Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное
учреждение высшего образования «Национальный исследовательский
университет ИТМО»

Факультет инфокоммуникационных технологий

Математическая лингвистика

Практическая работа №1

Выполнили:

студент группы К34422

Малаев Степан Геннадьевич

Проверил:

доцент практики, КТН

Болгова Екатерина Владимировна

Санкт-Петербург

2025

Цель работы. Ознакомиться с разделом квантитативной лингвистики и базовыми статистическими понятиями, научиться исследовать авторские инварианты.

Ход работы.

1. Выбор произведений одного писателя, упорядоченных хронологически, для создания текста автора.

Были выбраны произведения шотландского писателя и драматурга, Ирвина Уэлша:

- На игле (1993)
- Кошмары аиста Марабу (1995)
- Экстази (1996)

Данные произведения являются наиболее популярными работами писателя, позволяя захватить большую часть художественной ценности и приемов используемых и в остальных его произведениях.

Был осуществлен сбор всех произведений в разных текстовых форматах хранения (epub, pdf), и преобразованы в текст.

Предварительно, прежде чем приступить к анализу, текст был очищен используя следующие методы:

1. Нормализация
    1.1. Унификация регистра
    1.2. Удаление пустых символов
2. Очистка от символов
    2.1. Удаление цифр
    2.2. Удаление спецсимволов
    2.3. Удаление символов пунктуации
3. Удаление стоп-слов
4. Токенизация

2. Формирование ГЛС

После очистки данных, тексты всех произведений были хронологически отсортированы и объединены в единый корпус.

Общее количество токенов итогового ГЛС: 160663

3. Формирование ВЛС

Разобъем объединённый текст на выборки фиксированного объёма с заданным шагом.

В качестве параметром укажем:

- Размер сегмента: 1500
- Шаг: 1200

Количество сегментов: 133

4. Выбор и вычисление лингвистического параметра

В данном исследовании используется частота употребления отдельных слов как лингвистическая метрика, позволяющая проследить динамику эмоциональной и стилистической окраски текста по его сегментам.

Анализируется, как изменяется использование выбранных слов по мере продвижения по произведению, что дает представление о развитии нарратива и эмоциональном фоне.

В качестве главного параметра были выбраны слова:

- fuckin
- man
- bad
- good

Выбор этих слов обусловлен особенностями литературного стиля писателя. Его произведения характеризуются социальной направленностью к маргинальному обществу, сатирическими элементами, нецензурной лексикой и шотландскими диалектизмами.

Выбранные слова обладают ярко выраженной эмоциональной нагрузкой, что делает их особенно показательными для анализа. Отслеживая их относительную частоту в различных сегментах текста, можно выявить, как меняется эмоциональное состояние повествования, какие моменты усиливают эффект и каким образом автор через лексику подчеркивает особенности характеров и окружения своих героев.

