаживание исходных данных (binning).
2. Построение графика $p(x)x^{-\tau}$ в функции $\xi = \frac{x}{x_c(L)}$ для разных L с использованием эмпирической оценки показателя $\hat{\tau}$ (приблизительная оценка τ) - такой график (в случае истинности нулевой гипотезы) чтобы получить график с наклонный прямой участок и характерный нелинейный участок. Экстремум этой кривой, называемый ориентиром, соответствует x_c.
3. Измените τ, чтобы получить участок горизонтальной линии.

Если нулевая гипотеза H_0 верна, линейные участки графиков (для различных значений L) сливаются в единую горизонтальную прямую, тем самым доказывая, что данные соответствуют степенному закону (1). Этот подход подразумевает использование выборок разного размера. С этой целью из заданной выборки различных размеров L случайным образом генерируются подвыборки.

2.1.2 Подход Клаузета-Чализи-Ньюмана

Второй подход Колмогорова-Смирнова (КС) [119, p. 661-702] использует следующую функцию распределения с степенным законом:

$$p(x) = \begin{cases} q(x), & x \le x_{min} \\ ax^{-\tau}, & x > x_{min} \end{cases} \tag{2.1.3}$$

Константа нормировки $a = \frac{1}{\zeta(\tau, x_{min})}$, где

$$\zeta(\tau, x_{min}) = \sum_{i=1}^{n} (i + x_{min})^{-\tau} \tag{2.1.4}$$

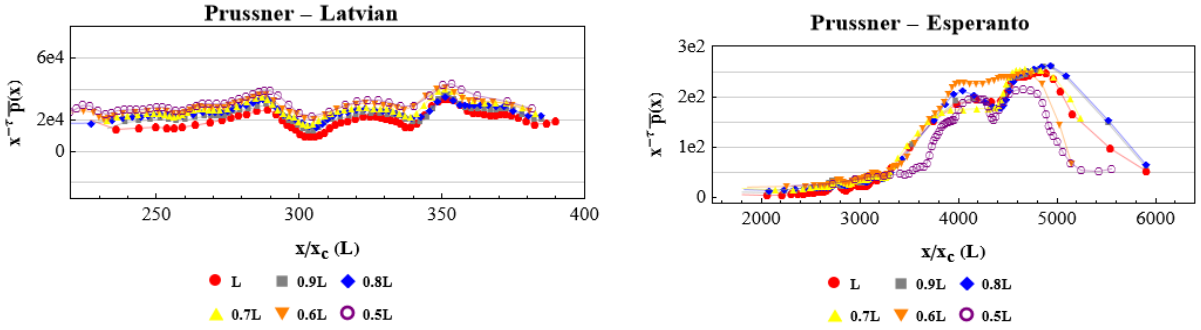
где $\zeta\ (\tau, x_{min})$ – представляет собой обобщенную дзета-функцию Гурвица;

$q(x)$ – это какое-то (не степенное) распределение для области, выходящей за пределы нашего рассмотрения. Оценивается показатель степенного закона τ для каждого значения нижнего порога, применяя принцип максимального правдоподобия:

$$\hat{\tau} = 1 + \mathrm{n}\left[\sum_{i=1}^{n} ln\ \frac{x_i}{x_{min}}\right]^{-1} \tag{2.1.5}$$

Статистика Колмогорова-Смирнова $S = sup_{x \geq x_{min}} |P(x) - \bar{P}(x)|$ рассчитывается для каждого $x_{\min}$. P(x) обозначает функцию распределения с оценённым значением τ; $\bar{P}(x)$ - эмпирическая функция распределения. Первый минимум (ближайший к x_{min}) статистики S даёт оценку x_{min}. Для проверки статистичнеской гипотезы о соответствии выборки степенному закону распределения генерирует синтетические выборки для проверки нулевой гипотезы. В частности, для распределения с степенным законом с параметрами, оцененными, как описано выше, мы генерируем несколько выборок. Для каждой из этих синтетических выборок применяем вышеупомянутую процедуру, чтобы получить оценки параметров и вычислить статистику Колмогорова-Смирнова. При этом, значение p представляет долю синтетических выборок (которые гарантированно подчиняются распределению с степенным законом), для которых эта статистика больше, чем для рассматриваемой выборки. Большое значение p указывает на то, что нулевая гипотеза должна быть отвергнута, в то время как малое значение p предполагает, что она верна. Чтобы определить количество синтетических наборов данных, эмпирическое правило состоит в том, чтобы сгенерировать не менее $\frac{1}{4}\epsilon^{-2}$ параметра для точности p-значения в пределах ϵ. Для точности до 2 десятичных разрядов установите $\epsilon = 0.01$, что требует около $N = 2500$ синтетических наборов.

В качестве иллюстрации, на рисунок 2.1 представлены результаты для двух языков: эсперанто (рисунок 2.1б) и латышского (рисунок 2.1а). Латышский язык удовлетворяет закону степенного распределения, в то время как эсперанто – нет. Результаты для других языков представлены в https://github.com/erbolova1983/Investigation-of-NL-structures.git.



а б

а – латышский текст подчиняется степенному распределению; б – эсперанто подчиняться гауссовскому распределению

Рисунок 2.1 – Одна шкала распределений а и б

На рисунке 2.1 маркеры представляют разные размеры словарей: красный цвет (с дисками) соответствует полному словарю размера L; серый цвет (с квадратами) – 0,9L; синий цвет (с ромбами) – 0,8L; желтый цвет (с

треугольниками) – 0,7L; оранжевый цвет (с перевернутыми треугольниками) – 0,6L; а фиолетовый цвет (с кружками) – 0,5L. Для лучшей наглядности кривые немного сдвинуты, чтобы избежать наложения, вызванного коллапсом данных. Метод коллапса данных оценил параметры степенного закона и нижние пределы для латышского языка как $x_{min} = 101$ и $\tau = 3.00$. В то же время метод Колмогорова-Смирнова (КС) дал $x_{min} = 103$ и $\tau = 2.98$. Аналогично, для эсперанто метод коллапса данных показал $x_{min} = 1001$ и $\tau = 1.60$, в то время как метод КС привел $x_{min} = 1043$ и $\tau = 5.25$. Вместе с тем, для эсперанто стандартная проверка на нормальность даёт результаты, подтверждающие, что распределение не соответствует нормальному закону.

2.2 Методы анализа хаотичности естественных языков в сематической траектории

Создание алгоритмов преобразования слов естественного языка в d-мерные вектора представления (эмбеддинги) позволило рассмотреть текст на естественном языке как d-мерный временной ряд – в [41, p. 113934] такие многомерные временные ряды получили название семантических траекторий. В [41, p. 113934] на материале нескольких языков было установлено, что подавляющее большинство семантических траекторий являются хаотическими. Рассмотрение множества семантических траекторий для данного языка (опираемся, в основном, на произведения национальной литературы) позволяет оценить хаотичность того или иного языка, что, в свою очередь, позволяет сравнить хаотичность различных языков, характеризующихся различными грамматическими и семантическими особенностями. В настоящей работе рассматривается семантические траектории (составленные, преимущественно, на основе корпусов национальных литератур) для 52 языков, принадлежащих 18 языковым семьям, что позволило провести сравнительный анализ их хаотичности[4]. Под хаотичностью языка понимаем среднюю хаотичность текстов литературных произведений, написанных на данном языке. Кроме теоретического интереса указанный анализ представляет и значительный прикладной интерес: указанные характеристики могут использоваться в задачах идентификации ботов [41, p. 113934] и задачах оценки качества перевода[5].

Процесс предварительной обработки состоял из несколько этапов. Во-первых, мы заменяли все небуквенные символы и приводили текст к нижнему регистру. Во-вторых, лемматизировали слова, т.е. привели их словарным формам. В-третьих, мы идентифицировали все именованные объекты, такие как имена, фамилии, названия организаций и географические названия. Эти объекты заменялись названиями соответствующих категорий, что позволяло улучшить

[4]Критерием выбора языка служило, во-первых, наличие достаточно обширной литературы в свободном доступе в интернете, во-вторых, мы старались выбирать языки таким образом, чтобы максимальным образом разнообразить их лингвистические характеристики.

[5]Полагаем, что перевод (не важно машинный или выполненный человеком) является тем более качественным, чем меньше отличие между относительным уклонением характеристик хаотичности текста от средних по соответствующему языку в языке оригинала и в языке перевода.

информативность текстов. Детальное описание библиотек, использованных при предварительной обработке языка (Приложении В).

2.2.1 Построение семантических траекторий текстов

Для построения семантической траектории [103, p. 20-26] текста необходимо получить векторные представления слов исследуемого языка. С этой целью, мы использовали метод сингулярного разложения TF-IDF (TF – term frequency, IDF – inverse document frequency) матрицы. А именно, формальная модель языка включает: корпус $\mathfrak{Z} = (\Omega_1, \cdots, \Omega_N)$ – набор текстов, словарь $\aleph = (\lambda_1, \cdots, \lambda_M)$ – уникальные слова языка. Для заданных $\mathfrak{Z}$ и $\aleph$ строится матрица весовых коэффициентов $W = (w_{ij})$, где каждый элемент вычисляется по формуле (2.2.1) $TF\,(i, j) \cdot IDF\,(i, \mathfrak{Z})$, где

$$TF(i,j) = \frac{n_{i,j}}{\sum_{t_{i'} \in \Omega_j} n_{i',j}}, IDF(i,\mathcal{J}) = \frac{N}{|\{\Omega_k \in \mathcal{J} : t_i \in \Omega_k\}|} \quad (2.2.1)$$

где $n_{i,j}$ - количество раз, когда слово t_i $(1 \leq \text{i} \leq \text{M})$ встречается в документе Ω_j $(1 \leq j \leq N)$.

Для матрицы W, построенной с помощью алгоритма TF-IDF, её сингулярное разложение даётся произведением:

$$W = U\Lambda V^T \quad (2.2.2)$$

Матрица W подвергается сингулярному разложению (SVD) с параметром $d \leq min(M, N)$ где U и V представляют ортонормированные матрицы $M \times d, N \times d$ сингулярных векторов, а Λ – диагональную матрицу $d \times d$ сингулярных значений. Данное разложение обеспечивает оптимальное приближение исходной матрицы в L_2-норме. Для численной реализации применялся модифицированный алгоритм Голуба-Кахана-Ланцоша [120], демонстрирующий повышенную вычислительную эффективность при работе с разреженными матрицами, реализованный в библиотеке SciPy языка Питон. d-мерный эмбеддинг [121] для слова d_i обозначается как слово $u_i\Lambda$ (где u_i – это i -я строка матрицы U). Существенным достоинством метода является то, что для получения эмбеддингов меньшего размера $k < d$ достаточно сократить исходные эмбеддинги размера d до значений размера k. Это позволяет значительно сократить вычислительные ресурсы, что важно для рассматриваемой задачи. Большинство других методов предполагают пересчет вложений для каждого значения d, в отличие от нашего подхода.

Для построения семантической траектории [41, p. 113934] для текста мы используем следующую процедуру. Сначала мы делим текст на n-граммы. n-грамма – это последовательность из n смежных слов в тексте, где n является гиперпараметром. Например, для текста "Lorem ipsum dolor sit amet" и $n = 2$ мы получим следующий список n -грамм: ("Lorem", "ipsum"), ("ipsum", "dolor"),

("dolor", "sit"), ("sit", "amet"). Во-вторых, мы создаем векторные представления для каждой n-граммы, которое представляет собой конкатенацию векторных представлений слов в ней. Таким образом, векторное представление имеет размер $n \times d$. Наконец, мы строим временной ряд, в котором каждое наблюдение – это векторное представление соответствующей n-граммы. Мы собираемся показать, что этот временной ряд является траекторией хаотической динамической системы в nd мерном семантическом пространстве.

2.2.2 Плоскость энтропии-сложности

Для определения хаотичности временного ряда и его отличия от простых детерминированных или стохастических рядов, Мартин, Пластино и Россо [59, p. 439-461] предложили подход, основанный на определении положения точки на плоскости "энтропия - сложность". Предлагаемый подход основан на расчете двух ключевых характеристик - энтропии и сложности - для временного ряда $\{x_i\}$. Полученные значения интерпретируются через сравнение точки с теоретически установленными границами (нижней и верхней границей), что позволяет классифицировать тип исследуемого временного ряда.

Алгоритм основан на анализе порядковых шаблонов [58, p. 174102] временного ряда $\{x_i\}$, которые представляют собой двоичные векторы, описывающие взаимное сравнение значений элементов ряда. В рамках алгоритма временной ряд делится на D – мерные векторы $s_i = (x_{i-(D-1)}, x_{i-(D-2)}, \dots, x_{i-1}, x_i)$, и вычисляются порядковые паттерны порядка n. Порядковый номер s_i – это перестановка $i_0, i_1, \dots, i_{D-1}$ множества $[0, 1, \dots, D-1]$ такая, что

$$x_{i-i_{D-1}} \leq x_{i-i_{D-2}} \leq \cdots \leq x_{i-i_0} \tag{2.2.3}$$

и $i_j < i_{j-1}$, если $x_{i-i_j} = x_{i-i_{j-1}}$.

Алгоритм позволяет оценить вероятность P_i появления i-й порядковой структуры на основе ее частоты в рассматриваемом временном ряду. Затем рассчитываются две основные хаотические статистики: сложность и энтропия. Нормированная энтропия Шеннона $H[P]$, связанная с распределением $P = \{P_j, j = 1, \dots, D!\}$ вычисляется в соответствии с формулами (2.2.4), (2.2.5). :

$$S[P] = -\sum_{i=1}^{n!} P_i \ln(P_i) \tag{2.2.4}$$

$$H[P] = \frac{S(P)}{S_{max}} \tag{2.2.5}$$

Понимание функции вероятностей P, связанной с временным рядом, является ключевым аспектом в определении меры сложности сложности $C[P]$:

$$C[P] = Q[P, P_e] \cdot H[P]. \tag{2.2.6}$$

где P_e - равномерное распределение, где $S_{max} = \ln(n!) = S[P_e]$ (энтропия достигает своего максимума именно при этом распределении.)

Сложность Q – это расхождение Дженсена-Шеннона между заданным и равномерным распределениями:

$$Q[P, P_e] = Q_0 \cdot \left(S\left[\frac{P+P_e}{2}\right] - \frac{1}{2}S[P] - \frac{1}{2}S[P_e]\right), \tag{2.2.7}$$

где Q_0 - константа нормализации.

Таким образом, семантическую траекторию текста можно представить, как точку на плоскости, которая называется плоскость "энтропия–сложность". Именно расположение временного ряда относительно теоретических границ определяет тип временного ряда [122]. Точки в нижнем правом углу соответствуют простым стохастическим процессам, точки в нижнем левом углу относятся к простым детерминированным процессам, а хаотические процессы расположены близко к верхней границе теоретического предела [121, p. 957-964].

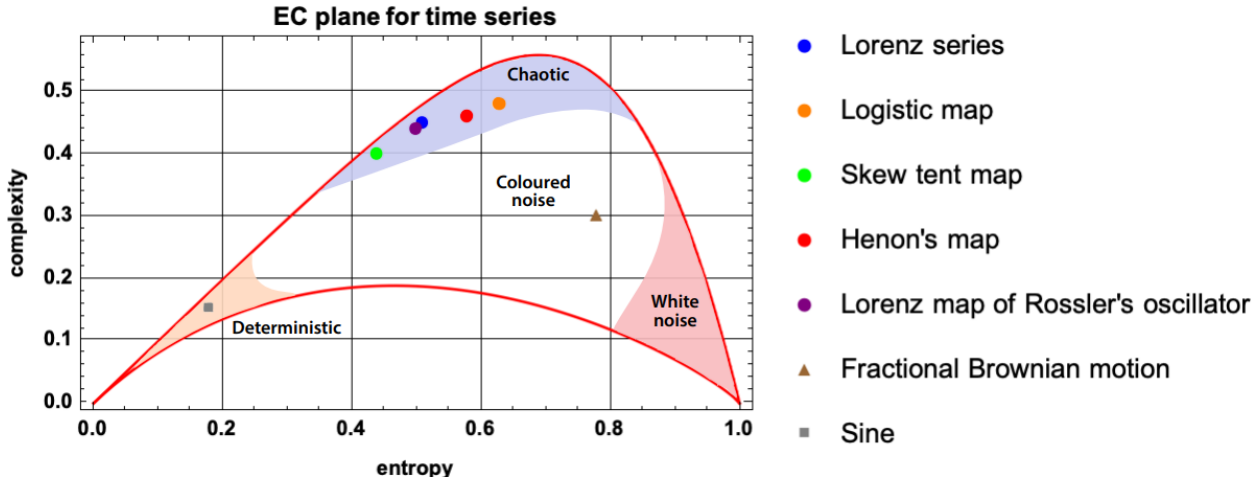


Рисунок 2.2 – Стохастический (треугольники), хаотический (диски) и простой детерминированный (квадраты) на плоскости энтропия–сложность

Примечание – Составлено по источнику [103, p. 22].

На рисунке 2.2 визуализированы верхние и нижние теоретические границы динамических режимов для одномерных временных рядов, включая области хаотических, детерминированных и стохастических процессов. Диаграмма отражает критические значения энтропии и сложности, разделяющие различные типы динамики, с выделением областей каждого класса процессов.

В рамках исследования, представленном в статье М.Т. Мартина, А. Пластино и О.А. Россо [122, p. 154102], были рассмотрены верхние и нижние теоретические границы для пары энтропий и сложности. В работах, таких как

[122, p. 154102; 123-125], продемонстрировано, что подавляющее большинство рядов, для которых известна хаотичность на основе теоретических соображений, соответственно попадает в хаотическую область на плоскости энтропий-сложности в соответствии со значениями пары энтропий и сложности.

В работе [56] в рамках широкомасштабного вычислительного эксперимента было установлено, что для всех естественных языков в подавляющем большинстве случаев семантические траектории представляют собой хаотические временные ряды. Это позволяет определять допустимые соотношения значений параметров n и d (количество слов в n-грамме и размерность пространства вложения, соответственно). Если значение параметра d слишком мало (при заданном n), мы получаем чисто случайные процессы (что для семантических траекторий, очевидно, невозможно); далее с увеличением d мы входим в область хаоса (корректные соотношения n и d); при дальнейшем увеличении d мы входим в область простых детерминированных процессов (что, очевидно, также не отвечает при роде рассматриваемых рядов).

2.3 Численные методы оценки внутренней размерности

Размерность данного множества можно определить различными способами, опираясь на различные математические дисциплины [64, с. 3-334]. В дальнейшем будем использовать следующие определения размерности геометрического объекта:

1. Топологическое (Лебегово покрытие) измерение метрического пространства [126] $(X;\rho)$, с функцией расстояния ρ, определяется как: $d_T = \inf\{i \in \mathbb{N} | \forall \epsilon > 0\ \exists C(X;\epsilon, i+1)\}$; $C(X;\epsilon, i+1)$ является конечным открытым $(i+1)$- множительным $\epsilon -$ покрытием X.

2. Размерность Хаусдорфа [13, p. e0164658] определяется как: $d_H = \sup\{\beta: \sup m(\epsilon,\beta) > 0 \,|\epsilon > 0\}$; $m(\epsilon,\beta) = \inf\sum_{\{A_i\}\in C(X;\epsilon)}(\operatorname{diam} A_i)^\beta$, $\operatorname{diam} A_i < \epsilon$.

Первое из приведенных выше определений не допускает нецелых значений размерности исследуемого геометрического объекта (алгоритмы получения оценки размерности для исследуемого геометрического объекта с нецелой размерностью обычно дают ближайшее целое число), второе – допускает.

Одной из ключевых целей настоящей оценить внутреннее размерности языковых объектов. Согласно концепции Пестова [62, p. 2959-2963] и Канца-Шрайбера [65, p. 3-366], внутренняя размерность d^*, определяется как функционал $d^*(X;\rho,V) \to \mathbb{R}$ отображающий пространства X (mm-пространство) (т.е. пространство X, заданное метрикой ρ и мерой V) в вещественное число. Данный функционал должен удовлетворять фундаментальным требованиям. Во-первых, аксиома концентрации (Леви): для параметризованного семейства пространств (X_d) одинаковой формы $(X;\rho,V)$, расхождение границы $\partial(X_d)$ стремится к бесконечности $\partial(X_d) \uparrow \infty$ тогда и только тогда, когда (X_d) образует семейство Леви. Во-вторых, аксиома непрерывности в смысле метрики концентрации, введённой М. Громовым [60, p. 4-584]: если последовательность пространств (X_d) сходится к пространству X в смысле

метрики концентрации Громова $d_{conc}(X_d, X) \to 0$, то значения функционала сходятся: $\partial(X_d) \to \partial(X)$.

В-третьих, аксиома нормировки (сфера): для стандартной d - мерной сферы $\mathbb{S}^d$ значение функционала имеет нормализацию $\partial(\mathbb{S}^d) = \Theta(d)$. Применительно к исследованию "языковых" множеств, ключевой задачей становится оценка их внутренних размерностии (включая топологические и хаусдорфову размерность) посредством анализа функционала d^*.

Формально необходимо:

– для заданного множества текстов естественного языка $\Im = (\Omega_1, \cdots, \Omega_N)$ построить множества $\aleph_n(d)$ для всех уникальных n -грамм, $n = 1..N, d = 1..D$ Здесь d – размерность пространства вложения; n – число слов в рассматриваемых n–граммах. При этом предполагается, что множество текстов $\Im$ представляет собой репрезентативную выборку текстов, написанных на соответствующем естественном языке;

– для каждого n, используя полученные множества $\aleph_n(d)$ для всех $d = 1..D$, вычислить оценку внутренней размерности $\hat{d}_i = f_i(\{\aleph_n(d)\})$ пространства этих n-грамм. Вычисление должно быть проведено с применением нескольких подходов, формализуемых как функции f_i;

– проверить гипотезу о фрактальности множеств n–грамм исследуемого естественного языка $\aleph_n(d)$, с помощью тех подходов f, которые позволяют получить нецелые значения $\hat{d}$;

– кластеризовать результаты значения размерности $\hat{d}$ рассматриваемых естественных языков с помощью двух разных методов (Wishart и K-Means). Провести сравнительный анализ полученных оценок внутренней размерности для различных языков, чтобы выявить возможные закономерности и различия, связанные с их структурными особенностями.

2.3.1 Геометрическое представление естественного языка

SVD. Для построения множеств точек $\aleph_n(d)$ по набору текстов естественного языка (контексту) $\Im$ использовали сингулярное разложение (SVD) матрицы совместной встречаемости [121, p. 957-964].

Контексту $\Im = (\Omega_1, \cdots, \Omega_L)$ и множеству слов $\aleph = (\lambda_1, \cdots, \lambda_M)$, мы строим $M \times L$ матрицу $S = (s_{ij})$, элементы которой вычисляются как:

$$s_{i,j} = (1 - \varepsilon_i) \frac{k_{i,j}}{\sum_{i' \in \Omega_j} k_{i',j}}. \tag{2.3.1}$$

где $k_{i,j}$ – количество вхождений слова λ_i в текст Ω_j;

ε_i – нормализованная энтропия слова λ_i в $\aleph$:

$$\varepsilon_i = -\frac{1}{\ln L} \sum_{j=1}^{L} \frac{k_{i,j}}{\tau_i} \ln \frac{k_{i,j}}{\tau_i}, \tag{2.3.2}$$

$$\tau_i = \sum_{j=1}^{L} k_{i,j}. \quad (2.3.3)$$

Следовательно, высокие значения матрицы указывают на слова, уникальные для конкретного текста (часто в нем встречающиеся), но редко встречающиеся в других контекстах рассматриваемого корпуса.

К полученной матрице S применяется метод сингулярного разложения (SVD) [127]. Дополнительные детали о методе получения векторных представлений с использованием SVD представлены в разделе 2.2.1.

Word2Vec (CBOW). Другой метод, который используем для получения векторных представлений, – это подход Word2Vec Continuous Bag of Words (CBOW) [126, p. 4-892].

Извлечение векторных представлений осуществляется на основе вероятностного распределения, сформированного в процессе обучения полносвязной нейронной сети. Ключевой принцип архитектуры CBOW (Continuous Bag-of-Words) реализуется последовательно: 1) итеративная обработка корпуса посредством контекстного окна фиксированного размера; 2) оптимизация логарифмической функции правдоподобия для предсказания целевого (центрального) слова по его окружающему контексту: вычислительная модель представляет собой однослойный перцептрон, на входной слой которого проецируются векторные эмбеддинги контекстных слов; 3) выходной слой генерирует вероятностное распределение, указывающее на принадлежность скрытого элемента словарной единице в рамках текущего окна. Данный подход позволяет восстанавливать слову по ее семантическому окружению.

Итогом обучения выступает матрица эмбеддингов W (формируемая весовыми коэффициентами скрытого слоя), где каждому слову поставлен в соответствие уникальный вектор-столбец. Фундаментальное достоинство алгоритма Word2Vec состоит в том, что линейные преобразования над полученными векторами демонстрируют смысловую согласованность [128], а метрика косинусного сходства коррелирует с уровнем семантической сходства. Параллельно, процедура обучения матрицы эмбеддингов Word2Vec характеризуется существенно меньшей вычислительной сложностью относительно альтернативных методов построения языковых моделей, что обусловило ее выбор в качестве вспомогательного инструментария в рамках данного исследования.

2.3.2 Оценка внутренней размерности. Оценка Швайнхарта

Для оценки внутренней размерности набора слов или n-грамм в рамках изучаемых естественных языков в работе задействованы два подхода. Оба метода используют в качестве основы графовое представление исследуемой выборки данных и вычисленное для него минимальное остовное дерево.

Выбор графовых аппроксиматоров обусловлен их устойчивостью к зашумленности данных [66, p. 263-276]. Принципиальное преимущество рассматриваемых алгоритмов обусловлено их методом выявления структурных зависимостей (локальных и глобальных): он базируется исключительно на определении частичного порядка для окрестностей графовых вершин. Такой подход кардинально ослабляет воздействие шума на выходные характеристики. Учитывая специфику исследуемых данных – тексты из открытых интернет-источников, сохраняющие существенную зашумленность даже после очистки – эта устойчивость приобретает ключевое значение. Ситуация осложняется неполнотой существующих моделей внутреннего устройства языка, что создаёт трудности в разграничении истинных структурных элементов и шумовых артефактов.

Оценка Швайнхарта. Первый метод опирается на теорему Швейнхарта [68, p. 107291], расширенную теоремой Стилом [129]:

Теорема (Schweinhart). Пусть на метрическом пространстве определена d-Альфорсова регулярная мера μ. Рассмотрим последовательность $\{x_l\}_{l\in\mathbb{N}}$, представляющую собой выборку независимых одинаково распределенных величин с мерой μ. На основе этой последовательности определено значение случайной величины E_α^0:

$$E_\alpha^0(x_1,\dots,x_\ell)=\sum_{e\in T_\ell(\{x_\ell\}_{\ell\in\mathbb{N}})}|e|^\alpha. \tag{2.3.4}$$

где $|e|, e\in T$ обозначает вес ребра e в минимальном остовном дереве $T(x_1,\dots,x_l)$, рассчитанный как Евклидова метрика между его вершинами. В случае $0<\alpha<d^*$:

$$const_1\le\frac{E_\alpha^0(x_1,\dots,x_\ell)}{\ell^{1-\alpha/d^*}}\le const_2, \tag{2.3.5}$$

$$\frac{ln\left(E_\alpha^0(x_1,\dots,x_\ell)\right)}{ln(\ell)}\to\frac{d^*-\alpha}{d^*} \tag{2.3.6}$$

при $\ell\to\infty$; $const_1$ *и* $const_2$ положительны и не зависят от ℓ.

Используя теорему Швайнхарта, вычисляется фрактальная размерность $\hat{d}_{\mathrm{Schw}}$ геометрического объекта, описываемой точечными множествами $\{x_\ell\}_{\ell\in\mathbb{N}}, x_\ell\in\mathbb{R}^d$. Процедура оценки включает: вычисление величин E_α^0 для различных ℓ с последующим построением модели линейной регрессии.

$$\ln\left(E_\alpha^0(x_1,\dots,x_\ell)\right)=\ln(\psi(\alpha,d^*))+\ln(\ell)\frac{d^*-\alpha}{d^*}, \tag{2.3.7}$$

Решение этого уравнения нелинейным методом наименьших квадратов (МНК) позволяет получить оценку внутренней (хаусдорфовой) размерности объекта как верхнего измерения коробки [68, p. 107291], используя $\frac{d^*-\alpha}{d^*} \approx \frac{\hat{d}_{\mathrm{Schw}}-\alpha}{\hat{d}_{\mathrm{Schw}}}$.

2.3.3 Оценка внутренней размерности. Байесовская оценка

Второй подход к оценке внутренней (топологической) [66, p. 263-276], использует случайные переменные, основанные на различных графовых репрезентациях данных: а именно, динамических графах k-соседей, минимальных остовных деревьях и диаграммах сфер влияния. В частности, для оценки внутреннего измерения мы используем сумму квадратов степеней вершин для минимального остовного дерева (Приложение Г).

На выборке $\{x_\ell\}_{\ell\in\mathbb{N}}$, из d-мерного пространства (X_d, ρ, V), оснащенного метрической функцией ρ и вероятностной мерой V, необходимо выполнить следующее:

1. Формировать граф данных с использованием взвешенных евклидовых расстояний и последующее построение его минимального остовного дерева $T(x_1, \dots, x_n)$. Статистический анализ фокусируется на распределении степеней вершин этого дерева. Согласно результатам [66, p. 263-276], а также более ранним работам. Согласно фундаментальному результату [129, p. 809-825], математическое ожидание степени вершины в минимальном остовном дереве (MST) обладает свойством инвариантности относительно объема выборки n. Критически важно, что для данных, генерируемых непрерывной мерой в $\mathbb{R}^d$, доля вершин степени j сходится почти, наверное, к предельному значению, определяемому исключительно парой (j, d). Более поздние исследования Брито и соавт. [66, p. 263-276] демонстрируют монотонную зависимость топологии MST от размерности: при ее возрастании наблюдается сопутствующий рост как числа терминальных вершин (степени = 1), так и узлов с высокой степенью. Эта закономерность обосновывает использование моментов распределения степеней MST порядка больше единицы в качестве релевантных дескрипторов для идентификации топологической размерности пространства:

$$M_\ell(d) \coloneqq \frac{1}{\ell} \sum_{x_i \in T(\{x_\ell\}_{\ell\in\mathbb{N}} \subset X_d)} \big(\deg(x_i)\big)^2, \tag{2.3.8}$$

2. Оценить параметры (при условии, что последовательность случайных переменных M_ℓ сходится, при $\ell \to \infty$, к случайной переменной с нормальным распределением [66, p. 263-276].

$$\hat{\mu}_i \coloneqq \frac{1}{K} \sum_{j=0}^{K} M_\ell \left(i; \{x_\ell\}_j \sim U(\mathbf{0_i}, \mathbf{1_i})\right), \tag{2.3.9}$$

$$\hat{\sigma}_i^2 \coloneqq \frac{\ell}{K-1}\sum_{j=0}^{K}\left(M_\ell\left(i;\{x_\ell\}_j{\sim}U(\mathbf{0_i},\mathbf{1_i})\right)-\hat{\mu}_i\right)^2. \quad (2.3.10)$$

K определяет число независимых выборок, где каждая представляет собой набор из ℓ точек, равномерно распределенных в единичном гиперкубе размерности i.

3. Оценить вероятность:

$$p\{\hat{d}=i\}=p(i|M'_\ell)=\frac{N\left(M'_\ell;\hat{\mu}_i,\frac{\hat{\sigma}_i^2}{\ell'}\right)}{\sum_j N\left(M'_\ell;\hat{\mu}_j,\frac{\hat{\sigma}_j^2}{\ell'}\right)}, \quad (2.3.11)$$

4. Вычисляет оценку внутренней размерности $\hat{d}_{BQY}$ как математическое ожидание $E[\hat{d}]$, округленное до ближайшего целого числа:

$$\hat{d}_{BQY}(M'_n)=\text{round}\left(\sum_i p(i|M'_\ell)*i\right). \quad (2.3.12)$$

Брито и др. Гаусса ([20] доказали, что $\hat{d}_{BQY}\to d_T$, почти наверняка, как $K,\ell,\ell'\to\infty$.

2.4 Методы топологического анализа данных и поиска персистентных гомологий

2.4.1 Семантическое пространство

Для построения элементов семантического пространства (эмбеддингов n-грамм) использовали модель CBOW [127, p. 2-22]. Выбираем эту модель благодаря наличию семантических характеристик векторных представлений, включая близость эмбеддингов слов, имеющих схожее значение (синонимов), а также возможность «семантической арифметики». (“woman” + “king” - “man” = “queen”). Крупномасштабное моделирование определило оптимальные размеры семантического пространства (число элементов в эмбеддингах), d = 100 (мы протестировали все возможные значения d от 1 до 200). Основной метод создания эмбеддингов для биграмм и триграмм заключается в конкатенации векторов входящих в них слов. Это приводит к тому, что размерность семантического пространства биграмм достигает 200 (исходный размер слова 100, умноженный на 2), а для триграмм – 300 (100 * 3).

Поскольку семантические свойства CBOW наиболее точно отражаются при использовании косинусного расстояния, именно она применяется во всех экспериментах. Для вычисления функции расстояния между n-граммами вводится следующая функция:

$$\rho((a_1, a_2, \dots, a_n), (b_1, b_2, \dots, b_n) = \frac{1}{n} \sum_{i \in \{1..n\}} \min_{j \in \{1..n\}} cosine(a_i, b_j) \quad (2.5.1)$$

Ключевая особенность данной метрики: для любых n-грамм, представляющих собой перестановки одних и тех же слов (например, "шагать весело" и "весело шагать"), расстояние будет нулевым.

2.4.2 Персистентная гомология

Данное исследование направлено на выявление семантических "слепых зон" языка – областей с низкой частотностью употребления языковых единиц (слов, биграмм и др.). Для этого в семантическом пространстве анализируются одномерные гомологии[6] первого порядка H_1, интерпретируемые как структурные пространства ("дырки"), в отличие от компонент связности нулевого порядка H_0. Методологической основой поиска послужило построение фильтрации Вьеториса-Рипса [130].

Определение. (Комплекс Вьеториса-Рипса): для заданного параметра ϵ комплексом Вьеториса-Рипса $VR(\epsilon)$ называется симплициальный комплекс, включающий в себя все симплексы σ, для которых максимальное расстояние между вершинами (диаметр) не превосходит ϵ: $VR(\epsilon) = \{\sigma | diam(\sigma) \leq \epsilon\}$.

Определение. Фильтрация Вьеториса-Рипса представляет собой монотонно возрастающую последовательность симплициальных комплексов $VR(\epsilon_1) \subseteq VR(\epsilon_2)$ ..., порождаемую ростом значений параметра $\epsilon_1, \epsilon_2, \dots$.

Посредством фильтрации Вьеториса-Рипса, управляемой возрастающим параметром ϵ, регистрируются критические точки появления ("рождения") и устранения ("смерти") классов гомологий различных размерностей, что позволяет выделить персистентные (устойчивые) топологические особенности. Конкретно, момент рождения n-мерной гомологии соответствует значению ϵ, на котором данная гомология впервые устраняется из-за образования нового (n+1)-мерного симплекса в комплексе.

2.4.3 Контуры “дырок” естественного языка

При анализе структуры естественного языка ключевая задача – не просто подсчёт оценки количества топологических «дыр» (чисел Бетти), но и определение их границ. Необходимо идентифицировать, какие именно единицы (слова или словосочетания) формируют контур каждой «дыры», располагаясь на её границе. Для решения этой задачи в настоящем исследовании применяется метод поиска представителей классов гомологий (homology representatives), разработанный Чуфаром и Вирком (гомологию [131, 132]). Данный подход позволяет выявить цепочки элементов, топологически обрамляющие дырки в векторном пространстве.

[6]Количество классов гомологий выражается через фундаментальную топологическую характеристику — числа Бетти, играющие ключевую роль в методах топологического анализа данных.

Алгоритм позволяет определить цепочки точек, которые оконтуривают дыры (границы всех дыр). Суть алгоритма состоит из двух отдельных фаз: 1) сокращение кобордной матрицы и выявление симплексов, которые способствуют расчету представительных циклов; 2) сокращение матрицы границ, сосредоточенное только на столбцах, выявленных на первой фазе. В основном, сокращенная кобордная матрица используется для определения исчезновения симплексов, игнорируя все остальные столбцы во время сокращения матрицы границ. Чуфар и Вирк сформулировали следующий алгоритм:

Входные данные: X – конечное метрическое пространство. Построить фиксированную инъективную фильтрационную функцию, связанную с фильтрацией Рипса для X. Сгенерировать матрицы кобордности (d_k).

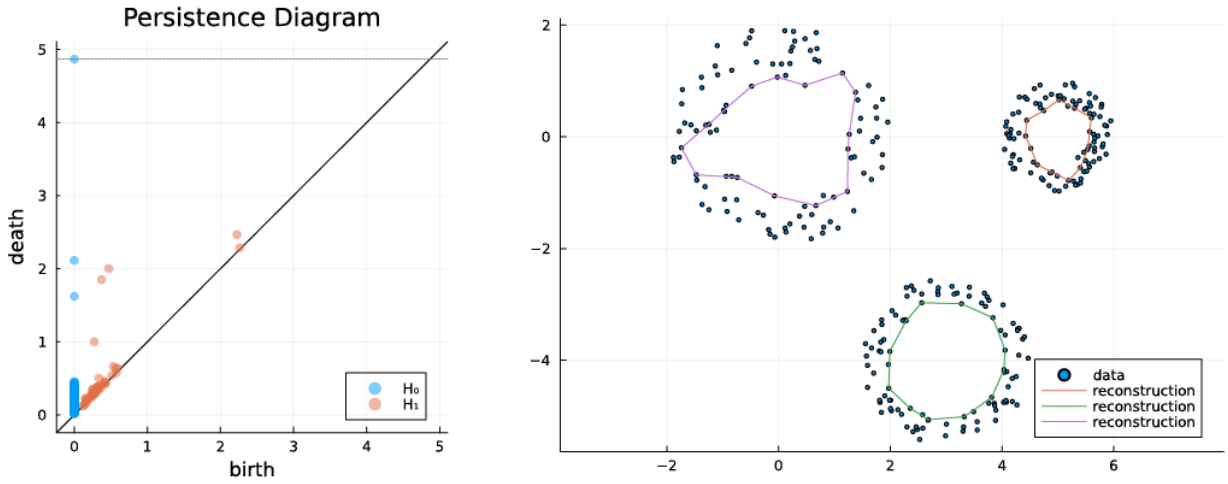
1) сократить каждую матрицу кобордности d_k. Согласно Теореме 1, можем извлечь гомологически мёртвые симплексы (т.е. ко-гомологически рождающиеся симплексы) и существенные симплексы;

2) для каждого k пусть D_k будет подматрицей матрицы границ гомологии ∂_k, состоящей из столбцов, соответствующих гомологически мёртвым симплексам и существенным симплексам. Сохраняем индексы симплексов для маркировки столбцов;

3) вычислить представления персистентной гомологии, сократив D_k. (Алгоритм сокращения по сути представляет собой процесс сокращения столбцов слева направо. Основная идея заключается в том, чтобы проверить, является ли граница добавленного симплекса σ_i (т.е. $Col\partial(i)$) гомологически тривиальной в K_{i-1} или нет, проверив, может ли она быть выражена как линейная комбинация предшествующих столбцов).

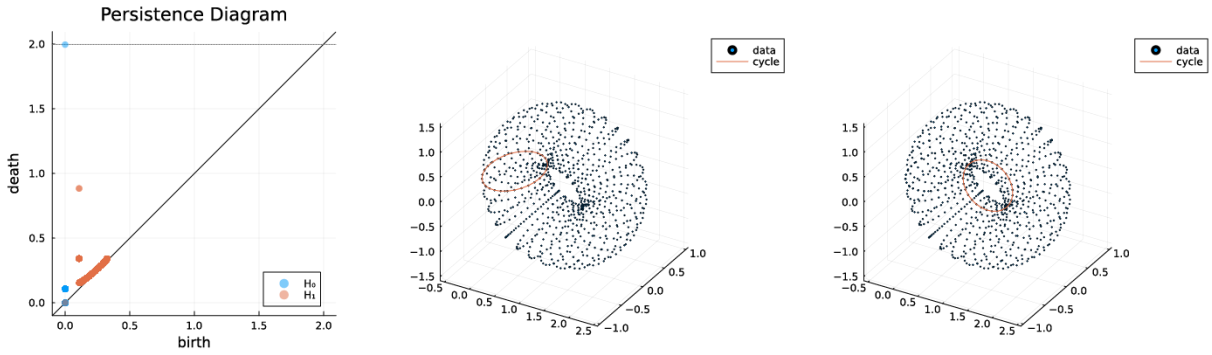
Выходные данные: вернуть представители гомологии из сокращённых форм матриц D_k.

