ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет гуманитарных наук

Программа подготовки бакалавров по направлению по направлению 45.03.03. Фундаментальная и прикладная лингвистика

Татаринов Максим Дмитриевич

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Оценка применимости метода детектирования семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений

Руководитель
канд. комп. н.
А. В. Демидовский

Научный консультант
канд. фил. н., доцент департамента фундаментальной и прикладной лингвистики
А. Ю. Хоменко

Нижний Новгород, 2024

Оглавление

Введение

На **актуальность** настоящей работы указывают следующие факторы. Во-первых, активное изучение темы автоматического определения семантических изменений. В последние годы в работах использовались различные методы, включая статические эмбеддинги, контекстуальные эмбеддинги и заканчивая генерацией определений с помощью языковых моделей в новейших исследованиях (Kutuzov, Øvrelid [и др.], 2018), (Rodina [и др.], 2020), (Giulianelli [и др.], 2023) . При этом, абсолютное большинство исследований, посвященных моделированию определений, проводятся с использованием материала английского языка (Gardner [и др.], 2022). Для русского языка вопрос анализа семантических изменений на основе автоматически сгенерированных определений недостаточно изучен. Во-вторых, неудовлетворительное качество традиционных методов для основных потенциальных пользователей таких технологий, таких как лексикографы, историки языка и социологов. Например, лексикографам недостаточно данных только о факте сдвига значения, им хотелось бы получать описания старых и новых значений слов в пригодной для чтения форме, возможно, даже с дополнительными пояснениями. Данная проблема может решаться моделированием определений с использованием языковых моделей, при использовании которых исследователи смогут получить более наглядные результаты (Giulianelli [и др.], 2023).

Целью настоящей работы является оценка применимости метода детектирования семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений.

Из поставленной цели были сформулированы следующие **задачи**:

1. Провести анализ существующей литературы и решений по задаче детектирования семантических изменений на основе генерируемых определений.
2. Собрать датасет словарей русского языка в качестве материала для обучения модели.
3. Обучить языковую модель на данных словарей для того, чтобы генерировать определения.
4. Создать алгоритм автоматического определения семантических сдвигов на основе векторного представления.
5. Провести анализ метрик и качества обученной языковой модели и сравнить их с существующими решениями.
6. Создать алгоритм визуализации результатов.
7. Провести качественный анализ результатов работы компьютерной программы.

Объектом исследования является метод детектирования семантических изменений слов.

Предметом исследования является применимость метода детектирования семантических изменений слов с использованием нейросетевой языковой модели на основе генерируемых определений.

Для решения поставленных задач были использованы следующие **методы**:

1. Метод анализа и синтеза для создания теоретической базы для данного исследования на основе литературы.
2. Компьютерный метод для написания алгоритмов программы и обучения модели.
3. Методы обработки естественного языка для предобработки текстов.

4. Методы глубокого обучения для алгоритма автоматического определения семантических сдвигов на основе их векторного представления.
5. Метод лексико-семантического анализа (используется при оценке визуализаций алгоритма).

Новизна настоящей работы состоит в том, что для детектирования семантических изменений значений слов применяется на материале русского языка и с использованием SOTA-моделей.

Практическая значимость данной работы заключается в том, что результаты настоящей работы можно применять для определения степени семантического сдвига лексем, с наличием визуализаций и определений для каждого выявленного значения, что может быть использовано в лексикологии, где необходимы актуальные данные для построения новых словарей (Giulianelli [и др.], 2023). Кроме того, модель, позволяющая автоматически генерировать качественные словарные определения, может быть полезна в таких задачах обработки естественного языка, как анализ тональности, машинный перевод и разграничение семантической неоднозначности (Gardner [и др.], 2022).

В качестве **материала исследования** используется диахронический корпус НКРЯ, охватывающий три периода (1700—1916, 1918—1991 и 1992—2016 годы) и имеющий в совокупности 250 миллионов словоупотреблений. Данный корпус выбран, поскольку датасет слов для валидации с изменившимся и неизменившимся значением, использующийся для оценки алгоритма, основан на данном корпусе (Kutuzov, Pivovarova, 2021). Корпус был получен по запросу к авторам НКРЯ.

В первой главе исследования будут рассмотрены теоретические аспекты автоматического выявления семантических изменений, включая обзор существующих на данный момент методов решения задачи автоматического детектирования семантических изменений. Во второй главе настоящей рабо-

ты будет дана общее описание предлагаемого подхода. Третья глава будет включать в себя описание имплементации предлагаемого подхода, включая вычислительные эксперименты. В заключительной главе работы будет описан качественный анализ результатов работы алгоритма.

Апробация работы. Основные положения настоящей работы были представлены на конференции (VIII Всероссийская научная студенческая конференция НИУ ВШЭ – Нижний Новгород «Цифровые технологии в современной молодежной науке», 17 апреля 2024 г., тема доклада: «Оценка применимости метода детектирования семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений»).

Глава 1. Теоретические аспекты автоматического выявления семантических изменений

1.1. Понятия и классификации

В рамках изучения исторических изменений в лексике языка или языков, лингвисты оперируют такими понятиями, как лексические изменения, семантические изменения, грамматикализация и лексическая замена.

Лексические изменения в широком смысле охватывают все виды диахронических преобразований в словарном составе языка, в то время как в более узком значении термин относится к устареванию форм в языке, а также появлению новых, таких как заимствованные слова и неологизмы (Tahmasebi [и др.], 2021).

Семантические изменения или семантический сдвиг являются особым случаем лексических изменений, когда существующая форма (лексема) приобретает или теряет конкретное значение, что приводит к увеличению или уменьшению полисемии (Tahmasebi [и др.], 2021).

Примером таких изменений может служить эволюция английских слов, когда ранее специализированное слово для обозначения определенного вида собаки стало общим термином (dog), в то время как более раннее общее слово для *собаки* — современный аналог которого *hound* — сейчас используется для обозначения специального вида собак.

Лексическая замена представляет собой явление, когда одно слово или выражение вытесняется другим, часто синонимичным, в языке. Например, в английском языке слово *happy* изначально означало ’быть удачливым’, но затем стало означать ’счастливый’. Обратный процесс описывается на примере слова *gay*, которое раньше означало ’счастливый’, а затем стало

использоваться исключительно для обозначения гомосексуальности. Этот процесс можно рассматривать как лексическую замену, где в контексте выражения счастья слово *gay* уступает место слову *happy* (Tahmasebi [и др.], 2021), (Periti [и др.], 2024).

Грамматикализация описывает особый вид семантических изменений, когда слова с полным значением превращаются в служебные слова и, в конечном итоге, в связанные грамматические морфемы. Примером может служить развитие глагольного аффикса *-ся* из безударного возвратного местоимения формы винительного падежа (Tahmasebi [и др.], 2021), (Майсак, 2016).

В рамках настоящей работы мы будем заниматься исследованием семантических изменений лексического значения слов, не касаясь иных явлений.

Чтобы подробно разобрать изменения семантики слова, полезно обратиться к типологии, разработанной американским лингвистом Леонардом Блумфилдом (Bloomfield, 1933), (Harris, 2014). Он выделил следующие типы:

1. Сужение значения (narrowing), при котором слово начинает употребляться в более узких сферах общения за счет конкретизации своего значения. Например, Old English mete, означавшее «еда», сузилось до современного английского meat, означающего «мясо».
2. Расширение значения (widening), при котором значение слова становится более общим, способным применяться во многих сферах общения. Например, слово dog в Middle English означало «собаку определенной породы», но теперь описывает «любую собаку».
3. Гипербола (hyperbole), при которой значение слова нарочно преувеличивается. Например, слово из допрефранцузского *ex-tonare, означавшее «ударить громом», преобразовалось в французское étonner, означающее «удивлять».