5. Построение графика эволюции параметров

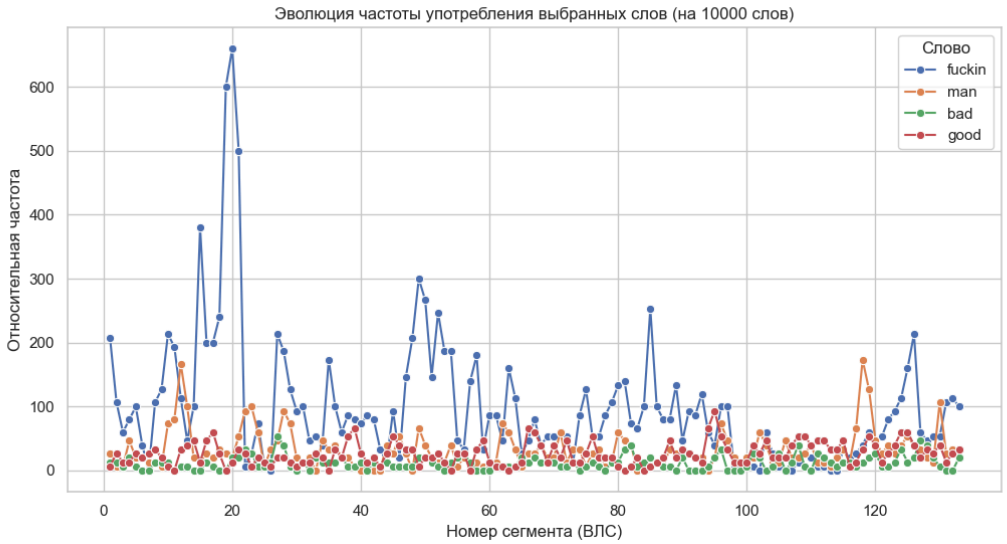

Рисунок 1 – График эволюция частоты параметров в сегментах.

Видно, что “fuckin” существенно доминирует по частоте и периодически даёт резкие всплески. Остальные слова употребляются реже и имеют более равномерное распределение.

6. Анализ лексического состава для нескольких выборок

Были взяты 3 сегмента: первый, средний и последний.

Для них будут построены 2 визуализации:

- Столбчатая диаграмма распределения частей речи
- График отношения частоты каждого слова к суммарной частоте своей части речи

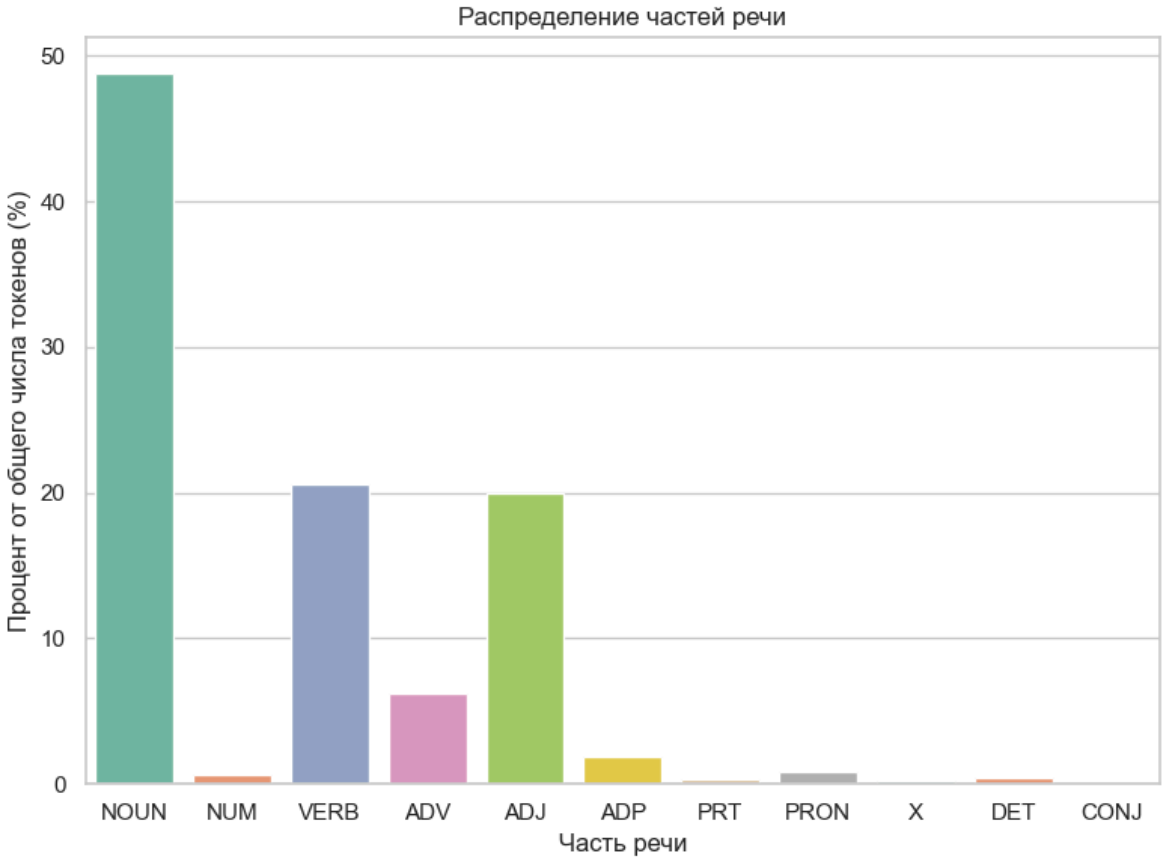

Рисунок 2 – График распределение частей речи для первого сегмента.