Рисунки 2.3, 2.4 демонстрируют визуализацию топологических «дыр» (гомологий первого порядка) и их границ в двумерном и трехмерном семантических пространствах. На двумерной проекции (рисунок 2.3) алгоритм идентифицирует три структурные «дырки», что подтверждается диаграммой персистентности – на ней наблюдаются три наиболее выраженных (персистентных) класса. В трехмерном случае (рисунок 2.4), для модели двумерного тора, корректно детектируются две гомологии первого порядка.



а б

Рисунок 2.3 – Диаграмма, отображающая персистентность классов гомологий (слева) и оконтуривания границ гомологий первого порядка в двумерном семантическом пространстве. (справа)



а б в

Рисунок 2.4 – Диаграмма, отображающая персистентность классов гомологий (слева) и оконтуривания границ гомологий первого порядка в трехмерном семантическом пространстве для двумерного тора (справа).

При анализе текстов естественного языка подавляющее большинство точек-эмбеддингов не имеют никакого отношения к текущей обрабатываемой “дырке”. Чтобы ускорить алгоритм, целесообразно разбить семантическое пространство на подобласти. Кластеризация K-Means [133] использовалась для разделения пространства на кластеры. Алгоритм строит кластерную структуру путем итеративного объединения наиболее близко расположенных групп кластеров; его следует применять многократно для получения сопоставимых кластеров (от пяти до десяти тысяч элементов в каждом кластере). Для того чтобы выделить дыры, принадлежащие крупномасштабной (грубозернистой) структуре языка (а не вызванные локальными неоднородностями выборки), сравниваем диаметры дыр с диаметрами групп синонимов для языка.

Задача идентификации ботов

Для решения задачи идентификации ботов в качестве классификаторов использовали характеристики распределений расстояний от n-грамма до

ближайшей “дырки”. Экспериментальная установка предполагает применение относительно простых классификаторов (Support Vector Classifier - SVC, Decision Tree - DT, Random Forest - RF) для идентификации текстового происхождения (человека или бота). Для обучения и последующего тестирования моделей был создан датасет на основе выводов четырех языковых моделей (mGPT, GPT-2, YaLM, LSTM). Критически важным элементом методологии является схема "обучение на одних генераторах – тестирование на других": из четырех LLM случайным образом выбирались две для формирования обучающей выборки, а оставшаяся пара использовалась исключительно для оценки (всего реализовано 6 сценариев). С целью исключения дисбаланса классов на этапе тестирования, для каждого запуска отбиралось идентичное число экземпляров: 600 текстов художественной литературы и по 300 текстов, созданных каждой из двух LLM, фигурирующих в тестовом наборе.

2.5 Методы идентификации ботов

Стремительная эволюция ИИ-технологий обусловила переход социума в фазу постправды – реальность, наблюдаемую уже сегодня. Генеративный ИИ (чат-боты, языковые модели), являясь катализатором этого перехода, оказывает деструктивное воздействие на языковую составляющую человеческой идентичности. Традиционно ее становление определялось корпусом текстов антропогенного происхождения (в подавляющей массе бенефициарных) – от фольклора до академических материалов. Грядущая тотальная автоматизация контент-генерации приведет к перманентному пребыванию новых поколений в среде, насыщенной синтетическими текстами ("неотличимыми от аутентичных"), производимыми ботами на порядки быстрее человека. Это формирует императив для разработки универсальных систем детекции ИИ-текстов.

Настоящее исследование развивает подходы, предложенные в работах [103, p. 20-26] и [134], где изучались фундаментальные структурные расхождения между сгенерированных различными ботами, и текстов, написанных людьми. В настоящей работе расширяем исследование и рассматриваем другие языки, а также другие модели текстовых ботов. Что наиболее важно, сосредотачиваемся на задаче поиска эффективных признаков, которые могут быть использованы для идентификации текстов, сгенерированных различными типами ботов (вместо построения модели классификации, которая может обнаружить только конкретного бота). По этой причине предлагаем модифицированное постановление задачи: для данного естественного языка отделить тексты, написанные людьми, от всех текстов, сгенерированных ботами. Чтобы проверить производительность соответствующего классификатора и его способность к обобщению, предлагаем случайным образом разделить набор ботов на ботах, использованных для построения классификатора, и тех, которые не использовались; последние используются для генерации текстов для тестового набора. Следовательно, для построения классификатора следует использовать наиболее общие характеристики семантического пространства. Кроме того,

необходимо проверить соответствующие гипотезы для нескольких языков, желательно из различных языковых групп и семейств.

Данное исследование включает извлечение признаков для текстов, написанных людьми, и текстов, сгенерированных ботами, посредством кластерного анализа (алгоритмы Wishart, K-Means и их нечеткие модификации) и нелинейного динамического анализа (меры энтропии и сложности). На этапе классификации намеренно применяются базовые алгоритмы (SVM, деревья решений, случайный лес) для детекции ботогенерированного контента. Для построения классификаторов рассматриваем следующие гипотезы:

1. В пространстве векторных представлений слов области, «населенные» ботами и «посещаемые» людьми совпадают – если что-то и отличается, так это то, что люди и боты используют один и тот же словарь естественного языка. Напротив, для пространства n-грамм можно выявить области людей и ботов: люди, как правило, создают неожиданные последовательности слов чаще, чем боты (люди пишут более изысканно, «острее»).

2. Смещение распределения n-грамм текста в сторону областей, плотно «заселенных» ботами, и отдаление от областей «обитания» человеческих n-грамм – ключевой признак синтетического происхождения контента.

3. Метрики компактности кластеров (индекс силуэта, диаметр) для n-грамм синтетических текстов существенно превосходят аналогичные показатели для текстов, написанных людьми. Статистический анализ подтверждает значимое различие характеристик кластеризации.

4. Анализ методом нечёткой кластеризации выявляет, что кластеры, соответствующие текстам людей, обладают менее выраженным ядром и большей размытостью, что противопоставляется резко очерченным кластерным ядрам, типичным для бот-текстов.

5. Между характеристиками нелинейной динамики семантических траекторий человеческих текстов и текстов ботов существуют статистически значимые различия.

В естественных науках принято различать два основных класса задач – прямые задачи и обратные задачи [135]. В рамках прямых задач предполагается решение некоторой чётко сформулированной задачи; в рамках обратных задач, напротив, предполагается “восстановление задачи по наблюдениям”, точнее определение по наблюдениям характеристик той системы, которая породила наблюдаемую выборку. По аналогии предлагаем разделить задачи обработки естественного языка на прямые и обратные, с парами типа «разработать бота–обнаружить бота», «перевести текст–автоматически оценить качество перевода» и т.д.

Исследуемое пространство Ω охватывает всю совокупность текстов на естественном языке, как созданных человеком, так и произведенных произвольными бот-системами. Оно структурируется на подмножество $A = \{\alpha_1, \dots, \alpha_a\}$[7], содержащее исключительно человеческие тексты, и подмножество M, представляющее собой объединение $M = \bigcup_{j=1}^{L} M_j$ всех текстов,

[7]Ср. др.-греч. ἄνθρωπος – человек

сгенерированных ботами, где $M_j = \{\mu_1, \dots, \mu_{m_j}\}$ – множество текстов, сгенерированных j-м ботом[8]. также рассматривается пространство ботов M: какие-то из ботов (не все) тексты которых участвовали в формировании пространства Ω. Требуется построить признаки $\Lambda = \{\lambda_1, \dots, \lambda_k\}$ и построить на их основе классификатор $R = R(\Lambda)$ с F1-оценкой классификации выше порога r^*.

Выборка человеческих текстов подвергается разделению на обучающую и тестовую части. Параллельно конструируются независимые обучающая и тестовая выборки для текстов, продуцированных ботами. Важным является процедура формирования обучающего и тестового множеств для текстов, сгенерированных ботами: не делим случайным образом на две части множество текстов, сгенерированных ботами, но делим случайным образом на две части множество ботов $\{M_j, j = 1..l\}$: тексты ботов, попавших в первое множество, используются для формирования классификатора, тексты ботов, попавших во вторую часть, – для его тестирования. При этом предполагается, что число текстов и распределение размеров текстов приблизительно одинаковы для частей обучающего множества, отвечающих людям и ботам; аналогичное предположение делается для тестового множества.

Блок-схема процесса классификации текста показана на рисунке 2.5.

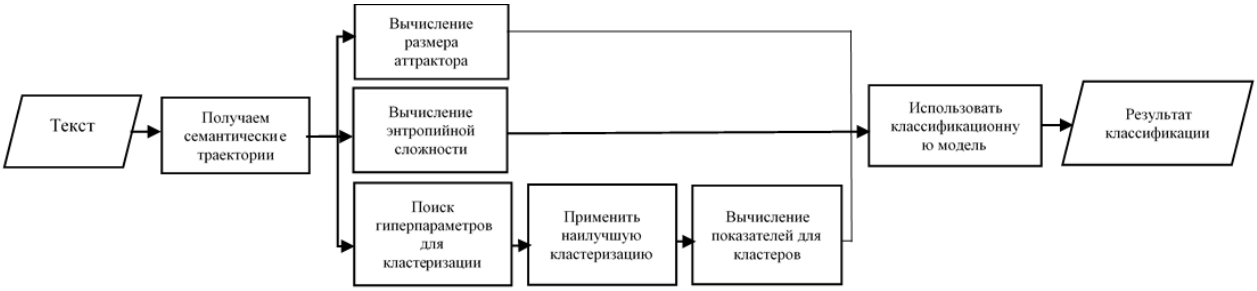


Рисунок 2.5 – Блок-схема методологии

2.5.1 Сбор и предварительная обработка данных

С целью построения универсальных моделей идентификации ботов, мы провели вычислительные эксперименты для различных языков, принадлежащих различным языковым группам и семьям (таблица 2.1).

Таблица 2.1 – Подробности о корпусах, написанных людьми

Язык	Группа	Семья	Число текстов	Средний размер текста, в словах
Русский	Восточнославянская	Индоевропейская	6,429	14,510
Английский	Германская	Индоевропейская	11,052	21,744
Немецкий	Германская	Индоевропейская	12,503	72,878
Вьетнамский	Вьетская	Австроазиатская	1,071	54,496
Французский	Романская	Индоевропейская	8,405	66,946

[8]Ср. др.-греч. μηχάνημα – машина

Для каждого языка в таблице 2.1 представлена информация о его языковой группе и языковой семье, количестве текстов, написанных людьми, в выборке; и среднем размере текста. Использовали корпусы текстов национальных литератур: полагаем, что художественные тексты обладают наивысшей степенью отражения языковой реальности, выступая оптимальным материалом для анализа лингвистических изменений и тенденций. Все тексты были собраны из открытых источников, таких как Проект “Гутенберг” и др. Для анализа отбирались тексты, длина которых превышала 100 слов. Каждый текст для каждого корпуса текстов подвергался токенизации и лемматизации. В рамках предобработки осуществляем маскировку местоимений, предлогов, числительных и имен собственных, применяя для их идентификации модели тегирования. Используем четыре типа ботов различной сложности для создания эффективного алгоритма для простых и сложных моделей: долгосрочная краткосрочная память (LSTM), GPT-2, многоязычный GPT (mGPT), «Еще одна языковая модель» (YaLM). Процедура генерации текстов, сгенерированных ботами, формализована в виде алгоритма и представлена в (Приложении Д).

Генерация текста. Модель LSTM обучалась на корпусах текстов, написанных людьми, с использованием следующих гиперпараметров: размер пакета – 16, длина последовательности – 256, число эпох – 10 000.

Модели GPT-2 применялись в виде предобученных версий, доступных в библиотеке Hugging Face. Для различных языков использовались соответствующие языковые модели GPT-2. Перечень моделей и количество обучаемых параметров приведены в (Приложении Д).

Модель GPT-2 для русского языка содержит большее число параметров по сравнению с моделями для других рассматриваемых языков. Модели с 124M и 356M параметрами демонстрировали повышенную склонность к повторяемости фрагментов текста, что ограничивало возможность генерации длинных последовательностей. В связи с этим для русского языка использовалась модель большего размера.

Модель mGPT содержит 1.4B параметров, модель YaLM – 1B параметров.

Процедура генерации была организована таким образом, чтобы распределение длины сгенерированных текстов соответствовало распределению длины текстов, написанных людьми (в логарифмическом масштабе). Подробные характеристики сгенерированных текстов приведены в (Приложении Д).

Получение векторных представлений слов (эмбеддингов)

В работе использовались различные способы получения векторных представлений слов. Первый подход – разложение по сингулярным значениям (SVD) матрицы между словами и текстами [121, p. 957-964]. Второй, Word2Vec [136], использует нейронную сеть для изучения словесных ассоциаций. Оба метода широко используются для изучения семантики текстов: и SVD, и Word2Vec эмбеддинги улавливают структурные связи между словами. Предоставляем в разделе 2.3 «Геометрическое представление естественного языка», где рассматриваются математические основы применения SVD и архитектуры Word2Vec (CBOW).

Получаем векторные представления для n-грамм, конкатенируя векторы слов, составляющих эту n-грамму. Таким образом, чтобы построить выборку векторных представлений n-грамм, необходимо: 1) собрать корпус текстов на естественном языке; 2) провести предварительную обработку; 3) создать словарь (множество слов с соответствующими векторами); 4) создать словарь n-грамм для всех n-грамм данного языка.

Семантическая траектория

Ключевым для методологии выступает концепт семантической траектории – многомерного временного ряда, возникающего при последовательном векторном моделировании слов текста [103, p. 20-26]. Важнейший вывод указанного исследования подтверждает статистически значимую хаотичность данных траекторий в корпусах английского и русского языков.

Характеристики для обнаружения текстовых ботов

В данном разделе представлен комплекс параметров, используемых алгоритмом для верификации происхождения текстовых данных (текста, написанные людьми и ботами). Поскольку сами классификаторы являются довольно простыми, эффективность представленных в работе алгоритмов обнаружения ботов в основном зависит от характеристик, которые они используют.

Плоскость энтропии-сложности

В исследовании Мартино, Пластино, Россо [122, p. 154102] представлен диагностический алгоритм, позволяющий надежно отделить хаотические ряды во временных рядах от детерминированных процессов, с другой стороны, от стохастических процессов. Метод использует энтропию и сложность временного ряда. Процедура расчета энтропии и сложности приведена в разделе 2.2.

Этот метод позволяет не только построить классификатор, но и установить значения числа слов в n-грамме, n, и размерности пространства вложения, d, при которых семантическая траектория отражает истинную «динамику» текста. При слишком малых значениях данных величин, семантическая траектория отображается в точку на плоскости энтропия – сложность, принадлежащую области чисто случайных процессов; при увеличении значений указанных величин точка смещается в область хаотических процессов, при дальнейшем увеличении точка перемещается в область простых детерминированных процессов. Согласно нашей гипотезе, подлинная динамика естественно языкового текста проявляется при таких комбинациях n и d, когда его семантическая траектория достигает в область хаоса. Теоретическое обоснование данного утверждения представлено в соответствующем исследовании [50, p. 1-20].

Гипотеза, которая проверяется в рамках данного подхода, формулируется следующим образом: “При определённых значениях числа слов в n-грамме n и размерности пространства вложения значения d координаты точек на плоскости “энтропия – сложность”, отвечающие семантическим траекториям текстов, написанных людьми, статистически значимо отличаются от координат точек, отвечающие семантическим траекториям текстов, сгенерированных ботами”.

Естественно, все пары n и d, для которых данная гипотеза верна, принадлежат области хаоса.

2.5.2 Оценка размерности аттракторов семантических траекторий

Этот подход также основывается на семантических траекториях текстов: для решения поставленной задачи оцениваются различные характеристики динамической системы (и ее аттрактора), которая генерирует наблюдаемые (многомерные) временные ряды (семантические траектории).

Гипотеза, которая проверяется в рамках данного подхода, формулируется следующим образом: “Значения координат точек на плоскости 'энтропия–сложность' определяются конкретными комбинациями параметров: длины n - граммы n и размерности пространства векторных представлений (эмбеддингов) d значения характеристик семантических траекторий текстов, написанных людьми, статистически значимо отличаются от характеристик семантических траекторий текстов, сгенерированных ботами”. В работе для решения поставленной задачи используем энтропию Реньи динамических систем для достижения поставленной цели [64, с. 3-334; 65, p. 3-368]. Методика вычисления обобщённой размерности и её применение к анализу текстовых данных подробно рассмотрены в работе [3, p. e2550].

2.5.3 Кластеризация n-грамм и показателей связности кластеров

В рамках этого раздела рассматриваются специфические особенности, обусловленные крупномасштабной, грубой топологией пространства семантических вложений. Ключевое наблюдение заключается в том, что данные различия (подробно обсуждаемые в последующих разделах) не проявляются для единичных слов $n = 1$ – что закономерно, поскольку и люди, и боты оперируют идентичным лексическим базисом – однако отчетливо фиксируются для биграмм, триграмм и языковых конструкций большей длины. ($n > 1$) – люди, как правило, создают более неожиданные, нетривиальные последовательности слов. Гипотезы, которые необходимо проверить в рамках этой модели, следующие:

1. Четкая кластеризация n-грамм выявляет статистически значимое превосходство в компактности кластеров сгенерированных текстов ботов над кластерами текстов, написанных людьми.
2. Нечёткая кластеризация n-грамм выявляет фундаментальное различие: для сгенерированных текстов ботов характерны кластеры со статистически значимо более выраженными ядрами и меньшие области нечёткости, чем тексты, составленные человеком.
3. Крупномасштабное моделирование позволяет выявить области семантического пространства, которые чаще «посещаются» людьми, и области, которые чаще «посещаются» ботами. Соответственно, ключевая гипотеза данного исследования формулируется следующим образом: «Вероятность появления n-грамм, ассоциированных с 'бот-доминантными' регионами семантического пространства, статистически значимо выше в текстах, сгенерированных ботами, по сравнению с текстами, написанными людьми».

Все вышеперечисленные предположения методологически опираются на процедуру кластеризации векторных представлений языковых n-грамм, реализуемую посредством того или иного алгоритмического подхода. Следует подчеркнуть, что кластеры, возникающие при кластеризации эмбеддингов *n*-грамм, могут обладать весьма прихотливыми формами; кроме того, общее число кластеров априори неизвестно. Это накладывает жёсткие требования на выбор алгоритма кластеризации. С одной стороны, он не должен требовать заранее известного количества кластеров, с другой стороны, он должен допускать кластера различной формы. В то время как для выполнения первого требования можно запустить алгоритм для различных заданных количеств кластеров (в разумных пределах); для выполнения второго требования следует использовать специфические алгоритмы кластеризации – необходимо четко понимать, что любой алгоритм кластеризации неявно определяет, что он «считает» кластером.

В рамках нашего исследования применяется ряд алгоритмов кластеризации. В их числе классический метод K-Means [132, p. 671-705], и его нечеткий вариант – алгоритм C-Means [136; 137]. Дополнительно используются алгоритм Wishart [138], а также его расширенная версия, оперирующая нечеткими числами, которая была представлена [138, p. 97; 139]. Алгоритм Wishart основывается на аппарате теории графов и концепции кластера как области сравнительно больших значений функции плотности вероятности: сочетание этих двух подходов позволяет выделять кластера практически любой структуры и самостоятельно, в процессе кластеризации определять их число. Подробное описание реализованных алгоритмов кластеризации приведено в работе автора [3, p. e2550].

При чёткой кластеризации текстов мы использовали следующие характеристики компактности чётких кластеров:

1. Количество уникальных элементов (без повторений) в кластере: n-грамма может встречаться в текстах, а значит и в кластере, много раз – однако мы считаем её только один раз. Нормировка признака выполняется путем деления на число элементов в наиболее крупном кластере для рассматриваемого разбиения данных.
2. Число уникальных элементов (без повторений) внутри кластера (характеристика) нормируется к нормированию по общему количеству уникальных векторов во всей выборке.
3. Число уникальных элементов (с повторениями) внутри кластера (характеристика), нормируется относительно размера максимального по наполненности кластера текущей кластеризации (размер определяется как общее число вхождений элементов).
4. Число элементов кластера (с повторами), нормируется к совокупному числу элементов (с повторами) в кластеризации.
5. Максимальное значение расстояния между любым элементом кластера и его элементом в центре кластера.
6. Среднее значение расстояния между любым элементом кластера и его элементом в центре кластера.

7. Максимальное значение расстояния между любым элементом кластера
8. Среднее значение расстояния между любым элементом кластера.

Во всех случаях используется евклидово расстояние. Подробное описание построения нечётких чисел и соответствующей метрики приведено в работе автора [3, p. e2550].

Указанные выше подходы подразумевают различные способы кластеризации данных, а следовательно, и различные кластеризации, полученные как при различных значениях гиперпараметров для одного и того же алгоритма, так и при использовании различных алгоритмов кластеризации. Для сравнения качества кластеризаций мы используем меры качества кластеризации [140].

Данные для обучения и тестирования

В соответствии с постановкой задачи, множество ботов M случайным образом разбивается на обучающую и тестовую выборки для классификаторов. Случайный отбор дал: GPT2 и YaLM для обучающих ботов (M_1, M_2), LSTM и mGPT для тестовых (M_3, M_4).

Размеры выборок едины для всех языков исследования: обучающая выборка содержит по 2000 текстов составленные людьми и 2000 бот-сгенерированных текстов, тестовая – по 600 текстов каждого типа. Распределение длин текстов между категориями статистически близко (Приложение Д).

Рассматриваются наиболее простые модели: метод опорных векторов, дерево решений и случайный лес. Гиперпараметры каждой из моделей подбираются с помощью 10-кратной кросс-валидации. Мы установили порог F1-меры ($r^* = 0.9$).

Выводы по второму разделу

Таким образом, результаты, представленные в данной главе, позволяют прийти к следующим выводам:

1. Представлены методы анализа крупномасштабной структуры естественного языка, сосредоточив внимание на статистических подходах и аналитических инструментах, используемых для исследования 52 языков, относящихся к 18 языковым семьям. Исследование охватывает широкий спектр языков, представляя 74,3% мировой языков, что подтверждает репрезентативность выборки. Ключевым аспектом анализа является понимание естественных языков как самоорганизующихся критических систем. Применение мощных статистических методов для проверки гипотез, связанных с степенными законами распределения, позволяет глубже осознать внутренние структуры и динамику языков. Различные подходы, включая метод коллапса данных и подход Клаузета-Чализи-Ньюмана, обеспечивают надежные оценки параметров распределений, таких как показатель степени и нижние пределы, что служит основой для классификации языков по их статистическим свойствам.

2. Примененные методы анализа хаотичности семантических траекторий, основанные на преобразовании слов естественного языка в d-мерные вектора

представления, позволяют оценить хаотичность текстов на 52 языках, что открывает новые возможности для сравнительного анализа языков с различными грамматическими и семантическими особенностями. Исследование показало, что подавляющее большинство семантических траекторий являются хаотическими, что подтверждает их природу как самоорганизующихся критических систем. Применение статистических методов, таких как анализ энтропии и сложности, позволяет классифицировать временные ряды и выявлять их хаотические свойства, что имеет значительный прикладной интерес в задачах идентификации ботов и оценки качества перевода.

3. Оценка внутренней размерности языковых объектов с использованием различных численных методов, таких как теорема Швайнхарта и подход Брито и др., позволяет глубже понять структурные особенности языков. Эти методы, основанные на графовых представлениях и минимальных остовных деревьях, обеспечивают устойчивость к шуму и позволяют выявлять фрактальные свойства n-грамм. Проведенный анализ внутренней размерности языков демонстрирует, что полученные оценки могут быть использованы для выявления закономерностей и различий в языковых структурах.

4. Методы топологического анализа данных и поиска персистентных гомологий, примененные для изучения семантического пространства, позволяют выявить семантические "слепые зоны" языка, а также определить контуры "дырок" в языковых структурах. Использование модели CBOW для построения эмбеддингов n-грамм и фильтрации Вьеториса-Рипса для анализа гомологий предоставляет инструменты для исследования топологических особенностей языков. Алгоритм поиска представителей классов гомологий позволяет идентифицировать границы "дыр", что имеет важное значение для понимания структуры естественного языка и может быть использовано в задачах идентификации ботов, основанных на характеристиках распределений расстояний от n-грамм до ближайших "дыр".

5. Методы идентификации ботов, основанные на анализе текстов, сгенерированных различными моделями, позволяют эффективно отделять тексты, написанные людьми, от текстов, сгенерированных ботами. Использование кластерного анализа и нелинейного динамического анализа для извлечения признаков, а также применение базовых алгоритмов классификации, таких как SVM и случайный лес, демонстрирует высокую эффективность в обнаружении синтетического контента. Исследование подтверждает, что характеристики семантического пространства и распределения n-грамм могут служить надежными индикаторами для идентификации ботов, что имеет важное значение в условиях растущей автоматизации контент-генерации.

3 СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ КРУПНОМАСШТАБНОЙ СТРУКТУРЫ ЕСТЕСТВЕННОГО ЯЗЫКА

3.1 Языковые характеристики

Чтобы сопоставить результаты с известными лингвистическими классификациями языков, рассмотрели следующие характеристики естественных языков: порядок слов, тип (локус) маркирования, выравнивание, морфологическую сложность и направленность заголовка [141]:

Порядок слов относится к расположению слов в предложении. Разные языки имеют различные модели порядка слов, такие как субъект-глагол-объект (SVO), субъект-объект-глагол (SOV), глагол-субъект-объект (VSO) или свободный порядок [141].

Тип (локус) маркировки относится к грамматической маркировке основных аргументов (например, подлежащего, дополнения) на глаголах или существительных. Языки используют различные стратегии для маркировки грамматических отношений [141; 142]:

– языки с маркированием главных элементов отмечают грамматические отношения на главном элементе фразы, обычно это глагол или существительное;

– языки с маркированием зависимого элемента маркируют грамматические отношения на зависимых элементах фразы, таких как подлежащее или дополнение;

– в языках с непоследовательным маркированием маркировка грамматических отношений может варьироваться в зависимости от таких факторов, как одушевленность, определенность или тематические роли;

– в языках с нулевым маркированием грамматические отношения не маркируются явно морфологически; вместо этого порядок слов, контекст и прагматические подсказки определяют роли участников в предложении.

Выравнивание в лингвистике относится к паттернам синтаксического и семантического выравнивания между различными частями предложения, такими как подлежащие и дополнения [143]:

Нейтральное выравнивание: в языках с нейтральным выравниванием минимальная или нет морфологической различимости между маркировкой подлежащих и дополнений.

Аккузативное выравнивание: в аккузативном выравнивании проводится различие между маркировкой подлежащего и дополнения.

Эргативное выравнивание: в эргативном выравнивании морфологическая маркировка различает подлежащее непереходного глагола и подлежащее переходного глагола.

Морфологическая сложность включает в себя различные типы, такие как синтетические, аналитические и изолирующие морфологические типы, которые представляют собой различные уровни морфологического богатства и сложности в языках [144]:

– синтетические языки демонстрируют развитую морфологию, где ключевые грамматические значения (время, аспект, наклонение, падеж, согласование) реализуются на уровне слова через изменение его формы;

– аналитические языки: эти языки имеют более низкий уровень морфологической сложности и передают грамматическую информацию преимущественно через вспомогательные глаголы, порядок слов и контекст;

– изолирующие языки: эти языки характеризуются самым низким уровнем морфологической сложности и имеют минимальную инфлекционную морфологию.

Направленность заголовка относится к направлению, в котором строятся фразы, при этом языки обычно отображают конструкции с заголовком в начале или в конце [141].

Характеристики языка представлены (Приложений В).

Для получения векторных представлений униграмм, биграмм и триграмм для 52 естественных языков мы формируем корпуса (национальной литературы), используя тексты, доступные для скачивания из открытых источников. В частности, для русского языка объем корпусов ($|\Im_1|$) составляет 6429 текстов, среди которых выявлено 103952 уникальных униграмм, 14775439 уникальных биграмм и 147533142 уникальных триграмм. Аналогично, для английского языка ($|\Im_2|$) количество текстов составляет 11052, что включает 94087 уникальных униграмм, 9490603 уникальных биграмм и 39173019 уникальных триграмм. Подобные данные также собраны для остальных 52 естественных языков.

Предварительная обработка корпуса текстов включала:

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

– токенизацию слов (преобразование имена собственных, разделители предложений, числительных);

– лемматизацию слов (для лемматизации русского языка использовался – natasha, для лемматизации английского – spacy); (полезно снижать вариативность языковых единиц за счёт приведения слов к начальной форме, то есть выделения корней).

3.1.1 Статистический анализ степенного распределения в языках

Представлены значения показателей степенной зависимости для всех 52 языков (Приложение В). Для почти всех языков оба статистических теста подтверждают, что их соответствующие распределения следуют степенному закону. В https://github.com/erbolova1983/Investigation-of-NL-structures.git представлены подробные результаты для всех 52 естественных языков. Важным является то, что существует хорошее согласие между оценками показателя распределения и нижним пределом, полученным этими методами (см. раздел 2.1). Показаны относительную ошибку в оценке показателя и качество соответствия по критерию Колмогорова-Смирнова (Приложение В). Эсперанто является единственным исключением (рисунок 2.1б): его распределение следует гауссовскому распределению ($p - value = 0.0005$), что обоснованно, поскольку это единственный искусственный язык в выборке. Очевидно, создание естественного языка – непростая задача. Проведённый анализ позволяет классифицировать языки на основе конечности или бесконечности

математического ожидания и дисперсии их соответствующих степенных распределений (таблица 3.1). Здесь большинство языков (63%) имеют бесконечное математическое ожидание и дисперсию; 18% показывают конечное математическое ожидание, но бесконечную дисперсию; и наконец, 19% языков имеют как конечное математическое ожидание, так и дисперсию. Для получения более значимых классификаций языков на основе указанных характеристик мы использовали деревья решений и алгоритм кластеризации на основе плотности, где были классифицированы четыре переменные: показатель τ и нижний предел, полученный из обоих статистических методов.

Таблица 3.1 – Математическое ожидание и дисперсия

Математическое ожидание	Дисперсия	Пример	Количество	Языки	Процент
Бесконечный	Бесконечный	$\tau \leq 2$	32	Белорусский, китайский, коптский, чешский, дхолуа, голландский, английский, французский, финский, немецкий, хинди, исландский, индонезийский, японский, латинский, малаялам, навахо, норвежский, оромо, персидский, польский, румынский, русский, сербский, осетинский, испанский, шведский, тагальский, тайский, тувинский, удмуртский, украинский	62.75
Конечный	Бесконечный	$2<t\leq3$	9	Атикмек, бенгальский, чеченский, кабильский, казахский, латышский, кечуа, узбекский, вьетнамский	17.65
Конечный	Бесконечный	$\tau > 3$	10	Амхарский, арабский, баскский, болгарский, панджаби, сингальский, суахили, табасаранский, татарский, турецкий	19.6

Деревья решений

В данном исследовании использован алгоритм деревьев решений [145]. Мы проводим его пять раз, используя различные языковые характеристики (Приложение В).

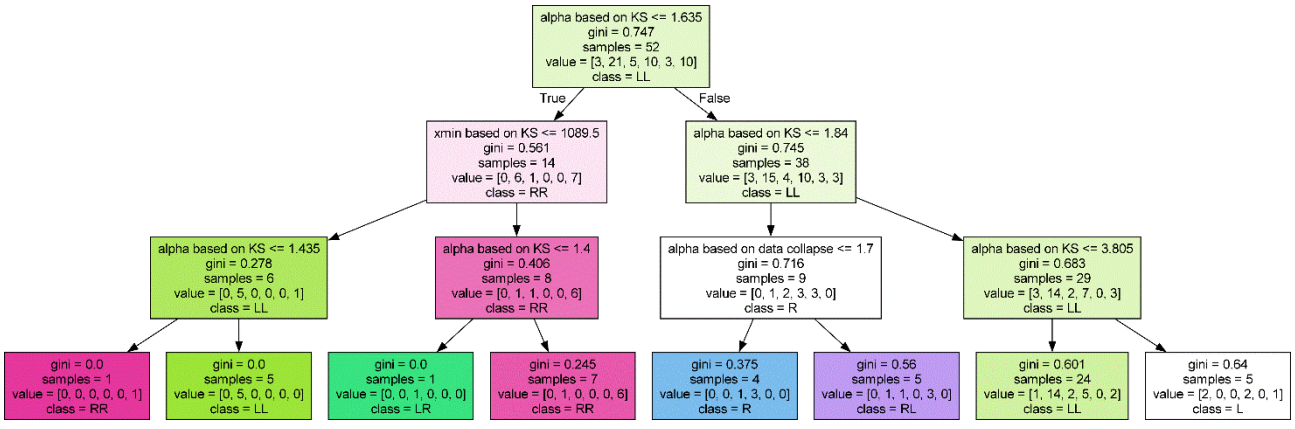


Рисунок 3.1 – Дерево решений, когда метка – направление головы

Примечание – Мы можем заметить, что корневой узел – это τ (показатели КС)

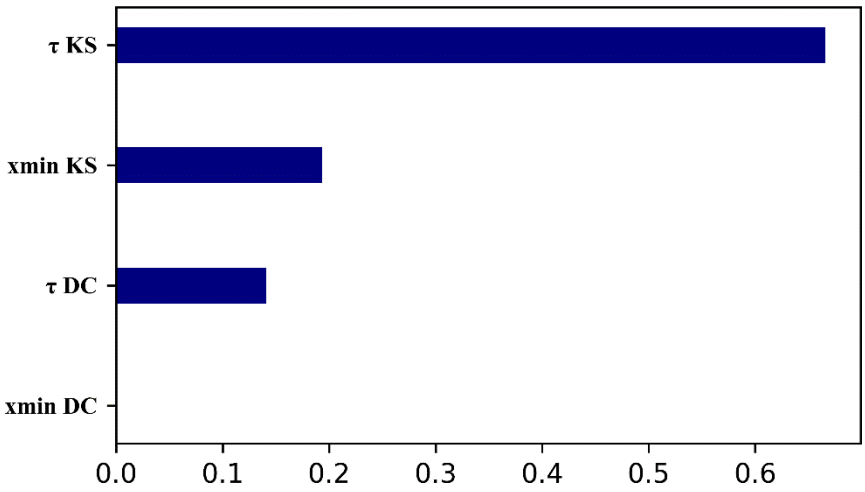


Рисунок 3.2 – Важность признаков, когда метка – направление головы

Примечание – Показатели τ (KS) имеют наивысшее значение

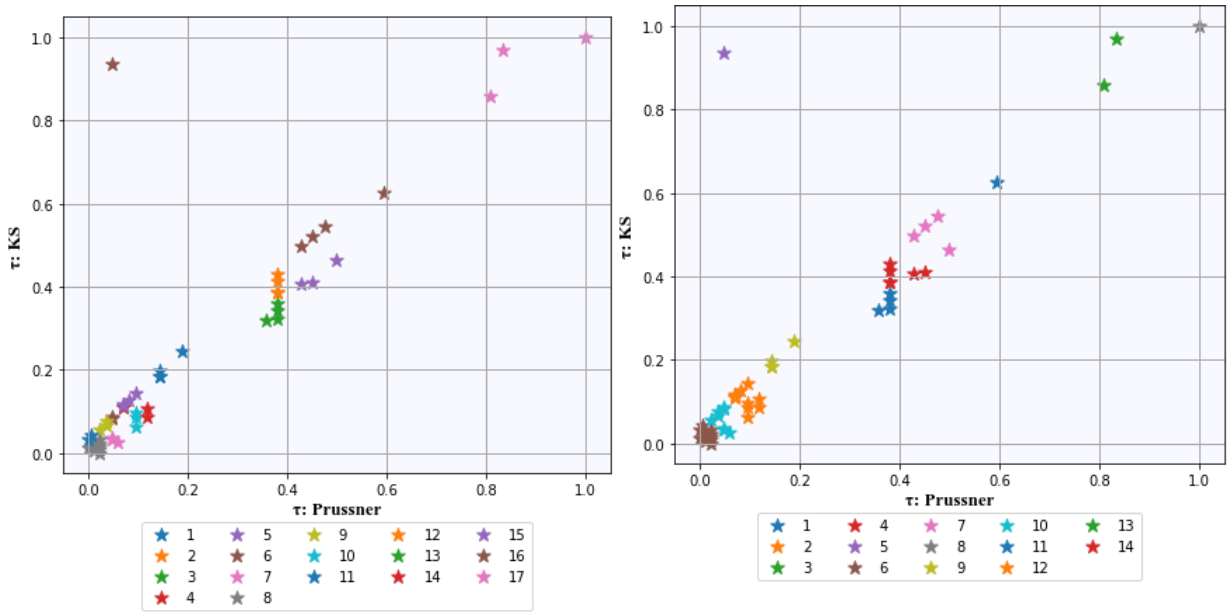
На рисунках 3.1, 3.2 представлены дерево решений и диаграмма важности признаков для метки направления головы (она, похоже, создаёт самое компактное дерево); результаты представлены другие языки https://github.com/erbolova1983/Investigation-of-NL-structures.git.

3.1.2 Кластеризация на основе статистических особенностей аналитических методов

Используемые в данном исследовании алгоритмы – это алгоритм K-means [132, p. 671-705] и алгоритм кластеризации Wishart [49, p. 8474-8477; 50, p. 1-20; 52, p. 3317-3321; 138, p. 97]. Последний метод не требует предварительного определения количества кластеров. Сначала группируем четыре признака: показатель τ и нижний порог, полученные из обоих статистических методов; затем два признака: признака (τ) обоих статистических методов; и, наконец, один признак: наивысшее значение показателя τ, полученное из каждого статистического метода отдельно. Чтобы определить оптимальные гиперпараметры для K-means и Wishart, проводим поиск по сетке, оценивая кластеризацию с помощью индексов силуэта [146], Дэвиса-Болдина [147] и Калинского-Харабаза [148].

Численные значения метрик и результаты подбора гиперпараметров приведены в (Приложении В).

Исходный код и результаты вычислительных экспериментов размещены в открытом репозитории: https://github.com/erbolova1983/Investigation-of-NL-structures. На рисунке 3.3 представлены оптимальная кластеризация на основе двух τ.



а б

а – Wishart: кластер 1 – Белорусский, Китайский, Кабильский, Тувинский; кластер 2 – Бенгальский, Латышский; кластер 3 – Японский; кластер 4 – Хинди; кластер 5 – Финский, Норвежский, Сербский, Украинский; кластер 6 – Голландский, Шведский; кластер 7 – Исландский, Навахо, Персидский; кластер 8 – Английский, Латинский, Малаялам, Румынский, Испанский, Тагальский, Тайский, Удмуртский; кластер 9 – Французский, Немецкий, Русский; кластер 10 – Чешский, Дхолуо, Оромо; кластер 11 – Индонезийский, Осетинский; кластер 12 – Атикамекв, вьетнамский; кластер 13 – Чеченский, Кечуа, Казахский, Узбекский; кластер 14 – Коптский, Польский; кластер 15 – Баскский, Панджабский, Татарский; кластер 16 – Амхарский, Арабский, Табасаранский, Турецкий, Эсперанто; кластер 17 – Болгарский, Сингальский, Суахили; б – K-means: кластер 1 – Чеченский, Казахский, Кечуа, Узбекский; кластер 2 – Финский, Хинди, Японский, Норвежский, Сербский, Украинский; кластер 3 – Суахили; кластер 4: Пенджабский, Татарский; кластер 5 – Эсперанто; кластер 6 – Английский, Индонезийский, Латинский, Малаялам, Осетинский, Румынский, Испанский, Тагальский, тайский, удмуртский; кластер 7 – Амхарский, Арабский, Баскский, Турецкий; кластер 8 – Болгарский; кластер 9 – Белорусский, Китайский, Кабильский, Тувинский; кластер 10 – Голландский, Французский, Немецкий, Исландский, Навахо, Персидский, Русский, Шведский; кластер 11 – Табасаранский; кластер 12 – Коптский, Чешский, Дхолуо, Оромо, Польский., кластер 13 – Сингальский; кластер 14 – Атикамекский, Бенгальский, Латышский, Вьетнамский

Рисунок 3.3 – Результаты кластеризации, полученные с помощью алгоритмов Wishart и K-means на данных[1]

Примечания:
1. [1] – (Приложение В).
2. Представлены результаты алгоритма Wishart справа, а результаты алгоритма K-means – слева.
3. Подфигуры соответствуют двум признакам: τ (KS, Pruessner)

Языковые кластеры

В основном, различные методы кластеризации и несколько вариантов наборов кластеризованных характеристик в основном приводят к одним и тем же

кластерам: кластеры, полученные алгоритмом Wishart, являются подкластером тех, что были получены методом K-means.