4. Литота (litotes), когда значение слова нарочно подвергается преуменьшению. Например, Old English cwellan, которое означало «убить», произошло от предшествующего *['kwalljan], означавшего «пытать».
5. Деградация значения (degeneration), когда значение слова приобретает более негативное значение в течение времени. Например, Old English cnafa, первоначально обозначало «мальчик, слуга», и превратилось в knave (лжец).
6. Возвышение значения (elevation), при котором значение слова приобретает более позитивное значение в течение времени. Например, knight произошло от Old English cniht, которое означало «мальчик, слуга».
7. Метафора (metaphor), заключающаяся в изменении значения на основе скрытого сравнения. Например, прагерманское *['bitraz], означавшее «колючий», превратилось в bitter, означающее «горький».
8. Перенос значения, или метонимия (metonymy), на основе смежности или близости ассоциативных связей. Например, Old French joue, означавшее «щека», стало означать «челюсть».
9. Синекдоха (synecdoche), когда значение слова представляет отношение части к целому или наоборот. Например, прагерманское *['tu:naz] означало «забор» и превратилось в английском в «небольшой город».

1.2. Семантическое описание слов

И.А. Стернин выделяет следующие принципы, применение которых необходимо в практике семантического описания (Стернин, Рудакова, 2017).

- Принцип неединственности метаязыкового описания ментальных единиц: семантика ментальных единиц может описываться разны-

ми метаязыками, и различия в этих описаниях требуют анализа и унификации, а не считаются ошибками.

- Принцип дополнительности семантических описаний: разные семантические описания языковых единиц дополняют друг друга и могут быть объединены в обобщающее описание.
- Принцип дополнительности словарных дефиниций: разные дефиниции лексической единицы в словарях отражают различные аспекты значения, и наиболее полное описание достигается их интеграцией.
- Принцип денотативной дифференциации значений: каждому уникальному денотату, обозначаемому словом, соответствует свое значение.

Итак, для наиболее полного описания значения слова на основе данных из словарей необходимо обобщить информацию из нескольких словарей. Для этого нужно собрать и объединить определения из различных словарей, относящихся к одному и тому же современному периоду, провести денотативную дифференциацию значений и описать смысловую структуру значений.

Так, алгоритм применения метода обобщения словарных дефиниций, по мнению И.А. Стернина, заключается в следующем:

- Выписываются значения слова из всех доступных словарей.
- Составляется единый список значений слова из разных словарей.
- Уточняется список значений по денотативному принципу. Если слово номинирует некий денотат, отличный от других денотатов, фиксируется отдельное значение.
- Анализируются примеры из словарных статей. Формулируются новые значения, если они выявляются только из примеров.
- Каждое значение представляется с дефинициями из разных словарей.

- Формулируется новый более развёрнутый вариант дефиниции, если требуется точность. Например, вместо синонимического ряда «юрисконсульт, адвокат» необходимо обобщение значений синонимов и формулировка семемы: ’специалист, защищающий чьи-либо интересы в суде, оказывающий юридические консультации; то же, что юрисконсульт, адвокат (разг.)’
- Обновляется состав семантемы при отсутствии некоторых значений в словарях.
- Местоимения в метаязыковых обозначениях заменяются на архисемы для унификации.
- Функциональные и стилистические пометы обобщаются в альтернативной форме.
- Актуализируются функциональные пометы, если они не соответствуют современному употреблению.
- Если значение устарело, добавляется помета <<устар>>.
- Территориальные семы обобщаются пометой <<обл.>> при указании конкретного региона.
- Приводится совокупность примеров употребления слова из разных словарей.
- Упорядочиваются значения многозначного слова от ядерных к периферийным.
- Все значения приводятся в обобщенном виде с одним примером употребления каждый.

1.3. Обзор существующих методов

1.3.1. Предыстория

Традиционно для изучения изменений семантики слов использовались ручные методы детального анализа текстов. Из существующих исследований истории значений слов в русском языке можно привести исследование

1500 слов и 5000 связанных с ними выражений В.В. Виноградова (Виноградов, Шведова, 1999), а также книгу «Два века в двадцати словах», в деталях описывающую историю значения двадцати слов.

В. В. Виноградов посвятил годы созданию монографии по истории русских слов и выражений. Он вручную собирал материал с 20-х годов, исследуя литературный язык, говоры, славянизмы, заимствования и профессионализмы. После его смерти жена и коллеги продолжили работу, систематизируя разрозненные записи, статьи и заметки. Трудоемкий процесс включал расшифровку и проверку данных, требуя значительных усилий из-за редкости и малодоступности источников.

Книга «Два века в двадцати словах» представляет собой исследование, посвященное изменениям значений 20 интересных с точки зрения их эволюции слов в русском языке на протяжении XIX и XX веков. Ее создание стало возможным благодаря использованию Национального корпуса русского языка (НКРЯ), который является огромным электронным хранилищем текстов с начала XVI века до наших дней. Однако, несмотря на наличие такого мощного инструмента, задача создания книги оставалась сложной и требовала участия множества специалистов. Так, в процессе работы авторы консультировались с ведущими лингвистами, обсуждали результаты на семинарах и получали ценные замечания от рецензентов, особую благодарность авторы выражают тем, кто помогал выверять графики, корректировать статьи и предоставлял организационную поддержку (Данова [и др.], 2018).

Таким образом, можно сказать, что исследования изменений семантики слов вручную представляют собой чрезвычайно трудоемкий, времязатратный и многогранный процесс, требующий значительных усилий и участия множества специалистов.

Хотя ручные методы продолжают применяться в лингвистике, появление цифровых корпусов и диахронических текстовых баз данных открыло новые перспективы для исследований. Цифровизация документов в различ-

ных областях не только облегчила доступ к текстам, но и позволила разработать полуавтоматические и автоматические методы анализа. Эти методы способны значительно расширить и углубить исследования изменений семантики слов, а также упростить их проведение (Tahmasebi [и др.], 2021).

1.3.2. Эмбеддинги

Определение эмбеддингов

Множество исследований по автоматическому анализу семантических изменений обращалось к векторным представлениям слов – эмбеддингам, или вложениям. Они представляют собой метод преобразования слов в численные векторы фиксированной размерности. Это позволяет моделям машинного обучения работать с текстом, который изначально представлен в виде строк, переведённых в числовую форму. Главное их преимущество в том, что близкие по смыслу слова получают близкие векторные представления (Jatnika [и др.], 2019).

Одним из наиболее популярных методов для создания эмбеддингов является алгоритм Word2Vec, предложенный командой Google (Mikolov [и др.], 2013).

Word2Vec предлагает два основных подхода – Continuous Bag of Words (CBOW) и Skip-Gram. CBOW предсказывает текущее слово по окружающим, тогда как Skip-Gram предсказывает окружающие слова по текущему слову.