На данном графике видно, что наиболее существенную долю в тексте сегмента занимают существительные, составляя около половины всех токенов. Это указывает на преимущественно номинативный характер отрывка, где акцент делается на именование объектов или понятий. Существенные доли также занимают глаголы и прилагательные, что отражает наличие описаний и некоторой динамики действий. Другие категории встречаются реже.

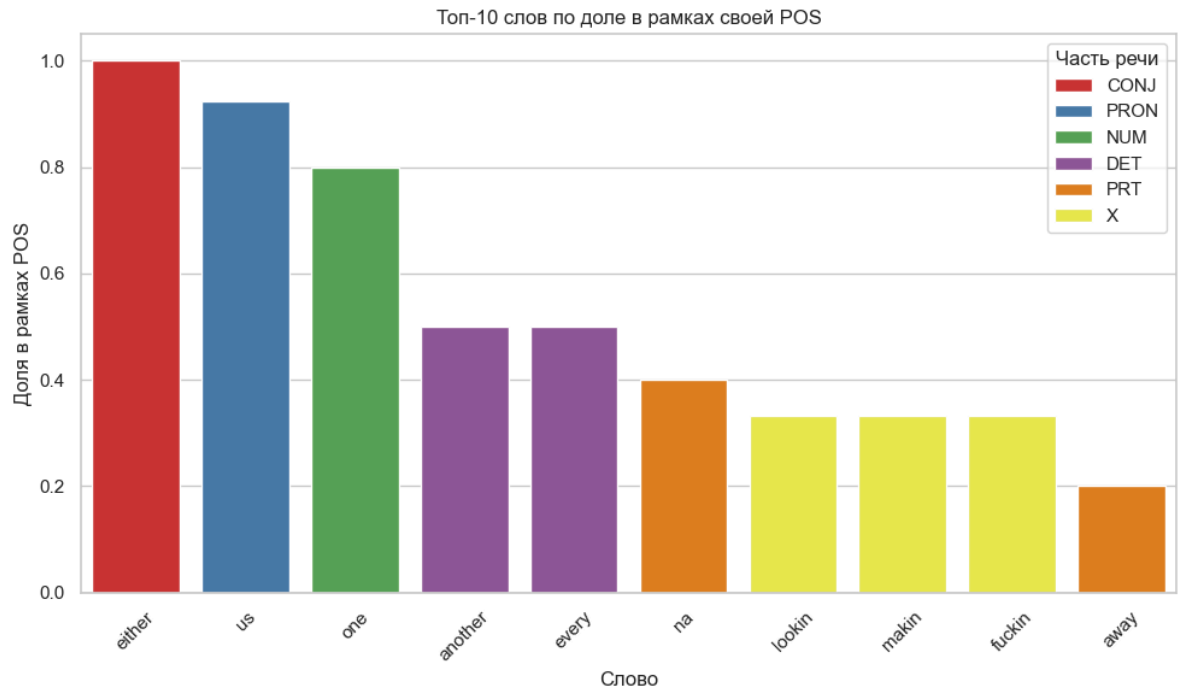

Рисунок 3 – График топ-10 слов по доле в рамках своей POS для первого сегмента.

Здесь представлены слова, которые доминируют в своих частях речи. Например, “either” практически полностью охватывает категорию союзов, а “us” преобладает среди местоимений. Аналогично “one” лидирует в разряде числительных, а “another” — среди детерминантов.