1. Язык Эсперанто выделяется как отдельный кластер. Этот результат предполагает, что эсперанто обладает уникальными статистическими характеристиками.

2. Белорусский, китайский, кабильский и тувинский языки были сгруппированы в нескольких симуляциях и обоими алгоритмами. Эта повторяющаяся кластеризация предполагает, что эти языки имеют сильные статистические характеристики, несмотря на их языковое разнообразие.

3. Аналогично, чеченский, казахский, узбекский и кечуа часто формировали кластер.

4. Кроме того, французский, немецкий и русский языки часто кластеризовались вместе, подчеркивая их сильные сходства.

Важно отметить, что полученная классификация значительно отличается от установленной генетической и грамматической классификации, что, вероятно, связано с тем, что для понимания языка в целом необходимо исследовать его семантические и культурные компоненты. Анализ показывает, что оба оцененных τ имеют наибольшую важность среди других признаков.

В настоящем исследовании гипотеза о том, что естественный язык представляет собой самоорганизованно-критичную систему, была подтверждена на значительном числе языков из 18 языковых семей.

В 62,75% случаев соответствующее степенное распределение (число слов в произвольно выбранном тексте для данного языка) обладает бесконечным математическим ожиданием и бесконечной дисперсией; в 17,65% – конечным математическим ожиданием, но бесконечной дисперсией; и в 19,6% – как конечным математическим ожиданием, так и дисперсией. Следует отметить, что распределение для эсперанто, искусственного языка, не удовлетворяет гипотезе; его распределение является нормальным.

Классификация языков по характеристикам указанных распределений, полученная различными методами, оказалась достаточно устойчивой, что свидетельствует о том, что самоорганизованная критичность предоставляет основу для новой классификации языков. Степенные показатели, похоже, являются наиболее значимой статистической особенностью [117, p. 1-13].

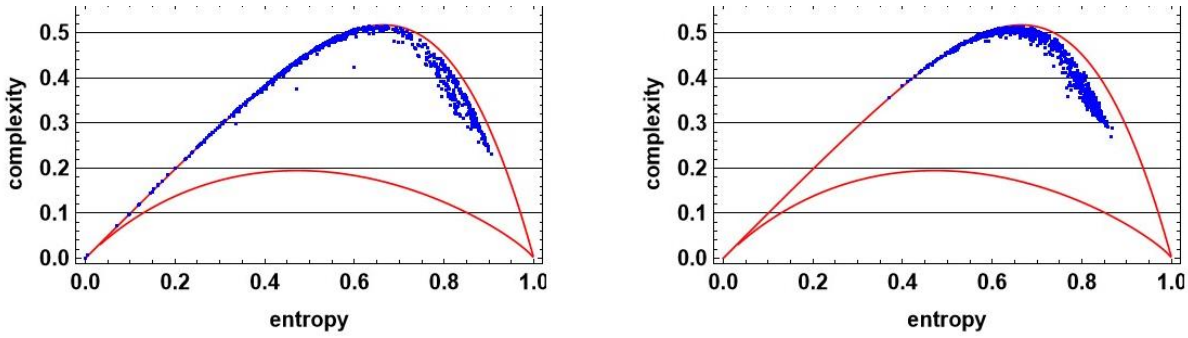
3.2 Хаотичность естественных языков

В настоящем разделе подробно рассмотрим результаты анализа корпусов национальной литературы на русском и английском языках. Характеристики языков представлены[9] (Приложении В). Результаты для других языков представлены https://github.com/erbolova1983/Investigation-of-NL-structures.git. Критическая зависимость эффективности методов наблюдается от размерности пространства вложений (d) и размера n-граммы (n). Фундаментальное предположение: параметры n и d адекватно отражают языковую динамику тогда

[9]Можно также посмотреть на TSNE-портреты языков в https://github.com/erbolova1983/The-chaos-of-natural-languages.git.

и только тогда, когда соответствующая им семантическая траектория попадает "область хаоса" пространства энтропия-сложность[10]. Детальный анализ данных аспектов представлен в исследовании [41, p. 113934; 56]. Моделирование динамики 52 языков было направлено на идентификацию диапазонов параметров, обеспечивающих попадание семантической траектории большинства литературных текстов в хаотическую область пространства энтропия-сложность.

В ссылке (https://github.com/erbolova1983/The-chaos-of-natural-languages.git) приведены минимальные, средние и максимальные значения энтропии и сложности для 52 языков (и названия соответствующих текстов).



а б

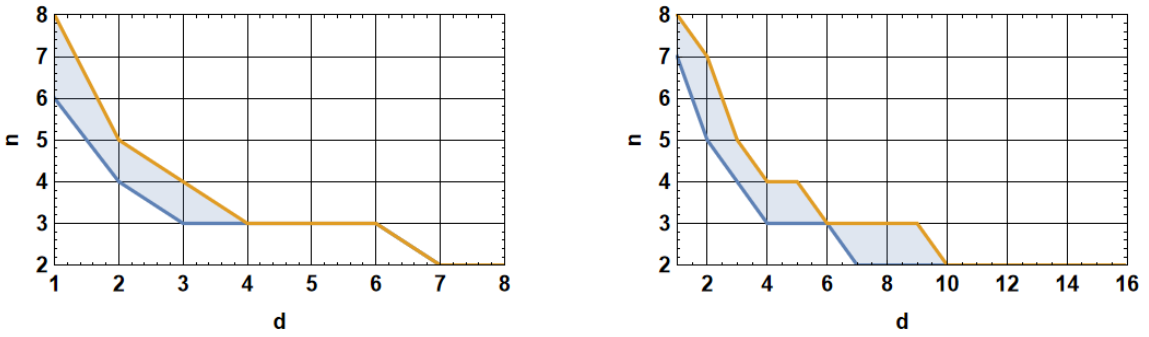
Рисунок 3.4 – Распределение семантических траекторий русской (а) и английской (б) художественной литературы в координатах энтропия–сложность (n = 3, d = 4)

На рисунке 3.4 представлены распределение точек в пространстве энтропия-сложность, отражающие семантические траектории художественных текстов на английском и русском языках. Явно видно, что результаты схожи при равных значениях n и d (с очень малыми значениями), что свидетельствует о том, что рисунок 2 (n=3, d=4) указывает на область хаотических процессов (оптимальные параметры).

На рисунке 3.5 представлены значения n и d для русского (рисунок 3.5а) и английского (рисунок 3.5б) языков, в которых большинство семантических траекторий литературных корпусов находятся в области хаотических процессов. Чтобы установить эти параметры, мы применили три описанных выше критерия. Синяя область обозначает допустимые значения d и n. Значения выше оранжевой линии соответствуют детерминированной области, тогда как значения ниже синей линии указывают на наличие шума. Аналогичные данные для других языков https://github.com/erbolova1983/Investigation-of-NL-structures.git. Однако из-за ограничений по размеру траекторий литературных шедевров, энтропийная сложность не может быть адекватно оценена при очень больших значениях d и

[10]Анализ рядов в рамках подхода плоскость “энтропия - сложность” позволяет отнести к одной из трёх категорий: простые детерминированные, чисто случайные (белые и цветные шумы) и хаотические процессы. Сложно себе представить, что автор литературного шедевра (даже и обычного текста) генерировал текст путём подбрасывания монетки или использования примитивного алгоритма – что оставляет нам только опцию хаотических процессов.

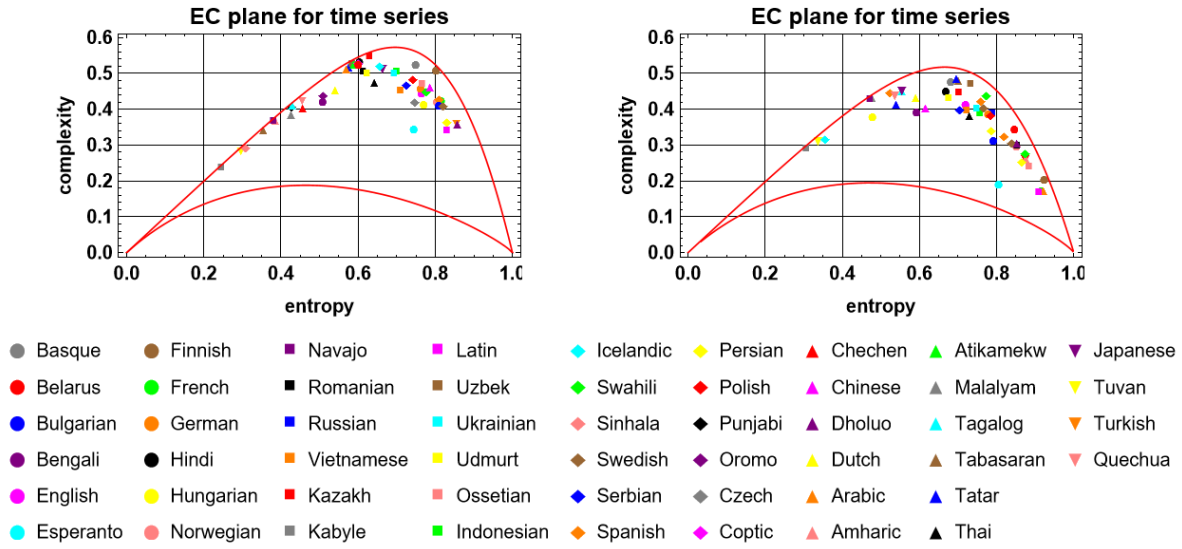
n, что определяет более высокую границу. Интересно, что области оптимальных значений могут существенно отличаться для разных языков. [3, p. e2550] связывают это с порядком слов в языке. Например, при d = 1: для румынского оптимального значения n составляют от 7 до 11, для русского и коптского – от 6 до 8, а для английского, японского и бенгальского – с 7 по 8. Согласно нашей гипотезе, подлинная динамика текста проявляется при таких комбинациях n и d, которые вызывают сдвиг семантической траектории в область хаоса.



а б

Рисунок 3.5 – Допустимые значения d, n для русского (а) и английского (б) языков. Оранжевыми линиями показаны границы для d и n, выше которых тексты попадают в детерминированную область; синими линиями показаны границы, ниже которых тексты попадают в шум

Примечание – Составлено по источнику [41, p. 113934]



а б

а – d=4, n=3; б – d=5, n=3

Рисунок 3.6 – Средние значения энтропии и сложности для разных языков

На рисунке 3.6 показаны средние значения энтропии и сложности для различных d и n. В плоскости энтропия-сложность различные языки

группируются при различных значениях d и n. В целом, эти кластеры относятся к языковым семьям. При d=5, n=3 индоевропейских языков попадают в одну область пространства, изображенную на рисунке 3.6а, 3.6б, а другие группы попадают в другую область пространства. Для различных значений d и n язык эсперанто «отдаляется» от нескольких языковых семей.

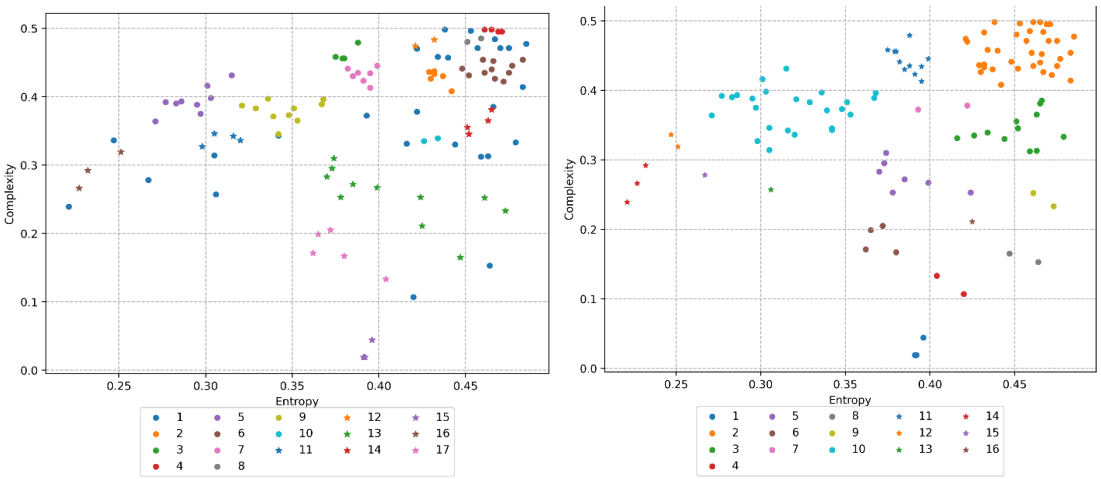
3.2.1 Кластеры пространства признаков

Чтобы кластеризовать пространство признаков, применяем следующую процедуру. Сначала определяем допустимые значения n и d (выделены зеленым цветом на рисунке 3 для английского и русского языков, а также для остальных 52 языков, отображенных в https://github.com/erbolova1983/Investigation-of-NL-structures.git). Далее мы выбираем значения энтропии E_i и сложности C_i, которые находятся в хаотической области, то есть в допустимом диапазоне. Кроме того, в этом интервале мы определяем максимальные значения энтропии E^* и сложности C^* и нормализуем $\hat{E}_i$, $\hat{C}_i$ согласно формуле (3.2.1):

$$\hat{E}_i = \frac{E_i - E^*}{E^*}, \ \hat{C}_i = \frac{C_i - C^*}{C^*} \tag{3.2.1}$$

В исследовании применяем несколько алгоритмов для кластеризации векторного пространства рассматриваемых языков, основываясь на анализе n-грамм. Векторные представления этих n-грамм создаются путем конкатенации векторов слов, входящих в них, что позволяет отразить структуру текстов. В ходе анализа используются два эффективных алгоритма кластеризации: Wishart [138, p. 98] и DBSCAN [149].

На рисунке 3.7 представлены различные методы кластеризации, такие как Wishart и DBSCAN, которые, несмотря на разнообразные характеристики, в основном показывают схожие результаты. Например, кластеры, созданные с помощью алгоритма Wishart, можно рассматривать как подкластер для кластеров, сформированных алгоритмом DBSCAN. В частности, языки, входящие в кластеры 3 (голландский, табасаранский, казахский, японский) и 7 (голландский, турецкий, чеченский, казахский, удмуртский) по алгоритму Wishart, можно считать подкластером для кластера 11 (голландский, турецкий, табасаранский, чеченский, казахский, удмуртский, японский) из результатов DBSCAN. Язык эсперанто выделяется в отдельный кластер в обоих методах. Основное различие между результатами двух алгоритмов составляет 8,5% в процентном соотношении [56].



а б

а – Wishart(n_cluster=17): кластер 1 – Английский, Русский, Норвежский, Украинский, Сербский, Суахили, Шведский, Шведский, Коптский, Коптский, Польский, Польский, Французский, Чешский, Осетинский, Оромо, Латинский, Тувинский, Сингальский, Навахо, Дхолуо; кластер 2 – Венгерский, Венгерский, Украинский, Румынский, Суахили; кластер 3 – Голландский, Табасаранский, Казахский, Японский; кластер 4 – Английский, Русский, Финский, Арабский; кластер 5: Бенгальский, Бенгальский, Оромо, Латинский, Панджаби, Узбекский, Хинди, Хинди; кластер 6 – Норвежский, Немецкий, Финский, Исландский, Исландский, Болгарский, Чешский, Французский, Арабский; кластер 7: Голландский, Турецкий, Чеченский, Казахский, Удмуртский; кластер 8 – Немецкий, Сербский; кластер 9 – Табасаранский, Чеченский, Японский; кластер 10 – Румынский, Белорусский; кластер11 – Навахо, Узбекский, Кечуа, Персидский; кластер 12 – Белорусский, Осетинский; кластер 13 – Тайский, Китайский, Индонезийский, Атикамекский, Тагальский, Тагальский, Дхолуо, Амхарский; кластер 14 – Немецкий, Испанский, Болгарский; кластер 15 – Эсперанто, Эсперанто; кластер 16 – Тувинский, Сингальский, Пунджабский; кластер 17 – Баскский, Вьетнамский, Индонезийский, Атикамекский; б – DBSCAN(n_cluster=16): кластер 1 – Эсперанто, Эсперанто; кластер 2 – Английский, Английский Русский, Русский, Норвежский, Немецкий, Немецкий, Сербский, Сербский, Венгерский, Венгерский, Финский, Финский, Французский, Украинский, Украинский, Румынский, Исландский, Исландский, Суахили, Суахили, Болгарский, Коптский, Коптский, Чешский, Чешский, Арабский, Арабский Белорусский, Осетинский, Осетинский: кластер 3 – Норвежский, Немецкий, Румынский, Испанский, Болгарский, Шведский, Польский, Польский, Французский, Белорусский; кластер 4 – Баскский; кластер 5 – Тайский, Китайский, Индонезийский, Атикамекский; кластер 6 – Вьетнамский, Индонезийский, Атикамекский; кластер 7 – Шведский, Голландский; кластер 8 – Дхолуо, Тагальский; кластер 9 – Дхолуо, Тагальский; кластер 10 – Бенгальский, Бенгальский, Оромо, Латинский, Табасаранский, Чеченский, Навахо, Малаялам, Персидский, Персидский, Панджаби, Кечуа, Кечуа, Кабильский, Татарский, Узбекский, Узбекский, Хинди, Хинди, Японский; кластер 11 – Голландский, Турецкий, Табасаранский, Чеченский, Казахский, Казахский, Удмуртский, Японский; кластер 12 – Оромо, Панджаби; кластер 13 – Латинский; кластер 14 – Тувинский, Сингальский, Сингальский; кластер 15 – Тувинский; кластер 16 – Амхарский

Рисунок 3.7 – Результаты кластеризации, полученные применением алгоритмов Wishart и DBSCAN

Примечания:

1. Данные, полученные из формул (6).
2. Результаты представлены рядом, с результатами алгоритма Wishart слева и результатами алгоритма DBCAN справа.
3. Подфигуры соответствуют двум признакам: $\hat{E}_i$ и $\hat{C}_i$

Кластеры языков. Чтобы изучить взаимосвязь между кластеризацией и различными языковыми характеристиками, мы применили значение V Крамера для измерения величины эффекта критерия независимости хи-квадрат. Наш анализ включал следующие четыре лингвистических характеристик: порядок слов, выравнивание, тип (локус) маркировки, морфологическая сложность.

Здесь также были вычислены значения корреляции между полученной кластеризацией и четырьмя языковыми характеристиками с помощью критерия Крамера V. Результаты представлены в Таблице 3.2, где указаны значения Крамера V для двух различных методов кластеризации: алгоритма Wishart и DBSCAN. Категориальным значениям каждой из языковых характеристик (word order, alignment, morphological complexity и locus of marking) были сопоставлены числовые. Например, для характеристики порядка слов использовались следующие метки: SVO – 1, SOV – 2, VSO – 3 и др. Такие же обозначения применялись ко всем перечисленным характеристикам. Критерий Крамера V позволяет измерить силу связи между двумя категориальными переменными. Значение V=0 указывает на отсутствие связи, а V=1 – на полную связь. В нашем анализе значения Крамера V, превышающие 0,7, свидетельствуют о наличии очень сильной связи между кластером и языковыми характеристиками [56].

Таблица 3.2. Значения V Крамера для ассоциаций между кластерами и языковыми характеристиками

Ассоциации между кластером и языковыми характеристиками			
Wishart		DBSCAN	
Комбинация признаков	V Крамера	Комбинация признаков	V Крамера
Кластер по порядку слов	0,791	Кластер по порядку слов	0,851
Кластер по выравниванию	0,534	Кластер по выравниванию	0,572
Кластер по морфологической сложности	0,470	Кластер по морфологической сложности	0,526
Кластер по типу маркировки	0,736	Кластер по типу маркировки	0,755

В таблице 3.2 представлено наиболее значительное значение Крамера Vдля метода Wishart было получено для характеристики порядка слов (0,791), что указывает на значительное влияние этой характеристики на формирование кластеров. Между тем, метод DBSCAN демонстрирует еще более высокое значение – 0,851, что подтверждает, что порядок слов является ключевым фактором в данной кластеризации. Для других языковых характеристик также наблюдаются интересные результаты. Например, для выравнивания значение Крамера V составило 0,534 для алгоритма Wishart и 0,572 для DBSCAN, что указывает на умеренную связь. В то же время, морфологическая сложность продемонстрировала более низкие значения V Крамера: 0,470 для Wishart и 0,526 для DBSCAN, что может свидетельствовать о менее значительном влиянии этой характеристики на формирование кластеров.

В таблице 3.3 представлены значения коэффициента Крамера V, которые отражают взаимосвязи между кластерами и комбинированными языковыми

характеристиками, полученными с использованием методов Wishart и DBSCAN. При наличии 4 различных признаков возможно образовать 7 уникальных комбинаций, исключая сравнение с одними и теми же признаками. Это стало основой для таблицы сопряженности, которая включает результаты для каждой комбинации. Некоторые сочетания языковых характеристик демонстрируют интересные закономерности. Например, взаимосвязь между порядком слов и типом маркировки продемонстрировала значение Крамера V равное 0,609 при использовании метода Wishart и 0,681 при применении DBSCAN, что указывает на умеренно сильную взаимосвязь. Также сочетание порядок слов и морфологическая сложность показало сопоставимые результаты: 0,629 для Wishart и 0,672 для DBSCAN, что подчеркивает значительное влияние этих характеристик на классификацию языков. Кроме того, связь между порядком слов и типом маркирования оказалась достаточно высокой: коэффициенты составили 0,616 для метода Wishart и 0,682 для DBSCAN. Эти данные подчеркивают важность учета указанных характеристик при анализе языковых кластеров.

Таблица 3.3 – Значения Крамера V для ассоциаций между кластерами и комбинациями языковых характеристик

Ассоциации между кластерами и комбинированными языковыми характеристиками			
Wishart		DBSCAN	
Комбинация признаков	V Крамера	Комбинация признаков	V Крамера
Кластер по порядку слов и выравниванию	0,609	Кластер по порядку слов и выравниванию	0,681
Кластер по порядку слов и морфологической сложности	0,629	Кластер по порядку слов и морфологической сложности	0,672
Кластер по порядку слов и типу маркировки	0,616	Кластер по порядку слов и типу маркировки	0,682
Кластер по выравниванию и морфологической сложности	0,478	Кластер по выравниванию и морфологической сложности	0,515
Кластер по выравниванию и типу маркировки	0,552	Кластер по выравниванию и типу маркировки	0,644
Кластер по морфологической сложности и типу маркировки	0,572	Кластер по морфологической сложности и типу маркировки	0,604

Интересно, что для комбинации выравнивания и морфологической сложности значение Крамера V было ниже: 0,478 для метода Wishart и 0,515 для DBSCAN, что указывает на менее выраженную зависимость. В целом, результаты таблицы подчеркивают значимость word order как ключевой характеристики, влияющей на формирование языковых кластеров, а также показывают, что сочетание различных языковых признаков может дать более полное представление о структурных особенностях языков и их классификации.

Данное исследование посвящено [56] сравнительному анализу хаотичности естественных языков через изучение семантических траекторий, полученных на основе данных из различных языковых семей. Анализ

семантических траекторий в 52 языках 18 различных языковых семей показал, что подавляющее большинство из них носят хаотического характер. Использование кластеризации (Wishart и DBSCAN) привело к обнаружению типологических кластеров, коррелирующих с распределением языков по параметрам энтропии и сложности, обосновывая применимость данных алгоритмов в анализе языковых характеристик.

В настоящей работе на материале значительного числа языков, принадлежащих разнообразным языковым семьям, рассматривается хаотические свойства семантических траекторий языка, т.е. предполагается, что источником хаоса является одновременно семантика и синтаксис[11].

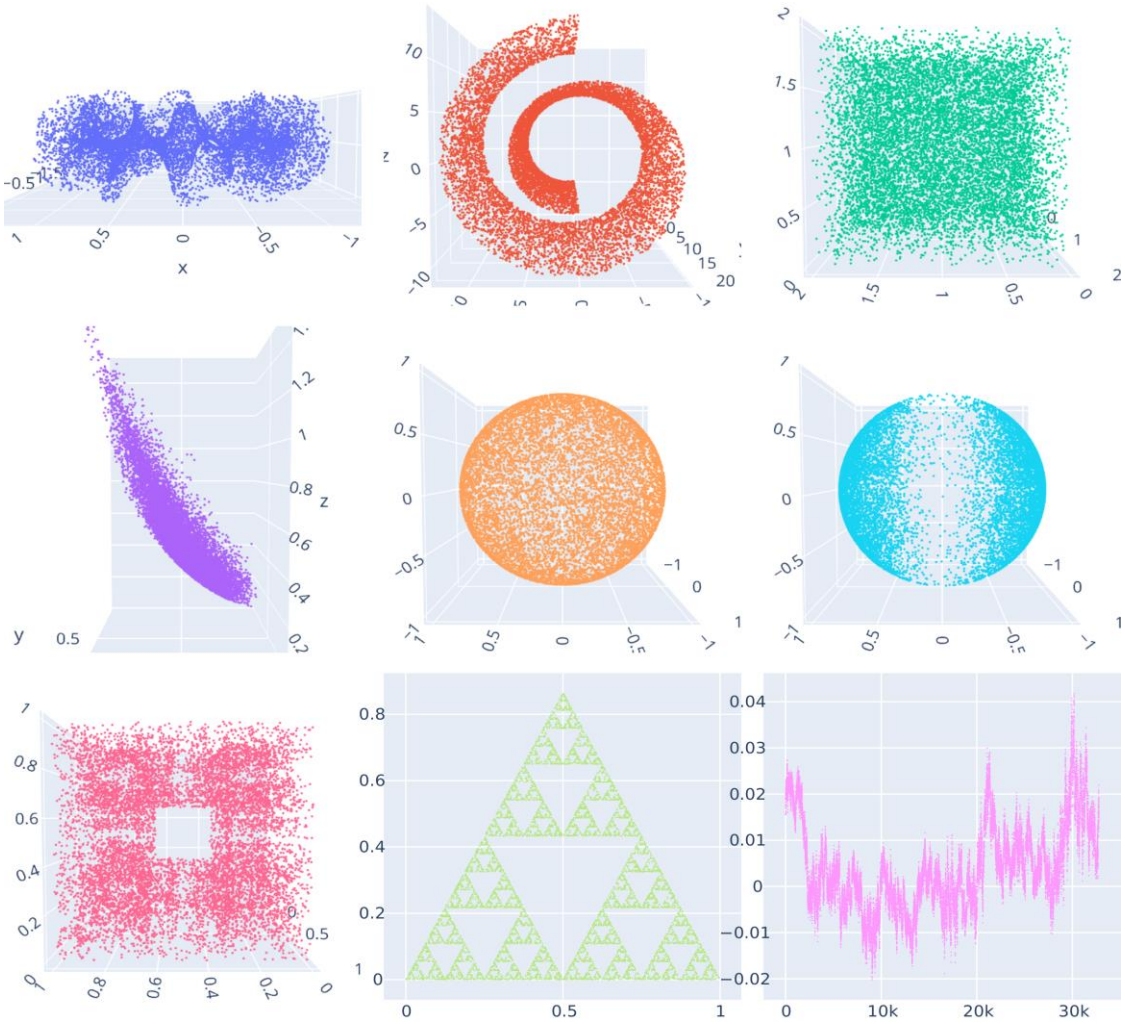
В ходе анализа было выявлено множество взаимосвязей между языковыми характеристиками, такими как порядок слов, тип выравнивания, морфологическая сложность и тип маркирования. Результаты анализа подтверждают теорию о хаотической природе семантических траекторий, основанных на литературных корпусах, и позволяют утверждать, что хаотичность играет важную роль в языковой динамике [56].

3.3 Внутренние размерности естественных языков

3.3.1 Внутренние размерности для стандартных многообразий и фрактальных множеств

На первом этапе выполнена верификация алгоритмов оценки внутренней размерности на геометрических объектах с априорно известными размерностями (многообразия, фрактальные множества; для списка объектов см. рисунок 3.8). Объекты были увеличены размерность в пространства различной размерности при объеме выборок 10^5 точек каждая. Согласно методу [66], синтез дополнительных координат осуществлялся применением полиномиальных/периодических операторов к исходным измерениям для контролируемого повышения размерности вложения.

[11]Разработка методологии для экспериментального разделения семантических и синтаксических детерминант хаоса представляет отдельное исследование.



а – многообразия (объекты с целой внутренней размерностью Хаусдорфа): лента Мёбиуса, швейцарский рулет, единичный куб (равномерное распределение точек); б – многообразия: параболоид, единичная сфера (равномерно), единичная сфера (бимодальное гауссово распределение на полюсах); Фракталы/Мультифракталы: губка Менгера, треугольник Серпинского, мультифрактальный сигнал (логнормальный каскад вейвлетов

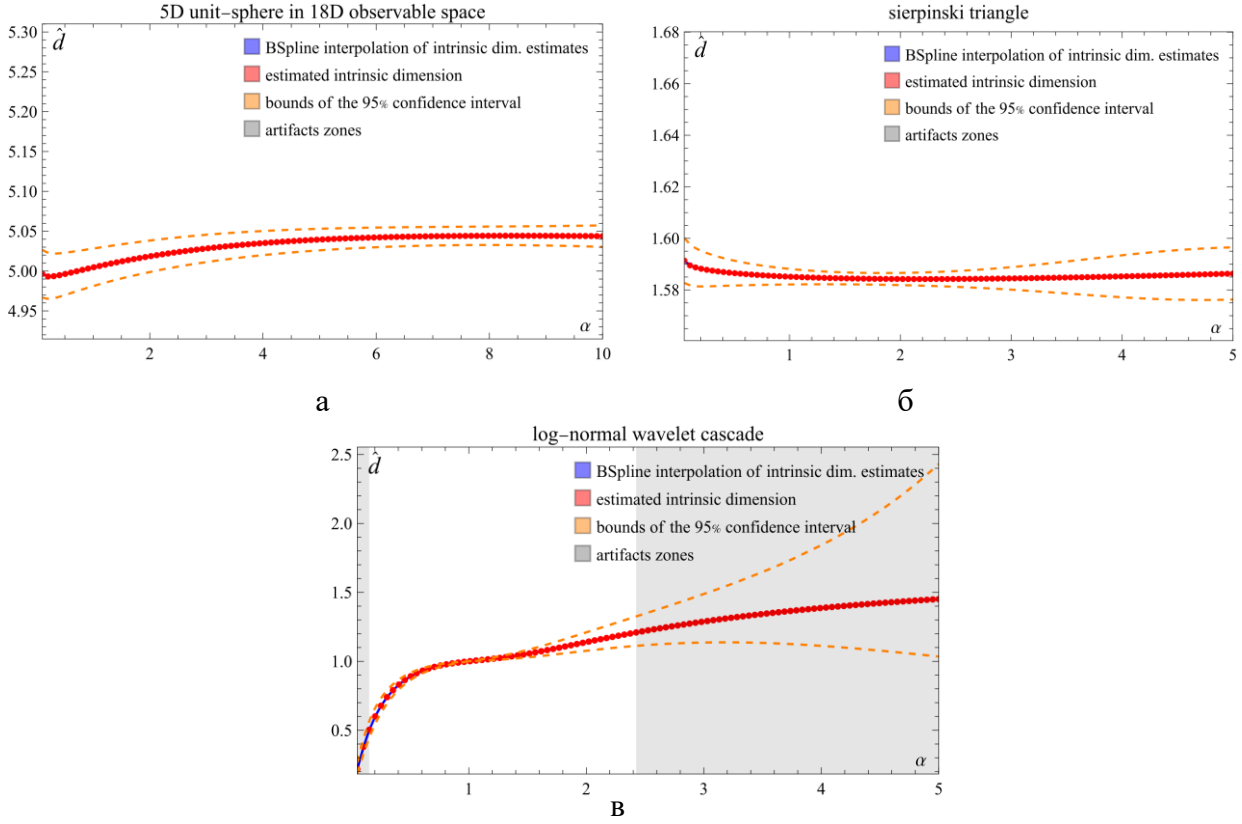
Рисунок 3.8 – Визуализация геометрических моделей для анализа внутренней размерности

Примечание – Составлено по источнику [8, p. e53227]

Алгоритма Швайнхарта. Данный метод предоставляет как точечные, так и интервальные оценки внутренней размерности. Ключевым аспектом применения является фильтрация результатов (адекватные и неадекватные оценки) на основе надежности: для анализа пригодны только те точечные оценки, чей доверительный интервал не выходит за рамки 10% от их величины (γ=10%). Остальные оценки, характеризующиеся чрезмерно широким интервалом неопределенности, признаются ненадежными и исключаются из рассмотрения. Эта дифференциация (условно обозначаемая на визуализациях как "белая" и "серая" области) позволяет выделить статистически значимые результаты. В представленных итогах используются исключительно оценки, прошедшие проверку на соответствие указанному критерию надежности. Важно

учитывать, что зависимость итоговой оценки размерности от параметра α требует анализа именно в контексте отобранных адекватных оценок.

Анализ влияния параметра α на оценку размерности. Первоочередной задачей является изучение вариативности получаемых результатов в зависимости[12] от величины α [2, p. 863360; 150].



для: а – единичной сферы (многообразие, $d_H = 5$, $d = 18$); б – треугольника Серпинского (фрактальное множество, $d_H = 1.58$, $d = 2$); в – каскада вейвлетов с логнормальным распределением[1] (мультифрактал)

Рисунок 3.9 – Зависимость оценок алгоритма Швайнхарта от параметра α

Примечания:
1. Составлено по источнику[39, p. 2-5]
2. Красные точки представляют собой оценки алгоритма, синяя линия – интерполяция B-spline зависимости оценок от α, оранжевая пунктирная линия обозначает верхние и нижние границы 95%-го доверительного интервала.
3. На рисунке 3.9 также визуализированы критерии информативности параметра α, такие как расхождение доверительных интервалов (серыми зонами)

На рисунок 3.9 отображены характерные графики функции $d_{\mathrm{Schw}}(\alpha)$. Каждый график включает как точечные оценки (сплошная синяя

[12]В качестве тестовых объектов были выбраны три принципиально различных типа множеств: стандартная единичная сфера (представитель гладких многообразий с целочисленной размерностью), классический треугольник Серпинского (пример фрактала с дробной размерностью) и реализация мультифрактала на основе логнормального вейвлет-каскада, характеризующегося нецелочисленной размерностной мерой.

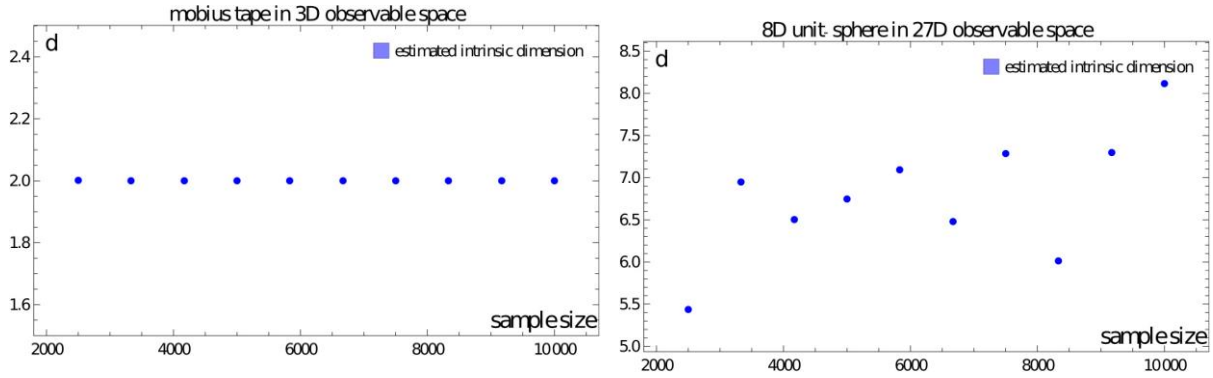
линия), так и границы 95% доверительных интервалов (пунктирные линии красного цвета). Тестирование алгоритма на полном наборе данных (рисунок 3.8) подтверждает его высокую точность: согласно (Приложение Г), средняя относительная погрешность определения внутренней размерности составляет менее 5% для всех рассмотренных объектов [2, p. 863360; 150, p. 68-75]. Моделирование на синтетических данных с известной размерностью демонстрирует универсальный принцип: для любого геометрического объекта (регулярного или фрактального) существует специфический диапазон значений α, обеспечивающий высокую точность оценки. За пределами этого диапазона (при экстремально малых или больших α) результаты становятся существенно недостоверными. Критически важным является следующее наблюдение: в оптимальном для оценки диапазоне α регрессионная зависимость характеризуется узким доверительным интервалом. Эта особенность позволяет сформулировать практические критерии выбора допустимых значений параметра α даже для объектов с неизвестной априори размерностью.

Обширные вычислительные эксперименты с синтетическими геометрическими структурами выявили устойчивые закономерности в поведении кривых. Для гладких многообразий и регулярных фракталов (рисунок 3.9а, 3.9б) характерно формирование горизонтального асимптотического плато, значение которого соответствует целочисленной или дробной размерности Хаусдорфа соответственно. Мультифрактальные множества, напротив, демонстрируют отчетливый дугообразный профиль зависимости. Критически важным общим наблюдением является неограниченное расширение доверительных интервалов оценки при экстремально больших значениях α (рисунок 3.9), что, согласно работе [68, p. 107291], указывает на нарушение условий применимости алгоритма. Анализ зависимости α для мультифракталов (рисунок 3.9в) выявляет специфические особенности: 1) присутствие выраженных экстремумов (максимума и минимума) или точки перегиба; 2) а также участок почти линейную зависимость. При этом установлено [67, p. 1-30; 68, p. 107291], что наблюдаемая линейность не отражает истинной связи размерности с параметром α, а является артефактом вычислительной методики. Данный интервал α сопровождается увеличением значение доверительного интервала.

Результаты второй тестовой серии, нацеленной на исследование устойчивости модели к шумовым воздействиям (Приложение Г), подтвердили ее робастность [2, p. 863360; 150, p. 68-75]. Помехи моделировались добавлением к координатам образцов случайной компоненты: либо равномерно распределенной $\sim U(0,1)$ либо изотропной гауссовой $\sim N(0,1 \cdot const_i)$. Алгоритм продемонстрировал стабильную работу в условиях обоих видов искусственно созданных помех (Приложение Г).

Алгоритм Брито. Результаты моделирования второго алгоритма показаны на рисунке 3.10 с помощью типичной диаграммы рассеяния. На оси абсцисс отображены порядки минимальных остовных деревьев, а на оси ординат – оценки внутренней топологической размерности. Результаты тестирования

алгоритма на синтетических геометрических объектах (Приложение Г) демонстрируют снижение точности по мере роста отношения внутренней топологической размерности к размерности пространства вложения. Хотя увеличение объема выборки (Приложение Г) способствует сходимости оценок и смягчает этот эффект, для конечных реальных выборок алгоритм применим преимущественно для верификации результатов, полученных первым методом [2, p. 863360; 150, p. 68-75].



а б

для: а – ленты Мёбиуса (многообразие, $d_T = 2$, $d = 3$); б – единичной сферы (многообразие, $d_T = 8$, $d = 27$)

Рисунок 3.10 – Зависимость оценок алгоритма Brito от параметра α

Примечание – По мере уменьшения соотношения между внутренней размерностью и размерностью пространства внедрения (embedding space) точность алгоритма также снижается, при прочих равных условиях

3.3.2 Внутренние размерности естественных языков

Для оценки внутренней размерности фрактальных структур языка были выбраны размерности вложения $d = \{5, 10, 15\}$. Это решение согласуется с работой [38, p. 113934], где показано, что при $n = 2$значения d свыше 15 не рекомендуются для пар "энтропия-сложность".