Метрики для измерения разницы между эмбеддингами

Для сравнения эмбеддингов используются различные метрики, которые измеряют степень их сходства или различия:

- Косинусное расстояние: Одна из наиболее популярных метрик, измеряющая косинус угла между двумя векторами. Если векторы направлены в одну сторону, косинусное расстояние близко к 0, если в противоположные – к 1.

$$\text{Косинусное расстояние} = 1 - \cos(\theta) = 1 - \frac{A \cdot B}{\|A\|\|B\|}$$

- Евклидово расстояние: Измеряет «прямую» дистанцию между двумя точками в пространстве.

$$\text{Евклидово расстояние} = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

Кластеризация эмбеддингов

Эмбеддинги также можно кластеризовать, группируя их по схожести. Это позволяет, например, выявлять схожие группы слов или документов.

Наиболее популярные алгоритмы кластеризации включают:

- K-means

  Алгоритм, который делит данные на K кластеров, минимизируя внутрикластерное расстояние. Данный алгоритм способен находить заранее установленное пользователем число кластеров.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

  Группирует точки по плотности, что позволяет работать с кластерами произвольной формы и игнорировать шум. Кроме того, DBSCAN способен выявлять неопределённое количество кластеров.

Основными типами векторных представлений, используемых для изучения семантических изменений, являются статические и контекстуализированные эмбеддинги, которые мы рассмотрим далее.

1.3.3. Статические эмбеддинги

До 2020 года в работах чаще используют статические эмбеддинги (Tahmasebi [и др.], 2021).

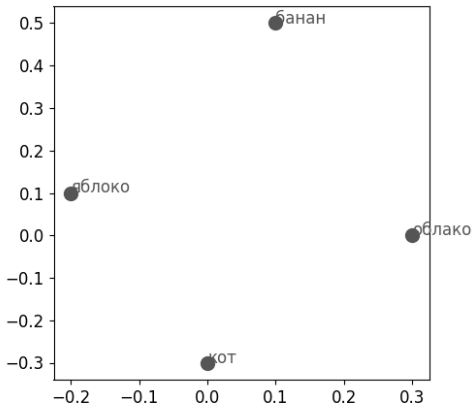
Статические эмбеддинги дают представление слова для всего корпуса, на котором модель была обучена (Tahmasebi [и др.], 2021). Например, если в корпусе использовалось 1000 слов, то пользователь получит 1000 векторов

– по одному вектору на каждое слово, который будет отражать усреднённое использование слова на протяжении всего корпуса.

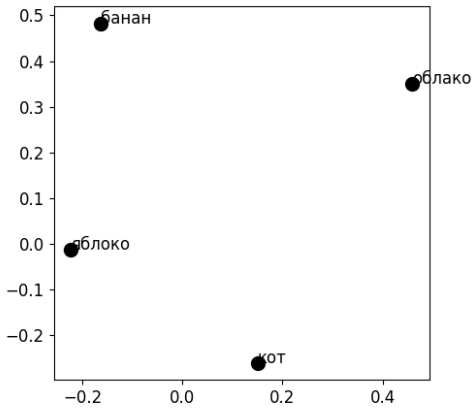
Для русского языка в качестве примера можно привести проект «Shiftry» (Kutuzov, Fomin [и др.], 2020), в котором для анализа семантических сдвигов использовались модели Word2Vec (Jatnika [и др.], 2019). Эти модели были обучены на обширном корпусе русскоязычных новостных текстов, охватывающем период с 2010 по 2020 годы, и позволили отследить диахронические изменения в употреблении слов. Поскольку при использовании статических эмбеддингов возможно производить только один вектор для одного слова в одном корпусе, корпус текстов был разделён по годам, позволив производить отдельные вектора слов для каждого года. В этом случае проявляется проблема. Хотя относительное положение эмбеддингов относительно друг друга для разных лет будет сохранено при условии сохранения значения, модели все же обучаются отдельно друг от друга и абсолютное положение векторов будет отличаться, поэтому до сравнений эмбеддингов из разных периодов, векторы необходимо «выправить», для чего в статье о проекте используется метод Прокруста.

Далее для подсчета степени семантического сдвига в статье использовалось косинусное расстояние между векторами из различных временных срезов. За некоторыми исключениями, исследования с использованием статических векторов придерживаются аналогичной проекту Shiftry методологии (Tahmasebi [и др.], 2021).

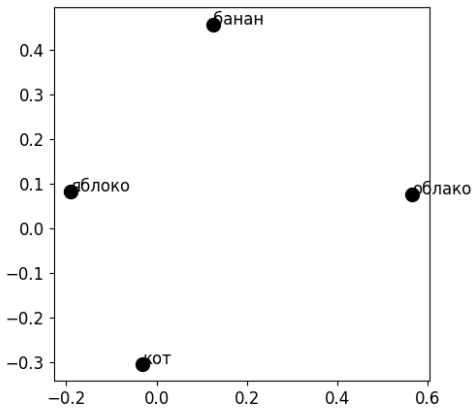
Пример работы вявления семантического изменения слова *облако* с использованием статистических эмбеддингов и метода Прокруста вы можете увидеть на графике снизу.



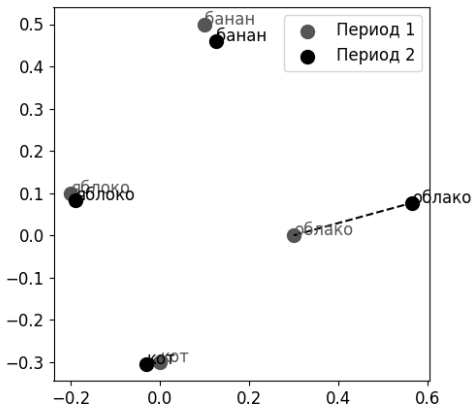
(а) Векторы для слов из эпохи 1.



(б) Векторы для слов из эпохи 2.



(в) Векторы эпохи 2 после применения метода Прокруста.



(г) Сравнение векторных пространств. Выявление сдвига слова *облако*.

Рис. 1.1.. Пример поиска семантических изменений для статических векторов с использованием метода Прокруста.

Статические эмбеддинги оставались наиболее актуальными вплоть до 2020, где показывали лучшие результаты в SemEval-2020 Task 1 (Schlechtweg [и др.], 2020). Они эффективно моделируют значение слов в зависимости от обучающего корпуса без опоры на объемные предобученные модели, превосходя по этому качеству модели, основанные на совстречаемости слов.

Среди недостатков статических эмбеддингов можно отметить:

- необходимость большого объема слов в корпусах для стабильности эмбеддингов;
- необходимость выравнивания моделей, обученных на отдельных наборах данных, соответствующим временным срезам, что может вносить шум;
- кроме того, такие модели моделируют только усреднённое значение слова на основе его употребления в корпусе, не позволяя различать отдельные значения слова.

1.3.4. Контекстуализированные эмбеддинги

1.3.4.1. Определение

Статические модели вложений слов присваивают каждому слову (лемме) один и тот же вектор независимо от контекста, в то время как современные достижения в области обработки естественного языка позволили разработать модели, обеспечивающие получение контекстуализированных представлений высокого качества. Данные модели отличаются тем, что на этапе вывода токенам присваиваются различные вложения в зависимости от их контекста, что позволяет различать отдельные значения одного слова.

Например, в работе Кутузова приводится наглядный пример работы контекстуальных эмбеддингов (Kutuzov, 2020). На графике снизу вы можете увидеть проекцию вложений ELMo для слова *cell* (клетка) в 2000 годы в английском языке. На визуализации видны три кластера, отражающие различные значения слова *cell* (клетка). Два кластера, расположенные слева, представляют традиционные значения: внизу биологическое значение, связанное с клетками живых организмов, и вверху тюремное, где *cell* означает камеру. Кластер, который занимает правую сторону графика и чётко отделён от левых, демонстрирует современное употребление слова *cell*, где оно используется в контексте мобильной связи.