Подобная концентрация говорит о том, что именно эти лексемы выполняют ключевую функцию в своих грамматических категориях в рамках выбранного сегмента.

Наличие разговорных или неформальных форм, часто попадающих в категорию “X” и “PRT”, указывает на стилистические особенности текста и подчёркивает его диалектный характер.

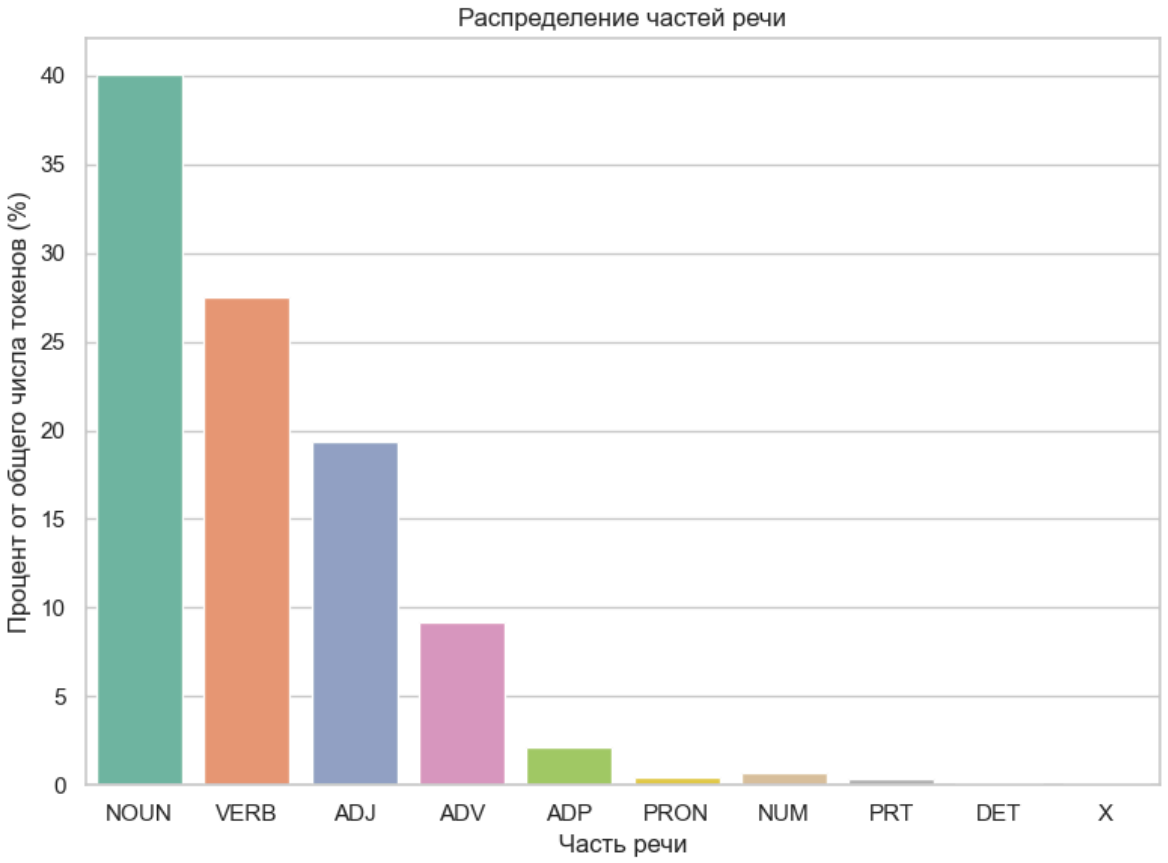

Рисунок 4 – График распределение частей речи для серединного сегмента.

Видно, что в данном сегменте доминируют существительные, занимающие около 40% всех токенов. Это может свидетельствовать о том, что текст акцентирован на описании объектов, персонажей или концепций, отражая номинативный характер повествования.