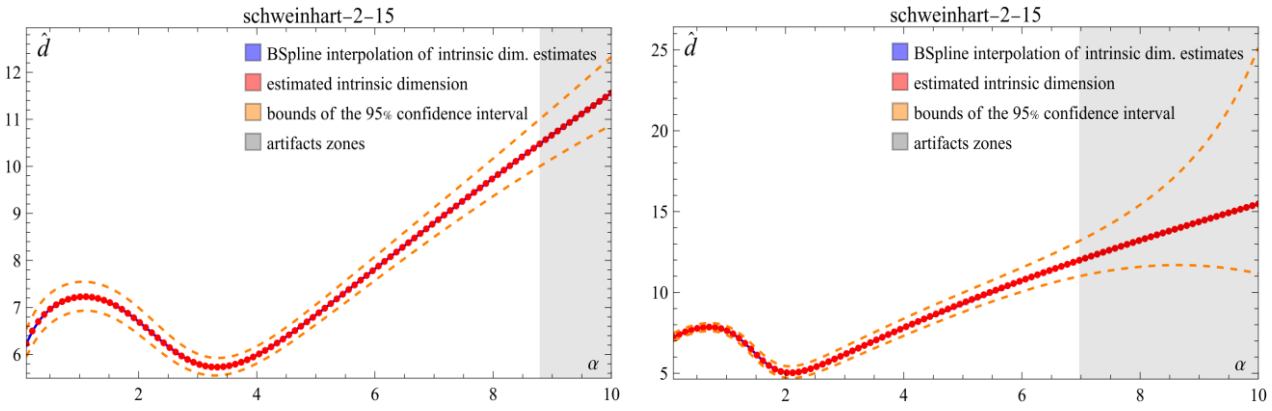
Для алгоритма Schweinhart, для данного языка

1) генерируем последовательность остовных деревьев, при n в диапазоне от 1e+5 до размера набора данных;

2) для каждого фиксированного n, при изменении α непрерывно от 10^{-4} до 10 с шагом $s \approx 0.1$ выполнялась оценка параметров регрессионной модели (4.2.4) с целью вычисления размерности $\hat{d}_{Schw}$;

3) исключаются наблюдения, не удовлетворяющие следующим критериям качества: 1) значение доверительного интервала для линии регрессии превышает порог γ; 2) доверительный интервал для нормированного параметра $(\hat{d} - \alpha)/\hat{d}$ велики допустимого значения γ;

4) выбираем минимальные и максимальные оценки $\hat{d}_{Schw}$ для всех допустимых α.

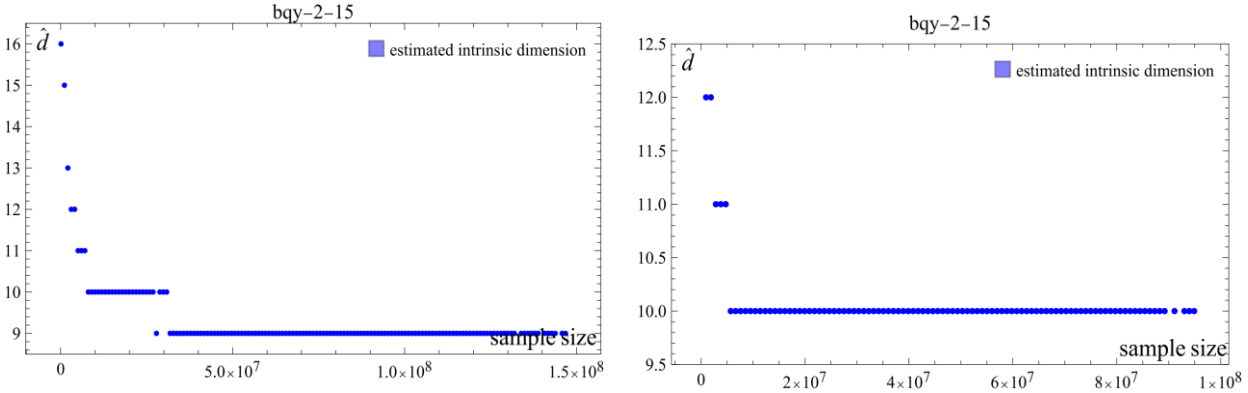


а б

для: а – SVD-векторные представления русской национальной литературы ($n = 2, d = 15$); б – SVD-векторные представления английской национальной литературы ($n = 2, d = 15$)

Рисунок 3.11 – Зависимость оценок алгоритма Schweinhart от параметра α

Примечания:
1. Красные точки представляют собой оценки алгоритма, синяя линия – интерполяция B-spline зависимости оценок от α, оранжевая пунктирная линия обозначает верхние и нижние границы 95%-го доверительного интервала.
2. На рисунке 3.11 также визуализированы критерии информативности параметра α, такие как расхождение доверительных интервалов (серыми зонами).
3. Поведение графиков для языков идентично поведению для многофрактальных структур



а б

для: а – SVD-векторные представления русской национальной литературы ($n = 2, d = 15$); б – SVD-векторные представления английской национальной литературы ($n = 2, d = 15$)

Рисунок 3.12 – Зависимость оценок алгоритма Brito от параметра α

В соответствии с рисунками 3.11, 3.12, для алгоритма Brito, для данного языка: 1) для всех $d = i \in 2..15$ генерировались выборки $\{x_1^j, \dots, x_{1e+6}^j\}_{j=1..100}, x \sim U(0_i, 1_i)$, где $0_i, 1_i$ и являются векторами нуля и единицы соответственно в i -мерном пространстве; 2) оцениваем $\hat{\mu}_i$ и $\hat{\sigma}_i^2$; 3) вычисляем оценки $\hat{d}_{BQY}(\{x_1, \dots, x_\ell\})$ для различных ℓ.

Таблица 3.4 – Оценка внутренней размерности для SVD представлений языков.

Язык	n	d	$\widehat{d}_{BQY}$	$\min \widehat{d}_{Schw}$	$\max \widehat{d}_{Schw}$	α
Русский	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
Английский	1	5	5	4.71	5.16	0-2.82
		10	8	6.63	7.89	0-4.02
		15	14-15	9.61	12.39	0-4.97
	2	5	5-6	5.23	7.82	0-5.62
		10	8-9	6.72	8.59	0-10
		15	14	8.61	9.85	0-10

Таблица 3.5 – Оценка внутренней размерности для CBOW представлений языков

Язык	n	d	$\widehat{d}_{BQY}$	$\min \widehat{d}_{Schw}$	$\max \widehat{d}_{Schw}$	α
Русский	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
Английский	1	5	4	4.49	4.92	0-0.6
		10	6-7	5.55	7.14	0-0.71
		15	8	6.01	9.26	0-0.71
	2	5	5	3.24	4.51	0-2.63
		10	7	3.86	5.80	0-1.61
		15	10	5.03	11.94	0-7

Результаты обобщены в таблицах 3.4 и 3.5, результаты по остальным языкам представлены в (Приложении Г). Репрезентативные зависимости для исследуемых оценок визуализированы на рисунках 3.11, 3.12 Аналогичные графики для методов SVD и CBOW (русский и английский языки) доступны в https://github.com/erbolova1983/Investigation-of-NL-structures.git соответственно для русского и английского языков, а для остальных языков в https://github.com/erbolova1983/Investigation-of-NL-structures.git); они проявляют типичную мультифрактальную зависимость (сравните с рисунком 3.9). Оценка внутренней размерности по алгоритму Швайнхарта даёт нецелые результаты, характерные для мультифракталов . Оцененные внутренние размерности для обоих языков выглядят достаточно устойчивыми: примерно одинаковые результаты получаются для различных размерностей встраиваемых пространств, методов построения встраиваний (SVD, CBOW) и методов оценки внутренней размерности. Полученные результаты подтверждают [2, p. 863360; 150, p. 68-75], что внутренняя размерность представляет собой инвариантную характеристику языка, геометрически соответствующую мультифрактальной структуре.

Принадлежность фазового пространства естественного языка классу мультифракталов даёт основания для следующих гипотез:

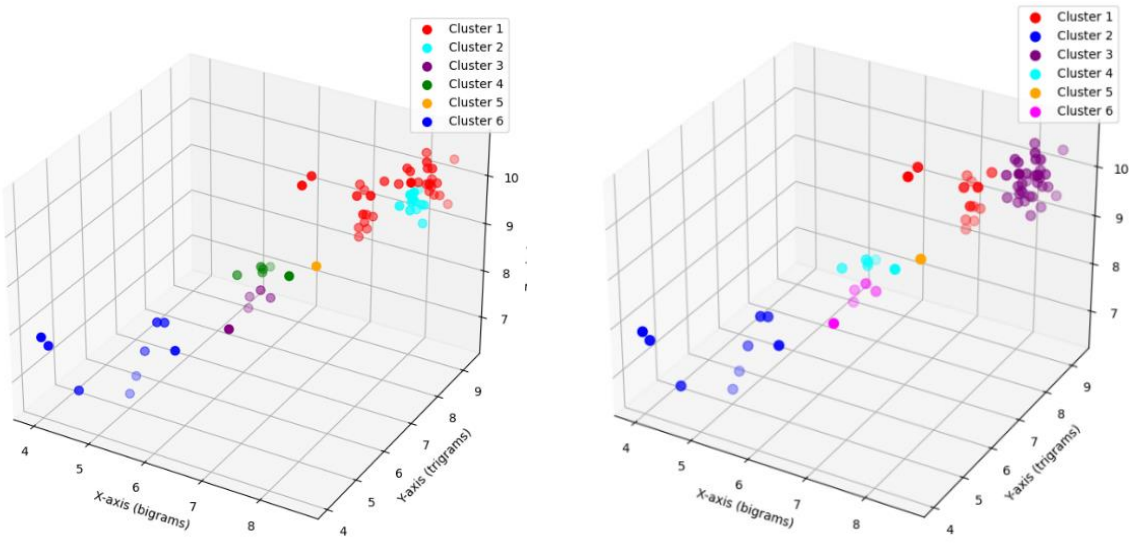
– язык функционирует как странный аттрактор в динамических системах [1, p. 1-6; 64, с. 125-126];

– тексты на естественном языке обладают свойствами хаотических рядов [1, p. 1-9; 41, p. 113934].

Экспериментально установленные верхние границы размерности (с контролем по Брито): 9 для русского языка и 10 для английского [151].

3.3.3 Кластеризация языковых представлений

Для анализа внутренних структур естественных языков выбрали данные в (Приложении Г), основанные на оценке внутренней размерности SVD-представлений. В данной работе были выбраны следующие параметры: $\mathrm{n} = 1,2,3$ (для униграмм, биграмм и триграмм соответственно) и пространства вложения $d = 15,$ с максимальными оценками $\max \hat{d}_{Schw}$ для допустимых значений α. Данные параметры обеспечили основу для генерации координат точек в трехмерном пространстве, где ось x представляет значения униграмм, ось y – значения биграмм, а ось z – значения триграмм. В процессе анализа используются два эффективных алгоритма кластеризации: Wishart [138, p. 97] и K-Means [147, p. 224-226]. На рисунке 3.13 представлены результаты методов кластеризации, такие как Wishart и K-Means.



а б

а – Wishart (n_cluster=6): кластер 1 – Амхарикский, Арабский, Армянский, Ассирийский, Бамбарский, Белорусский, Болгарский, Английский, Эрзянский, Французский, Финский, Хинди, Венгерский, Исландский, Индонезийский, Итальянский, Японский, Коми-зырянский, Латинский, Латышский, Норвежский, Осетинский, Персидский, Польский, Пенджабский, Румынский, Русский, Сербский, Словацкий, Шведский, Тагальский, Удмуртский, Украинский, Идиш; кластер 2 – Бартанги, Бенгальский, Чешский, Датский, Голландский, Немецкий, Литовский, Португальский, Сингальский, Словенский, Тайский, Вьетнамский; кластер 3 – Казахский, Кыргызский, Татарский, Турецкий, Тувинский, Узбекский; кластер 4 – Абхазский, Баскский, Чеченский, Суахили, Табасаранский; кластер 5 – Эсперанто; кластер 6 – Атикамекский, Коптский, Дхолуо, Кабильский, Корейский, Малаялам, Навахо, Кечуа, Тибетский; б – K-Means (n_cluster=6): кластер 1 – Армянский, Бамбарский, Белорусский, Французский, Хинди, Индонезийский, Японский, Латышский, Норвежский, Тагальский, Украинский, Идиш; кластер 2 – Атикамекский, Коптский, Дхолуо, Кабильский, Корейский, Малаялам, Навахо, Кечуа, Тибетский; кластер 3 – Амхарский, Арабский, Ассирийский, Бартанги, Бенгальский, Болгарский, Чешский, Датский, Голландский, Английский, Эрзянский, Финский, Немецкий, Венгерский, Итальянский, Исландский, Коми-Зирийский, Латинский, Литовский, Осетинский, Персидский, Польский, Португальский, Пенджабский, Румынский, Русский, Сербский, Сингальский, Словацкий, Словенский, Шведский, Тайский, Вьетнамский, Удмуртский; кластер 4 – Абхазский, Баскский, Чеченский, Суахили, Табасаранский; кластер 5: Эсперанто; кластер 6: Казахский, Кыргызский, Татарский, Турецкий, Тувинский, Узбекский

Рисунок 3.13 – Результаты кластеризации, полученные применением алгоритмов Wishart и K-Means

Примечания:
1. Данные, полученные из максимальной оценки max $\hat{d}_{schw}$ языков.
2. Результаты представлены рядом, с результатами алгоритма Wishart слева и результатами алгоритма K-Means справа)

Таким образом, результаты кластеризации показывают не только внутренние структуры языков, но и их взаимосвязи. Наблюдается интересные тенденции в расположении языков в кластерах. Например, языки, относящиеся к одной семье, чаще всего группируются вместе, что подтверждает существующие лингвистические гипотезы о родстве. Однако это также подводит нас к

следующему этапу анализа: изучению коэффициента соседства языков как по их территории, так и по языковым семьям и типу грамматических отношений. Указанную кластеризацию можно сравнить со стандартными лингвистическими гипотезами:

По ареальной гипотезе. Ареальная группировка языков выявляет особые объединения – языковые союзы. Они представляют собой совокупности родственных или неродственных идиомов, сформировавшихся в пределах общего географического пространства под влиянием длительных контактов их носителей. Классическими примерами таких конвергентных зон являются Балканский, Поволжский (Волжско-Камский), Мезоамериканский и Гималайский (Центрально-Азиатский) союзы. Примечательно, что современная лингвистическая картина мира и её исторические истоки (синхрония и диахрония) демонстрируют в ареальном разрезе значительные расхождения. Это обусловлено миграционными процессами, в ходе которых языки, сложившиеся в едином очаге, распространились на обширные территории, удалённые от первоначальной зоны формирования.

Полученные нами данные в целом не противоречат ареальному принципу группировки языков. В частности, белорусский язык (кластер 1), имеющий пять соседей (Россия, Украина, Польша, Литва и Латвия), является примером языкового кластера, где наблюдаются прочные связи с языками государств-соседей. Финно-угорский и балканский языковые союзы также продемонстрировали завидную последовательность, все языки которого сосредоточились в первом кластере вместе со славянскими языками. Сходные результаты показали и языки тюркского союза, находящиеся в кластере 3. Однако, языки кавказского союза (армянский, осетинский, абхазский, чеченский) оказались в разных, причем далеко не соседних кластерах. Первые два принадлежат первому кластеру, в то время как последние – к четвертому, соседствуя с баскским и суахили, оба из которых относятся к разным языковым союзам: изолированным и африканским соответственно. Полученные данные показывают, что средний коэффициент соседства по регионам языков составляет 0.62 (метод Wishart) и 0.61 (метод K-Means), что указывает на высокую степень территориальной близости и взаимосвязи между языками в этом кластере.

По генетической (генеалогической) гипотезе. Согласно этому принципу, языки классифицируются в соответствии с принадлежностью к языковой семье (группе родственных языков), которые имеют общие языковые особенности (произношение, словарный запас, грамматика) и произошли от общего предка (протоязыка). Нами также был проведен анализ на основе родственных связей в рамках языкового семейства.

Судя по полученным данным [2, p. 863360; 150, p. 68-75], алтайские языки продемонстрировали наивысшую степень согласованности, будучи сгруппированы в кластеры 3 (Wishart) и 6 (K-Means) соответственно. То же самое можно сказать о семитских, австронезийских языках и языках Юго-Восточной Азии, входящих в кластер 1. Наименее последовательной языковой семьей в нашей выборке является индоевропейская, которая распространилась в кластере

1 и 2 (Wishart), также кластер 1 и 3 (K-Means). Это вполне может быть объяснено ностратической гипотезой в макрокомпаративной лингвистике, предполагающей, что несколько языковых семей, включая индоевропейские, уральские и алтайские, имеют общий прародительский протоностратический язык. Целью было выявить более глубокие взаимосвязи между, казалось бы, не связанными языками и понять происхождение самого человеческого языка. С этой целью сторонники теории приводят фонетические сходства, общие грамматические структуры и словарный запас индоевропейских, уральских и алтайских языков в качестве доказательств, подтверждающих их утверждения. Выявляя систематические соответствия в звучании и значении между этими языками, они утверждают, что эти особенности указывают на общий язык предков. Кроме того, они изучают древние тексты и надписи, чтобы выявить лингвистические закономерности, которые могут указывать на исторические связи. Стоит отметить, что баскский язык, как языковой изолят, и кавказские языки (хотя это понятие не подразумевает их генетического родства) находятся в одном кластере. Это наводит на мысль о том, что более ранняя, но неподтвержденная гипотеза о родстве баскского языка со всеми тремя группами кавказских языков все еще может быть верна.

В результате, согласно полученным данным, средний коэффициент соседства по семействам языков составляет 0.74 (Wishart) и 0.76 (K-Means), что также свидетельствует о значительной степени родства и взаимосвязи между языками в данном кластере.

По типу грамматических отношений Wishart-0,87 и K-Means-0,87. Типологическая классификация языков, основанная на морфологическом принципе, позволяет лучше понять, как устроены грамматические системы различных языков и как они выражают грамматические значения. Внутренняя размерность языков, связанная с типами грамматических отношений, проявляется также в различных способах выражения грамматических значений, что позволяет выделить несколько основных типов языков. Языки традиционно подразделяются на аналитические, синтетические, изолирующие, агглютинативные, флективные и инкорпорирующие, в зависимости от того, как в них выражаются грамматические отношения, такие как падеж, число, время и т.д. Поэтому целесообразно кратко охарактеризовать каждый из них:

– в аналитических языках грамматические отношения выражаются в основном с помощью порядка слов и служебных слов (предлогов, частиц и т.д.), а каждое слово обычно имеет одну четкую грамматическую функцию (например, английский, китайский);

– грамматические отношения в синтетических языках выражаются внутри слова, с помощью аффиксов и флексий, а одно слово может выражать несколько грамматических значений одновременно (например, русский, латынь);

– в изолирующих языках отсутствуют морфологические показатели грамматических отношений, а слова не изменяются. При этом грамматические отношения выражаются в основном порядком слов и служебными словами (например, вьетнамский, тайский, бирманский);

– грамматические значения в агглютинативных языках выражаются с помощью присоединения к корню слова последовательных аффиксов, каждый из которых выражает одно определенное грамматическое значение, например в турецком и японском;

– флективные языки выражают грамматические значения с помощью флексий, которые часто передают несколько грамматических значений одновременно;

– в инкорпорирующих языках слова могут состоять из нескольких морфем, которые объединяются в одно слово, выражающее целое предложение (языки коренных народов Северной Америки).

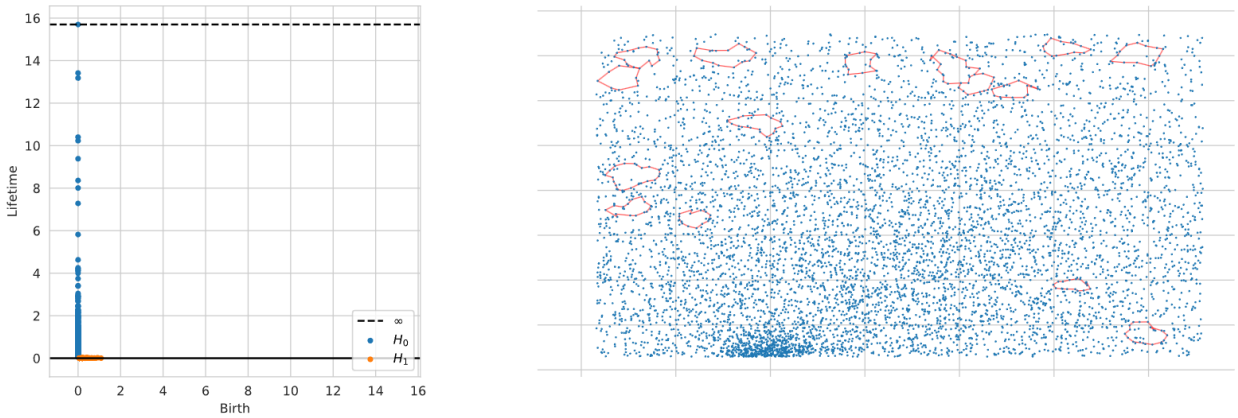
Из этого следует, что тип языка влияет на его внутреннюю структуру [2, p. 863360; 150, p. 68-75]. Естественные языки имеют разную степень морфологической сложности, при этом синтетические языки (агглютинативные и флективные) обычно более морфологически сложны, чем аналитические. Это связано с тем, что в них больше грамматических показателей, выраженных внутри слова, а не в отдельности.

3.4 Топологическая структура естественного языка

Для генерации векторных представлений n-грамм (1-3 порядка) в русском и английском языках были сконструированы корпуса на основе текстов национальной литературы из открытых источников. Для русского языка ($|\Im_1| =$ 6429 текстов) идентифицировано 103 952 уникальных униграмм, 14 775 439. Английский корпус ($|\Im_2|$) = 11 052 текста) содержит 94 087 униграмм, 9 490 603 биграммы. Для обеспечения вычислительной осуществимости задачи формируется случайная подвыборка данных: 10% от общего корпуса слов и 1% от всех встречающихся биграмм (или триграмм). С целью достижения структурной сопоставимости между текстами бота и человека, языковым моделям задаётся генерация фрагментов длиной ровно тысяча слов. Генерация начинается с того же слова, что и соответствующий человеческий текст, и завершается идентичным конечным словом. Перед анализом все текстовые данные (как человеческие, так и текст, написанные ботом) проходят стандартизированную предобработку: токенизацию, лемматизацию и подмену вспомогательных частей речи специальными токенами. Инструментарий: natasha для русского языка, spaCy для английского.

3.4.1 Выделение гомологии на участке векторного пространства

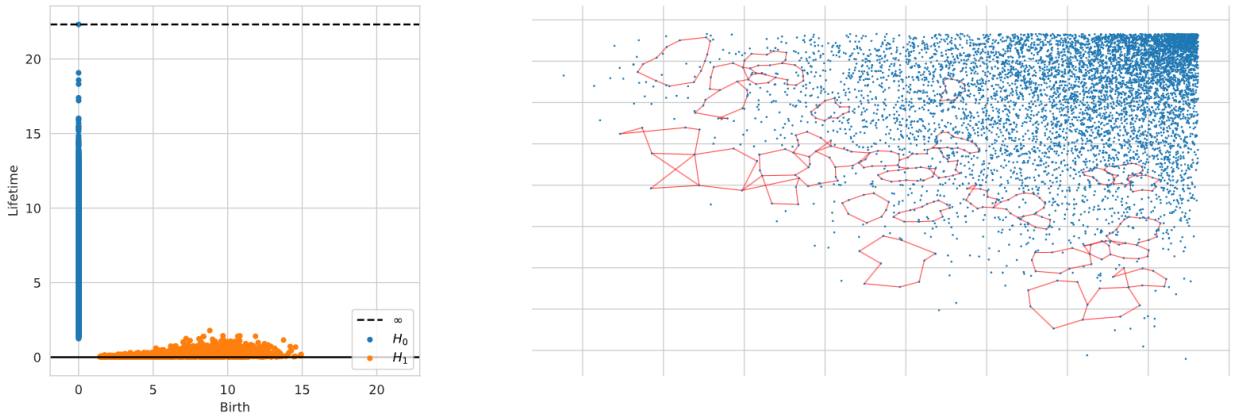
Для исследования структуры семантического пространства был выполнен предварительный анализ: указанный алгоритм применен к репрезентативной подвыборке данных $d = 100$. С целью визуализации многомерные данные были спроецированы на двумерную плоскость посредством метода главных компонент (PCA). Результаты масштабного вычислительного эксперимента свидетельствуют о более высокой плотности распределения объектов в центральной области облака по сравнению с областями на "окраине". Данная особенность проявляется в уменьшении диаметра топологических "дырок" в центре и их более раннем исчезновении на диаграммах персистентности (рисунки 3.14, 3.15, 3.16).



а б

а – диаграммы персистентности гомологий нулевого и первого порядка; б – границы наиболее персистентных диаграмм

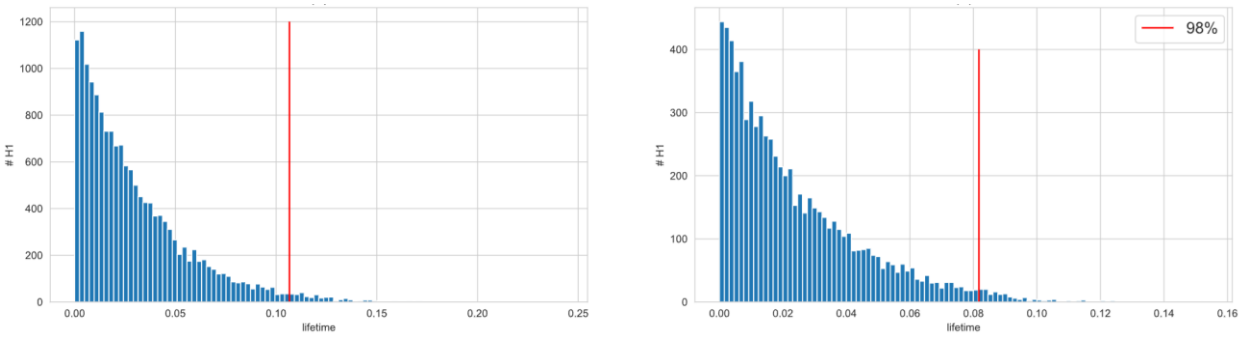
Рисунок 3.14 – Гомологии в центральной области языка (двумерная проекция PCA)



а б

а – диаграммы персистентности гомологий нулевого и первого порядка; б – границы самых персистентных диаграмм

Рисунок 3.15 – Гомологии на крае языка (двумерная проекция PCA)



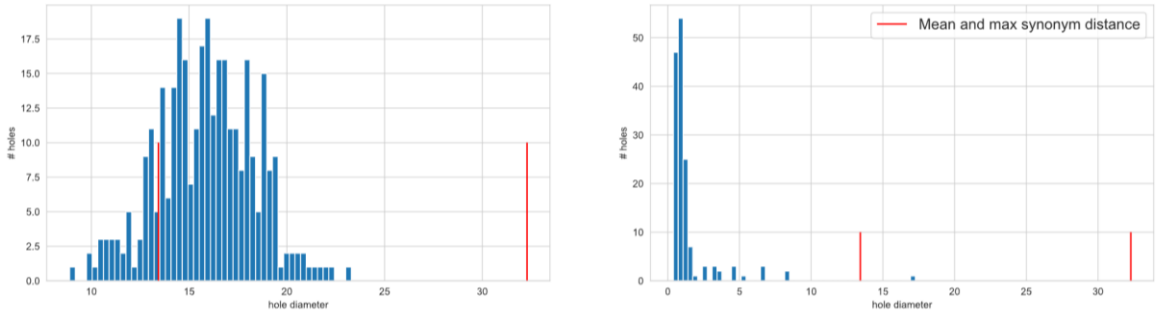
а б

а – на крае; б – в центральной области

Рисунок 3.16 – Распределение персистентности гомологии первого порядка

Примечание – Красными вертикалями отмечены 96% квантили

Рисунок 3.17 показывает соответствующие распределения (в центральной области продолжительность жизни составляет 0.15; тогда как на "окраине"пространства гомологии живут до – 0.24). Наконец, рисунок 3.18 иллюстрирует соотношение между постоянством гомологий и диаметрами дырок. Очевидно, значения для дырок в центральной области пространства слов меньше. В дальнейшем мы изучаем дыры в центральной части и на окраинах отдельно и используем для них разные пороговые значения.

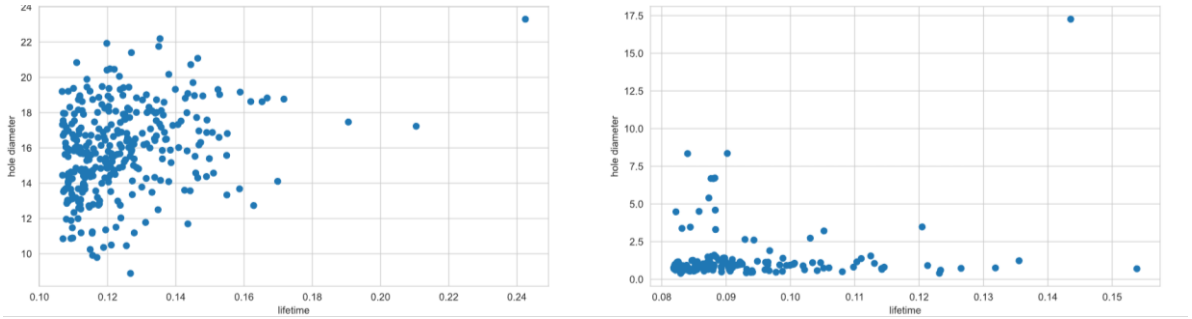


а б

а – на крае; б – в центральной области

Рисунок 3.17 – Распределение диаметров гомологий первого порядка

Примечание – Красными вертикалями отмечены среднее и максимальное значения диаметров синонимичных групп



а б

а – на крае; б – в центральной области

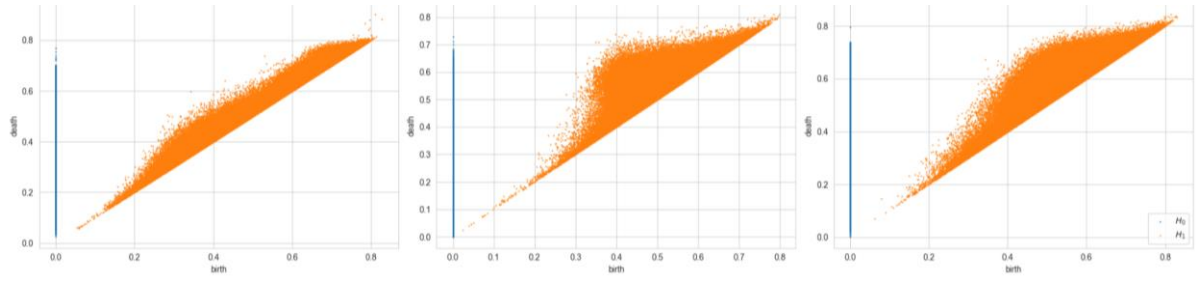
Рисунок 3.18 – Распределение диаметров первого порядка гомологии

Примечание – Красными вертикалями отмечены среднее и максимальное значения диаметров синонимичных групп

3.4.2 Выделение гомологий первого порядка

Для определения топологических структур (нулевого и первого порядков гомологий) в векторных представлениях слов, биграмм и триграмм на материале русского, английского, казахского и киргизского языков использовался ранее описанный алгоритм [152]. Результирующие диаграммы устойчивости (персистентности) визуализированы на рисунках 3.19, 3.20, 3.21, 3.22. Методология включала последовательные этапы: пространство признаков подвергалось кластеризации, затем для каждого полученного кластера

независимо рассчитывались гомологии, и на заключительном этапе осуществлялся сводный анализ собранных данных. Количественные характеристики гомологий первого порядка систематизированы по языкам в таблице 3.6.

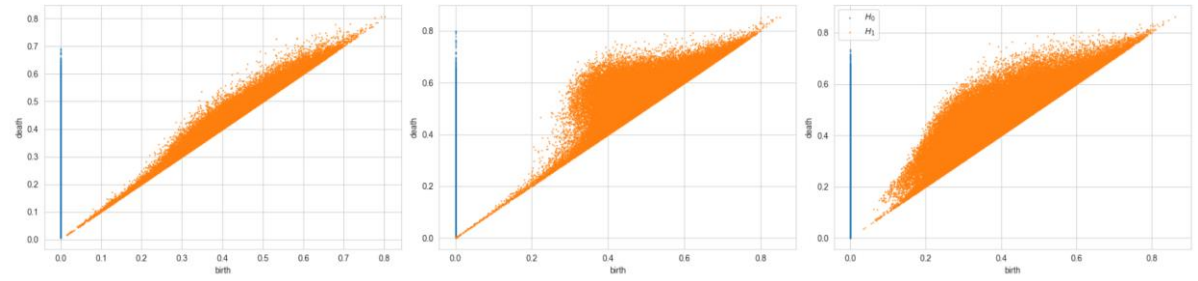


а б в

а – слов; б – биграмм; в – триграмм русского языка

Рисунок 3.19 – Диаграммы устойчивости в пространстве векторов для

Примечание – Гомологии нулевого порядка обозначены синим цветом, первого порядка – оранжевым

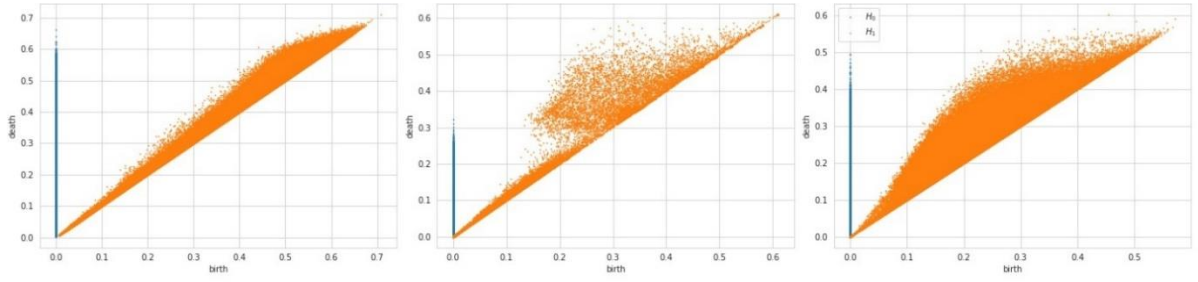


а б в

а – слов; б – биграмм; в – триграмм английского языка

Рисунок 3.20 – Диаграммы устойчивости в пространстве векторов для

Примечание – Гомологии нулевого порядка обозначены синим цветом, первого порядка – оранжевым

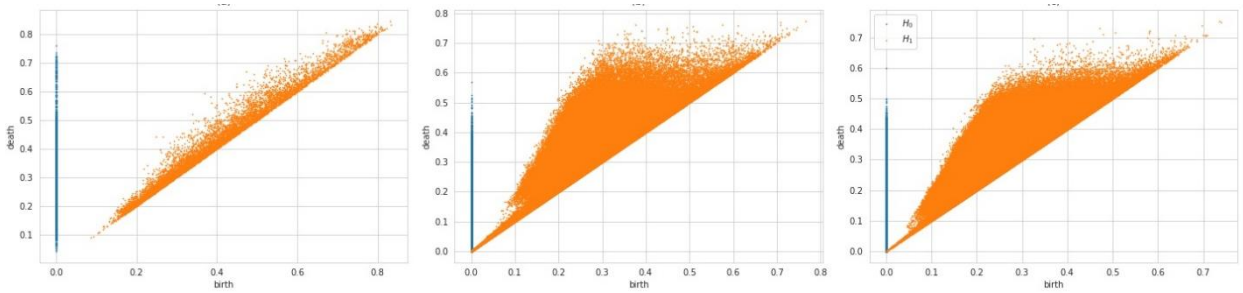


а б в

а – слов; б – биграмм; в – триграмм казахского языка

Рисунок 3.21 – Диаграммы устойчивости в пространстве векторов для

Примечание – Гомологии нулевого порядка обозначены синим цветом, первого порядка – оранжевым



а б в

а – слов; б – биграмм; в – триграмм киргизского языка

Рисунок 3.22 – Диаграммы устойчивости в пространстве векторов для

Примечание – Гомологии нулевого порядка обозначены синим цветом, первого порядка – оранжевым

Таблица 3.6 – Число гомологий первого порядка в векторных пространствах

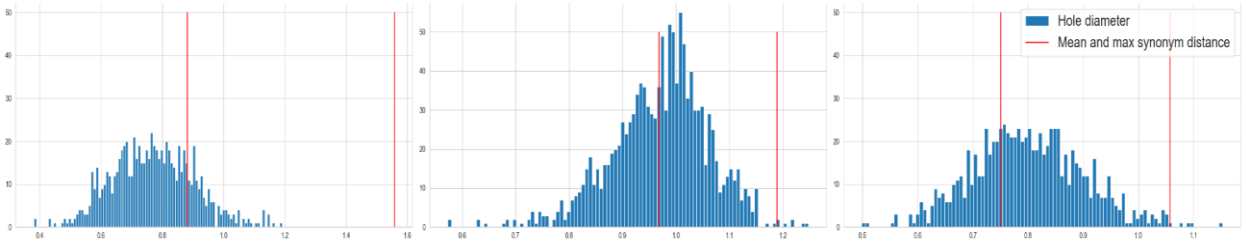
Языки	Слова	Биграммы	Триграммы
Русский	153813	225488	227724
Английский	77477	127436	110472
Казахский	51676	125705	136889
Киргизский	47805	108277	184974

В соответствии с ожиданиями, преобладающая доля гомологических циклов характеризуется минимальной персистентностью и ранним исчезновением. Их проекции локализованы у границы $birth = death$ на соответствующих диаграммах устойчивости (см. визуализации для русского, англ., каз., кирг. языков: рисунки 3.19, 3.22). Объектом дальнейшего исследования выступают наиболее устойчивые гомологии, репрезентируемые точками, максимально удаленными от диагональной линии.

Диаграммы персистентности обнаруживают общие черты для всех четырех языков (русский, английский, казахский, киргизский) на каждом уровне представления данных (слова, биграммы, триграммы): в пространствах слов гомологии первого порядка демонстрируют пониженную устойчивость (сужение оранжевой полосы, концентрация точек у диагонали $birth = death$, свидетельствующая о быстром затухании), при анализе биграмм и триграмм установлено, что существенная часть топологических признаков характеризуется высокой устойчивостью, что визуализируется значительным удалением соответствующих точек от диагонали [153].

Отбор наиболее персистентных гомологий

Идентификация значимых гомологий первого порядка осуществлялась путем фильтрации по 0.99 квантилю персистентности. Каждой отобранной топологической особенности (топологической пустоте) сопоставлен характеристический диаметр – максимум расстояний между узлами соответствующего симплициального комплекса. На рисунках 3.23, 3.24, 3.25, 3.26 отражены распределения данных диаметров для всех анализируемых языков.