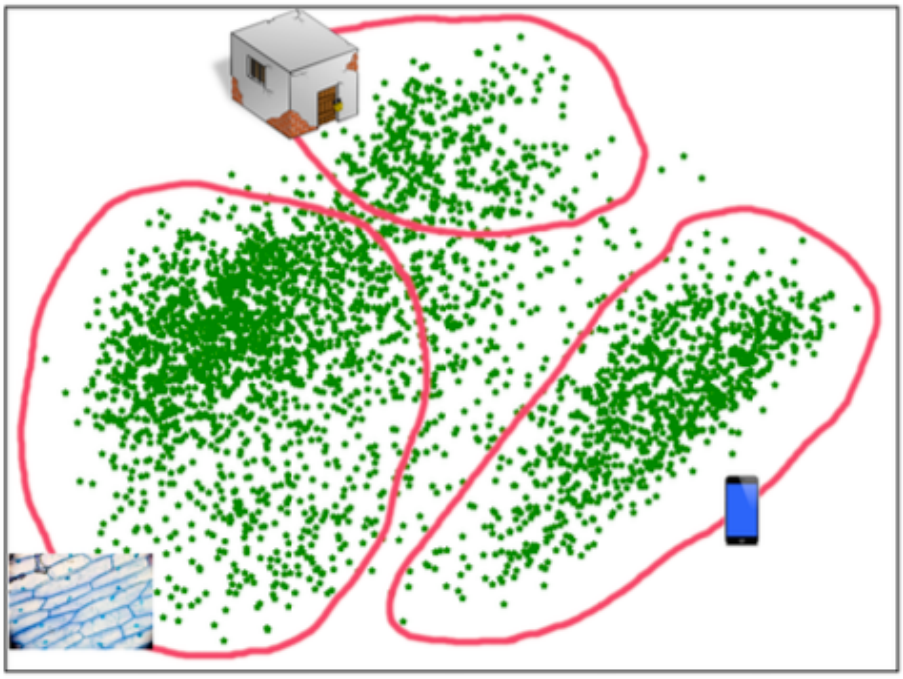


Рис. 1.2.. Проекция репрезентаций использований слова *cell* в 2000-ые годы

Применение контекстуализированных векторных представлений задало новый стандарт для высококачественных, чувствительных к контексту представлений в обработке естественного языка. В статье, где исследователи использовали предварительно обученные модели BERT и ELMo, настроенные на полном корпусе Русского национального корпуса, было обнаружено, что эти модели показывают значительную корреляцию с человеческими оценками при определении диахронического семантического изменения слов в русском языке (Rodina [и др.], 2020).

1.3.4.2. Соревнование по выявлению семантических изменений Rushifteval

Одни из последних работ по теме автоматического выявления семантических сдвигов для русского языка были написаны в рамках соревнования RuShiftEval, прошедшего в 2021 году (Kutuzov, Pivovarova, 2021). В ходе него участники должны были рассмотреть три исторических периода русского языка и общества: предсоветский (1700-1916), советский (1918-1990) и постсоветский (1992-2016). Исследование базировалось на наборе данных RuShiftEval, который состоит из 111 русских существительных (99 в тесто-

вом наборе и 12 в наборе для разработки), вручную аннотированных по степени изменения их значения в трех парах временных периодов.

Аннотаторам предлагалась задача, которую можно свести к оценке семантической связь значений целевого слова в парах предложениях из разных временных периодов. Оценки (от 1 до 4) отражали степень семантического родства между значениями слова, где 1 обозначало отсутствие связи между значениями, а 4 – их совпадение. Затем индивидуальные оценки усреднялись, формируя общую меру семантической родственности между употреблениями слова в разные временные периоды. Такая задача как правило называется Word-in-Context или WiC.

Вы можете увидеть пример вхождения в датасет Rushifteval снизу в таблице. Оценка аннотатора 1 означает разное значение, когда оценка 4 – одинаковое.

Таблица 1.1.. Пример вхождения в DWUG

Слово	радикал
Предложение 1	А вот социалисты и наши ***радикалы*** -- это совсем другого подбора.
Предложение 2	При некоторых условиях при отдельных элементарных реакциях возникают сразу два ***радикала,*** что приводит к разветвлению цепи.
Среднее	1.0
Аннотатор 1	1
Аннотатор 2	1
Аннотатор 3	1

Для каждого из 99 целевых русских слов участники должны были представить три значения, соответствующих семантическому изменению в упомянутых парах временных периодов. Эти значения использовались для

построения трех ранжирований: RuShiftEval-1 (изменение значений между досоветским и советским периодом), RuShiftEval-2 (изменение значений между советским и постсоветским периодом) и RuShiftEval-3 (изменение значений между досоветским и постсоветским периодом). В качестве метрики оценки использовалась ранговая корреляция Спирмена между ранжированием слов, сгенерированным системой, и золотым ранжированием, полученным в ручной аннотации. После этого бралась средняя оценка между ранжированиями.

Победители вышеупомянутого соревнования (команда GlossReader) указывают, что проблемой в существующих решениях являлось то, что эмбеддинги несут в основном информацию о форме слова, а не значении (Rachinskiy, Arefyev, 2021). Чтобы решить это, они дообучали модель XLM-R на задаче генерации эмбеддингов, максимально близким к таким, какие получены на соответстующим использованиям слов словарным определениям (Conneau [и др.], 2019).

При дообучении их система включает в себя два отдельных энкодера на основе XLM-R: Энкодер контекстов для кодирования предложения с целевым словом и энкодер глоссов для кодирования определения слова. Система оценивает возможные значения смысла слова путём сравнения векторных представлений слова и его определений. При этом для обучения использовались данные только по английскому языку, но модель также показала хорошие результаты для русского языка.

Далее, исследователи получали эмбеддинги контекстов слов с помощью дообученного энкодера контекстов, высчитывали расстояние с помощью различных метрик расстояния, самым эффективным из которых были евклидово расстояние с нормализацией, после чего логистическая регрессия приводила значения к формату в датасете, то есть к значениям от 1 до 4.

Авторы статьи предоставляют доступ к части исходного кода их исследования (URL: https://github.com/myrachins/RuShiftEval).

Так, были опубликованы следующие компоненты:

1. Код, предназначенный для генерации прогнозов на основе заранее вычисленных эмбеддингов, полученных с использованием модели.
2. Код для оценки результатов.

В соответствии с инструкциями, данными авторами, мы запустили доступный код, в следствие чего были полуены высокие результаты, совпадающие с тем, что сообщают авторы в своей работе:

Таблица 1.2.. Коэффициэнты корреляции

Пары периодов	Коэффициент корреляции
Среднее	0.8021
pre-Soviet:Soviet	0.7808
Soviet:post-Soviet	0.8032
pre-Soviet:post-Soviet	0.8223

Среди недостатков работы можно отметить неспособность модели корректно выявлять значения тех слов, которые отличаются от ближайших аналогов в английском, например, «пионер», связанный с коммунистической идеологией и не соответствующий в полной мере слову «scout».

Кроме того, команда DeepMistake представила решение, занявшее в соревнование второе место (Arefyev [и др.], 2021). Однако, они смогли доработать его и повысить результаты до первого уже после окончания соревнования.

Исследователи обучали модель XLM-R на обширном многоязычном датасете Word-in-Context, а затем дообучали ее на наборе данных RuSemShift для настоящей задачи.