Следующими по значимости идут глаголы, указывая на присутствие динамики действий, и прилагательные, обеспечивающие описательность. Относительно невысокая доля наречий и местоимений может говорить о том, что автор предпочитает более прямое, предметно ориентированное изложение с меньшим количеством указаний на говорящего или на абстрактные характеристики действий. Подобная конфигурация частей речи может быть обусловлена как стилистическими предпочтениями автора, так и спецификой выбранного фрагмента, нужно уточнять из контекста.

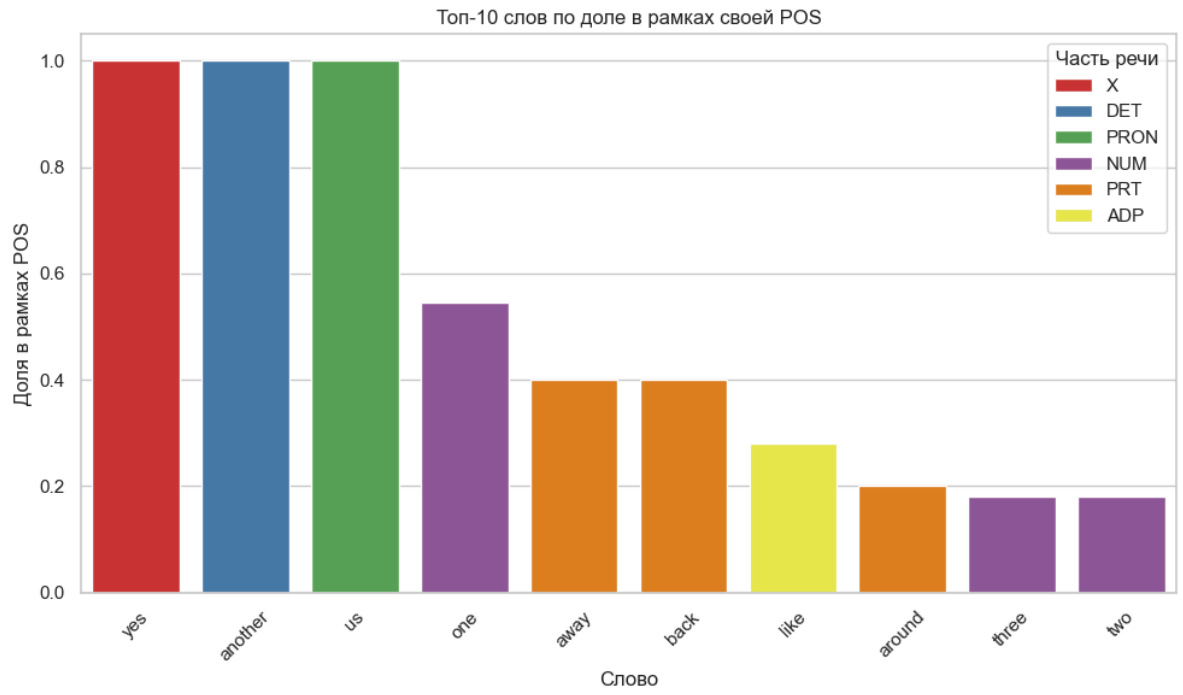

Рисунок 5 – График топ-10 слов по доле в рамках своей POS для серединного сегмента.

Здесь выделяются несколько слов, занимающих ведущие позиции в своих категориях.

Так, “yes” доминирует в категории “X”, что может говорить о разговорном стиле, а “another” преобладает среди детерминаторов, указывая на повторяющиеся конструкции “another thing”, “another time” и подобные. “Us” как ведущая форма среди местоимений говорит о частом обращении к коллективному субъекту. Преобладание слов типа “away”, “back”, “like”, “around” в других категориях может отражать склонность автора к динамичным описаниям движения, указаниям на местоположение или разговорным оборотам. Подобная картина зачастую возникает в нарративе, где автор активно использует повседневную лексику, фразовые глаголы и маркеры неформального общения.