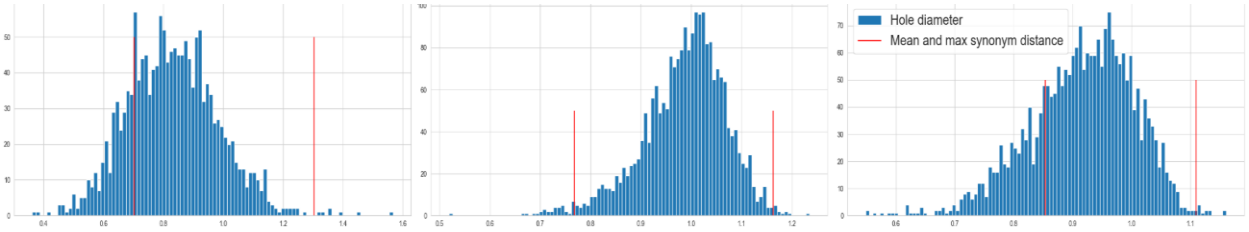


а б в

а – уровень слов; б – уровень биграмм: в – уровень триграмм

Рисунок 3.23 – Распределение характеристических диаметров устойчивых гомологий H1 для английского языка

Примечание – Пунктирные линии красного цвета показывают среднее и максимальное значения диаметров внутри кластеров синонимичных элементов

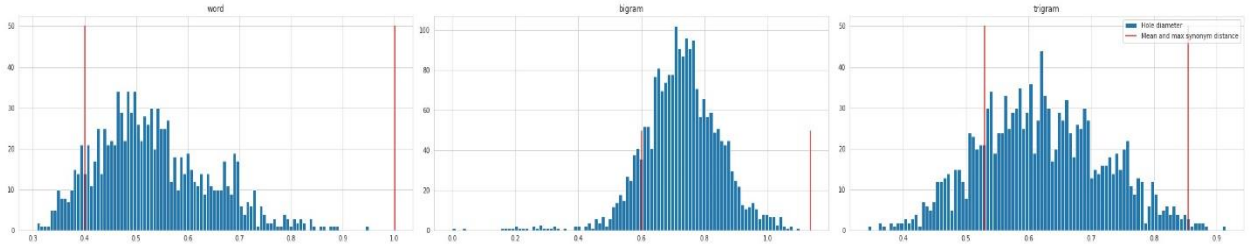


а б в

а – уровень слов; б – уровень биграмм: в – уровень триграмм

Рисунок 3.24 – Распределение характеристических диаметров устойчивых гомологий H1 для русского языка

Примечание – Пунктирные линии красного цвета показывают среднее и максимальное значения диаметров внутри кластеров синонимичных элементов

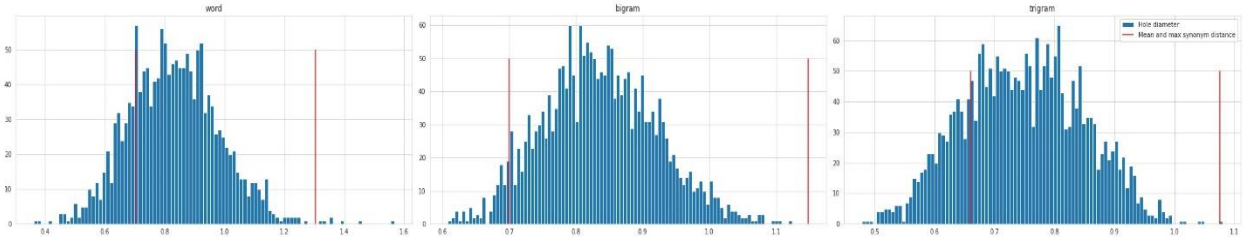


а б в

а – уровень слов; б – уровень биграмм: в – уровень триграмм

Рисунок 3.25 – Распределение характеристических диаметров устойчивых гомологий H1 для казахского языка

Примечание – Пунктирные линии красного цвета показывают среднее и максимальное значения диаметров внутри кластеров синонимичных элементов



а б в

а – уровень слов; б – уровень биграмм: в – уровень триграмм

Рисунок 3.26 – Распределение характеристических диаметров устойчивых гомологий H1 для киргизского языка

Примечание – Пунктирные линии красного цвета показывают среднее и максимальное значения диаметров внутри кластеров синонимичных элементов

Топологические особенности минимального диаметра типично возникают вокруг синонимических кластеров вследствие геометрической близости их элементов в векторном пространстве. Интерпретация таких "дыр" как "слепых зон" языка проблематична, поскольку закономерно воспроизводят результаты семантической кластеризации. Данное соображение мотивировало разработку процедуры выборке гомологий, основанной на сопоставлении их метрических характеристик (диаметров) с аналогичными параметрами для групп синонимов.

Критерии синонимии устанавливались по следующим лексикографическим источникам: русский («Словарь синонимов русского языка», ред. Л.Г. Бабенко, 2011), английский (модель WordNet из библиотеки NLTK), казахский («Словарь синонимов», ред. С.А. Байзакова, 2007) и киргизский (Жапаров Ш., Сейдакматов К., Сыдыкова Т., «Словарь синонимов кыргызского языка», Бишкек, 2015). Из словарей извлекаются группы синонимов (напр., «ИЗГИ́Б, изви́в, изви́лина, изворо́т, излом, излу́чина»), для которых определяются диаметры. Ключевой этап – установление нового порога отбора гомологий: им становится наибольший из полученных диаметров. Сохраняются только те "дыры", которые превышают размер любой синонимической группы. При обработке n-грамм синонимическая группа формируется как исчерпывающее множество комбинаций синонимов входящих в n-грамму слов. Например, для биграммов «зыбкий берег» группа объединяет варианты вроде «зыбучий берег», «зыбучее взморье», порожденные синонимами каждого компонента («ЗЫ́БКИЙ, зыбу́чий...» и «БЕ́РЕГ, взмо́рье...»). Для каждой такой группы n-грамм, как и в случае слов, вычисляется диаметр по заданной метрике (раздел 2.4). Максимальный диаметр среди всех групп синонимичных n-грамм впоследствии используется как порог для отсечения гомологий в соответствующем n-граммном векторном пространстве.

Результаты фильтрации представлены на рисунках 3.23, 3.24, 3.25, 3.26, где крайние правые красные линии обозначают пороговые значения. Количественное распределение релевантных топологических структур по языкам и уровням представлены в таблице 3.7.

Таблица 3.7 – Результаты количественных распределений топологических структур по языкам и уровням в векторных пространствах

Языки	Униграммы	Биграммы	Триграммы
Русский	7	12	12
Английский	4	5	7
Казахский	4	6	8
Киргизский	4	5	7

Важно отметить, что состав этих дырок неоднороден. Приведем контуры некоторых из них:

– смешливый, черноглазый, чернобровый, молодица, баять, вишь, сякой, срамить, бранить, сердиться, конфузиться, сконфузить, смущенный, растерянный, озадаченный, ошарашивать, оторопеть, озадаченно, многозначительно, благожелательно, доброжелательно, дружелюбный, приветливый, добродушный, простоватый, разбитной, бойкий;

– ладиться княжество, петровский империя, китайский государство, немецкий власть, латвийский власть, говорить могущество, столько могущество, столько теплота, сколько любовь, ревнивость любовь, выглядеть любовь, выглядеть жизнь, выглядеть либо, озорница либо, эмоциональный либо, эмоциональный будущее, руководимый будущее, многий будущее, многий предок, всеведущий потомок;

– домой нива свой, домой город наступление, домой командующий фронт, прозывать командующий фронт, бывший офицер, вместе полк товарищ, свой отряд товарищ, свой ординарец оренбургский, свой нива полагаться, свой лионский землячка.

Как видно из примеров, объекты, формирующие аномалии (отдельные слова, биграммы или триграммы), включают в себя синонимические пары или группы, но не ограничиваются исключительно синонимами; «линейная комбинация» несинонимических элементов такой границы даёт значения, которые недоступны в соответствующем языке, тем самым выдавая его дыры, слепые зоны [151, p. 4142-4163]. Списки слов/двухсловных/трехсловных сочетаний границ доступны по адресу https://github.com/quynhud/stb-tda/tree/main/hole_contours.

Полученные данные относительно группировки языковых единиц на границах дырок в анализируемых языках позволяют сделать некоторые выводы. В русском языке на уровне слов (униграмм) группируются в основном синонимические ряды слов (преимущественно прилагательные и наречия: *дружелюбный (–о), приветливый, добродушный и т.д),* описывающие характер, эмоции и поведение людей. Однако в английском языке наряду с синонимами “embarrassed” – “awkward” и др., на границах дыр можно выделить и антонимические пары слов “delicate” - “rude”, “embarrassed” – “unabashed”, ‘sun’ – ‘moon’. Что касается глагольной лексики, то она представлена в основном глаголами, характеризующими эмоциональное состояние человека. Это даёт нам основание говорить о концентрации эмотивной лексики на границах дырок (*'бранить', 'сердиться', 'конфузиться', 'сконфузить', 'смущенный',*

'растерянный', 'озадаченный', 'ошарашивать', 'оторопеть', 'озадаченно'), часть которой может содержать лексические единицы, описывающие «уникальные» для конкретного языка эмоциональные состояния (например, «тосковать» в русском языке). Схожая ситуация наблюдается и в английском языке, а также в алтайских языках (казахский, киргизский). Примечательно, что эмотивная лексика преобладает во всех n-граммах (словах, биграммах и триграммах) анализируемых языков.

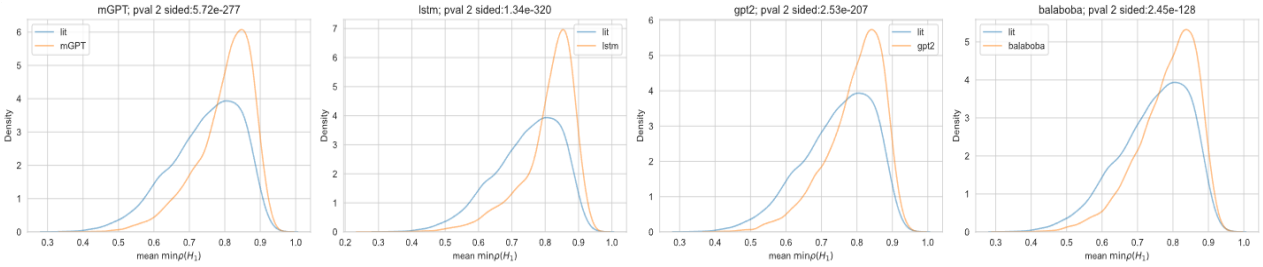
На уровне биграмм показательной является группировка лексики по принципу «гипероним - гипоним» (*власть, мощь – империя, государство, княжество; емкость – бутыль, банка, контейнер, пробирка*). Также выделяются единицы языка, относящиеся к одной лексико-семантической группе (например, одежда – *рубаха, куртка, комбинезон;* еда – *овощ, огурец, десерт).* В целом, следует отметить, что очертания дырок формируются системой взаимосвязанных слов, объединенных общим значением или концептом, где слова могут быть разных частей речи и находиться в различных отношениях (синонимия, антонимия, гипонимия и т.д.).

Пространственный анализ текстов относительно выделенных гомологий

Для выявления различий в распределении языковых единиц между корпусами художественной литературы и искусственно сгенерированных текстов относительно обнаруженных гомологий ("дырок") для каждого слова (лексемы) в каждом корпусе были вычислены следующие пространственные метрики:

1. Евклидово расстояние до центра каждой гомологии.
2. Минимальное расстояние до границы каждой гомологии.
3. Максимальное расстояние до границы каждой гомологии.
4. Расстояние до ближайшей гомологии (минимальное среди расстояний до всех гомологий).

На рисунках 3.27, 32.8, 3.29 представлены усредненные показатели минимальной и максимальной дистанции до слов-гомологов в векторной модели семантического пространства русского языка.



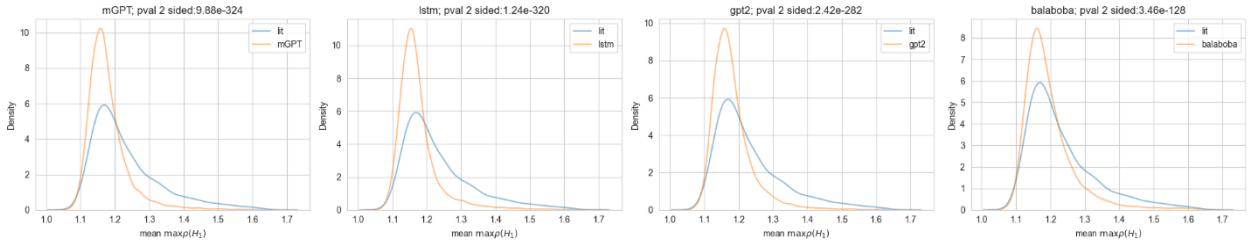
а б в г

а – mGPT; б – LSTM; в – GPT-2: г – YaLM

Рисунок 3.27 – Распределение усреднённых минимальных расстояний до гомологий первого порядка в векторном пространстве слов русского языка

Примечание – Синим показано распределение для текстов литературы, оранжевым – ботов

Рисунок 3.27 (распределение для слов) выдаёт тот факт, что распределение для человеческих текстов (синий цвет) статистически значительно смещено относительно распределения для текстов ботов (оранжевый). То же самое верно и для распределений ближайших гомологий (рисунок 3.29).

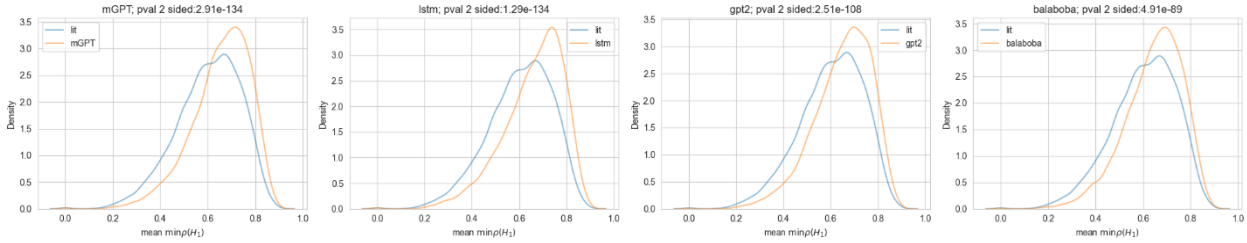


а б в г

а – mGPT; б – LSTM; в – GPT-2: г – YaLM

Рисунок 3.28 – Распределение усреднённых максимальных расстояний до гомологий первого порядка в векторном пространстве слов русского языка

Примечание – Синим показано распределение для текстов литературы, оранжевым – ботов

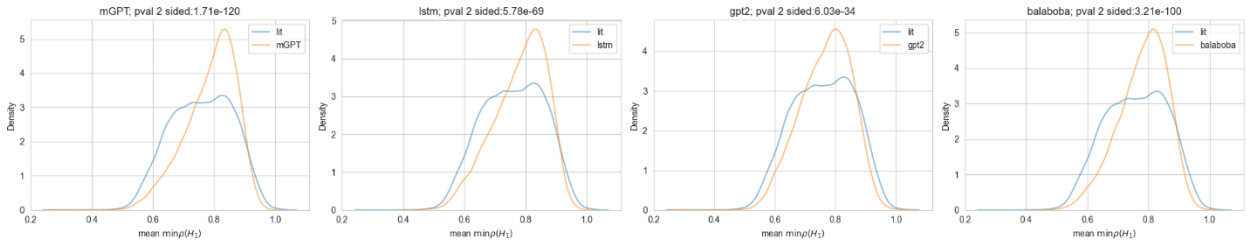


а б в г

а – mGPT; б – LSTM; в – GPT-2: г – YaLM

Рисунок 3.29 – Распределение расстояний до ближайших гомологий первого порядка в векторном пространстве слов русского языка

Примечание – Синим показано распределение для текстов литературы, оранжевым – ботов

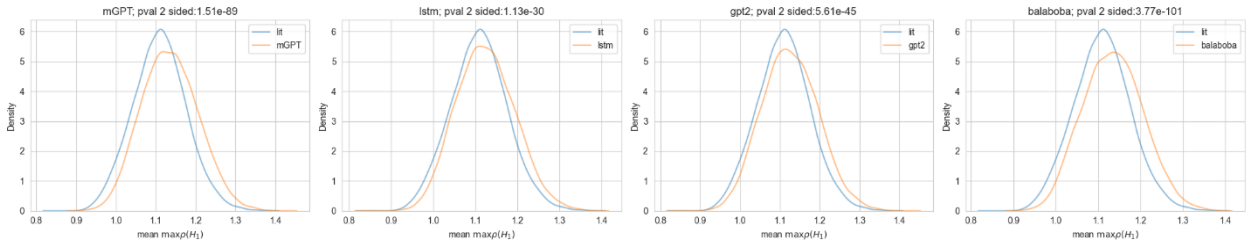


а б в г

а – mGPT; б – LSTM; в – GPT-2: г – YaLM

Рисунок 3.30 – Распределение усреднённых минимальных расстояний до гомологий первого порядка в векторном пространстве слов английского языка

Примечание – Синим показано распределение для текстов литературы, оранжевым – ботов

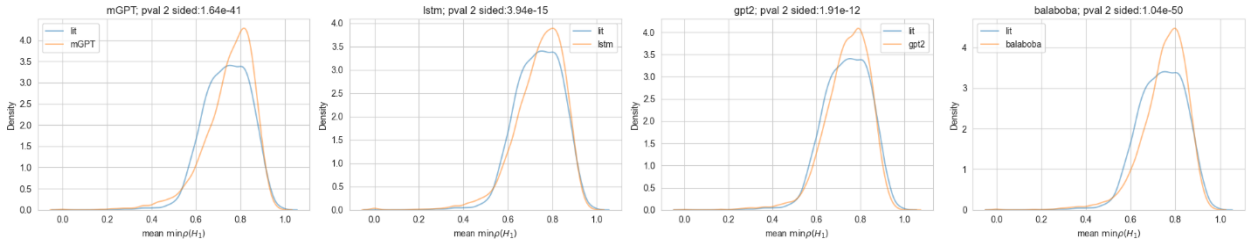


а б в г

а – mGPT; б – LSTM; в – GPT-2: г – YaLM

Рисунок 3.31 – Распределение усреднённых максимальных расстояний до гомологий первого порядка в векторном пространстве слов английского языка.

Примечание – Синим показано распределение для текстов литературы, оранжевым – ботов



а б в г

а – mGPT; б – LSTM; в – GPT-2: г – YaLM

Рисунок 3.32 – Распределение расстояний до ближайших гомологий первого порядка в векторном пространстве слов английского языка

Примечание – Синим показано распределение для текстов литературы, оранжевым – ботов

Рисунки 3.30, 3.31, 3.32 показывают аналогичные распределения для английского языка. В общем, слова и n-граммы текстов ботов статистически находятся дальше от границ гомологий (границ языка), чем тексты людей. В рамках решения задачи идентификации ботов сформулировали статистическую гипотезу о том, что в среднем расстояние (как его не измерять) между текстом и ближайшей гомологией для текстов людей статически не отличается от соответствующей характеристики для текстов ботов, и для верификации гипотезы о различии лингвистических профилей был применен непараметрический метод Колмогорова-Смирнова [154]. Анализ проводился попарно между характеристиками корпуса художественной литературы и текстов, составленных каждым исследуемым ботом. Во всех случаях полученные *p-value* оказались ниже критического порога *0.05*. Это позволяет статистически обоснованно опровергнуть нулевую гипотезу о единой природе распределений сравниваемых выборок. Таким образом, выбранные гомологии демонстрируют диагностический потенциал для дифференциации аутентичных и искусственно созданных текстов [3, p. e2550].

Ключевым индикатором также выступает распределение слов по ближайшим гомологическим центрам ("дыркам"). Наблюдается выраженная дивергенция: в художественной литературе доминирует привязка к 5-ой дырке (визуализация контуров: https://github.com/quynhu-d/stbtda/blob/main/hole_contours/RU/ru_word_hole_contours.txt). Напротив, генеративные модели (mGPT, LSTM, GPT-2, YaLM) проявляют устойчивую ориентацию на 1-ую дырку (40, 44, 38, 32% слов соответственно). При этом доля лексем, ассоциированных с 5-ой дыркой, в сгенерированных корпусах резко сокращена (8-10%), что формирует дополнительный контрастный маркер для детекции синтетического текста.

Переработанный список характеристик для идентификации синтетических текстов:

1. Среднее расстояние до гомологических центров: для каждой отдельной гомологии вычисляется среднее значение расстояний всех лингвистических единиц текста (слов/биграмм/триграмм) до её центра.
2. Интегральный показатель удалённости от гомологий: Усреднённое расстояние всех лингвистических объектов текста (слов/биграмм/триграмм) до центров *всех* гомологий (двойное усреднение: по объектам и по гомологиям).
3. Средняя минимальная дистанция до гомологии: для каждой гомологии рассчитывается среднее значение *минимальных* расстояний, на которые к её центру приближаются лингвистические объекты текста.
4. Обобщённая близость к гомологическим кластерам: Усреднённое значение *минимальных* расстояний всех лингвистических единиц текста до *любого* из центров гомологий (усреднение по объектам и по множеству гомологий).
5. Средняя максимальная удалённость от гомологии: для каждой гомологии определяется среднее значение *максимальных* расстояний, на которые от её центра отстоят лингвистические объекты текста.
6. Интегральный показатель максимальной дистанции: Усреднённое значение *максимальных* расстояний всех лингвистических единиц текста до центров *всех* гомологий (усреднение по объектам и гомологиям).
7. Профиль распределения объектов по ближайшим гомологиям: Доля лингвистических единиц текста (слов/биграмм/триграмм), для которых та или иная конкретная гомология является ближайшей (определяется через минимальное расстояние).

Для каждого анализируемого текста формируется набор из *(n_holes + 1) * 4* признаков, при этом *n_holes* соответствует количеству выявленных гомологий.

Стоит отметить, что величины, проанализированные в [62, p. 2450083], также могут использоваться в качестве таких характеристик.

3.4.3 Классификация: люди и боты

В рамках широкомасштабного исследования для классификации текстов применяются базовые алгоритмы: метод опорных векторов (SVM), деревья решений и ансамбли случайных лесов. Стратегия обучения и оценки строго соответствует постановке задачи: обучающая выборка формируется случайным

образом из текстов, созданных двумя разными генеративными моделями, тогда как тестирование осуществляется на материалах, составленные двумя другими моделями. Оптимизация гиперпараметров реализуется методом 10-кратной перекрестной проверки. Для SVM варьируется параметр регуляризации C в диапазоне от 1e-5 до 10. При решающих деревьях решений и случайных лесов исследуется глубина деревьев (от 3 до 15 уровней) и минимальное количество объектов, требуемое для образования листа (от 1 до 4).

Результаты классификации (accuracy, выборки сбалансированы) представлены в таблицах 3.8 (русский) и 3.9 (английский). Несмотря на тестирование на данных *новых*, незнакомых моделей ботов, все алгоритмы показали выше 0.71. Наилучшие усредненные показатели среди трех моделей по всем шести комбинациям выборок: русский язык (слова - 0.86, биграммы - 0.82, триграммы - 0.88) и английский язык (слова - 0.87, биграммы - 0.93, триграммы - 0.89). При этом для английского языка, за редким исключением, модель случайного леса продемонстрировала превосходство над остальными классификаторами [3, p. e2550; 155].

Таблица 3.8 – Значения accuracy классификации текстов русского языка

Мо дели	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
	M_1, M_2	M_3, M_4	M_1, M_3	M_2, M_4	M_1, M_4	M_2, M_3	M_2, M_3	M_1, M_4	M_2, M_4	M_1, M_3	M_3, M_4	M_1, M_2
Слова												
SVC	0.93	0.916	0.947	0.895	0.939	0.952	0.983	0.743	0.963	0.779	0.971	0.729
DT	0.979	0.908	0.932	0.926	0.98	0.928	0.994	0.774	0.982	0.868	0.964	0.779
RF	0.998	0.895	1	0.921	0.974	0.945	0.998	0.756	1	0.795	0.992	0.774
Биграммы												
SVC	0.962	0.937	0.975	0.852	0.986	0.79	0.972	0.714	0.974	0.727	0.988	0.703
DT	0.995	0.886	0.922	0.876	0.981	0.828	0.99	0.705	1	0.712	1	0.708
RF	1	0.929	1	0.917	0.995	0.911	1	0.726	1	0.737	0.999	0.712
Триграммы												
SVC	0.963	0.934	0.971	0.924	0.964	0.885	0.961	0.718	0.965	0.721	0.961	0.705
DT	0.987	0.886	0.98	0.883	0.987	0.859	0.989	0.822	0.992	0.927	0.992	0.737
RF	0.993	0.943	1	0.947	0.997	0.879	1	0.755	1	0.737	1	0.735

Примечания:
1. SVC – метод опорных векторов.
2. DT – решающее дерево.
3. RF – случайный лес.
4. Обозначения моделей: M_1 – LSTM, M_2 – YaLM, M_3 – GPT-2, M_4 – mGPT

Таблица 3.9 – Значения accuracy классификации текстов английского языка

Моде ли	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
	M_1, M_2	M_3, M_4	M_1, M_3	M_2, M_4	M_1, M_4	M_2, M_3	M_2, M_3	M_1, M_4	M_2, M_4	M_1, M_3	M_3, M_4	M_1, M_2
1	2	3	4	5	6	7	8	9	10	11	12	13
Слова												
SVC	0.879	0.922	0.919	0.903	0.878	0.771	0.96	0.844	0.958	0.84	0.972	0.803
DT	0.99	0.932	0.983	0.889	0.973	0.768	0.983	0.85	0.949	0.828	0.986	0.84
RF	0.998	0.935	0.983	0.92	0.98	0.773	0.998	0.857	0.989	0.861	0.997	0.852

Продолжение таблицы 3.9

1	2	3	4	5	6	7	8	9	10	11	12	13
Биграммы												
SVC	0.947	0.94	0.972	0.925	0.964	0.868	0.963	0.941	0.959	0.94	0.968	0.878
DT	0.975	0.935	1	0.914	0.989	0.866	0.977	0.879	1	0.918	0.993	0.866
RF	0.99	0.964	0.999	0.929	0.997	0.856	0.99	0.93	1	0.963	0.996	0.896
Триграммы												
SVC	0.922	0.886	0.952	0.83	0.968	0.841	0.906	0.755	0.893	0.764	0.935	0.722
DT	0.969	0.878	1	0.813	0.941	0.793	0.957	0.819	0.976	0.902	0.982	0.884
RF	0.976	0.914	1	0.883	0.994	0.747	0.993	0.928	1	0.909	0.993	0.87

Примечания:
1. SVC – метод опорных векторов.
2. DT – решающее дерево.
3. RF – случайный лес.
4. Обозначения моделей: M_1 – LSTM, M_2 – YaLM, M_3 – GPT-2, M_4 – mGPT

В работе в рамках естественнонаучного подхода [3, p. e2550; 155, с. 281-311] рассматривается естественный язык, а именно наиболее базовые топологические свойства Хаилонакеи (совокупности языковых фрактальных структур для эмбеддингов слов, биграмм, триграмм и т.д.). В основе данного исследования лежит применение методов топологического анализа данных (ТАД) для декомпозиции семантических пространств естественного языка. Ключевым инструментом выступило вычисление персистентных гомологий с последующей оконтуриванием их границ. Анализ сфокусировался на гомологиях 1-го порядка, где был проведен строгий отбор структур, релевантных макромасштабной организации языка (английского и русского), с отсевом артефактов, порожденных нерепрезентативностью выборки. Интересно, что в обоих языках количество таких значимых топологических инвариантов оказалось ограниченным. Результирующие контуры очерчивают специфические «слепые зоны» языка – области, лишенные униграмм, биграмм или триграмм, что указывает на концептуальные области, не репрезентированные в данной языковой системе.

Полученные топологические особенности легли в основу нового подхода к детекции автоматически сгенерированных текстов. Для корпусов разнородной природы (художественная литература и составленные тексты моделей mGPT, GPT-2, LSTM, YaLM) были рассчитаны дистанционные метрики, отражающие удаленность языковых единиц (слов, биграмм, триграмм) от границ выявленных персистентных гомологий. На этих признаках обучались классификационные модели в рамках эксперимента: обучение проводилось на данных от одного подмножества генеративных моделей, а валидация – строго на сгенерированных текстах иных ботов. Суть метода – сравнение усредненных дистанций языковых элементов до ближайшей семантической «дыры» в человеческих и синтетических текстах. Разработанные классификаторы демонстрируют статистически значимое отличие и высокую эффективность со средней точностью (average accuracy) свыше 0.8.

В рамках решения задачи идентификации ботов, для текстов различной природы: текстов художественной литературы и текстов, сгенерированных моделями mGPT, GPT-2, LSTM, YaLM, были посчитаны характеристики, основанные на расстоянии входящих в тексты слов/биграмм/триграмм до выделенных персистентных гомологий. На указанных признаках были обучены модели классификации с нестандартной постановкой – модели обучались на текстах, сгенерированных одним набором ботов, а тестировались на текстах, сгенерированных оставшимися ботами. Для этой цели мы оцениваем средние расстояния от слов, биграмм и триграмм текста до границ ближайшей «дыры» как для текстов, написанных людьми, так и для текстов, сгенерированных ботами, и строим классификаторы. Классификаторы показывают сравнительно хорошие результаты: средняя точность превышает 0.8.

Выводы по третьему разделу

Опираясь на материал, изложенный в данной главе, можно прийти к следующим выводам:

1. Подтверждена гипотезу о том, что естественные языки являются системами самоорганизующейся критичности по широкому спектру языков с замечательным исключением: эсперанто, искусственный язык, который демонстрирует гауссовское распределение. Эта находка подчеркивает уникальность естественных языков и выделяет стабильность классификаций, основанных на характеристиках их сложных систем. Более того, результаты открывают путь к переопределению классификации языков, которая учитывает динамические свойства самоорганизующейся критичности. Дополнительно, исследование показывает, что статистические характеристики степенных распределений – в частности, параметры степенного закона – могут служить надежными метриками для понимания и классификации языков за пределами их традиционных классификации.

2. Результаты анализа семантических траекторий в 52 языках 18 различных языковых семей подтверждают теорию о хаотической природе языков, показывая, что подавляющее большинство из них обладает хаотическими свойствами. Кластерный анализ выявил типологические кластеры, которые коррелируют с распределением языков по параметрам энтропии и сложности. Это подчеркивает важность хаоса как фактора, влияющего на языковую динамику и структуру. Установленные взаимосвязи между языковыми характеристиками, такими как порядок слов, выравнивание, морфологическая сложность и тип маркировки, подчеркивают необходимость глубокого анализа междисциплинарных аспектов языков, что может иметь практическое применение, в том числе в области выявления ботов и оценки качества перевода.

3. Результаты исследования внутренней размерности естественных языков выявил их мультифрактальную природу, что отражает сложность и разнообразие структур ествественного языка. Оценка внутренней размерности, выполненная с использованием алгоритмов Швайнхарта и Брито, продемонстрировала, что размерности языков являются инвариантными

характеристиками, устойчивыми к изменениям в методах извлечения векторных представлений и параметрах оценки. Полученные результаты показывают, что внутренние размерности языков существенно отличаются от теоретически ожидаемых значений. Данные о взаимосвязях между языками, основанные на кластерном анализе, подтверждают существующие лингвистические гипотезы о родственных языках и их территориальных взаимосвязях, что может способствовать переосмыслению классификаций языков с учётом их динамических свойств и морфологической структуры.

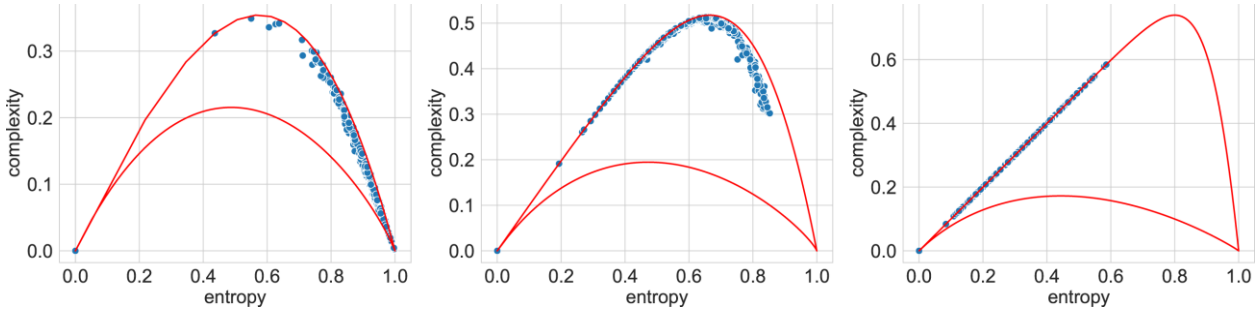
4. Применён естественнонаучный подход для анализа топологических свойств языковых структур на основе концепции Хаилонакеи, с фокусом на фрактальных структурах для униграмм, биграмм, триграмм и т. д. Использованы методы топологического анализа данных (TDA) для декомпозиции семантических пространств естественного языка, особенно с расчётом персистентных гомологий первого порядка, что позволяет выявить масштабные характеристики естественныхо языков, исключая артефакты выборки. Исследование показало ограниченное количество значимых топологических инвариантов и специфические «слепые зоны» языка, указывая на области концептуального отсутствия в языковых системах. В экспериментах была проведена аналитика текстовых корпусов, включая художественную литературу и тексты, сгенерированные моделями mGPT, GPT-2, LSTM и YaLM. Рассчитаны метрики, отражающие удалённость языковых единиц от границ персистентных гомологий, на основе которых обучены классификационные модели, продемонстрировавшие высокую точность (средняя точность более 0.8) в различении текстов, написанных людьми, и сгенерированных ботами.

4 ИДЕНТИФИКАЦИЯ БОТОВ

4.1 Результаты классификации текстов на естественном языке

Выбор значений размерности пространства вложения d и числа слов в n-грамме n имеет определяющее значение для эффективности работы рассматриваемых алгоритмов. Как отмечалось ранее, в качестве критерия выбора значений указанной пары параметров рассматриваем попадание точек с координатами энтропия и сложность для большинства семантических траекторий корпуса литературных текстов в “область хаоса“ на плоскости энтропия-сложность. В ходе широкомасштабного вычислительного эксперимента [3, p. e2550; 56; 155, с. 281-311; 156] для всех рассматриваемых языков были установлены значения указанных параметров, при которых подавляющее большинство литературных текстов оказывается в области хаоса.

В качестве иллюстрации на рисунке 4.1 приведены точки на плоскости энтропия-сложность, отвечающие семантическим траекториям корпуса литературных текстов русского языка: видно, что рисунок 4.1 выявляет следующие закономерности: для $n = 4, d = 1$ подавляющая часть точек соответствует простым случайным процессам (экстремально малые величины). При $n = 3, d = 4$ большинство точек попадает в область хаотических процессов, что свидетельствует о достижении оптимальных параметров. Параметры $n = 5, d = 4$ смещают распределение точек преимущественно в область простых детерминированных систем; данные на других языках приведены в (Приложении Д).



а б в

а – $n = 4, d = 1$; б – $n = 3, d = 4$; в – $n = 5, d = 4$

Рисунок 4.1 – Плоскость энтропии-сложности: точки, соответствующие текстам русской литературы

Используем малые и большие значения параметров d и n, при которых текстовые данные демонстрируют свойства, близкие к случайному процессу, признаны недостаточно релевантными для моделирования их истинной природы. Эти пороговые значения устанавливают нижний предел применимости метода. Семантические траектории становятся информативными лишь при превышении данного порога. Однако конечный объем литературных

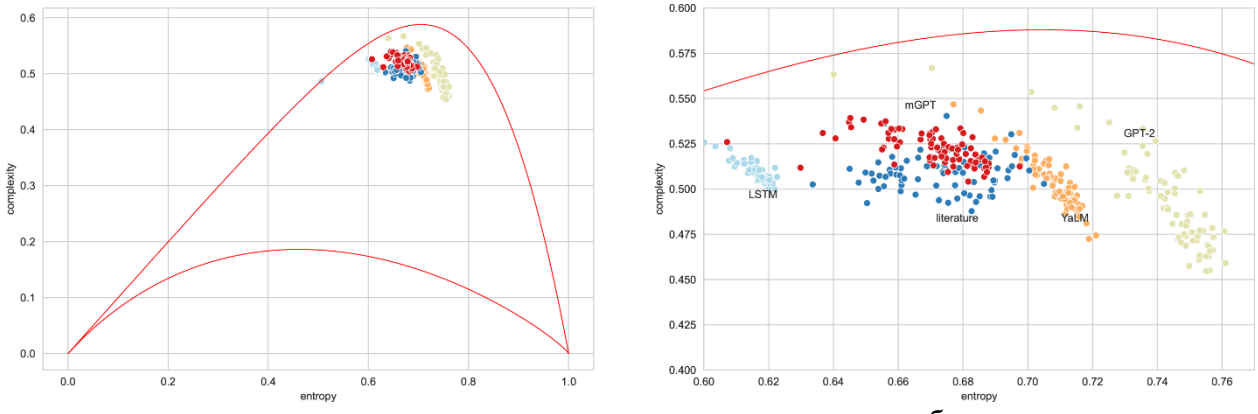
произведений накладывает естественное ограничение сверху, делая некорректной оценку энтропии и сложности для экстремально больших d и n.

Отметим, что области оптимальных значений могут существенно отличаться для разных языков, например: для вьетнамского в рассматриваемую область входят более длинные последовательности слов; при $d = 1$ оптимальные значения n от 10 до 14, тогда как для русского – от 6 до 8, а для английского – от 7 до 8. Мы связываем это с порядком слов в языке (свободный, фиксированный, промежуточные варианты).

Критически важно, что диапазоны оптимальных параметров d и n варьируются в зависимости от языка. Например, анализ показывает, что для вьетнамского языка (характеризующегося относительно свободным порядком слов) значимы более длинные последовательности (n от 10 до 14 при $d = 1$), в то время как для русского (от 6 до 8) и английского (от 7 до 8) оптимальные длины n существенно ниже. Эта зависит от порядка слов в определенном языке (промежуточный/фиксированный/свободный порядок слов).

4.1.1 Классификация на основе положения точки на плоскости “энтропия - сложность”

На рисунке 4.2 представлены характерный вид множества точек на плоскости энтропия-сложность, отвечающих семантическим траекториям литературных текстов. Рисунок соответствует вьетнамскому языку: здесь синие точки отвечают литературным текстам, точки других цветов - текстам, сгенерированных рассматриваемыми ботами – рисунок позволяет сделать вывод о принципиальной возможности разделить тексты с использованием этих признаков.



а б

а – полная плоскость; б – увеличенная часть

Рисунок 4.2 – Плоскость энтропии-сложности: точки, соответствующие текстам на вьетнамском языке для допустимых значений $d = 3, n = 4$

В таблице 4.1 приведены результаты применения классификаторов, опирающегося на данные характеристики, на тестовых выборках для всех рассматриваемых в работе языков. Любопытно, что оптимальные модели

отличаются для разных языков – для русского и французского языков наибольшее качество классификации достигается при использовании метода опорных векторов, для английского, немецкого и вьетнамского – случайный лес.

Таблица 4.1 – Значения F1-score для классификаторов на основе энтропии-сложности

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
SVM	*0.89*	0.64	0.97	*0.95*	0.95
DT	0.76	0.81	0.97	0.59	0.96
RF	0.78	*0.82*	*0.98*	0.87	*0.97*
Примечания: 1. SVM – Support Vector Machine (метод опорных векторов). 2. DT – Decision Tree (решающее дерево). 3. RF – Random Forest (случайный лес)					

При этом значения F1-score оптимальных моделей выше 0.8, что указывает на то, что даже при обучении на текстах одного набора ботов (GPT2 и YaLM), модель имеет обобщающую способность и отделяет литературные тексты от текстов, сгенерированных LSTM и mGPT, прежде не встречавшихся при обучении классификатора.

4.1.2 Классификация по характеристикам семантических траекторий

На рисунке 4.3 представлен характерный вид множества значений обобщённой энтропии ($q = 0..20$) для вьетнамского языка. Множества значений для других языков и для других значений параметра q может приведены в (Приложении Д).