Исследователи выложили исходный код полностью и предлагают возможность воспроизвести их результат (URL: https://github.com/Daniil153/

DeepMistake). Значения метрик, сообщенные исследователями, воспроизводятся.

Таблица 1.3.. Коэффициэнты корреляции с использованием IsoReg

Пары периодов	Коэффициент корреляции
Среднее	0.8494
pre-Soviet:Soviet	0.8563
Soviet:post-Soviet	0.841
pre-Soviet:post-Soviet	0.8511

Среди недостатков статьи можно выделить то, что авторы не предоставляют возможность визуализации или интерпретации результатов, кроме непосредственно получившегося значения метрики.

Тем не менее, применимость таких методов была подвергнута сомнению в работе Giulianelli et al. (Giulianelli [и др.], 2023), где ставится под сомнение широкая практичность ранее упомянутых подходов. Они утверждают, что такие методы практически неинтерпретируемы, поскольку они не дают описаний значений слов, а лишь бинарные результаты наличия или отсутствия семантического изменения. Исследование, которое в наибольшей степени занимается этой проблемой, - это GlossReader, где исследователи предлагают способ визуализации и интерпретации результатов. Однако у этого метода есть свои недостатки, обсуждаемые выше, а также необходимость определять заранее значения, для которых будет строиться визуализация. Учитывая эти факты, новый подход, включающий моделирование определений, вызывает интерес для задачи обнаружения семантических изменений.

1.3.5. Моделирование определений

Моделирование определений (также генерация определений) описывается исследователями как «задача генерации определений слов в формате,

который может быть прочитан людьми, как те, что можно найти в словарях», где на вход подается целевое слово, пример его использования, а на выход ожидается сгенерированное определение (Giulianelli [и др.], 2023). Вы можете увидеть пример в таблице ниже.

Таблица 1.4.. Пример моделирования определений

Пример использования	‘Примерно половина солдат в наших стрелковых взводах были призывниками, которых мы обучали около шести недель.’
Целевое слово	призывник
Сгенерированное определение	‘Человек, который подлежит призыву в вооруженные силы’

Начало интереса к моделированию определений как теме исследования в области обработки естественного языка можно отнести к работе Noraset et al. (Noraset [и др.], 2016). Они были среди первых, кто исследовал потенциал использования векторных представлений слов для автоматической генерации определений. Изначально была поставлена упрощенная задача с моносемантическими словами, которые, как правило, имеют одно значение и, следовательно, одно определение.

Однако оставалась нерешенной проблема многозначных слов. Gadetsky et al. выделили важное условие для моделирования определений: необходимость контекста для точного захвата нюансов языка (Gadetsky [и др.], 2018). В своем исследовании они включили примеры предложений для предоставления контекста модели, что оказалось решающим шагом в возможности модели справляться с полисемией и улучшении ее производительности.

Несмотря на достижения, сделанные вышеупомянутыми исследователями, область моделирования определений все еще сталкивалась с значительными проблемами. Huang et al. выявили наличие таких проблем, как проблема слов вне словаря, когда модели сталкиваются с трудностями в работе со словами, не встречавшимися во время обучения, а также проблемы избыточной и недостаточной специфичности в определениях (Huang [и др.], 2021). Исследователи сообщают: «Избыточно специфичные определения представляют узкие значения слов, в то время как недостаточно специфичные определения представляют общие и нечувствительные к контексту значения.» Huang et al. решили эти проблемы, используя предварительно обученную модель энкодера-декодера, а именно Text-to-Text Transfer Transformer (T5), и ввели механизм ранжирования, предназначенный для тонкой настройки специфичности генерируемых определений. Метод был протестирован на стандартных наборах данных для оценки и показал значительное улучшение по сравнению с предыдущими методами.

Самой актуальной работой по теме использования сгенерированных большими языковыми моделями определений для автоматического выявления семантических изменений является статья Giulianelli et al. (Giulianelli [и др.], 2023).

Авторы определяют задачу генерации определений следующим образом: для заданного слова w и примера использования s (предложения, содержащего w) необходимо сгенерировать определение d на естественном языке, которое будет грамматически корректным и точно передавать значение слова w в контексте его использования. Для генерации определений они используют модель Flan-T5, версию трансформера T5, большую генеративную языковую модель, дополнительно обученную на 1,8 тысячах задач по обработке естественного языка.

Для дообучения модели авторы используют три датасета, каждый из которых содержит определения слов, сопровождаемые примерами употреб-

ления: WordNet, данные Оксфордского словаря и CoDWoE, основанный на определениях и примерах, извлеченных из Викисловаря.

Для демонстрации работы со сгенерированными определениями авторы работы используют датасет, в котором слова представлены в графах диахронного использования слов (Diachronic Word Usage Graphs, DWUG), взвешенных, ненаправленных графах, узлами которых служат примеры использования слов, а веса рёбер отражают семантическую близость пар употреблений. DWUG созданы на основе многоэтапного процесса человеческой аннотации, в ходе которого аннотаторы оценивали семантическую связность пар употреблений слов по 4-балльной шкале по схожей схеме с датасетом сорвенования Rushifteval.

Прежде всего, авторы исследования проводят анализ корреляции между близостью пар слов в DWUG и контекстуальными эмбеддингами токенов, эмбеддингами предложений примеров использования, а также сгенерированными определениями и эмбеддингами, полученными на основе них. Результаты показали, что сгенерированные определения обладают более высокой степенью корреляции с данными из DWUG, чем традиционно полученные эмбеддинги.

Кроме того, исследователи обнаружили, что эмбеддинги определений образуют более плотные и четко определенные кластеры по сравнению с традиционными эмбеддингами, что делает их подходящими для представления значений слов.

Далее авторы исследовали возможность присваивать кластерам, полученным на основе данных из DWUG, соответствующие им определения. Для обобщения определений в одном кластере авторы использовали самое прототипическое из них. Они представляли все определения с помощью их эмбеддингов предложений и выбирали в качестве прототипичного определение, эмбеддинг которого наиболее похож на среднее значение всех эмбеддингов в кластере.

Авторы приходят к выводу, что сгенерированные определения слов могут играть роль семантического представления слов, аналогичному традиционным эмбеддингам. Они находят большие языковые модели достаточно развитыми для генерации определений простым промптом. При этом полученные таким образом определения превосходят по качеству традиционные эмбеддинги и являются более наглядными.

1.4. Метрики оценки качества сгенерированных определений

Когда мы запускаем модель на тестовой части датасета, мы можем сравнить каждое полученное определение с «эталонным определением» из датасета.

Одним из основных способов оценки сгенерированных определений являются метрики сходства строк (Gardner [и др.], 2022).

Примером такой метрики является BLEU (Bilingual Evaluation Understudy). BLEU — это стандартный алгоритм, используемый для оценки машинных переводов (Papineni [и др.], 2002). BLEU рассчитывается как точность n-грамм, то есть отношение правильных n-грамм к общему числу n-грамм в выходной строке. Недостатком BLEU является то, что он оценивает только совпадение n-грамм. Формула для расчета BLEU выглядит следующим образом:

$$\text{BLEU} = \exp\left(\text{BP} + \sum_{n=1}^{N} w_n \log p_n\right)$$

где BP — это brevity penalty, который вычисляется как:

$$\text{BP} = \min\left(1 - \frac{L_r}{L_c}, 0\right)$$

L_r — длина эталонного определения, L_c — длина сгенерированного определения, w_n — веса для n-грамм, $p_n = \frac{\text{Число пересечений n-грамм}}{\text{Общее число n-грамм в сгенерированном тексте}}$ — точность n-грамм,.