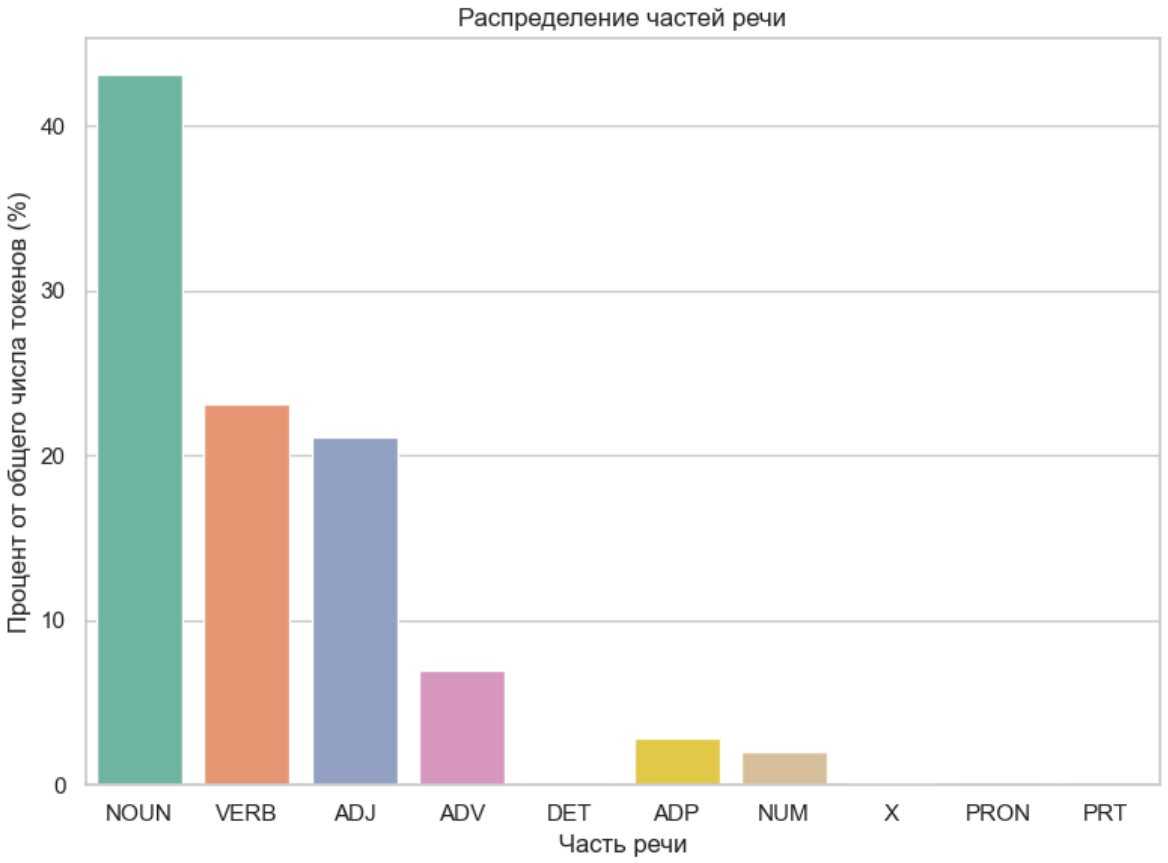

Рисунок 6 – График распределение частей речи для последнего сегмента.

В данном сегменте наибольшую долю занимают существительные, что обычно указывает на ориентацию текста на описание объектов, персонажей или понятий. Относительно высокий процент прилагательных свидетельствует о склонности автора к детальному описанию характеристик и состояний. При этом глаголы занимают меньшее место, чем в некоторых других фрагментах, предполагая не столь активное повествование, а скорее внимание к статическим деталям или рассуждениям. Небольшие доли наречий и местоимений могут отражать ограниченность экспрессивных и указательных конструкций, что, в совокупности с высокой номинализацией, создаёт впечатление повествования, фокусирующегося на сущностях и их качествах, а не на действиях или говорящем субъекте.