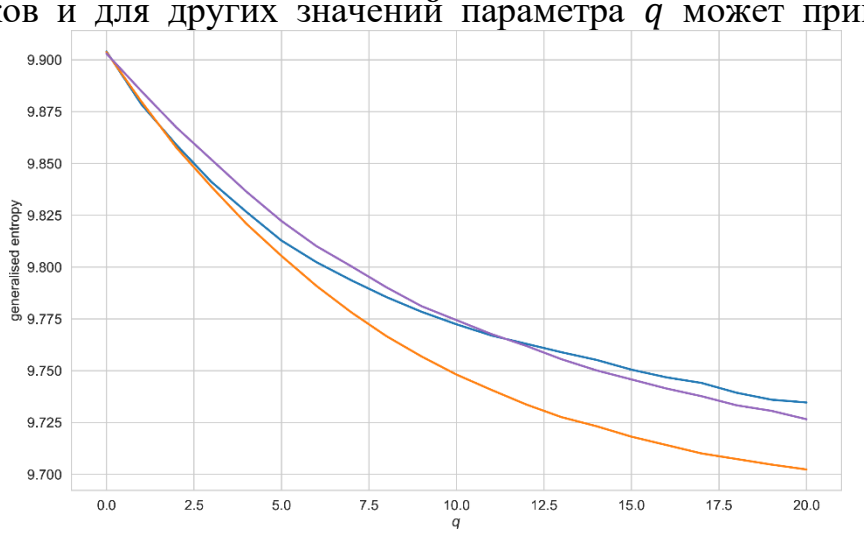


Рисунок 4.3 – Обобщенные значения энтропии для текстов на вьетнамском языке, q q в диапазоне от 0 до 20

Примечание – Синяя линия относится к литературным текстам, пурпурная линия - к текстам, сгенерированным GPT-2, оранжевая - к текстам, сгенерированным LSTM

Таблица 4.2 – Значения F1-score для классификаторов с семантическими характеристиками траекторий

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
Support Vector Machine					
SVD	0.52	0.50	0.59	0.75	0.52
CBOW	0.00	0.59	*0.77*	0.00	*0.70*
Skip-Gram	0.00	0.00	0.65	0.59	0.50
Decision Tree					
SVD	0.66	0.78	0.65	0.85	0.67
CBOW	0.02	0.58	0.80	0.00	0.62
Skip-Gram	0.00	0.00	0.62	0.66	0.59
Random Forest					
SVD	0.68	0.79	0.63	0.86	0.68
CBOW	0.05	0.56	0.81	0.00	0.68
Skip-Gram	0.00	0.00	0.68	0.74	0.60

В таблице 4.2 представлены результаты вычислительного эксперимента для данного типа признаков, другие показатели точности классификаторов приведены в (Приложении Д). Полученные результаты позволяют сделать вывод, что признаки данного типа дают существенно худшие результаты, чем другие классы признаков. Тем не менее, даже худшие результаты для немецкого и вьетнамского языков здесь не опускаются ниже 0.5. Для русского, английского и французского языков лучшими оказались решающие деревья, обученные на SVD-эмбеддингах. Для немецкой наилучшей модели также оказалось решающее дерево, однако, обученное на CBOW-векторах. Заметим, что, как и в случае с энтропией-сложности решающее дерево нельзя назвать универсально оптимальной моделью – для вьетнамского языка решающее дерево переобучается, и качество классификации выше при применении метода опорных векторов (на эмбеддингах CBOW).

4.1.3 Подход, основанный на кластеризации n-грамм

В таблице 4.3 приведены результаты моделирования для признаков, основанных на методе кластеризации Wishart. Оценки точности классификаторов представлены в разделе «Кластеризация n-грамм и меры когезии кластеров», а также в (Приложении Д). Для большинства языков метод опорных векторов (SVM) демонстрирует более высокие показатели качества по сравнению с деревом решений; однако для русского языка дерево решений показывает лучшие результаты. Выбор оптимального пространства вложения также зависит от языка. В таблице 4.4 приведены соответствующие результаты для признаков, основанных на кластеризации K-Means; показатели точности классификаторов приведены в (Приложении Д). Аналогично, SVM показывает лучшие результаты, чем дерево решений для большинства языков. Однако для русского и вьетнамского языков решающее дерево показывает значительно лучшие F1-оценки. Лучшее вложение для кластеризации K-Means также нельзя выбрать. Таким образом, невозможно выбрать конкретное вложение или конкретную архитектуру для классификации с использованием кластеризации n-грамм текстовых векторов. Интересно, что для разных языков лучшие результаты показывали разные комбинации типа вложения и варианта классификатора, и, что самое важное, вариант, который является наилучшим для

данного языка, показывает плохие результаты для языков других языковых семей.

Таблица 4.3 – Значения F1-score для классификаторов на основе кластеризации Wishart

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
Support Vector Machine					
SVD	0.41	0.75	0.92	0.93	0.57
CBOW	0.53	0.87	0.60	0.57	0.76
Skip-Gram	0.61	0.81	0.58	0.96	0.82
Decision Tree					
SVD	0.61	0.70	0.81	0.84	0.76
CBOW	0.81	0.75	0.69	0.57	0.32
Skip-Gram	0.79	0.68	0.58	0.64	0.28
Random Forest					
SVD	0.48	0.77	0.89	0.48	0.45
CBOW	0.81	0.83	0.64	0.58	0.28
Skip-Gram	0.78	0.76	0.52	0.66	0.31

Таблица 4.4 – Значения F1-score для классификаторов на основе кластеризации K-Means

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
SVM					
SVD	0.59	0.78	0.92	0.58	0.70
CBOW	0.21	0.95	0.82	0.84	0.44
Skip-Gram	0.40	0.86	0.85	0.85	0.26
Decision Tree					
SVD	0.33	0.80	0.87	0.35	0.72
CBOW	0.74	0.87	0.47	0.45	0.27
Skip-Gram	0.84	0.82	0.32	0.71	0.19
Random Forest					
SVD	0.50	0.92	0.86	0.83	0.66
CBOW	0.18	0.90	0.38	0.93	0.17
Skip-Gram	0.78	0.86	0.62	0.63	0.22

В ходе широкомасштабного вычислительного эксперимента установлено, что результаты классификации зависят от используемого алгоритма кластеризации. В таблице 4.5 представлены результаты классификации при применении алгоритмов Wishart, K-Means и их нечётких модификаций – Fuzzy Wishart и C-Means; соответствующие показатели точности приведены в (Приложении Д).

В качестве признаков использовались усреднённые по всем кластерам значения внутрикластерных расстояний (признаки 5–8, см. раздел «Кластеризация n-грамм и меры когезии кластеров»).

Таблица 4.5 – Значения F1-score для классификаторов на основе внутрикластерных расстояний

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
SVM					

Wishart	0.53	0.64	0.54	0.35	0.68
Fuzzy Wishart	0.49	0.84	0.50	0.89	0.60
K-Means	0.95	0.80	0.51	0.63	0.65
C-Means	0.93	0.76	0.52	0.47	0.54
Decision Tree					
Wishart	0.53	0.64	0.54	0.35	0.68
Fuzzy Wishart	0.49	0.84	0.50	0.89	0.60
K-Means	0.95	0.80	0.51	0.63	0.65
C-Means	0.93	0.76	0.60	0.47	0.54
Random Forest					
Wishart	0.55	0.72	0.71	0.61	0.67
Fuzzy Wishart	0.69	0.85	0.89	0.93	0.81
K-Means	0.98	0.86	0.61	0.51	0.70
C-Means	0.95	0.78	0.60	0.67	0.72

В целом наилучшие результаты при формировании признаков для классификации продемонстрировал алгоритм Wishart как в чёткой, так и в нечёткой постановке.

Введение нечёткости в алгоритм Wishart существенно повышает качество классификации на выделенных признаках: наилучшие результаты для немецкого, французского и вьетнамского достигаются на моделях случайного леса (random forest), обученных именно на признаках, выделенных из нечёткой модификации кластеризации Wishart. При этом и для русского и английского языков также видно улучшение качества модели по сравнению со случаем, когда применяется классический алгоритм Wishart. Наиболее это ярко выражено в случае французского языка – применение обычного Wishart давало значение F1=0.61, нечёткой модификации – F1 = 0.93 (увеличение на 0.12).

Единая классификационная модель

Рассмотрена классификационная модель, основанная на совокупности всех ранее описанных признаков. В таблице 4.6 приведены результаты классификации соответствующих моделей; показатели точности представлены в (Приложении Д).

Таблица 4.6 – Значения F1-score для классификаторов на основе всех описанных методов в совокупности

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
SVM	0.82	0.98	0.63	0.82	0.74
DT	0.76	0.85	0.55	0.59	0.62
RF	0.86	0.86	0.57	0.68	0.66

Примечания:
1. SVM – Support Vector Machine (метод опорных векторов).
2. DT – Decision Tree (решающее дерево).
3. RF – Random Forest (случайный лес)

Вычислительный эксперимент показал, что простое объединение признаков не всегда приводит к повышению качества классификации.

Например, для немецкого и вьетнамского языков оптимальные значения F1-меры для объединённого классификатора оказались ниже 0.75, в сравнении с

классификатором, основанном на энтропии-сложности (F1 = 0.98 для немецкого и 0.97 для вьетнамского). С другой стороны, значительно повысилось качество классификации английских текстов – значение F1 поднялось до 0.98 по сравнению с максимальным 0.87 классификатора, основанного на характеристиках кластеров. При этом, было установлено, что добавление одних признаков снижают дифференцирующую способность других признаков, а значит, и модели в целом.

Выводы по четвертому разделу

Опираясь на материал, изложенный в данной главе, можно прийти к следующим выводам:

1. Сформулирована задача разделения текстов всех ботов от текстов всех людей – как представляется, такая постановка задачи является более разумной – чем задача идентификации отдельного бота, сколь бы эффективным он не был. Обучающая и тестовая выборка здесь формировались путём случайного разделения текстов ботов и людей, но самих ботов и людей, таким образом в тестовой выборке оказываются тексты тех ботов (и людей), которые отсутствовали в выборке обучающей.

2. В ходе широкомасштабного вычислительного эксперимента было установлено, что задача разделения текстов может быть успешно решена, однако оптимальные алгоритмы классификации и признаки варьируются в зависимости от языковой семьи. Использование простейших классификаторов позволило определить сравнительную эффективность различных признаков. Наилучшие результаты на тестовых выборках были достигнуты следующими методами: для русского языка – случайный лес с внутрикластерными расстояниями K-Means (F1 = 0.98); для английского языка – метод опорных векторов с комбинацией всех признаков (F1 = 0.98); для немецкого языка – случайный лес с энтропией-сложностью (F1 = 0.98); для французского языка – метод опорных векторов с внутрикластерными расстояниями и усреднёнными координатами центров кластеров (F1 = 0.96); для вьетнамского языка – случайный лес с энтропией-сложностью (F1 = 0.97).

3. Качество классификации и выбор признаков: несмотря на отсутствие ботов в тестовой выборке и использование простейших классификаторов, разумный выбор признаков для классификации позволил достичь качества классификации свыше 96% для языков различных языковых семей. Моделирование показало, что механическое сочетание признаков не всегда приводит к улучшению качества классификации, подчеркивая важность тщательного отбора характеристик.

ЗАКЛЮЧЕНИЕ

Анализ экспериментальных исследований подтвердил гипотезу о том, что естественные языки функционируют как самоорганизованно-критичные системы, что наблюдается в широком диапазоне языков (52 языка из 18 различных языковых семей), за исключением эсперанто – искусственного языка. Это подчеркивает уникальность естественных языков и указывает на устойчивость классификаций, основанных на характеристиках сложных систем. Кроме того, исследование демонстрирует, что статистические характеристики степенных распределений, включая параметры степенного закона, могут выступать в качестве надежных метрик для понимания и классификации языков, выходя за пределы традиционных подходов.

Проведенный анализ семантических траекторий подтверждает хаотическую природу языков, демонстрируя, что большинство из них обладает хаотическими свойствами. Кластерный анализ выявил типологические кластеры, коррелирующие с распределением языков по параметрам энтропии и сложности. Это подчеркивает значимость хаоса как фактора, влияющего на языковую динамику и структуру.

Методы исследования внутренней размерности геометрических объектов выявил мультифрактальную природу естественных языков, отражающую сложность и разнообразие структур языка. Оценка внутренней размерности, проведённая с использованием алгоритмов Швайнхарта и Брито, показала, что размерности языков являются инвариантными характеристиками, устойчивыми к изменениям в методах извлечения векторных представлений и параметрах оценки. Данные о взаимосвязях между языками, полученные в результате кластерного анализа, коррелируют (хотя и не всегда) существующие лингвистические гипотезы о родственных языках и их территориальных взаимосвязях.

Применение методов топологического анализа данных для декомпозиции семантических пространств естественного языка, показало ограниченное количество значимых топологических инвариантов и определенные «слепые зоны» в языковых системах, что указывает на концептуальные пробелы.

При анализе задач классификации текстов ботов и людей, был получен высокий уровень эффективности алгоритмов, варьирующихся, впрочем, в зависимости от языковой семьи. Исследование показало, что использование простейших классификаторов и разумный отбор признаков позволили достичь качества классификации свыше 96% для различных языков: для русского и немецкого языков применялся случайный лес с энтропией-сложностью, для английского – метод опорных векторов, а для французского и вьетнамского языков также использовался случайный лес и метод опорных векторов с оптимальными комбинациями признаков. Результаты подтверждают, что механическое сочетание признаков не всегда ведет к улучшению качества, что подчеркивает важность тщательного выбора характеристик для классификации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Gromov V.A., Migrina A.M. A language as a Self-Organized Critical System // Complexity. – 2017. – Vol. 2017. – P. 1-7.

2 Gromov V.A., Borodin N.S., Yerbolova A.S. A language and its dimensions: intrinsic dimensions of language fractal structures // Complexity. – 2024. – Vol. 2024, Issue 1. – P. 863360.

3 Gromov V.A., Dang Q.N., Yerbolova A.S. et al. Spot the bot: the inverse problems of NLP // Peer J Computer Science. – 2024. – Vol. 10. – P. e2550.

4 Allott N., Lohndal T., Rey G. Synoptic introduction // In book: A Companion to Chomsky. – New Jersey: Wiley Blackwell, 2021. – P. 1-17.

5 Chomsky N. Knowledge of language Its nature, origin, and use. – NY.: Praeger, 1986. – 307 p.

6 Hernandez-Fernáandez A., Garrido J.M., Luque B. et al. Linguistic laws in catalan // In book: Quantitative Approaches To Universality and Individuality in Language. – Berlin, 2022. – Vol. 75:49-62.

7 Torre I.G., Luque B., Lacasa L. et al. On the physical origin of linguistic laws and lognormality in speech // R. Soc. open sci. – 2019. – Vol. 6, Issue 8. – P. 191023.

8 Baixeries J., Elvevag B. et al. The evolution of the exponent of zipf's law in language ontogeny // Plos One. – 2013. – Vol. 8, Issue 3. – P. e53227.\

9 Wang Y., Liu H. Revisiting Zipf's law: a new indicator of lexical diversity // In book: Quantitative Approaches to Universality and Individuality in Language. – Berlin; Boston: De Gruyter Mouton, 12023. – P. 193-202.

10 Debowski L.Information theory meets power laws: stochastic processes and language models. – Hoboken: J. Wiley, 2021. – 384 p.

11 Tanaka-Ishii K. Statistical Universals of Language: Statistical Universals of Language: Mathematical Chance vs. Human Choice. – Ed. 1st. – Cham: Springer, 2021. – 244 p.

12 Tanaka-Ishii K., Aihara S. Computational constancy measures of texts–yule's k and renyi's entropy // Computational Ling. – 2015. – Vol. 41, Issue 3. – P. 481-502.

13 Tanaka-Ishii K., Bunde A. Long-range memory in literary texts: on the universal clustering of the rare words // Plos One. – 2016. – Vol. 11, Issue 11. – P. e0164658.

14 Tanaka-Ishii K., Takahashi S. A comparison of two fluctuation analyses for natural language clustering phenomena–taylor vs. ebeling & neiman methods // Fractals. – 2021. – Vol. 29, Issue 02. – P. 2150033.

16 Semple S. et al. Linguistic laws in biology // Trends in Ecology & Evolution. – 2022. – Vol. 37, Issue 1. – P. 53-66.

15 Dębowski Ł. A Simplistic Model of Neural Scaling Laws: Multiperiodic Santa Fe Processes // https://arxiv.org/pdf/2302.09049v1. 10.10.2025.

17 Garg M., Kumar M. The structure of word co-occurrence network for microblogs // Physica A: Statistical Mechanics and its Applications. – 2018. – Vol. 512. – P. 698-720.

18 Garg M., Gupta A.K., Prasad R. Graph Learning and Network Science for Natural Language Processing (1st ed.). – Boca Raton: CRC Press, 2022. – 256 p.

19 Koltsova O.Y., Koltcov S.N., Nikolenko S.I. Communities of co-commenting in theRussian LiveJournal and their topical coherence // Internet Research. – 2016. – Vol. 26, Issue 3. – P. 710-732.

20 Saddy D., Uriagereka J. Measuring language // International Journal of Bifurcation and Chaos. – 2004. – Vol. 14. Issue 02. – P. 383-404.

21 Krivochen D.G. Mixed computation: Grammar up and down the Chomsky Hierarchy // Evolutionary Linguistic Theory. – 2021. – Vol. 3, Issue 2. – P. 215-244.

22 Krivochen D.G. On Phrase Structure building and labeling algorithms: towards a non-uniform theory of syntactic structures // The Linguistic Review. – 2015. – Vol. 32, Issue 3. – P. 515-572.

23 Uriagereka J. On the Emptiness of'Design'Polemics // Natural Language & Linguistic Theory. – 2000. – Vol. 18, Issue 4. – P. 863-871.

24 Krivochen D.G. Language, chaos and entropy: A physical take on bio linguistics // An International Journal of Theoretical Linguistics. – 2014. – Vol. 6. – P. 27-74.

25 Mirkin B.G. Clustering for data mining: A data recovery approach. – London: Chapman & Hall/CRC, 2012. – 266 p.

26 Cancho R.F. et al. The small world of human language // Proc Biol Sci. – 2001. – Vol. 268, Issue 1482. – P. 2261-2265.

27 Ferrer I., Cancho R. The variation of Zipf's law in human language // Eur. Phys. J. – 2005. – Vol. 44. – P. 249-257.

28 Ribeiro L.C., Bernardes A.T., Mello H. On the fractal patterns of language structures // Plos One. – 2023. – Vol. 18, Issue 5. – P. e0285630.

29 Ma Q., Xinxin W. What Is Language Complexity? // Macrolinguistics. – 2019. – Vol. 7, Issue 11. – P. 1-29.

30 马庆株. 系统理论框架下的语言课题, 2012 (Ma Q. Linguistics Topics under System Theories. – Shanghai).

31 Larsen-Freeman D. Chaos/complexity science and second language acquisition // Applied Linguistics. – 1997. – Vol. 18, Issue 2. – P. 141-165.

32 Шелестюк Е.В., Щетинкина Е.А. Стохастичность и энтрипия в лингвистике // Вестник Челябинского государственного университета. – 2023. – №2(472). – С. 150-165.

33 Arikawa K. Frustrated Nonphases as Catalysts for Phases - How Graph Theory Calculates Optimal Balance in Linguistic Fibonacci Trees and Graphs // Journal of Cognitive Science. – 2020. – Vol. 21, Issue 2. – P. 253-384.

34 Marcolli M. et al. Mathematical Structure of Syntactic Merge. – Cambridge: MIT Press, 2025. – 414 p.

35 Medeiros D. et al. The Golden phrase: steps to the physics of language // In book: Language, Syntax, and the Natural Sciences. – Cambridge, 2018. – P. 333-350.

36 Orus R., Martin R., Uriagereka J. Mathematical foundations of matrix syntax // https://arxiv.org/abs/1710.00372. 10.10.2025.

37 Piattelli-Palmarini M., Vitiello G. Third factors in language design: some suggestions from Quantum Field Theory // In book: Cambridge Companion to Chomsky. – Ed. 2nd. – Cambridge: Cambridge University Press, 2017. – P. 134-152.

38 Baglioni M., Macedo J., Renso C. et al. An ontology-based approach for the semantic modelling and reasoning on trajectories // Procced. internat. conf. on conceptual modelling. – Barcelona, 2008. – P. 344-353.

39 Oueslati W., Sami O., Bahri A. et al. Generic Semantic Trajectory Data Modelling Approach based on Ontologies // Journal of Information & Knowledge Management. – 2024. – Vol. 23, Issue 06. – P. 2450083.

40 Wu X., Liu Y., Zhao X. et al. STKST-I: An Efficient Semantic Trajectory Search by Temporal and Semantic Keywords // Expert Systems with Applications. – 2023. – Vol. 225. – P. 120064.

41 Gromov V.A., Dang Q.N. Semantic and sentiment trajectories of literary masterpieces // Chaos Solitons Fractals. – 2023. – Vol. 175. – P. 113934.

42 Krivochen D.G. The search for Minimal Search: A graph-theoretic approach // Biolinguistics. – 2023. – Vol. 17. – P. e9793.

43 Bradley E., Kantz H. Nonlinear time-series analysis revisited // Chaos. – 2015. – Vol. 25, Issue 9. – P. 097610.

44 Yao T.L., Liu H.F., Xu J.L. et al. Lyapunov-exponent spectrum from noisy time series // International J of Bifurcation and Chaos. – 2013. – Vol. 23, Issue 06. – P. 1350103.

45 Sahbani M., Das S., Green J.R. Classical Fisher information for differentiable dynamical systems // Chaos. – 2023. – Vol. 33, Issue 10. – P. 1-13.

46 Papana A., Kugiumtzis D. Evaluation of mutual information estimators for time series // International Journal of Bifurcation and Chaos. – 2009. – Vol. 19, Issue 12. – P. 4197-4215.

47 Marwan N., Romano M.C., Thiel M. et al. Recurrence plots for the analysis of complex systems // Physics reports. – 2007. – Vol. 438, Issue 5-6. – P. 237-329.

48 Gao Z.K., Li S., Dang W.D. et al. Wavelet multiresolution complex network for analyzing multivariate nonlinear time series // International Journal of Bifurcation and Chaos. – 2017. – Vol. 27, Issue 08. – P. 1750123.

49 Gromov V.A., Shulga A.N. Chaotic time series prediction with employment of ant colony optimization // Expert Systems with Applications. – 2012. –Vol. 39, Issue 9. – P. 8474-8478.

50 Gromov V.A., Baranov P.S. Prediction after a Horizon of Predictability: Nonpredictable Points and Partial Multi-tep Prediction for Chaotic Time Series // https://doi.org/10.1155/2023/6689371. 10.10.2025.

51 Gromov V.A., Borisenko E.A. Predictive clustering on non-successive observations for multi-step ahead chaotic time series prediction // Neural Computing and Applications. – 2015. – Vol. 26. – P. 1827-1838.

52 Gromov V.A., Konev A.S. Precocious identifica tion of popular topics on Twitter with the employment of predictive clustering // Neural Computing and Applications. – 2016. – Vol. 28, Issue 11. – P. 3317-3322.

53 Gromov V.A., Beschastnov Y.N., Tomashchuk K.K. Generalized relational tensors for chaotic time series // Peer J Computer Science. – 2023. – Vol. 9. – P. e1254.

54 Gromov V.A., Tomashchuk K.K., Rukavishnikov A. Multi-Step-Ahead Prediction of Chaotic Time Series: Self-Healing Algorithm for Restoring Values at Non-Predictable Points // Qubahan Academic Journal. – 2024. – Vol. 4, Issue 3. – P. 763-781.

55 Xie W.J., Han R.Q., Zhou W.X. Time series classification based on triadic time series motifs // International Journal of Modern Physics B. – 2019. – Vol. 33, Issue 21. – P. 1950237.

56 Yerbolova A.S., Tomashchuk K.K., Kogan A.S. et al. Relative Chaoticity of Natural Languages // https://doi.org/10.1155/cplx/5519690. 10.03.2026.

57 Tian Z. Preliminary research of chaotic characteristics and prediction of short-term wind speed time series // International Journal of Bifurcation and Chaos. – 2020. – Vol. 30, Issue 12. – P. 2050176.

58 Bandt C., Pompe B. Permutation entropy: a natural complexity measure for time series // Physical review letters. – 2002. – Vol. 88, Issue 17. – P. 174102.

59 Martin M.T., Plastino A., Rosso O.A. Generalized statistical complexity measures: Geometrical and analytical properties // Physica A: Statistical Mechanics and its Applications. – 2006. – Vol. 369, Issue 2. – P. 439-462.

60 Amigó J.M., Keller K., Unakafova V. On entropy, entropy-like quantities, and applications // In book: Frontiers in Entropy Across the Disciplines. – Singapore, 2022. – P. 197-231.

61 Roy O., Campbell-Cousins A., Carrasco J.S.F. et al. Graph Permutation Entropy: Extensions to the Continuous Case, A step towards Ordinal Deep Learning, and More // https://arxiv.org/abs/2407.07524. 10.10.2025.

62 Pestov V. Intrinsic dimension of a dataset: what properties does one expect? // Procced. internat. joint conf. on Neural Networks. – Orlando, 2007. – P. 2959-2964.

63 Gromov M. Metric Structures for Riemannian and Non-Riemannian Spaces. – Ed. 1st. – Boston, 2007. – 586 p.

64 Малинецкий Г.Г., Потапов А.Б. Современные проблемы нелинейной динамики. – М., 2000. – 336 с.

65 Kantz H, Schreiber T. Nonlinear Time Series Analysis. – Ed. 2nd. – Cambridge: Cambridge University Press, 2003. – 369 p.

66 Brito M.R., Quiroz A.J., Yukich J.E. Intrinsic dimension identification via graph-theoretic methods // Journal of Multivariate Analysis. – 2013. – Vol. 116. – P. 263-277.

67 Adams H., Aminian M., Farnell E. et al. A Fractal Dimension for Measures via Persistent Homology // Topological Data Analysis: procced. Abel symp. – Cham: Springer, 2020. – P. 1-31.

68 Schweinhart B. Fractal dimension and the persistent homology of random geometric complexes // Advances in Mathematics. – 2020. – Vol. 372, Issue 1. – P. 107291.

69 Sole R.V. The small world of human language // Proceedings of the Royal Society B. – 2001. – Vol. 268, Issue 1482. – P. 2261-2265.

70 Costa J.A., Girotra A., Hero AO. Estimating Local Intrinsic Dimension with k-Nearest Neighbor Graphs // Procced. 13th Workshop on Statistical Signal Processing. – Bordeaux: IEEE, 2005. – P. 417-422.

71 Farahmand A.M. et al. Manifold-Adaptive Dimension Estimation // Proceed. of the 24th International Conference on Machine Learning, Corvalis: Association for Computing Machinery. – Oregon, 2007. – P. 265-272.

72 De Santis E., De Santis G., Rizzi A. Multifractal Characterization of Texts for Pattern Recognition: on the Complexity of Morphological Structures in Modern and Ancient Languages // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – Vol. 45, Issue 8. – P. 10143-10160.

73 Beuria J. Persistent homology of collider observations: When (w) hole matters // Physics Letters B. – 2023. – Vol. 846. – P. 138188.

74 Horak D. et al. Persistent homology of complex networks // https://iopscience.iop.org/article/10.1088/1742-5468/2009/03/P03034. 10.10.2025.

75 Myers A. et al. Persistent homology of complex networks for dynamic state detection // Physical Review E. – 2019. – Vol. 100, Issue 2. – P. 022314.

75 Xu X. et al. Finding cosmic voids and filament loops using topological data analysis // Astronomy and Computing. – 2019. – Vol. 27. – P. 34-52.

77 Bermejo R. et al. Topological bias: How haloes trace structural patterns in the cosmic web // Monthly Notices of the Royal Astronomical Society. – 2024. – Vol. 529, Issue 4. – P. 4325-4353.

78 Skaf Y. et al. Topological data analysis in biomedicine: A review // Journal of Biomedical Informatics. – 2022. – Vol. 130. – P. 104082.

79 Meng Z. et al. Weighted persistent homology for biomolecular data analysis // Scientific reports. – 2020. – Vol. 10, Issue 1. – P. 2079.

80 Dey T.K., Mandal S. Protein classification with improved topological data analysis // Procced. 18th internat. Workshop on Algorithms in Bioinformatics (WABI 2018). – Helsinki, 2018. – P. 6:1-6:13.

81 Corcoran P. et al. Topological data analysis for geographical information science using persistent homology // International Journal of Geographical Information Science. – 2023. – Vol. 37, Issue 3. – P. 712-745.

82 Caputi L. et al. Promises and pitfalls of topological data analysis for brain connectivity analysis // NeuroImage. – 2021. – Vol. 238. – P. 118245.

83 Yoo J. et al. Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages // Journal of neuroscience methods. – 2016. – Vol. 267. – P. 1-13.

84 Zhu X. Persistent homology: An introduction and a new text representation for natural language processing // IJCAI. – 2013. – Vol. 2013. – P. 1953-1959.

85 Elyasi N. et al. An introduction to a new text classification and visualization for natural language processing using topological data analysis // https://arxiv.org/abs/1906.01726. 10.10.2025.

86 Savle K. et al. Topological data analysis for discourse semantics? // Proceed. of the 13th internat. conf. on computational semantics-student papers. – Gothenburg, 2019. – P. 34-43.

87 Tymochko S.J. et al. Con connections: Detecting fraud from abstracts using topological data analysi // Procced. 20th IEEE internat. conf. on Machine Learning and Applications (ICMLA). – Pasadena, 2021. – P. 403-408.

88 Rathore A. Topological Data Analysis and Visualization for Interpretable Machine Learning: dis. … doc. of philos. in comp. – Salt Lake City, 2023. – 155 p.

89 Daya A.A., Salahuddin M.A., Limam N. et al. A graph-based machine learning approach for bot detection // Procxced. IFIP/IEEE sympos. on integrated network and service management (IM). – Washington, 2019. – P. 144-152.

90 Mesnards N. et al. Detecting bots and assessing their impact in social networks // Operations Research. – 2021. – Vol. 70, Issue 1. – P. 1-22.

91 Li S., Zhao C., Li Q. et al. Botfinder: a novel framework for social bots detection in online social networks based on graph embedding and community detection // World Wide Web. – 2022. – Vol. 26. – P. 1-17.

92 Grover A., Leskovec J. node2vec: scalable feature learning for networks // Proceed. of the 22nd ACM SIGKDD internat. conf. on knowledge discovery and data mining (KDD'16). – NY., 2016. – P. 855-864.

93 Pham P. et al. Bot2vec: a general approach of intra-community oriented representation learning for bot detection in different types of social networks // Information Systems. – 2021. – Vol. 103. – P. 101771.

94 Feng S., Wan H., Wang N. et al. Botrgcn: twitter bot detection with relational graph convolutional networks // Proceed. of the 2021 IEEE/ACM internat. conf. on advances in social networks analysis and mining (ASONAM '21). –NY., 2021. – P. 236-239.

95Fu C., Shi S., Zhang Y. et al. Squeezegcn: adaptive neighborhood aggregation with squeeze module for twitter bot detection based on gcn // Electronics. – 2023. – Vol. 13. – P. 56.

96 Liu F. et al. Segcn: a subgraph encoding based graph convolutional network model for social bot detection // Scientific Rep. – 2024. – Vol. 14, Issue 1. – P. 4122.

97 Latah M. Detection of malicious social bots: a survey and a refined taxonomy // Expert Systems with Applications. – 2024. – Vol. 151. – P. 113383.

98 Garcia-Silva A. et al. An empirical study on pretrained embeddings and language models for bot detection // Proceed. of the 4th workshop on representation learning for NLP (RepL4NLP-2019). – Florence, 2019. – P. 148-155.

99 Garcia-Silva A. et al. Understanding transformers for bot detection in twitter // https://arxiv.org/abs/2104.06182. 10.10.2025.

100 Kang A.R., Kim H.K., Woo J. Chatting pattern-based game bot detection: do they talk like us? // KSII Transactions on Internet & Information Systems. – 2012. – Vol. 6, Issue 11. – P. 4-7.

101 Cardaioli M. et al. It's a matter of style: detecting social bots through writing style consistency // Procced. internat. conf. on computer communications and networks (ICCCN). – Piscataway, 2021. – P. 1-9.

102 Chakraborty M., Das S., Mamidi R. Detection of fake users in twitter using network representation and nlp // Procced. 14th internat. conf. on communication systems & NETworkS (COMSNETS). – Piscataway: IEEE, 2022. – P. 754-758.

103 Gromov V., Dang Q.N. Spot the bot: distinguishing human-written and bot-generated texts using clustering and information theory techniques // Procced. Internat. conf. on pattern recognition and machine intelligence. – Cham, 2023. – P. 20-27.

104 Monica C., Nagarathna N. Detection of fake tweets using sentiment analysis // SN Computer Science. – 2020. – Vol. 1, Issue 1. – P. 89.

105 Uymaz H.A., Metin S.K. Vector based sentiment and emotion analysis from text: a survey // Engineering Applications of Artificial Intelligence. – 2022. – Vol. 113. – P. 104922.

106 Heidari M. et al. An empirical study of machine learning algorithms for social media bot detection // Procced. IEEE internat. IOT, electronics and mechatronics conf. (IEMTRONICS). – Piscataway, 2021. – P. 1-5.

107 Liao W. et al. Multi-level graph neural network for text sentiment analysis // Computers & Electrical Engineering. – 2021. – Vol. 92. – P. 107096.

108 Lin S.-Y., Kung Y.-C., Leu F-Y. Predictive intelligence in harmful news identification by bert-based ensemble learning model with text sentiment analysis // Information Processing & Management. – 2022. – Vol. 59, Issue 2. – P. 102872.

109 Galgoczy M.C., Phatak A., Vinson D. et al. (Re) shaping online narratives: when bots promote the message of president trump during his first impeachment // Peer J Computer Science. – 2022. – Vol. 8. – P. e947.

110 Lira D.B. et al. Combining clustering and classification algo- rithms for automatic bot detection: a case study on posts about Covid-19 // Procced. 17th Brazilian sympos. on information systems. – NY., 2021. – P. 1-7.

111 Mu Y., Aletras N. Identifying twitter users who repost unreliable news sources with linguistic information // Peer J Comp. Science. – 2020. – Vol. 6. – P. e325.

112 Ren Y., Ji D. Neural networks for deceptive opinion spam detection: an empirical study // Information Sciences. – 2017. – Vol. 385. – P. 213-224.

113 Altmann E.G., Cristadoro G., Esposti M.D. On the origin of long-range correlations in texts // Proceedings of the National Academy of Sciences of the United States of America. – 2012. – Vol. 109, Issue 29. – P. 11582-11587.

114 Altmann E.G., Gerlach M. Statistical laws in linguistics // In book: Creativity and Universality in Language. – Cham: Springer, 2016. – P. 7-26.

115 Brown P.F. et al. An estimate of an upper bound for the entropy of English // Computational Linguistics. – 1992. – Vol. 18, Issue 1. – P. 31-40.

116 Browse by language family – ethnologue // https://www.ethnologue.com. 10.07.2024.

117 Gromov V.A., Zvorykina E.I., Beschastnov Y.N., Sohrabi M. Data-Driven Approach for Identifying State of Hemodialysis Fistulas: Entropy-Complexity and Formal Concept Analysis // Proceed. of the 11th internat. conf. on Analysis of Images, Social Networks and Texts (AIST 2023). – Yerevan, 2023. – P. 1-14.

118 Pruessner G. Self-organised criticality: Theory, models, and characterisation. – NY.: Cambridge University Press, 2012. – 494 p.

119 Clauset A., Shalizi C.R., Newman M.E.J. Power- Law distributions in empirical data // SIAM Review. – 2009. – Vol. 51, Issue 4. – P. 661-703.

120 Golub G., Kahan W. Calculating the singular values and pseudo-inverse of a matrix // Journal of the Society for Industrial and Applied Mathematics. – 1965. – Vol. 2, Issue 2. – P. 205-224.

121 Bellegarda J.R. Globally optimal training of unit boundaries in unit selection text-to-speech synthesis // IEEE transactions on audio, speech, and language processing. – 2007. – Vol. 15, issue 3. – P. 957-965.

122 Rosso O.A., Larrondo H., Martin M.T. et al. Distinguishing noise from chaos // Physical Review Letters. – 2007. – Vol. 99, Issue 15. – P. 154102.

123 Barbosa K., Frery A.C., Cavalcanti G.D. Analysis of signals from air conditioner compressors with ordinal patterns and machine learning // Journal of Low Frequency Noise, Vibration and Active Control. – 2024. – Vol. 44, Issue 1. – P. 21-38.

124 Boaretto B.R., Macau E.E., Masoller C. Characterizing the spike timing of a chaotic laser by using ordinal analysis and machine learning // Chaos. – 2024. – Vol. 34, Issue 4. – P. 043108.

125 Martinuzzi F., Mahecha M.D., Camps-Valls G. et al. Learning extreme vegetation response to climate drivers with recurrent neural networks // Nonlinear Processes in Geophysics. – 2024. – Vol. 31, Issue 4. – P. 535-557.

126 Čech E. Topological spaces. – Ed. 1st. – Prague, 1966. – 894 p.

127 Kalman D. A singularly valuable decomposition: the SVD of a matrix // The college mathematics journal. – 1996. – Vol. 27. – P. 2-23.

128 Mikolov T. Efficient estimation of word representations in vector space // https://arxiv.org/abs/1301.3781. 10.10.2025.

129 Steele J.M., Shepp L.A., Eddy W.F. On the Number of Leaves of a Euclidean Minimal Spanning Tree // J. of Applied Probability. – 1987. – Vol. 24. – P. 809-826.

130 Edelsbrunner H., Harer J.L. Computational topology: an introduction. – NY., 2022. – 241 p.

131 Čufar M., Virk Z. Fast computation of persistent homology representatives with involuted persistent homology // https://arxiv.org/pdf/2105.03629. 10.102025.

132 Obayashi I. Stable volumes for persistent homology // Journal of Applied and Computational Topology. – 2023. – Vol. 7, Issue 4. – P. 671-706.

133 MacQueen J. Some methods for classification and analysis of multivariate observations // Proceed. of the 5th Berkeley sympos. on mathematical statistics and probability. – Oakland (CA), 1967. – P. 281-297.

134 Gromov V.A., Kogan A.S. Spot the bot: coarse-grained partition of semantic paths for bots and humans // Procced. internat. conf. on pattern recognition and machine intelligence. – Cham: Springer, 2023. – P. 348-355.

135 Groetsch C.W., Groetsch C. Inverse problems in the mathematical sciences. – NY., 1993. – 84 p.

136 Mikolov T. Model Architectures // Efficient estimation of word representations in vector space // https://arxiv.org/abs/1301.3781. 10.10.2025.

137 Bezdek J.C., Ehrlich R., Full W. Fcm: the fuzzy c-means clustering algorithm // Computers & Geosciences. – 1984. – Vol. 10, Issue 2-3. – P. 191-203.

138 Wishart D. Numerical Classification Method for deriving Natural Classes // Nature. – 1969. – Vol. 221. – P. 97-98.

139 Novak V., Perfilieva I., Mockor J. Mathematical principles of fuzzy logic. – NY., 2012. – 320 p.

140 Xiong H, Li Z. Clustering validation measures // In book: Data clustering. – London, 2018. – P. 571-606.

141 Wals Online // https://wals.info/. 01.07.2024.

142 Nichols J. Head-marking and dependent-marking grammar // Language. – 1986. – Vol. 62, Issue 1. – P. 56-119.

143 Blake B.J. Case. – Cambridge, 2001. – 248 p.

144 Haspelmath M. et al. Understanding morphology. – NY., 2010. – 224 p.

145 Furnkranz J. Decision tree // In book: Encyclopedia of Machine Learning and Data Mining. – NY., 2017. – P. 330-335.

146 Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. – 1987. – Vol. 20. – P. 53-65.

147 Davies D.L., Bouldin D.W. A cluster separation measure // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1979. – Vol. PAMI-1, Issue 2. – P. 224-227.

148 Calinski T., Harabasz J. A dendrite method for cluster analysis // Communication in Statistics. – 1974. – Vol. 3, Issue 1. – P. 1-27.

149 Ester M., Kriegel H.P., Sander J. et al. A density-based algorithm for discovering clusters in large spatial databases with noise // KDD'96: proceed. of the 2nd internat. conf. on Knowledge Discovery and Data Mining. – Portland, 1996. – P. 226-231.

150 Yerbolova A.S., Kurmashev I.G. Identification of Intrinsic Dimensionality Patterns in Semantic Spaces of Natural Languages Using Graph Algorithms // Eastern-European Journal of Enterprise Technologies. – 2026. – Vol. 1/2, Issue 138. – P. 68-76.

151 Arneodo A/, Bacry E/, Muzy J-F. Random cascades on wavelet dyadic trees // Journal of Mathematical Physics. – 1998. – Vol. 39. – P. 4142-4164.

152 Kuznetsov S.O., Gromov V.A., Borodin N.S. et al. Formal concept analysis for evaluating intrinsic dimension of a natural language // Procced. 10th internat. conf. on Pattern Recognition and Machine Intelligence. – Kolkata, 2023. – P. 331-339.