Например, для текстов *Кошка залезла на шкаф* и *Кошка залезла на стол* будет рассчитываться так:

Для простоты рассмотрим биграммы более подробно, а остальные параметры сразу укажем.

1. Биграммы текста 1: *Кошка залезла*, *залезла на*, *на шкаф*

   Биграммы текста 2: *Кошка залезла*, *залезла на*, *на стол*
2. Считаем точность биграмм:

   Совпавшие биграммы: *Кошка залезла*, *залезла на*

   Точность биграмм (p2): 2/3
3. Штраф за краткость (BP):

   Длина эталонного текста (Lr): 4 слова

   Длина перевода (Lc): 4 слова

   BP = min(1, Lr / Lc) = min(1, 4/4) = 1
4. Итоговый расчет BLEU:

   Для униграмм (p1): 0.75

   Для триграмм (p3): 0.5

   Для четырехграмм (p4): 0.0
5. Подставляем значения (вместо логарифма 0.0 подставим 0.1, чтобы избежать деления на ноль):

$$\text{BLEU} = 1 \cdot \exp\left(\frac{1}{4}(\log 0.75 + \log 0.6667 + \log 0.5 + \log 0.1)\right)$$

Итак, BLEU score для *Кошка залезла на шкаф* и *Кошка залезла на стол* составляет примерно 59.46 (от 100).

Еще одной популярной метрикой является ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation). ROUGE измеряет совпадение n-грамм между эталонным и кандидатом на определение (Lin, 2004). ROUGE-L — это модифицированная версия ROUGE, которая использует наибольшую общую подпоследовательность для измерения сходства между двумя определениями. Преимуществом ROUGE-L является то, что он автоматически

определяет самые длинные последовательные общие n-граммы. Формула для расчета ROUGE-L выглядит следующим образом:

$$\text{ROUGE-L} = \frac{LCS(X, Y)}{L_r} \tag{1.1}$$

где $LCS(X, Y)$ — длина наибольшей общей подпоследовательности между строками X и Y, а L_r — длина эталонного определения.

Например, для *Кошка залезла на шкаф* и *Кошка залезла на стол* ROUGE-L составит 0.75 (от 1).

Тем не менее, у таких метрик есть недостаток. Они анализируют не семантику слов, а только буквальное совпадение n-грамм между ними.

Примером более продвинутой метрики является BERTScore. BERTScore (Bidirectional Encoder Representations from Transformers) — это метрика, которая вычисляет оценку сходства между кандидатом и эталонным определением на основе предварительно обученных контекстуальных эмбеддингов из BERT (Zhang [и др.], 2020). BERTScore вычисляет точность (precision), полноту (recall) и F1-меру. Формулы для расчета BERTScore выглядят следующим образом:

$$\text{Precision} = \frac{1}{|C|} \sum_{x \in C} \max_{y \in R} \text{sim}(x, y) \tag{1.2}$$

$$\text{Recall} = \frac{1}{|R|} \sum_{y \in R} \max_{x \in C} \text{sim}(x, y) \tag{1.3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1.4}$$

где C — множество токенов кандидата, R — множество токенов эталона, и $\text{sim}(x, y)$ — косинусное сходство между эмбеддингами токенов x и y.

Например, для текстов *выложил документ в удалённое хранилище*, *выгрузил файл в облако* значение BERT-F1 -- 0.7621 (от 1), несмотря на непосредственное совпадение лишь одного слова.

1.5. Классификация ошибок сгенерированных определений

В работах Huang et al. и Noraset et al. проводится оценка выборки сгенерированных определений, для которой они предсталяют классификацию ошибок, часто встречаемых в сгенерированных определениях (Noraset [и др.], 2016), (Huang [и др.], 2021). Далее представлена классификация ошибок на основе классификации Huang et al. Исключением является «Избыточность или чрезмерное использование общих фраз», которую опустили в исследовании из-за её редкости. Данный тип ошибки будет описан нами.

Таким образом, ошибками являются:

1. **Избыточное конкретизирование**: Определение даёт слишком узкое описание слова.

   Пример: *дуновение* - ’движение неприятного запаха через воздух’. Слово *дуновение* имеет в Большом толковом словаре С. А. Кузнецова определение *«Лёгкий порыв ветра; движение воздуха.»* (Кузнецов, 1998). Ошибка заключается в том, что определение ограничивает значение слова компонентом ’неприятный запах’, тогда как оно имеет более широкое значение.

2. **Недостаточное конкретизирование**: Определение даёт слишком общее или неполное описание.

   Пример: *капитан* - ’член команды’. Слово *дуновение* имеет в Большом толковом словаре С. А. Кузнецова определение *«Командир, начальник судна.»* Ошибка заключается в отсутствии указания на ключевые семантические компоненты руководящей роли капитана, что приводит к недостаточной конкретизации значения слова.

3. **Самореференция**: Определение содержит само слово или его производные.

   Пример: *самосознание* - ’состояние, при котором у человека присутствует самосознание’. Ошибка заключается в том, что в определении не раскрывается семантическая сущность слова, так как для описания значения используется само же слово.

4. **Неправильная часть речи**: Модель даёт такое определение, которое относится к лексеме другой части речи.

   Пример: *стекло* - ’переместиться вниз, сбежать (о жидкости)’ для контекста *После урока стекло оказалось сломанным.* Верным определением здесь было бы ’изделие из этого материала, предмет’

5. **Противоположное значение**: Определение выражает смысл, противоположный истинному значению слова.

   Пример: *внутрь* - ’ненаправленный в центр’.

6. **Близкая семантика**:

   Определение верно передает только отдельные стороны денотата, но упускает ключевые.

   Пример 1: *машина* - ’устройство с автоматическими функциями’.

   Пример 2: *милый* - ’имеющий признаки ребёнка’.

7. **Некорректность**: Определение полностью ошибочно и не соответствует значению слова.

   Пример 1: *первый* - ’следующий после всех остальных в списке предметов’.

   Пример 2: *упаковать* - ’сделать внезапный звук’.

8. **Избыточность или чрезмерное использование общих фраз**: Определение содержит повторы или избыточные формулировки.

   Пример 1: *пропан* - ’горючий газ, используемый для горения газа’.

   Пример 2: *спутник* - ’тот, кто совершает путь, путь вместе с кем-л.’.

9. **Корректность определения**: Определение точно и в нужной степени полно передает значение слова и не содержит какие-либо из вышеописанных ошибок.

   Пример: *винодельня* - ’заведение, помещение для изготовления вина’.

Выводы

В последние годы возврос интерес к полуавтоматическим и автоматическим подходам выявления семантических изменений, основанным на векторных представлениях слов (эмбеддингах). Статические эмбеддинги, обучаемые на корпусах текстов, позволяют выявлять семантические сдвиги, однако имеют ограничения, связанные с необходимостью выравнивания моделей и неразличением значений отдельного слова. Контекстуализированные эмбеддинги, генерируемые современными языковыми моделями, показывают более высокую точность в задачах обнаружения семантических изменений, поскольку учитывают контекст употребления слов.

Новым перспективным направлением является использование моделирования определений слов на основе генеративных больших языковых моделей. Сгенерированные определения демонстрируют более высокую корреляцию с данными о семантической близости слов, чем традиционные эмбеддинги, и могут служить семантическим представлением слов, превосходящим по качеству векторные репрезентации.