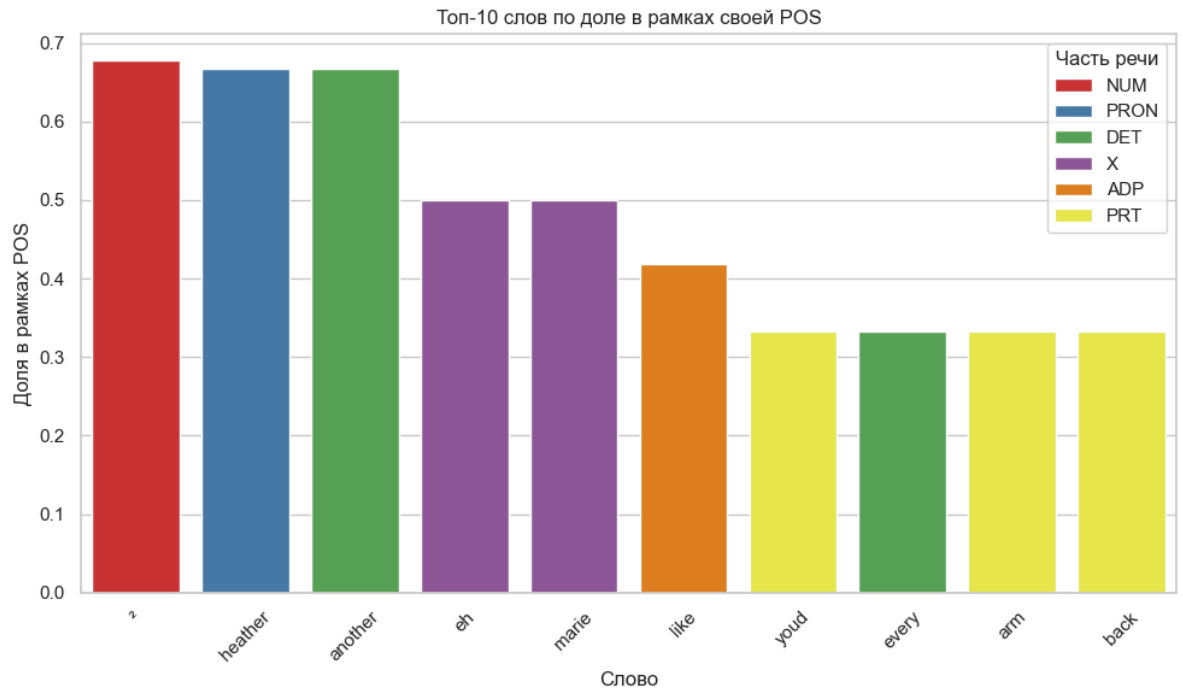

Рисунок 7 – График топ-10 слов по доле в рамках своей POS для последнего сегмента.

Здесь доминируют слова вроде “’s”, “heather”, “another”, “eh” и “mine”, которые занимают ведущие позиции в своих категориях.

Высокая доля “’s” может свидетельствовать о частотном использовании разговорных сокращений или притяжательных форм, отражая более неформальный стиль.

“Heather” и “mine” указывают на персонализированную манеру изложения, а слова типа “another”, “every” демонстрируют обобщающее или уточняющее описание. Появление “eh” в категории “X” может отражать вставные междометия или характерные элементы живой речи. Такие формы, как “youd” (you would) и “am”, говорят о склонности к использованию личных форм глаголов и сокращений.

Все вместе это указывает на текст, где разговорные конструкции, имена собственные и местоименные формы играют заметную роль, подчёркивая, возможно, диалогический характер последнего сегмента.

Вывод.

В ходе исследования были выполнены все этапы анализа лингвистических характеристик текстов Ирвина Уэлша.

Комплексный подход, от предварительной очистки и сегментации до частотного и морфологического анализа, позволил выявить ключевые особенности лексики и стиля данного писателя, а также продемонстрировал, как статистические и визуальные методы помогают интерпретировать текстовые данные.

Ссылки.

Репозиторий с исходным кодом, использованный в данной лабораторной работе: [ссылка]