153 Gromov V.A., Dang Q.N., Yerbolova A.S. et al. ALanguage and Its Holes: The First Order Homologies of the Large-scale Geometrical Structure of a Natural Language // Complexity. – 2025. – Issue 1. – P. 9659172.

154 Смирнов Н.В. Приближение законов распределения случайных величин по эмпирическим данным // УМН. – 1944. – №10. – С. 179-206.

155 Громов В.А., Бородин Н.С., Ерболова А.С. и др. Поймай бота: крупномасштабная структура естественного языка // Проблемы цифровой реальности. Проектирование будущего: тр. 7-й междунар. конф. – М.: ИПМ им. М.В. Келдыша, 2024. – С. 281-312.

156 Ерболова А.С., Громов В.А., Аканова А.С. и др. Семантические методы выявления текстов, сгенерированных системами искусственного интеллекта // Вестник Алматинского университета энергетики и связи. – 2026. – Т. 5, №72. – С. 309-318.

ПРИЛОЖЕНИЕ А

Акты внедрения

УТВЕРЖДАЮ
ВрИО заместителя начальника Национального университета обороны Республики Казахстан по научной работе – начальник Военного научно-исследовательского центра
д.в.н., профессор, генерал-майор Ж. Ахметов
«11» 02 2025 г.

АКТ
о внедрении научно-исследовательских результатов диссертационной работы
Ерболовой Асель Серикановны
на тему: «Исследование структур естественного языка в задаче идентификации ботов»

Мы, нижеподписавшиеся, представители Национального университета обороны Республики Казахстан составили настоящий акт о том, что результаты диссертационного исследования на тему: «Исследование структур естественного языка в задаче идентификации ботов», полученные докторантом PhD по специальности 8D06101 – «Информатика, вычислительная техника и управление» Ерболовой А.С., а также разработанная в рамках диссертационной работы классификационная модель для идентификации текстов, написанных естественным языком и сгенерированных ботами (авторы: Ерболова А.С., Курмашев И.Г.), основанная на анализе самоорганизованно-критичных характеристик естественного языка, хаотичности семантических траекторий, внутренних размерностей языковых структур и топологическом анализе семантического пространства, внедрены в деятельность лаборатории исследования, проектирования и разработки программного обеспечения управления информационных технологий Военного научно-исследовательского центра Национального университета обороны Республики Казахстан.

Полученные научные результаты предназначены для решения задач идентификации ботов и автоматически сгенерированных текстов в цифровом информационном пространстве, а также для анализа текстового контента с целью выявления автоматизированной и манипулятивной информационной активности на основе структурных, статистических, геометрических и динамических характеристик естественного языка.

В процессе внедрения установлено, что разработанные в диссертационной работе модели анализа семантических траекторий, фрактальных языковых структур и топологических особенностей семантических пространств позволяют эффективно различать тексты, созданные человеком, и тексты, сгенерированные интеллектуальными

системами, за счёт выявления различий в статистических закономерностях, внутренней размерности и хаотических свойствах языковых данных.

Использование результатов диссертационного исследования способствует повышению уровня информационной безопасности, обеспечивает поддержку принятия решений при мониторинге цифрового контента, а также позволяет повысить эффективность защиты информационных систем и цифровых платформ от автоматизированных информационных воздействий.

Разработанные в рамках диссертационного исследования классификационные модели и программные решения обладают высокой точностью идентификации текстов благодаря применению современных подходов анализа данных, нелинейной динамики, машинного обучения и технологий искусственного интеллекта, что обеспечивает их практическую значимость в процессе проверки научных работ на предмет использования заимствованного материала и технологий искусственного интеллекта без ссылок на них диссертационными советами Национального университета обороны Республики Казахстан.

Председатель комиссии

к.в.н. асс.профессор , полковник запаса **А. Мукишов**

Члены комиссии:

PhD, полковник **Н. Асилов**

PhD, полковник **М. Джакипбеков**

PhD, полковник запаса **Е. Бекишов**

УТВЕРЖДАЮ

Заведующий лабораторией анализа семантики центра языковых и семантических технологий департамента анализа данных и искусственного интеллекта Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики», д.ф.-м.н., профессор

_____ В.А. Громов

«20» февраля 2026 г.

М.П.

АКТ

о внедрении научно-исследовательских результатов диссертационной работы Ерболовой Асель Серикановны

По теме «Исследование структур естественного языка в задаче идентификации ботов»

Настоящий акт подтверждает, что результаты диссертационного исследования по теме «Исследование структур естественного языка в задаче идентификации ботов», выполненного докторантом PhD по специальности 8D06101 – «Информатика, вычислительная техника и управление» Ерболовой А.С., внедрены в научно-исследовательскую деятельность лаборатории анализа семантики центра языковых и семантических технологий департамента анализа данных и искусственного интеллекта факультета компьютерных наук Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики».

В рамках диссертационной работы разработаны модели анализа естественного языка как самоорганизованно-критичной системы, методы исследования семантических траекторий текстов как многомерных хаотических временных рядов, алгоритмы оценки внутренней размерности языковых фрактальных структур и топологического анализа семантических пространств, а также классификационная модель для идентификации текстов, написанных людьми и сгенерированных автоматизированными системами.

Полученные результаты используются в научных исследованиях лаборатории, связанных с анализом нелинейной динамики языковых процессов, изучением внутренней геометрии семантических пространств и разработкой новых подходов к построению языковых моделей. Разработанные методы способствуют развитию направлений, связанных с обучением на многообразиях, анализом больших языковых моделей и созданием семантических технологий.

В ходе внедрения установлено, что предложенные модели и алгоритмы обеспечивают эффективное выявление структурных и динамических различий между текстами, созданными человеком и генеративными моделями, что имеет значение для задач анализа цифрового контента, детектирования ботов и решения обратных задач обработки естественного языка.

Использование результатов диссертационного исследования расширяет инструментарий лаборатории в области топологического анализа данных, теории сложных систем и анализа хаотических временных рядов, а также способствует развитию совместных научных проектов и публикационной активности в международных рецензируемых изданиях.

Младший научный сотрудник
лаборатории анализа семантики  К.К. Томащук

УТВЕРЖДАЮ
Проректор
Научной работе и инновациям
Карагандинского университета
Казпотребсоюза,
д.э.н., профессор
_____ М.Р. Сихимбаев
_____ 2026 г.
М.П.

АКТ ВНЕДРЕНИЯ (ИСПОЛЬЗОВАНИЯ)
результатов научно-исследовательской работы, выполненной в рамках диссертационной работы PhD докторанта
Северо-Казахстанского университета имени М. Козыбаева
Ерболовой Асель Серикановны
в учебный процесс

Мы, нижеподписавшиеся, Тажбаев Н.М. – директор центра управления цифровой трансформации бизнес- процессов, к.э.н., доцент

Тен Т.Л. – зав. кафедрой "Цифровой инженерии и IT аналитики", д.т.н., профессор.

Курмашев И.Г. – кандидат технических наук, ассоциированный профессор Северо-Казахстанского университета имени М. Козыбаева, научный консультант внедряемых результатов, составили настоящий АКТ ВНЕДРЕНИЯ (ИСПОЛЬЗОВАНИЯ) результатов научно-исследовательской работы, выполненной докторантом Ерболовой А.С. в рамках диссертационной работы на тему: «Исследование структур естественного языка в задаче идентификации ботов».

Основные результаты работы: разработана модель естественного языка как самоорганизованно-критичной системы и выполнен сравнительный анализ языков на основе степенных распределений и статистических критериев согласия; разработана модель анализа семантических траекторий текстов как многомерных хаотических временных рядов с использованием мер энтропии и сложности; разработана модель анализа языковых фрактальных структур, включающая оценку внутренней размерности и топологических особенностей семантического пространства; разработана классификационная модель для идентификации текстов, написанных людьми и сгенерированных ботами; разработан и реализован программный модуль анализа текстовых данных, предназначенный для интеграции в учебные дисциплины по анализу данных и искусственному интеллекту.

Указанная работа внедрена (использована) в учебный процесс в 2025 году в рамках образовательной программы 6В06101 – «Информационные системы», 6В06102 "Вычислительная техника и программное обеспечение", 6В06103 "IT аналитика", 6В06104 "Цифровой дизайн и мультимедиа" в следующих лекционных и практических курсах:

– «Модели и методы управления IT-проектами»;
– «Представление базы знаний ИС»;
-«Разработка программных приложений для бизнес анализа»;
– «Технология разработки программных приложений»;
– «Системы искусственного интеллекта».

Наименование объекта и предмета внедрения (использования) результатов научно-исследовательской работы докторанта:

объект внедрения – текстовые корпуса естественного языка и цифровой контент, содержащий тексты, созданные человеком и автоматизированными системами (ботами).

предмет внедрения – методы анализа крупномасштабной структуры естественного языка, алгоритмы оценки внутренней размерности и топологической структуры семантических пространств, а также классификационные модели идентификации ботов.

Эффект от внедрения (использования) результатов: научно-технические результаты по проблеме анализа структур естественного языка, математического моделирования семантических пространств и идентификации автоматически сгенерированного контента имеют важное значение при обучении студентов методам анализа данных, машинного обучения и искусственного интеллекта. Использование полученных результатов способствует формированию у обучающихся практических навыков построения классификационных моделей, анализа текстовых данных и разработки интеллектуальных информационных систем. Применение результатов диссертационной работы в учебном процессе позволяет повысить качество подготовки бакалавров по образовательным программам 6B06101 – «Информационные системы», 6B06102 "Вычислительная техника и программное обеспечение", 6B06103 "IT аналитика", 6B06104 "Цифровой дизайн и мультимедиа".

Директор центра управления
цифровой трансформации
бизнес- процессов,
к.э.н., доцент Тажбаев Н.М.

КАДРЛЫҚ ҚЫЗМЕТ
КАДРОВАЯ СЛУЖБА

Зав. кафедрой "Цифровой инженерии и IT аналитики",
д.т.н., профессор Тен Т.Л.

ПРИЛОЖЕНИЯ А Б

Свидетельство об авторском праве

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

СВИДЕТЕЛЬСТВО

О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ

№ 66441 от «19» января 2026 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):
ЕРБОЛОВА АСЕЛЬ СЕРИКАНОВНА

Вид объекта авторского права: **программа для ЭВМ**

Название объекта: **Анализ оценки внутренней размерности семантических пространств естественных языков**

Дата создания объекта: **16.01.2026**

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

Подписано ЭЦП

С. Ахметов

ПРИЛОЖЕНИЕ В

Корпус языковых данных и результаты статистического анализа

Таблица В.1 – Набор данных (Datasets). Языковые семьи и объем текстов 52 языков

Языки	Языковая семья /группа	Население	Кол.текстов	Библиотека
1	2	3	4	5
Амхарский	Семитский / Эфиопский	32 млн.	1320	SparkNLP
Арабский	Семитский /Центрально-семитский	310 млн.	5000	Farasa
Атикемек	Алгонкинский	6000	7000	manual preprocessing
Баскский	Изолированный язык	750000	10052	SparkNLP
Белорусский	Индоевропейский / Восточнославянский	6.7 млн.	10311	manual preprocessing
Бенгальский	Индоевропейский / Бенгальский Аассамский	230 млн.	29745	manual preprocessing
Болгарский	Индоевропейский /Южнославянский	9 млн.	103853	spacy
Чеченский	Нахско-Дагестанский /Вайнахский	1.5 млн.	20254	manual preprocessing
Китайский	Китайско-тибетский	918 млн.	8358	spacy
Коптский	Афроазиатский, Египетский	-	2000	manual preprocessing
Чешский	Индоевропейский, Славянский, Западнославянский	10.7 млн.	10000	spacy
Дхолуа	Нило-сахарский/ Нилотский	4.2 млн.	7000	manual preprocessing
Голландский	Индоевропейский / Германский	24 млн.	15861	spacy
Английский	Индоевропейский / Германский	375 млн.	11046	spacy
Французский	Индоевропейский / Италийский	77 млн.	1568	spacy
Финский	Уральский / Финно-угорский	5.5 млн.	4064	uralicNLP
Немецкий	Индоевропейский / Германский	76 млн.	12503	spacy
Хинди	Индоевропейский / Индо-иранский	341 млн.	1042	StanfordNLP
Исландский	Индоевропейский / Скандинавский	314000	20000	Huspacy
Индонезийский	Австронезийский / Малайский	43 млн.	3264	Nefnir
Японский	Японский-Рюкю	130 млн.	14498	PySastrawi
Кабильский	Афроазиатский / Северный берберский	_	6377	MeCab
Казахский	Алтайский / Кипчакский	18 млн.	2090	manual preprocessing
Латинский	Индоевропейский / Латинско-фалисканский	_	7820	KazNLP
Латышский	Индоевропейский / Балтийский	1.5 млн.	152983	spacy
Малаялам	Дравидийский / Тамильский-канадский	38 млн.	3463	LibIndic
Навахо	Атабаскский / Южный атбаскский	170000	500	manual preprocessing
Норвежский	Индоевропейский / Германский	5 млн.	4125	spacy
Оромо	Афроазиатский / Кушитский	35 млн.	1000	manual preprocessing
Персидский	Индоевропейский / Иранский	70 млн.	1334	Hazm
Польский	Индоевропейский / Славянский / Западнославянский	45 млн.	10000	spacy
Панджаби	Индоевропейский / Индо-иранский	125 млн.	2884	manual preprocessing
Кечуа	Кечуанский	8-10 млн.	1000	manual preprocessing
Румынский	Индоевропейский / Италийский	3-5 млн.	2380	spacy
Русский	Индоевропейский / Восточнославянский	260 млн.	6429	natasha
Сербский	Индоевропейский / Южнославянский	12 млн.	5615	Stanza

Продолжение таблицы В.1

1	2	3	4	5
Осетинский	Индоевропейский / Индо-иранский	50000	1201	manual preprocessing
Сингальский	Индоевропейский / Индоарийский	16 млн.	51143	spacy
Испанский	Индоевропейский / Романский	460 млн.	8505	spacy
Суахили	Нигеро-конголезский / Бенуэ-конголезский	50-100 млн.	1514	manual preprocessing
Шведский	Индоевропейский / Скандинавский	10 млн.	4741	spacy
Табасаранский	Северо-восточный кавказский / Лезгинский	125000.	2337	manual preprocessing
Тагальский	Австронезийский / Центральнофилиппинский	28 млн.	1000	spacy
Татарский	Алтайский / Кипчакский	7 млн.	5071	manual preprocessing
Тайский	Тай-кадайский / Тайский	60 млн.	5881	PyThaiNLP
Турецкий	Алтайский / Юго-западный (огузский)	70-80 млн.	1308	Zeyrek
Тувинский	Алтайский / Саянский	200000	5336	manual preprocessing
Удмуртский	Уральский / Пермский	300000	1731	uniparser-udmurt
Украинский	Индоевропейский / Восточнославянский	33 млн.	10052	spacy
Узбекский	Алтайский / Тюркский / Карлукско-хорезмский	32 млн.	10534	manual preprocessing
Вьетнамский	Австроазиатский / Вьетский	90 млн.	1070	pyvi
Эсперанто	Искусственный язык	-	1175	esperanto-analyzer

Таблица В.2 – Лингвистические характеристики 52 языков

Языки	Порядок слов	Выравнивание	Тип маркировки	Морфологическая сложность
1	2	3	4	5
Амхарский	SOV	Inconsistent	Accusative	Synthetic
Арабский	SVO	Inconsistent	Accusative	Synthetic
Атикемек	SVO	Dependent	Accusative	Synthetic
Баскский	SOV	Inconsistent	Ergative	Synthetic
Белорусский	SVO	Dependent	Accusative	Synthetic
Бенгальский	SOV	Dependent	Accusative	Synthetic
Болгарский	SVO	Dependent	Accusative	Synthetic
Чеченский	SOV	Dependent	Ergative	Synthetic
Китайский	SVO	Dependent	Neutral	Isolating
Коптский	SVO	Dependent	Accusative	Analytic
Чешский	SVO	Dependent	Accusative	Synthetic
Дхолуа	VSO	Dependent	Accusative	Analytic
Голландский	SOV	Inconsistent	Accusative	Analytic
Английский	SVO	Dependent	Accusative	Analytic
Французский	SVO	Inconsistent	Accusative	Synthetic
Финский	SVO	Inconsistent	Accusative	Synthetic
Немецкий	SVO	Dependent	Accusative	Analytic
Хинди	SOV	Inconsistent	Accusative	Analytic
Исландский	SVO	Dependent	Accusative	Synthetic
Индонезийский	SVO	Zero	Accusative	Synthetic
Японский	SOV	Dependent	Neutral	Synthetic
Кабильский	SVO	Dependent	Neutral	Synthetic
Казахский	SOV	Dependent	Accusative	Synthetic
Латинский	SOV	Dependent	Accusative	Synthetic
Латышский	SVO	Dependent	Accusative	Synthetic
Малаялам	SOV	Dependent	Ergative	Synthetic
Навахо	SOV	Dependent	Accusative	Synthetic
Норвежский	SVO	Dependent	Accusative	Analytic

Продолжение таблицы В.2

1	2	3	4	5
Оромо	SOV	Dependent	Accusative	Synthetic
Персидский	SOV	Inconsistent	Accusative	Synthetic
Польский	SVO	Dependent	Accusative	Synthetic
Панджаби	SOV	Dependent	Accusative	Synthetic
Кечуа	SOV	Inconsistent	Accusative	Synthetic
Румынский	SVO	Dependent	Accusative	Synthetic
Русский	SVO	Dependent	Accusative	Synthetic
Сербский	SVO	Dependent	Accusative	Synthetic
Осетинский	SOV	Dependent	Accusative	Synthetic
Сингальский	SOV	Dependent	Accusative	Synthetic
Испанский	SVO	Inconsistent	Accusative	Synthetic
Суахили	SVO	Inconsistent	Accusative	Synthetic
Шведский	SVO	Dependent	Accusative	Analytic
Табасаранский	SOV	Dependent	Ergative	Synthetic
Тагальский	VSO	Inconsistent	Neutral	Analytic
Татарский	SOV	Dependent	Accusative	Synthetic
Тайский	SVO	Inconsistent	Neutral	Analytic
Турецкий	SOV	Inconsistent	Accusative	Synthetic
Тувинский	SOV	Dependent	Accusative	Synthetic
Удмуртский	SOV	Dependent	Accusative	Synthetic
Украинский	SVO	Dependent	Accusative	Synthetic
Узбекский	SOV	Inconsistent	Accusative	Synthetic
Вьетнамский	SVO	Zero	Neutral	Isolating
Эсперанто	Free	RI13	Accusative	Synthetic

Таблица В.3 – Оценка параметров степенного закона и нижние пределы для естественных языков

Языки	Оценки на основе коллапса данных		Оценки на основе критерия Колмогорова-Смирнова	
	x_{min} *(nw)*	T *(dl)*	x_{min} *(nw)*	T *(dl)*
1	2	3	4	5
Амхарский	8020	3,30	8758	3,55
Арабский	201	3,20	200	3,45
Атикемек	51	3,00	50	3,10
Баскский	1001	3,50	1014	3,30
Белорусский	1001	2,00	1178	2,15
Бенгальский	3001	3,00	3040	2,98
Болгарский	99	5,60	101	5,52
Чеченский	201	2,90	201	2,71
Китайский	1001	2,00	1153	2,15
Коптский	101	1.90	103	1.75
Чешский	1002	1.80	1032	1.79
Дхолуа	183	1.80	177	1.75
Голландский	1001	1,60	1046	1,75
Английский	1003	1,50	1661	1,39
Французский	12192	1,50	12522	1,62
Финский	10015	1,80	12218	1,98
Немецкий	10132	1,55	11909	1,70
Хинди	401	1,70	438	1,85
Исландский	101	1,60	144	1,54
Индонезийский	1001	1,40	1505	1,53

[13]RI означает регулярный и преднамеренный.

Продолжение таблицы В.3

1	2	3	4	5
Японский	1001	1,70	1013	1,85
Кабильский	101	2,20	102	2,40
Казахский	1005	3,00	1021	2,81
Латинский	1001	1,40	1051	1,44
Латышский	101	3,00	103	2,98
Малаялам	1006	1,50	999	1,46
Навахо	900	1.65	900	1.50
Норвежский	10006	1,70	10251	1,87
Оромо	190	1.80	186	1.65
Персидский	4002	1,60	4170	1,52
Польский	1055	1,90	1046	1,70
Панджаби	101	3,20	102	3,07
Кечуа	101	3,00	101	2.88
Румынский	801	1,50	688	1,43
Русский	1001	1,55	1330	1,66
Сербский	501	1,75	592	1,91
Осетинский	3037	1,43	3045	1,57
Сингальский	1001	4,90	1141	5,39
Испанский	5003	1,45	5396	1,41
Суахили	1001	4,80	1014	4,93
Шведский	5003	1,60	5622	1,73
Табасаранский	401	3,90	421	3,97
Тагальский	1044	1,50	1128	1,48
Татарский	501	3,29	503	3,08
Тайский	1002	1,45	1253	1,49
Турецкий	20009	3,40	23678	3,64
Тувинский	101	2,00	103	2,20
Удмуртский	1004	1,50	999	1,52
Украинский	10023	1,70	10766	1,86
Узбекский	201	3,00	203	2,72
Вьетнамский	35039	3,00	35000	3,17

Таблица В.4 – Оценка параметров степенного закона и нижние пределы для искусственного языка

Языки	Оценки на основе коллапса данных		Оценки на основе критерия Колмогорова-Смирнова	
	x_{min} *(nw)*	T *(dl)*	x_{min} *(nw)*	T *(dl)*
Эсперанто	-	-	-	-

Таблица В.5 – Результаты тестирования на соответствие требованиям для теста Колмогорова-Смирнова и процент ошибок для двух аналитических методов

Языки	(Прюсснер – КС) / Прюсснер, % ≤ 10%	p-value ≥ 0.1
Амхарский	8	True[14]
Арабский	8	True
Атикемек	3	True
Баскский	6	True
Белорусский	8	True
Бенгальский	1	True
Болгарский	7	True

[14]Значения True, когда p-значение ≥ 0.05, в случае p-значение принимает множество различных значений для 52 языков в диапазоне [0.5 - 1].

Продолжение таблицы В.5

1	2	3
Чеченский	8	True
Китайский	8	True
Коптский	8	True
Чешский	1	True
Дхолуа	9	True
Голландский	3	True
Английский	7	True
Французский	8	True
Финский	10	True
Немецкий	10	True
Хинди	9	True
Исландский	4	True
Индонезийский	9	True
Японский	9	True
Кабильский	9	True
Казахский	6	True
Латинский	3	True
Латышский	1	True
Малаялам	3	True
Навахо	9	True
Норвежский	10	True
Оромо	8	True
Персидский	5	True
Польский	4	True
Панджаби	4	True
Кечуа	4	True
Румынский	5	True
Русский	7	True
Сербский	9	True
Осетинский	10	True
Сингальский	10	True
Испанский	3	True
Суахили	3	True
Шведский	8	True
Табасаранский	2	True
Тагальский	1	True
Татарский	6	True
Тайский	3	True
Турецкий	7	True
Тувинский	10	True
Удмуртский	1	True
Украинский	9	True
Узбекский	9	True
Вьетнамский	6	True

Таблица В.6 – Результаты тестирования на соответствие требованиям для теста м и процент ошибок для двух аналитических методов

Языки	(Прюсснер – КС) / Прюсснер % ≤ 10%	p-value ≥ 0.1
Эсперанто	228%	False (0.0005)

Таблица В.7 – Значения показателей кластеризации и количество кластеров для четырех симуляций в алгоритме K-means

Симуляция	Коэффициент силуэта	Индекс Дэвиса–Боулдина	Индекс Калинского–Харабаша
Симуляция 1	0.65	0.27	315.76
Число кластеров	6	7	14
Симуляция 2	0.77	0.26	772.28
Число кластеров	4	4	14
Симуляция 3	0.76	0.30	1453.51
Число кластеров	3	3	14
Симуляция 4	0.82	0.24	3940.26
Число кластеров	3	14	14

Таблица В.8 – Значения метрик кластеризации и количество кластеров для четырех симуляций в алгоритме Wishart

Симуляция	Коэффициент силуэта	Индекс Дэвиса-Боулдина	Индекс Калиньки-Харабаша
Симуляция 1	0.42	0.97	17.74
Число кластеров	2	13	2
Симуляция 2	0.63	0.75	83.37
Число кластеров	2	17	2
Симуляция 3	0.58	0.49	274.11
Число кластеров	4	18	18
Симуляция 4	0.78	0.40	166
Число кластеров	2	2	2

ПРИЛОЖЕНИЕ Г

Результаты оценки внутренней размерности и тестирования алгоритмов

Таблица Г.1 – Абсолютная процентная ошибка (APE) для оценки Schweinhart внутренней размерности для различных геометрических объектов

α	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 Г.2 – Влияние шума на результаты алгоритма Schweinhart

Noise type	Noise params	$\boldsymbol{min\,\hat{d}_{Schw}}$	$\boldsymbol{max\,\hat{d}_{Schw}}$	$\boldsymbol{\alpha}$
Gaussian (on coordinates)	$\mu = 0 \;\; \sigma = 0.001$	1.97	2.02	0-3.4
	$\mu = 0 \;\; \sigma = 0.01$	1.96	2.05	0-5
	$\mu = 0 \;\; \sigma = 1$	3.07	3.47	0-4.23
Uniform	$p = 0.05$	2.07	2.91	0-2.13
	$p = 0.2$	2.01	3.04	0-5

Таблица Г.3 – Абсолютная процентная ошибка (APE) для оценки Brito внутренней размерности Lebesgue для различных геометрических объектов

α	*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

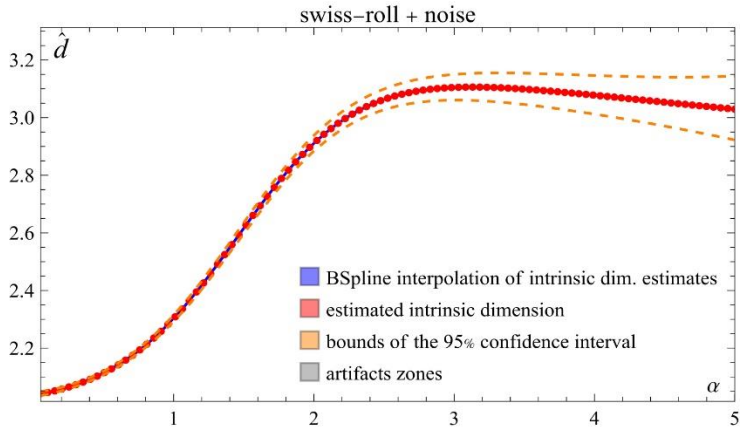


Рисунок Г.1 – Зависимость оценок алгоритма Schweinhart от параметра α для многообразия, а швейцарского рулета $d_H = 2, d = 3$ с добавленным белым шумом в соответствии с 20% от всех точек в наборе данных

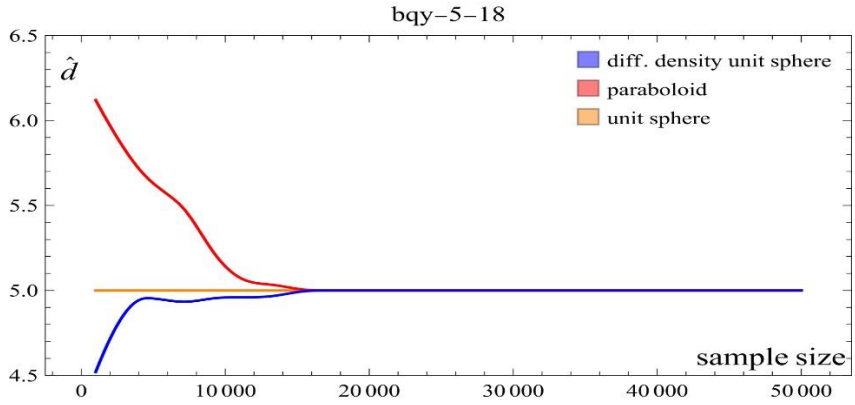


Рисунок Г.2 – Зависимость оценок алгоритма Brito от размера выборки для различных многообразий с $d_T = 2$ and $d = 9$

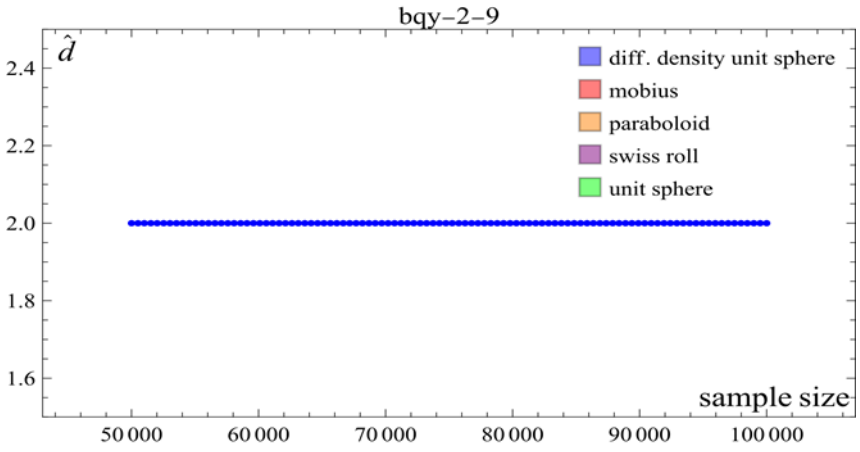


Рисунок Г.3 – Зависимость оценок алгоритма Brito от размера выборки для различных многообразий с $d_T = 5$ and $d = 18$

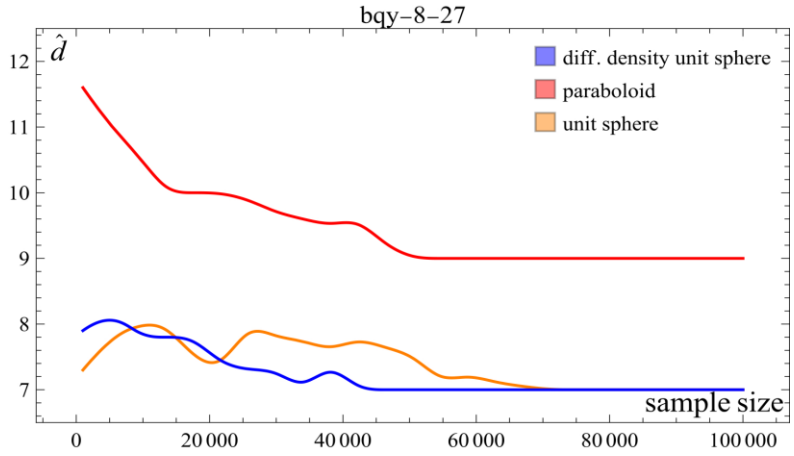


Рисунок Г.4 – Зависимость оценок алгоритма Brito от размера выборки для различных многообразий с $d_T = 8$ and $d = 27$

Таблица Г.4 – Оценка внутренней размерности SVD-представлений унограмм.

Языки	d	d^_BQY	max (d^_Schw)	min (d^_Schw)	α
1	2	3	4	5	6
Абхазский	5	4-5	4.96	4.54	0.1-1.51
	10	8-9	7.45	5.10	0.1-1.71
	15	11-12	8.34	5.03	0.1-1.91
Амхарский	5	7-8	5,58	5,20	0.1-1.11
	10	9-10	8.25	6.52	0.1-2.12
	15	13-14	10.30	10,06	0.1-1.01
Арабский	5	4-5	6.69	5.28	0.1-1.41
	10	8-9	8.18	7.65	0.1-1.51
	15	13-15	10.49	9.72	0.1-1.61
Армянский	5	4-6	5.59	5.51	0.1-1.01
	10	7-8	7.15	5.90	0.1-1.11
	15	9-13	9.67	8.76	0.1-1.21
Ассирийский	5	5-7	7,22	5,33	0.1-1.01
	10	8-10	8,68	7,15	0.1-1.01
	15	15,00	10,64	8,13	0.1-1.01
Атикемек	5	3	4,09	1,74	0.1-1.31
	10	4-5	5.94	2.65	0.1-1.41
	15	7	6,86	4,08	0.1-1.41
Бамана	5	5-7	5.52	4.72	0.1-1.01
	10	8-10	8.73	7.29	0.1-1.01
	15	11-15	10.26	8.41	0.1-1.01
Бартангский	5	4-6	5,74	4,42	0.1-1.01
	10	7-8	7,64	6,25	0.1-1.01
	15	9-12	9,77	7,88	0.1-1.01
Баскский	5	4-5	6,28	5,05	0.1-1.41
	10	8-9	7,93	6,18	0.1-1.61
	15	11-13	8.86	8.44	0.1-1.61
Белорусский	5	4	5.67	4.59	0.1-1.62
	10	6-7	8.55	7.45	0.1-1.62
	15	9-10	9.85	8.14	0.1-1.62
Бенгальский	5	4-5	5.48	5,01	0.1-1.11
	10	7-9	8.31	7.95	0.1-1.11
	15	8-10	9.80	8.67	0.1-1.31
Болгарский	5	4-5	5.37	5,09	0.1-1.21
	10	10	8.66	8.54	0.1-1.51
	15	14-15	10.51	9.93	0.1-1.61

Продолжение таблицы Г.4

1	2	3	4	5	6
Чеченский	5	5-7	6,65	5,84	0.1-1.01
	10	8	7,26	6,87	0.1-1.01
	15	9	8.71	6,57	0.1-1.01
Китайский	5	4-5	5.43	4.95	0.1-1.11
	10	7-9	8.73	8.27	0.1-1.01
	15	10-14	10,06	9.26	0.1-1.01
Коптский	5	4-6	6,08	5.36	0.1-1.21
	10	7-8	6.96	6.68	0.1-1.31
	15	9-10	7.60	7.38	0.1-1.51
Чешский	5	4-5	6.33	5.15	0.1-1.11
	10	8-9	8.48	8.14	0.1-1.01
	15	11-13	9.85	9.39	0.1-1.01
Датский	5	4-6	5.46	4.55	0.1-1.21
	10	8-9	7.30	5.86	0.1-1.21
	15	11-13	9.78	8.94	0.1-1.21
Дхолуа	5	4-6	5.73	4.42	0.1-1.01
	10	7-8	6,05	5.31	0.1-1.41
	15	10-12	6.51	5.41	0.1-1.51
Голландский	5	4-5	4.93	4.54	0.1-1.41
	10	8-9	7.44	5.10	0.1-1.71
	15	11-12	9.81	6,53	0.1-1,91
Английский	5	4	6.90	5.35	0.1-0.90
	10	6-9	8.42	6.93	0.1-0.70
	15	10-11	10.39	8.81	0.1-0.70
Эрзянский	5	5-6	6.92	5.72	0.1-1.01
	10	7-8	8.61	7.58	0.1-1.01
	15	9-11	10.05	9,07	0.1-1.01
Эсперанто	5	4-6	5.82	5,06	0.1-1.01
	10	8-9	7,03	6.34	0.1-1.01
	15	8-9	8.47	8.17	0.1-1.01
Эстонский	5	5-6	6.27	5.21	0.1-1.41
	10	7-8	7.25	6.85	0.1-1.41
	15	12-14	10.52	8.61	0.1-1.41
Французский	5	4	5.82	3.84	0.1-1.41
	10	5-6	6,78	4.98	0.1-1.41
	15	7-9	9.47	7.91	0.1-1.71
Финский	5	4-5	6.41	5.38	0.1-1.41
	10	6-9	8.12	5.88	0.1-1.51
	15	10-11	10.51	9.97	0.1-1.51
Немецкий	5	5-8	6.68	5.64	0.1-1.01
	10	8-9	8.73	6.16	0.1-1.11
	15	10-12	9.72	6.60	0.1-1.11
Хинди	5	4-5	6.48	6.42	0.1-1.11
	10	6-7	7.55	7.13	0.1-1.21
	15	11-15	10.51	10.05	0.1-1.21
Венгерский	5	4-5	7.41	6,01	0.1-1.41
	10	5-8	7.72	6.28	0.1-1.11
	15	10-11	10.43	9.79	0.1-1.11
Исландский	5	4	5.58	5.35	0.1-1.21
	10	6-8	10.38	9.48	0.1-1.21
	15	10-12	10.50	9.86	0.1-1.21
Индонезийский	5	4-5	5.75	4.93	0.1-1.41
	10	7-9	8.52	7.44	0.1-1.41
	15	10-12	10.04	8.43	0.1-1.41

Продолжение таблицы Г.4

1	2	3	4	5	6
Итальянский	5	5-7	6.57	6,05	0.1-1.31
	10	8-10	8.67	8,08	0.1-1.31
	15	11-13	9.56	8.64	0.1-1.31
Японский	5	4	4.76	3.66	0.1-1.51
	10	5-6	7.37	6.74	0.1-1.51
	15	8-10	9.29	7.52	0.1-1.61
Кабильский	5	4-5	7.00	6.29	0.1-1.01
	10	7-9	7.65	7.24	0.1-1.01
	15	7-9	7.92	7.88	0.1-1.01
Казахский	5	4	4.78	3,70	0.1-1.41
	10	6	6.39	5,08	0.1-1.51
	15	7-8	8.78	5.77	0.1-1.61
Коми-зырянский	5	6-7	7.38	5.77	0.1-4.04
	10	8-9	8.20	7.65	0.1-4.04
	15	10-15	10.19	9.78	0.1-4.04
Корейский	5	3-4	6.18	3.52	0.1-1.51
	10	5-6	7,03	4.13	0.1-1.51
	15	8-9	7,90	7.61	0.1-1.51
Кыргызский	5	4-6	5,06	3.78	0.1-1.31
	10	6-8	6.92	6.76	0.1-1.31
	15	9-10	8.85	8.51	0.1-1.31
Латинский	5	4-5	6,03	5,05	0.1-2.02
	10	6-8	9.13	7.93	0.1-2.02
	15	9-13	10.45	8.48	0.1-2.02
Латышский	5	4-5	6.64	5.16	0.1-1.51
	10	8-9	9.77	7.91	0.1-1.41
	15	11-14	10.15	8.63	0.1-1.41
Литовский	5	5-6	6.18	5.58	0.1-1.01
	10	8-10	9.26	7.99	0.1-1.01
	15	12-15	9.69	9.16	0.1-1.01
Малаялам	5	3-4	4.52	3.64	0.1-1.31
	10	5-6	6.39	4.86	0.1-1.41
	15	7	7.87	5.32	0.1-1.41
Навахо	5	4-5	5.20	4.35	0.1-1.01
	10	6-7	6.60	5.30	0.1-1.01
	15	8-11	6.78	6.24	0.1-1.01
Норвежский	5	4	4.46	3.91	0.1-1.41
	10	5-7	7.33	6.41	0.1-1.41
	15	8-10	9.42	8.23	0.1-1.51
Осетинский	5	4-5	5.21	4.88	0.1-1.31
	10	6-8	7,07	6.88	0.1-1.31
	15	9-11	10.20	9.10	0.1-1.31
Персидский	5	5-7	6.29	5,88	0.1-1.01
	10	8-9	8.53	7.49	0.1-1.01
	15	10-12	9.95	8.50	0.1-1.01
Польский	5	4-5	5.94	4.81	0.1-1.11
	10	7-9	8.24	7,07	0.1-1.11
	15	10-12	9.92	7.23	0.1-1.11
Португальский	5	5-6	6.25	5.10	0.1-1.11
	10	7-10	8.45	6.85	0.1-1.11
	15	11-13	10.02	7.18	0.1-1.11
Панджаби	5	4-6	5,67	4,06	0.1-1.21
	10	6-8	8.69	6.26	0.1-1.21
	15	9-13	9.68	7.37	0.1-1.21