Тем не менее, моделирование определений остаётся недостаточно исследованной темой в контексте семантических изменений, особенно на материале русского языка.

Вместе с тем, существует несколько способов оценки сгенерированных определений: от метрик сходства текста до классификаций для качественного анализа.

Глава 2. Предлагаемый подход

Методология настоящей работы заключается в следующем.

В данном исследовании применяется метод моделирования определений слов, который характеризуется как задача генерации определений в формате, доступном для чтения людьми, аналогично тем, что представлены в словарях. Входными данными для модели служат целевое слово w и пример его использования в предложении s, на выходе генерируется определение d.

Этап 1: Обучение

Пусть имеется датасет $D = \{(w_i, s_i, d_i)\}_{i=1}^{N}$, где w_i — это слово, s_i — пример его использования в предложении, а d_i — определение данного слова. Целью первого этапа является обучение генеративной большой языковой модели. Модель M обучается на данных словаря, содержащих определения слов и их контекст использования, с целью генерации определений слов, аналогичных тем, что можно найти в словарях. Формально, модель M обучается минимизировать функцию потерь L, определенную как:

$$L(M) = \sum_{i=1}^{N} \text{loss}(M(w_i, s_i), d_i)$$

где loss — это функция потерь, измеряющая расхождение между сгенерированным определением и эталонным определением.

Этап 2: Тестирование

На втором этапе проводится тестирование обученной модели по ряду метрик. Этот этап включает:

1. Генерацию определений для тестовой выборки $D_{\text{test}} = \{(w_j, s_j, d_j)\}_{j=1}^{M}$ и последующую оценку качества сгенерированных определений $\hat{d}_j = M(w_j, s_j)$ относительно эталонных определений d_j. Для оценки качества используются меры сходства текста, которые оценивают

формальную схожесть или семантику сравниваемых текстов. Формально, метрика metric определяется как:

$$\text{metric} = \frac{1}{M} \sum_{j=1}^{M} \text{similarity}(\hat{d}_j, d_j)$$

где similarity — функция, измеряющая сходство между сгенерированным и эталонным определениями.

2. Проверку модели на датасете, посвященном детектированию семантических изменений и содержащих оценки аннотаторов между парами вхождений.

Для каждой пары использований слов в датасете генерируются определения $(\hat{d}_{k1}, \hat{d}_{k2})$, которые затем векторизуются $(\vec{d}_{k1}, \vec{d}_{k2})$. Полученное значение расстояния между векторизованными определениями $\text{dist}(\vec{d}_{k1}, \vec{d}_{k2})$ сравнивается с оценками аннотаторов в датасете.

Этап 3: Визуализация

Для визуализации семантических изменений слов полученные с помощью модели определения векторизуются с помощью векторизатора V.

$$\mathbf{v}_i = V(d_i)$$

Определения, имеющие семантически близкие значения, группируются с использованием алгоритма кластеризации C.

$$\{K_1, K_2, \ldots, K_m\} = C(\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\})$$

где m - количество кластеров.

Для каждого кластера K_j выбирается прототипическое определение $\hat{d}_{\text{proto}}$, которое определяется как наиболее близкое к центру кластера. Затем создаются столбчатые диаграммы, отражающие частоту употреблений различных значений слова во времени, с обозначением категорий цветовой градацией и легендой.

Этап 4: Качественный анализ

Проводится качественный анализ результатов, полученных с помощью визуализаций, на наборе слов $\{w_k\}_{k=1}^{P}$. Полученные определения $\{\hat{d}_k\}_{k=1}^{P}$ сравниваются с таковыми из словарей, а статистическая информация о частоте использования тех или иных значений сравнивается с информацией из лексикографических изданий.

Глава 3. Имплементация автоматического выявления семантических изменений

3.1. Обучение языковой модели на данных тезауруса

В качестве модели была выбрана FRED-T5-1.7B, являющаяся одной из новейших языковых моделей, выпущенных SberDevices и обученных с нуля на материале русского языка (Zmitrovich [и др.], 2023). Для выбора модели мы использовали бенчмарк для оценки продвинутого понимания русского языка «RussianSuperGLUE» (Shavrina [и др.], 2020). В бенчмарке присутствуют шесть групп задач, охватывая общую диагностику языковых моделей и различные лингвистические задачи: понимание здравого смысла, логическое следование в естественном языке, рассуждения, машинное чтение и знания о мире. FRED-T5-1.7B занимает самое высокое место в лидерборде данного бенчмарка, со значением 0.762, уступая лишь результатам выполнения данных заданий людьми со значением 0.811, что свидетельствует о ее способности к выдающемуся языковому пониманию и анализу. Таким образом, FRED-T5-1.7B представляется наиболее подходящей языковой моделью для задачи генерации определений.

Одной из ключевых особенностей модели FRED-T5-1.7B является наличие денойзеров. Денойзеры — это специальные механизмы, задача которых состоит в очистке текста от шума, то есть в восстановлении удаляемых или искажаемых частей текста. В модели используется семь различных денойзеров, каждый из которых выполняет уникальную функцию в процессе обучения. Основные задачи денойзеров включают в себя восстановление удаленных участков текста, а также продолжение текстовых последовательностей.

В настоящей работе при работе с моделью используется денойзер, помеченный спецтокеном «<LM>», который задействован в задаче продолжения последовательности текста.

В качестве материала, используемого для обучения модели, выступил датасет, включающий в себя материал из из МАС – «Малого академического словаря» (Евгеньева, 1981-1984). Материал Малого академического словаря был получен с помощью скраппера, написанного на языке Python. В загруженном наборе данных в каждом вхождении присутствовали идентификатор статьи, лексема, про которую написана данная статья, а также определения с примерами использования.

Таблица 3.1.. Информация о лексеме из МАС

Лексема	**Определения и примеры использования**
прозябнуть	**сильно озябнуть, промёрзнуть:** *Я* и без того прозяб, инстинкт тянет меня согреться, а какой-то нелепый долг повелевает лезть в холодную воду. Усталость возьмет свое, тогда можно жестоко прозябнуть и опасно заболеть. **прорасти:** Сперва надо его в землю посадить, потом ожидать, покуда в нем произойдет процесс разложения, потом оно даст росток, который прозябнет, в трубку пойдет, восколосится и т. д.

Полученный материал был очищен от вхождений, не имеющих при себе примеров использования, информативных определений, например, «Состояние по знач. глаг. линять», или не содержащих определений вовсе, а также имеющие такие определения, которые представляют грамматическую информацию о слове вместо лексического значения, например, «наречие к

причастию приглашающий». На выходе было получено 122 тысяч 350 вхождений.

Примеры и слова были отформатированы под формат запроса модели. В начале после слова «Контекст» шел пример использования слова, после чего шла фраза «Определение слова», в которую включалось само слово. Таким образом, на вход модель принимает лексему и контекст, в которой она употреблялась, а на выход ожидается сгенерированное определение.

Таблица 3.2.. Пример отформатированного запроса модели

Поле	**Значение**
input_text	<LM>Контекст: ”Усталость возьмет свое, тогда можно жестоко прозябнуть и опасно заболеть.” Определение слова ”прозябнуть”:
target_text	Сильно озябнуть, промёрзнуть.