Продолжение таблицы Г.4

1	2	3	4	5	6
Кечуа	5	3-4	3.88	2.16	0.1-1.21
	10	5-8	5.66	2.43	0.1-1.41
	15	9-10	7.31	3,06	0.1-1.41
Румынский	5	4-7	6.14	5,01	0.1-1.11
	10	8-9	8.29	7.67	0.1-1.51
	15	10-11	10.15	9.73	0.1-1.51
Русский	5	4-5	7.24	5.48	0.1-1.01
	10	8-10	9.22	8.49	0.1-1.01
	15	11-12	10.62	9.64	0.1-1.01
Сербский	5	4-5	6.70	5.47	0.1-1.61
	10	9-11	9.81	8.92	0.1-1.11
	15	13-15	10.24	9.94	0.1-1.01
Сингальский	5	4-5	6,27	5,18	0.1-1.31
	10	7-8	8,33	6,86	0.1-1.31
	15	10-12	9,48	7,27	0.1-1.31
Словацкий	5	4-5	5.37	4.94	0.1-1.11
	10	6-8	8.41	6.40	0.1-1.11
	15	9-12	10.12	8.39	0.1-1.11
Испанский	5	5-6	6,51	5,72	0.1-1.11
	10	7-9	9.17	8.47	0.1-1.11
	15	10-13	9.83	9.18	0.1-1.11
Словенский	5	4-5	5.57	5,07	0.1-1.21
	10	8-9	8.38	7.23	0.1-1.21
	15	10-14	9.94	8.12	0.1-1.21
Суахили	5	4-5	6.41	4.73	0.1-1.41
	10	7-9	7.47	6.30	0.1-1.41
	15	10-12	7.91	6.73	0.1-1.41
Шведский	5	6-7	6.83	4.86	0.1-1.61
	10	8-9	9.65	6.44	0.1-1.71
	15	10-11	10.21	8.13	0.1-1.71
Табасаранский	5	4-5	5.23	5.11	0.1-1.01
	10	6-8	7.34	6,96	0.1-1.11
	15	9-11	8.71	6.41	0.1-1.51
Тагальский	5	4-5	6.79	5,02	0.1-1.11
	10	6-8	9.27	7.97	0.1-1.11
	15	9-12	10.17	9.61	0.1-1.11
Татарский	5	4	6.77	3.56	0.1-1.61
	10	5-6	7.60	4.78	0.1-1.61
	15	8-10	8.19	5.76	0.1-1.61
Тайский	5	4-5	5.82	5.32	0.1-1.01
	10	7-9	8.53	7.49	0.1-1.01
	15	10-12	9.94	8.49	0.1-1.01
Тибетский	5	4-6	6.28	4.12	0.1-1.21
	10	7-8	7.17	5.13	0.1-1.21
	15	9-11	7,69	6,33	0.1-1.21
Турецкий	5	4-5	5.45	3.40	0.1-1.41
	10	6	6.76	4,07	0.1-1.51
	15	9-10	8.61	5.16	0.1-1.61
Тувинский	5	5-6	6.80	5.32	0.1-1.01
	10	7	7.71	6,07	0.1-1.01
	15	8-10	8.12	6.58	0.1-1.01

Продолжение таблицы Г.4

1	2	3	4	5	6
Удмуртский	5	4-5	5.78	5.51	0.1-1.81
	10	6-7	8.43	8.14	0.1-1.21
	15	8-10	10.21	9.57	0.1-1.21
Украинский	5	4-5	5.97	4.11	0.1-1.71
	10	7-8	7.96	6,09	0.1-2.02
	15	9-12	10.21	7.46	0.1-2.02
Узбекский	5	4-5	5.79	4.11	0.1-1.61
	10	6-7	7,08	5.82	0.1-1.61
	15	9-10	7.95	6.45	0.1- 1.71
Вьетнамский	5	4	6.92	4,08	0.1-1.51
	10	6-7	8,00	5.22	0.1-1.51
	15	9-11	10.04	6.41	0.1-1.51
Идиш	5	5-6	6.50	5.22	0.1-1.01
	10	7-9	8.49	8.26	0.1-1.01
	15	10-13	9.65	9.53	0.1-1.01

Таблица Г.5 – Оценка внутренней размерности SVD-представлений биграмм

Языки	d	d^_BQY	max (d^_Schw)	min (d^_Schw)	α
1	2	3	4	5	6
Абхазский	5	6-8	5,06	4.71	0.1-4.34
	10	10-11	6.19	5.38	0.1-4.54
	15	12-15	6.49	5.53	0.1-3.63
Амхарский	5	6-8	5,62	5.44	0.1-4.23
	10	9-10	7,17	5.88	0.1-5.85
	15	12-15	7,90	6.46	0.1-6.66
Арабский	5	5-6	5.91	5.42	0.1-2.82
	10	7-10	7.52	6,98	0.1-4.04
	15	11-15	8.41	7.66	0.1-5.85
Армянский	5	6	5.38	5.28	0.1-3.03
	10	7-8	6.43	6.20	0.1-3.03
	15	9-11	7.71	7.33	0.1-3.13
Ассирийский	5	6-8	4.27	4,01	0.1-3.01
	10	10-12	6,08	5.62	0.1-3.01
	15	14-15	8.27	8,01	0.1-1.01
Атикемек	5	3-4	2.53	2.29	0.1-1.91
	10	5-6	3.17	2.77	0.1-2.32
	15	10-12	4.53	4,05	0.1-3.23
Бамана	5	5-6	5,09	4.83	0.1-3.34
	10	7-10	6.64	6.57	0.1-3.34
	15	14-15	8.10	7.77	0.1-3.47
Бартангский	5	6-7	5.10	4.34	0.1-3.23
	10	8-9	6.87	5.52	0.1-3.23
	15	10-14	8.36	7.14	0.1-2.82
Баскский	5	6-7	5,24	4.89	0.1-2.92
	10	9-10	6,41	5.80	0.1-4.64
	15	12-15	7,03	6.15	0.1-5.65
Белорусский	5	5,00	5,17	5.00	0.1-4.14
	10	7-8	6,47	5.39	0.1-5.05
	15	12-15	7,79	5.14	0.1-5.05
Бенгальский	5	5-7	5.18	4.40	0.1-3.93
	10	9-10	7.38	5.72	0.1-4.04
	15	11-14	8.11	5.89	0.1-4.04

Продолжение таблицы Г.5

1	2	3	4	5	6
Болгарский	5	6-7	5.64	4.44	0.1-3.53
	10	8-12	7.49	4.65	0.1-4.34
	15	13-15	8.42	5.45	0.1-4.64
Чеченский	5	5-6	5.85	5.58	0.1-2.22
	10	7	5,90	5.63	0.1-2.22
	15	8	6.22	5.70	0.1-2.52
Китайский	5	5-6	5.77	5.25	0.1-4.94
	10	7-8	7,29	4,94	0.1-8.28
	15	9-10	9,22	5,85	0.1-9.09
Коптский	5	5-9	4,74	4,70	0.1-2.02
	10	13	5.30	5,17	0.1-2.02
	15	15	5.74	5.50	0.1-2.02
Чешский	5	5-6	5.35	5.18	0.1-2.52
	10	10-12	7.27	5.25	0.1-4.14
	15	14-15	8.25	5.31	0.1-6.26
Датский	5	5-8	5.61	5.32	0.1-3.53
	10	9	6.99	6.44	0.1-3.13
	15	10-15	8.41	8.12	0.1-3.13
Дхолуа	5	6-8	4,12	3.96	0.1-1.91
	10	11-13	4.78	4.44	0.1-2.32
	15	15	5.00	4.54	0.1-2.32
Голландский	5	6-7	5,04	3.90	0.1-4.64
	10	10-12	7.27	5.18	0.1-6.46
	15	14-15	8.26	5.65	0.1-7.27
Английский	5	5-6	5.44	4.93	0.1-3.03
	10	8-9	6.46	5.13	0.1-3.23
	15	13-15	8.56	5.95	0.1-3.53
Эрзянский	5	6-8	6.38	6.25	0.1-3.53
	10	9-10	7.25	6.91	0.1-3.53
	15	13-15	8.46	8.39	0.1-3.53
Эсперанто	5	6-7	4.93	4,09	0.1-3.23
	10	8	5.93	4,56	0.1-3.73
	15	11-14	6.93	5,06	0.1-3.93
Эстонский	5	5-6	5.71	5.47	0.1-5.05
	10	7-8	7,08	6.42	0.1-5.05
	15	12-15	8.42	7.37	0.1-5.05
Французский	5	5	4,55	2.97	0.1-3.53
	10	7-8	5.86	3.17	0.1-4.04
	15	12-13	7.66	4,96	0.1-4.24
Финский	5	6-7	5.34	5.30	0.1-2.52
	10	8-12	7.82	6.61	0.1-5.15
	15	13-15	8.49	7.26	0.1-5.35
Немецкий	5	6-9	6.71	6.15	0.1-3.23
	10	10	7.82	6.77	0.1-3.83
	15	11-13	8.26	6.83	0.1-3.63

Продолжение таблицы Г.5

1	2	3	4	5	6
Хинди	5	6	5.16	5,02	0.1-2.62
	10	8-9	6.34	5.53	0.1-3.83
	15	14-15	7,02	5.82	0.1-3.73
Венгерский	5	8-9	5.14	05.02	0.1-2.82
	10	13	6.31	5.52	0.1-3.63
	15	14-15	6.98	5.82	0.1-3.73
Исландский	5	5-6	5.23	5.00	0.1-3.83
	10	9-10	7,32	6.45	0.1-4.04
	15	14-15	8.15	7,13	0.1-4.04
Индонезийский	5	5-6	5.19	4.86	0.1-3.59
	10	9-10	7.26	6.28	0.1-5.05
	15	13-15	7.58	5.82	0.1-6.36
Итальянский	5	5-7	5.50	5,05	0.1-2.72
	10	8-10	7,06	6.81	0.1-2.92
	15	14-15	8.52	8.27	0.1-3.13
Японский	5	5-6	4.80	3.69	0.1-2.92
	10	8-10	6.10	4.47	0.1-3.43
	15	12-14	7.67	6,04	0.1-3.53
Кабильский	5	6-9	4,15	04.05	0.1-1.61
	10	11-12	5,07	4.69	0.1-1.81
	15	13-15	5.30	4.82	0.1-2.02
Казахский	5	4	4.16	3.86	0.1-3.03
	10	5-6	5.34	4.95	0.1-4.04
	15	8-10	6.52	6,08	0.1-5.05
Коми-зырянский	5	6-8	6.25	5.94	0.1-6.26
	10	9-11	7.41	6.17	0.1-6.55
	15	12-15	8.45	8.37	0.1-7.97
Корейский	5	5-6	3.32	3,07	0.1-2.12
	10	7-8	4,09	3.50	0.1-2.33
	15	14-15	5.39	4.32	0.1-2.52
Кыргызский	5	4-6	4.73	3.74	0.1-3.63
	10	7-8	6.22	5,09	0.1-3.93
	15	9-11	6.51	5.19	0.1-4.04
Латинский	5	5-8	6.45	6.30	0.1-5.05
	10	9	7.18	7.13	0.1-7.07
	15	12-15	8.20	8.19	0.1-8.08
Латышский	5	6-8	5.26	5.11	0.1-3.23
	10	9-10	7,05	6.89	0.1-2.32
	15	12-15	7.65	6.45	0.1-2.32
Литовский	5	7-9	6.81	4.80	0.1-3.73
	10	10-14	7.64	5.80	0.1-3.73
	15	15	8.31	7.30	0.1-3.83
Малаялам	5	3-4	2.95	2.87	0.1-3.03
	10	5	3.68	3.55	0.1-3.38
	15	7-9	3.97	3.92	0.1-3.83

Продолжение таблицы Г.5

1	2	3	4	5	6
Навахо	5	3-5	3.81	3.77	0.1-2.72
	10	6	4,10	4,05	0.1-2.72
	15	7-10	5,01	4.95	0.1-2.82
Норвежский	5	4	4.14	3.44	0.1-3.33
	10	6-9	5.40	3.66	0.1-4.14
	15	11-13	7.75	6.24	0.1-6.06
Осетинский	5	6-9	5.67	5.44	0.1-4.14
	10	10-11	7.33	6.10	0.1-4.54
	15	12-14	8.55	7.30	0.1-4.74
Персидский	5	5-6	5.18	4.92	0.1-9.39
	10	8-10	7.17	6.28	0.1-6.76
	15	11-15	8.51	7,12	0.1-7.77
Польский	5	6-7	5.52	4.95	0.1-4.24
	10	10-11	7.51	5.88	0.1-6.56
	15	14-15	8.57	6.60	0.1-7.57
Португальский	5	6-8	6.14	6,03	0.1-3.33
	10	10-12	7,01	6.55	0.1-3.33
	15	13-15	8.24	7.78	0.1-3.73
Панджаби	5	6-7	5.97	5.70	0.1-4.57
	10	9-10	7.19	6.96	0.1-5.15
	15	11-13	8.64	8.15	0.1-5.35
Кечуа	5	4-8	4.10	4.00	0.1-2.32
	10	9-12	4.81	3.79	0.1-2.52
	15	14-15	5.13	4,09	0.1-2.72
Румынский	5	5-8	5,02	4.30	0.1-3.33
	10	9-10	6.51	5.19	0.1-3.13
	15	12-14	8.39	7,07	0.1-3.93
Русский	5	5-6	5,13	3,93	0.1-3.43
	10	9-10	7.54	4.82	0.1-3.53
	15	14-15	8.40	7.20	0.1-3.53
Сербский	5	5-6	5.51	5.15	0.1-2.92
	10	8-10	7.48	7.22	0.1-3.43
	15	12-15	8.32	7,07	0.1-4.14
Сингальский	5	6-7	6,09	5.26	0.1-4.14
	10	10-12	7.39	7.35	0.1-4.94
	15	14-15	8.45	7.91	0.1-6.16
Словацкий	5	5-7	6.96	5.14	0.1-6.44
	10	8-9	7.76	5.62	0.1-6.44
	15	10-15	8.47	7.25	0.1-6.56
Испанский	5	5-7	5.87	5.73	0.1-4.64
	10	8-9	7,56	6,54	0.1-7.36
	15	9-15	8.76	5.98	0.1-8.08
Словенский	5	508,00	5.27	4.89	0.1-3.93
	10	10-11	7.53	6.66	0.1-4.54
	15	13-15	8.29	6.94	0.1-5.15

Продолжение таблицы Г.5

1	2	3	4	5	6
Суахили	5	6-7	4.95	4.69	0.1-2.62
	10	8-10	6,04	5.86	0.1-3.93
	15	11-14	6,35	5.95	0.1-4.64
Шведский	5	5-6	5.17	4.17	0.1-5.65
	10	10-13	8.24	7.15	0.1-6.46
	15	13-15	8.43	6.92	0.1-7.47
Табасаранский	5	5-6	5,02	4.99	0.1-3.03
	10	7	6.12	6,04	0.1-3.03
	15	8-10	6.53	6.50	0.1-3.13
Тагальский	5	5-7	6.55	6.37	0.1-2.82
	10	8	7,01	6.93	0.1-2.92
	15	9	7.53	7.43	0.1-3.33
Татарский	5	5-6	4.70	4,29	0.1-3.03
	10	7-8	5.61	5,05	0.1-3.73
	15	9-10	6.63	5.64	0.1-4.94
Тайский	5	5-7	5.18	4.89	0.1-3.53
	10	10-12	7.23	5.93	0.1-5.25
	15	14-15	8.22	6.23	0.1-6.26
Тибетский	5	5-7	3.37	3.30	0.1-2.91
	10	8-9	3.62	3.54	0.1-2.91
	15	10-11	4,06	4,04	0.1-3.23
Турецкий	5	5	4.60	3.81	0.1-3.83
	10	7-8	5.92	3.94	0.1-4.03
	15	10-11	6.31	3.00	0.1-4.34
Тувинский	5	5	4.69	4.29	0.1-1.71
	10	6-9	5.67	5.30	0.1-1.71
	15	10-12	6.62	6,09	0.1-1.71
Удмуртский	5	5-8	5.56	5.49	0.1-3.93
	10	9	7.57	5.94	0.1-4.24
	15	10-12	8.47	8.40	0.1-4.44
Украинский	5	5-8	5.51	5.13	0.1-4.14
	10	10-11	7.18	6.73	0.1-5.05
	15	12-15	7.87	6.84	0.1-5.45
Узбекский	5	5	4.56	4.32	0.1-3.63
	10	6	5.43	4.72	0.1-4.74
	15	7-10	6.32	5.48	0.1-5.05
Вьетнамский	5	5-6	4.79	4.12	0.1-3.73
	10	7-10	6.30	4.78	0.1-4.84
	15	12-15	8.27	6.77	0.1-4.74
Идиш	5	5-7	5.22	5,09	0.1-2.72
	10	10-11	7,05	5.77	0.1-2.65
	15	12-15	7.81	7.68	0.1-2.52

Таблица Г.6 – Оценка внутренней размерности SVD-представлений триграмм.

Языки	d	dˆ_BQY	max (dˆ_Schw)	min (dˆ_Schw)	α
1	2	3	4	5	5
Абхазский	15	14	6.65	6.46	0.1-5.25
Амхарский	15	12-15	8.12	7,06	0.1-5.05
Арабский	15	12-15	8.60	8.45	0.1-3.93
Армянский	15	10-13	7.90	7.74	0.1-2.62
Ассирийский	15	12-15	8.86	8.69	0.1-1.91
Атикемек	15	13-14	4.23	3.30	0.1-2.42
Бамана	15	11-15	8.22	8.21	0.1-3.37
Бартангский	15	10-14	8.53	8.11	0.1-3.13
Баскский	15	12-15	6.57	6.53	0.1-4.44
Белорусский	15	13-15	7,60	7.43	0.1-3.63
Бенгальский	15	13-15	8.25	6.61	0.1-4.04
Болгарский	15	12-15	7.97	7,04	0.1-5.45
Чеченский	15	6-8	6,37	5.67	0.1-2.82
Китайский	9-10	9,08	9,48	8,94	0.1-5.15
Коптский	15	15	5.19	5.19	0.1-6.06
Чешский	15	14-15	8.39	4.69	0.1-4.94
Датский	15	10-15	8.52	7.97	0.1-5.15
Дхолуа	15	12-15	5,02	4.82	0.1-1.31
Голландский	15	15	8.51	8,02	0.1-5.65
Английский	15	13-15	9.29	5.97	0.1-1.91
Эрзянский	15	11-15	8.58	8.57	0.1-3.33
Эсперанто		11-13	7.66	5.89	0.1-3.73
Эстонский	15	11-15	8.56	8.54	0.1-4.94
Французский	15	10-14	7.73	7.10	0.1-3.63
Финский	15	12-15	8.60	8.48	0.1-5.75
Немецкий	15	11-13	8.32	6.81	0.1-3.33
Хинди	15	13-15	7.34	6.53	0.1-4.24
Венгерский	15	15	7,09	6.65	0.1-3.33
Исландский	15	15	8.41	7.00	0.1-4.84
Индонезийский	15	13-15	8,08	6.35	0.1-6.36
Итальянский	15	14-15	8.61	8.19	0.1-3.33
Японский	15	14-15	7.72	7.34	0.1-4.24
Кабильский	15	14-15	5.41	4.93	0.1-2.12
Казахский	15	8-10	6.67	5.89	0.1-5.05
Коми-зырянский	15	14-15	8.58	8.55	0.1-9.49
Корейский	15	13-15	5.49	05.05	0.1-3.03
Кыргызский	15	10-11	6.62	6,03	0.1-6.46
Латинский	15	11-15	8.42	8,02	0.1-9.19
Латышский	15	12-15	7.66	7.00	0.1-2.12
Литовский	15	14-15	8.46	6.32	0.1-4.94
Малаялам	15	7-9	4,07	4,01	0.1-2.12
Навахо	15	7-10	5.22	5.20	0.1-3.63
Норвежский	15	11-13	7.84	6.95	0.1-5.85
Осетинский	15	12-15	8.67	7.23	0.1-7.27
Персидский	15	12-15	8.66	8.54	0.1-4.64
Польский	15	14-15	8.74	8.68	0.1-4.64
Португальский	15	12-15	8.35	8.33	0.1-2.32
Панджаби	15	11-13	8.91	8.87	0.1-3.93
Кечуа	15	14-15	5.30	4.67	0.1-1.61
Румынский	15	11-14	8.62	7.43	0.1-3.53
Русский	15	14-15	8.59	7.96	0.1-4.34
Сербский	15	11-15	8.47	7.52	0.1-3.83
Сингальский	15	14-15	8.39	8.36	0.1-6.96
Словацкий	15	10-15	8.64	8.55	0.1-6.26
Испанский	15	10-15	8,84	7,82	0.1-6.06
Словенский	15	12-15	8.36	8.31	0.1-4.84
Суахили	15	10-14	5.84	5.78	0.1-3.73

Продолжение таблицы Г.6

1	2	3	4	5	6
Шведский	15	13-15	8.74	7.77	0.1-7.37
Табасаранский	15	7-10	6.90	6.87	0.1-4.14
Тагальский	15	8-9	8,02	7.98	0.1-2.62
Татарский	15	9-10	6.72	6.41	0.1-4.24
Тайский	15	15	8.71	8.64	0.1-4.24
Тибетский	15	10-11	4.13	4,11	0.1-3.23
Турецкий	15	10-11	6.62	2.38	0.1-7.07
Тувинский	15	10-12	6.70	6.44	0.1-2.52
Удмуртский	15	10-12	8.58	7.65	0.1-4.04
Украинский	15	12-15	7.69	6.34	0.1-5.45
Узбекский	15	8-10	6.56	6.38	0.1-4.24
Вьетнамский	15	13-15	8.43	7.33	0.1-7.57
Идиш	15	12-15	7.92	7.46	0.1-7.67

ПРИЛОЖЕНИЕ Д

Экспериментальные результаты генерации текстов и их классификации

Таблица Д.1 – Модели GPT-2 и количество параметров

Язык	Модель (Hugging Face)	Количество параметров
Русский	ai-forever/rugpt3large_based_on_gpt2	760M
Английский	openai-community/gpt2	124M
Немецкий	dbmdz/german-gpt2	124M
Вьетнамский	NlpHUST/gpt2-vietnamese	124M
Французский	dbddv01/gpt2-french-small	124M

Таблица Д.2 – Средняя длина сгенерированных текстов (в словах)

Язык	LSTM	GPT-2	mGPT	YaLM
Русский	12185	6287	13097	12405
Английский	36721	32682	2807	3953
Немецкий	56423	42439	30650	23252
Французский	10813	55387	33897	16236
Вьетнамский	13475	10460	11479	10105

Алгоритм 1. Алгоритм генерации текста
Вход:
D – корпус текстов, написанных людьми (фиксированный язык).
bot – модель генерации текста.
l – максимальное число слов, генерируемых за один шаг.
n – число генерируемых текстов.
Выход:
D_{bot} – корпус текстов, сгенерированных ботом.
$D_{bot} \leftarrow \emptyset$
for $m \leftarrow 1 \ldots n$ do
$d \leftarrow random\_choice(D)$
$i \leftarrow 1$
$d_m \leftarrow \emptyset$
while $i < \ell(d)$ do
$r \leftarrow bot(d[i], l)$
$d_m \leftarrow d_m \cup r$
$i \leftarrow i + \ell(r)$
$D_{bot} \leftarrow D_{bot} \cup d_m$
return D_{bot}

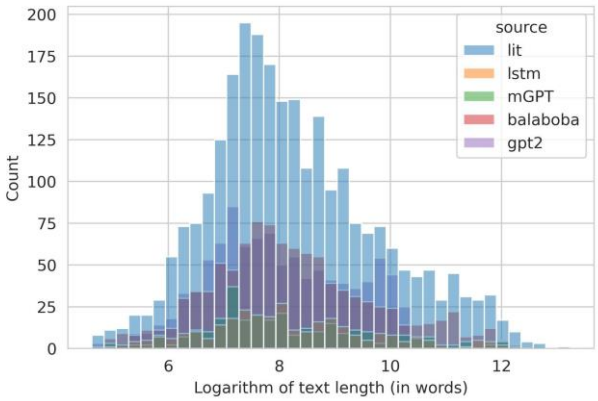
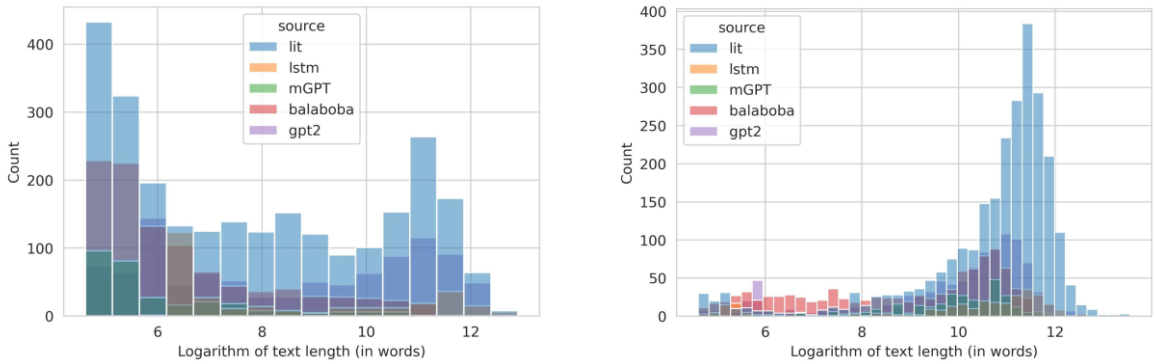


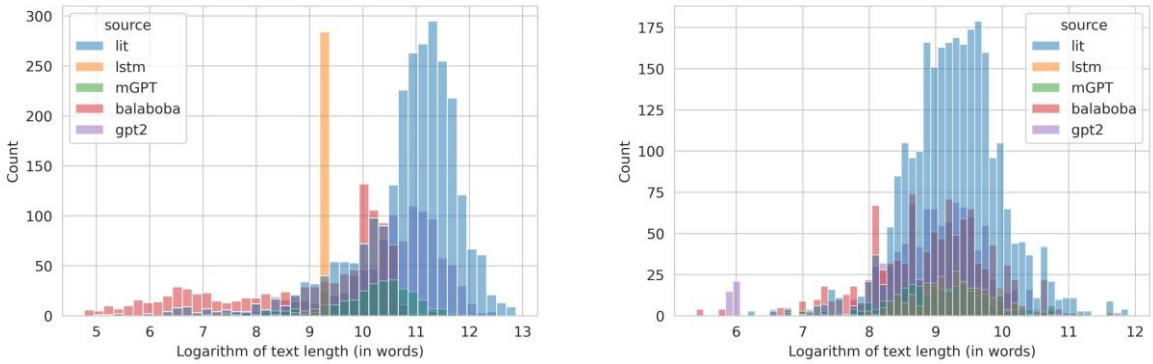
Рисунок Д.1 – Распределение длины слов (логарифмическая шкала) для текстов на русском языке



а б

а – для текстов на английском языке; б – для текстов на немецком языке

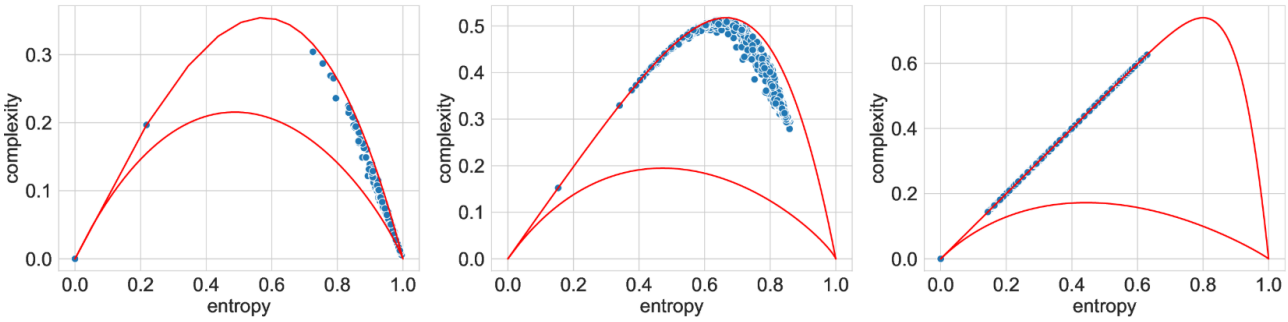
Рисунок Д.2 – Распределение длины в словах (логарифмическая шкала)



а б

а – для текстов на французском языке; б – для текстов на вьетнамском языке

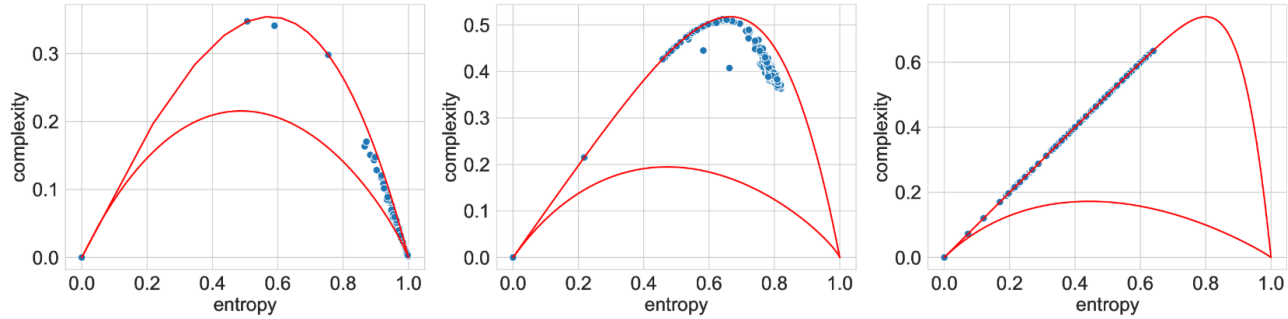
Рисунок Д.3 – Распределение длины в словах (логарифмическая шкала)



а б в

а – $n = 3, d = 4$; б – $n = 4, d = 1$; в – $n = 5, d = 4$

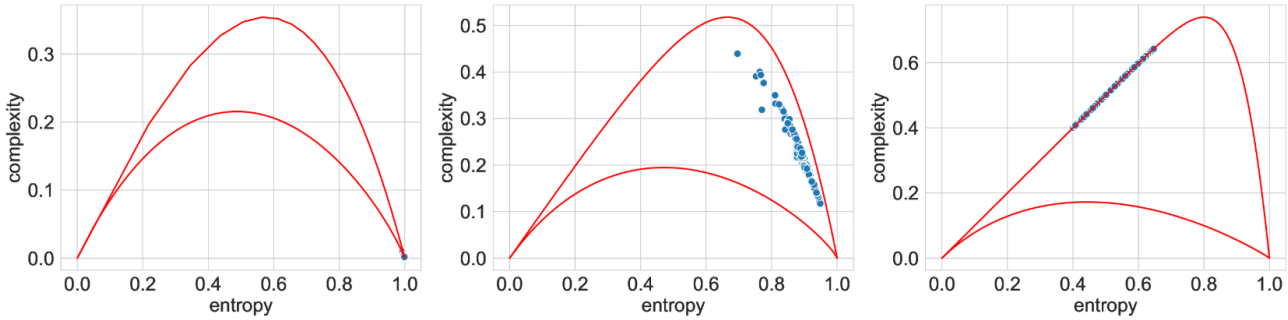
Рисунок Д.4 – Плоскость энтропии-сложности: точки, соответствующие текстам английской литературы



а б в

а – $n = 4, d = 1$; б – $n = 3, d = 4$; в – $n = 5, d = 4$

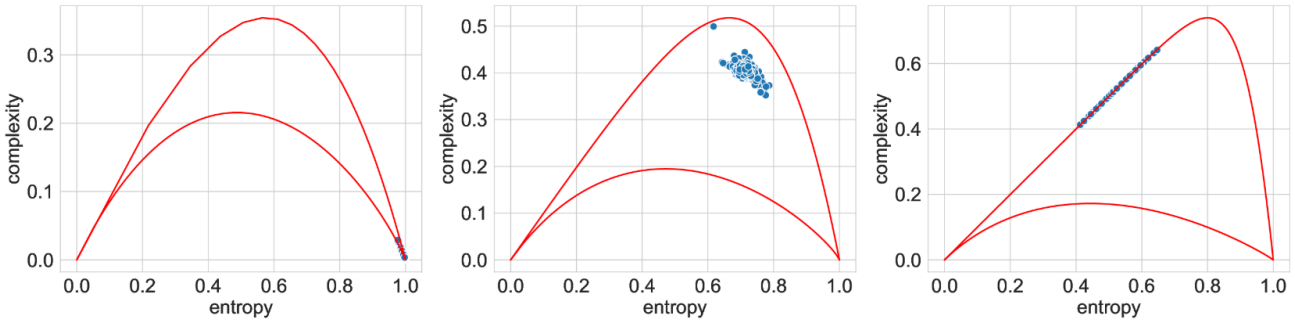
Рисунок Д.5 – Плоскость энтропии-сложности: точки, соответствующие текстам немецкой литературы



а б в

а – $n = 4, d = 1$; б – $n = 3, d = 4$; в – $n = 5, d = 4$

Рисунок Д.6 – Плоскость энтропии-сложности: точки, соответствующие текстам французской литературы



а б в

а – $n = 4, d = 1$; б – $n = 3, d = 4$; в – $n = 5, d = 4$

Рисунок Д.7 – Плоскость энтропии-сложности: точки, соответствующие текстам вьетнамской литературы

Таблица Д.8 – Значения показателей оценки точности для единого классификационного моделя Support Vector Machine, DT, и Random Forest являются синонимами SVM

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
SVM	0.88	0.64	0.97	0.96	0.95
DT	0.77	0.82	0.97	0.64	0.95
RF	0.78	0.83	0.98	0.87	0.97

Таблица Д.9 – Значения показателей точности для классификаторов с семантическими характеристиками траекторий

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
Support Vector Machine					
SVD	0.54	0.42	0.59	0.74	0.50
CBOW	0.50	0.63	0.77	0.50	0.73
Skip-Gram	0.50	0.50	0.65	0.71	0.63
Decision Tree					
SVD	0.61	0.78	0.58	0.85	0.67
CBOW	0.50	0.58	0.81	0.50	0.59
Skip-Gram	0.50	0.50	0.58	0.74	0.60
Random Forest					
SVD	0.64	0.79	0.68	0.86	0.67
CBOW	0.51	0.61	0.82	0.50	0.68
Skip-Gram	0.50	0.50	0.63	0.78	0.62

Таблица Д.10 – Значения показателей точности для классификаторов на основе кластеризации Wishart

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
1	2	3	4	5	6
Support Vector Machine					
SVD	0.60	0.75	0.91	0.93	0.57
CBOW	0.68	0.88	0.69	0.69	0.80
Skip-Gram	0.71	0.82	0.69	0.95	0.84
Decision Tree					
SVD	0.68	0.74	0.88	0.86	0.80

Продолжение таблитцы Д.10

1	2	3	4	5	6
CBOW	0.84	0.80	0.75	0.69	0.59
Skip-Gram	0.73	0.73	0.70	0.74	0.58
Random Forest					
SVD	0.65	0.80	0.90	0.65	0.64
CBOW	0.84	0.85	0.73	0.70	0.58
Skip-Gram	0.82	0.80	0.67	0.74	0.59

Таблица Д.11 – Значения показателей классификаторов, основанные на кластеризации K-Means

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
Support Vector Machine					
SVD	0.70	0.72	0.91	0.58	0.69
CBOW	0.55	0.95	0.84	0.86	0.62
Skip-Gram	0.62	0.86	0.86	0.86	0.56
SVD	0.60	0.82	0.88	0.59	0.77
CBOW	0.79	0.87	0.65	0.64	0.57
Skip-Gram	0.86	0.84	0.59	0.77	0.55
Random Forest					
SVD	0.66	0.92	0.88	0.85	0.74
CBOW	0.55	0.90	0.61	0.93	0.54
Skip-Gram	0.82	0.87	0.72	0.73	0.56

Таблица Д.12 – Значения показателей оценки точности для классификаторов на основе внутрикластерных расстояний

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
Support Vector Machine					
Wishart	0.59	0.63	0.50	0.50	0.67
Fuzzy Wishart	0.49	0.66	0.50	0.88	0.60
K-Means	0.50	0.80	0.51	0.63	0.65
C-Means	0.92	0.75	0.52	0.47	0.54
Decision Tree					
Wishart	0.56	0.71	0.72	0.64	0.65
Fuzzy Wishart	0.70	0.85	0.86	0.92	0.88
K-Means	0.97	0.86	0.63	0.70	0.67
C-Means	0.93	0.78	0.68	0.64	0.73
Random Forest					
Wishart	0.55	0.73	0.71	0.61	0.67
Fuzzy Wishart	0.70	0.85	0.89	0.93	0.81
K-Means	0.98	0.87	0.61	0.51	0.70
C-Means	0.95	0.78	0.60	0.67	0.72

Таблица Д.13 – Значения показателей оценки точности для единого классификационного моделя Support Vector Machine, DT, и Random Forest являются синонимами SVM

Модели	Русский	Английский	Немецкий	Французский	Вьетнамский
SVM	0.82	0.98	0.63	0.82	0.74
DT	0.98	0.88	0.90	0.86	0.72
RF	0.99	0.91	0.91	0.86	0.72

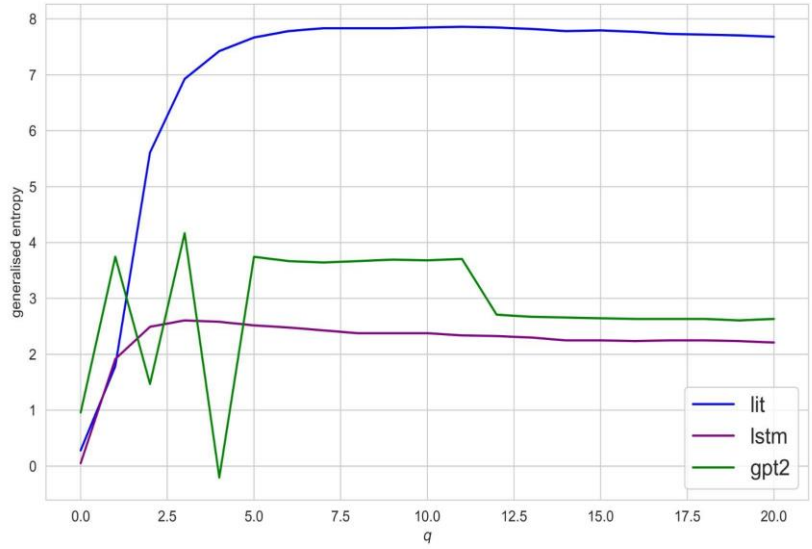


Рисунок Д.8 – Обобщённые значения энтропии для текстов на русском языке, q в диапазоне от 0 до 20

Примечание – Синяя линия относится к литературным текстам, зеленая линия – к текстам, сгенерированным GPT-2, фиолетовая – к текстам, сгенерированным LSTM

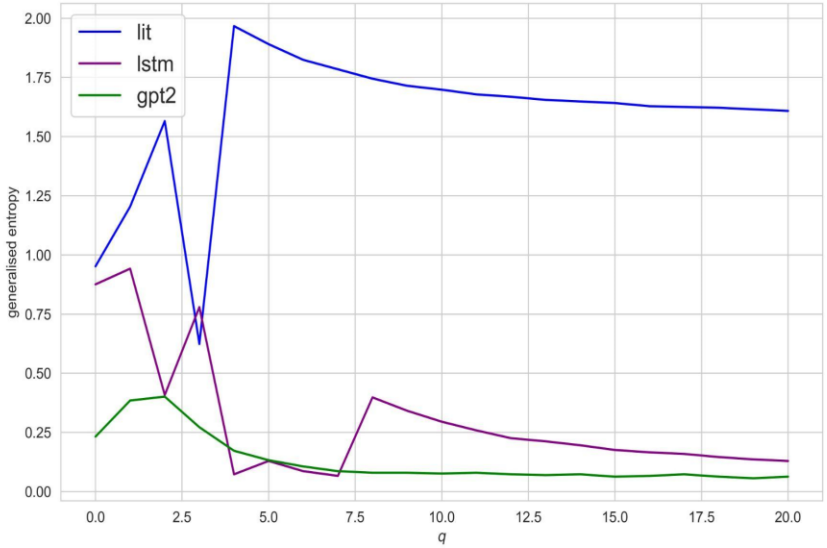


Рисунок Д.9 – Обобщённые значения энтропии для текстов на английском языке, q в диапазоне от 0 до 20

Примечание – Синяя линия относится к литературным текстам, зеленая линия – к текстам, сгенерированным GPT-2, пурпурная – к текстам, сгенерированным LSTM