FRED-T5-1.7B была дообучена на полученном из «Малого академического словаря» материале в течение трёх эпох с линейным шагом обучения 0.001, размером батча 16 и оптимизатором Adafactor на одной видеокарте RTX 3090. Для ускорения обучения и экономии видеопамяти использовалась технология LoRa со следующими параметрами: r – 32, alpha – 64, dropout – 0.1, что позволило уменьшить количество обучаемых параметров до 14155776 (0.8% от общего числа параметров), сэкономить используемую видеопамять и ускорить обучение. В качестве метрики потерь (лосса) используется кросс-энтропия. Более подробный обзор гиперпараметров модели, а также хода ее обучения доступен в приложении Б.

3.2. Тестирование модели метриками сходства строк

Для оценки качества обучения модели используются метрики BLEU и ROUGE-L, которые оценивают формальную схожесть текста: BLEU оценивает точность совпадений n-грамм в сгенерированном тексте по сравне-

нию с эталонным текстом (Papineni [и др.], 2002), а ROUGE-L измеряет схожесть между сгенерированным текстом и эталонным текстом на основе наибольшей общей последовательности слов (Lin, 2004). Также использовалась метрика BERT-F1, учитывающая семантику сравниваемых текстов благодаря использованию контекстуальных эмбеддингов при подсчете значения метрики (Zhang [и др.], 2020). Использование нескольких метрик позволяет получить более полную картину качества модели, поскольку каждая из них оценивает разные аспекты сгенерированного текста. Как традиционные BLEU и ROUGE-L, так и более современный BERT-F1 активно используются в задачах обработки естественного языка, в том числе в задачах генерации текста. Так, в обзорной статье по моделированию определений утверждается, что на момент выпуска статьи BLUE использовался в 9 научных публикациях, ROUGE-L и BERTScore – в 3 (Gardner [и др.], 2022). Кроме того, данные метрики используются и в более новых работах. Так, в настоящей статье результаты данных метрик будут сравниваться с таковыми из статьи Giulianelli M. et al., где сообщаются результаты трёх вышеперечисленных метрик при обучении модели T5 для задаче генерации определений на английском языке (Giulianelli [и др.], 2023).

В данной работе использовались версии этих инструментов, взятые из библиотеки evaluate (,)

Таблица 3.3.. Результаты дообучения FRED-T5-1.7B на датасете MAC (больше – лучше)

Метрика	Значение
BLEU	11.02%
ROUGE-L	29.36%
BERT-F1	75.22%

Тестирование показывает низкие значения метрик BLEU и ROUGE-L, что говорит о том, что модель формулирует определения не так, как они на-

писаны в тестовой выборке. Тем не менее, это не означает некорректность генерируемых определений. Судя по результатам метрики BERT-F1, можно сказать, что семантически сгенерированные определения совпадают с таковыми из тестовой выборки. Такие результаты можно объяснить тем, что модель выдаёт семантически верные, но иначе сформулированные определения. Например, для слова *отощалый* в контексте «Иван Бедный сидел в развалившейся лачуге, худой, отощалый.» ожидаемым определением являлось ’ставший тощим, отощавший’, но моделью было сгенерировано ’сильно исхудавший от недоедания’. Оба данных определения корректны и описывают того, кто стал худым, но при этом между ними не совпадает ни единого токена. Метрики BLEU и ROUGE-L для такой пары показывают 0% совпадения. Только BERT-F1 вовзращает 70.39% совпадения, поскольку данная метрика вместо совпадения последовательности сравниваемых определений использует векторизатор и сравнивает расстояние между полученными вложениями, которое обозначает семантическую близость двух текстов.

Следует сказать, что предварительный анализ полученных определений показал наличие определений, имеющих ошибку самореференции. Например, для контекста *[Пантелей Прокофьевич] ничего не мог сделать, чтобы восстановить в семье прежний порядок.* и целевого слова *восстановить* изначально было сгенерировано «привести в прежнее состояние; восстановить», что содержит повторение целевого слова, а результаты метрик были ниже представленного ранее: 10.64% для BLEU, 29.09% для ROUGE-L, и 75.19% для BERT-F1.

Мы смогли значительно снизить число определений, страдающих от проблемы самореференции добавлением в исключения токенов, используемых для кодирования целевого слова.

Так, перед генерацией для каждого примера мы получаем все возможные варианты написания целевого слова с помощью модуля pymorphy3, после чего кодируем используемым для инференса токенизатором и добавляем

получившиеся токены в аргумент «bad_words_ids», что позволило избавиться от такого рода ошибок и повысило результаты. Например, для контекста *[Пантелей Прокофьевич] ничего не мог сделать, чтобы восстановить в семье прежний порядок.* и целевого слова *восстановить* после правки было сгененировано «привести в прежнее состояние, положение».

3.3. Тестирование модели на материале соревнования Rushifteval

С помощью модели были получены определения для тестовой части датасета соревнования Rushifteval.

Для векторизации сгенерированных определений использовалась модель paraphrase-multilingual-mpnet-base-v2 (URL: https : / / huggingface . co / sentence-transformers/paraphrase-multilingual-mpnet-base-v2). Данная модель является одним из лидеров по задаче семантической схожести текстов (URL: https://github.com/avidale/encodechka). Далее векторы были нормализированы, после чего расстояние между векторным представлением определений считалось с помощью косинусного расстояния. Результат приводился в формат значений датасета с помощью линейной регрессии, тренированной на датасете Rusemeval. Так, значения косинусного расстояния от 0 до 2, где 0 обозначает идентичность векторов, а 2 – их противоположность, преобразованы в значения от 1 до 4, соответствующие оценкам аннотаторов, где 1 – противоположные значения, а 4 – идентичные.

Таблица 3.4.. Коэффициэнты корреляции с использованием LinReg

Пары периодов	Коэффициент корреляции
Среднее	0.7156
pre-Soviet:Soviet	0.7056
Soviet:post-Soviet	0.7251
pre-Soviet:post-Soviet	0.7160

В качестве попытки улучшить результаты был дообучен векторизатор paraphrase-multilingual-mpnet-base-v2 на материале Rusemshift (аналог Rushifteval с другими лексемами), который рекомендуется авторами Rushifteval для улучшения результатов. Параметры дообучения векторизатора доступны в Приложении В.

Таблица 3.5.. Коэффициэнты корреляции с использованием LinReg и дообученного векторизатора

Пары периодов	Коэффициент корреляции
Среднее	0.8002
pre-Soviet:Soviet	0.7843
Soviet:post-Soviet	0.8139
pre-Soviet:post-Soviet	0.8023

Таким образом, благодаря дообучению векторизатор была улучшена производительность алгоритма на более чем 8%.

Рассмотрим данные результаты в сравнении с аналогами из соревнования Rushifteval.

Таблица 3.6.. Результаты алгоритма в сравнении с результатами команд Rushifteval.

Команда	**досоветский: советский**	**советский: постсовет-ский**	**досоветский: постсовет-ский**	**Среднее**
DeepMistake (после соревнования)	**0.863**	**0.854**	**0.834**	**0.850**
GlossReader	0.781	0.803	0.822	0.802
FRED-T5-FN с дообучением векторизатора	**0.784**	**0.814**	**0.802**	**0.800**
DeepMistake	0.798	0.773	0.803	0.791
vanyatko	0.678	0.746	0.737	0.720
FRED-T5-FN	**0.706**	**0.725**	**0.716**	**0.716**
aryzhova	0.469	0.450	0.453	0.457
Discovery	0.455	0.410	0.494	0.453
UWB	0.362	0.354	0.533	0.417
dschlechtweg	0.419	0.373	0.383	0.392
jenskaiser	0.430	0.310	0.406	0.382
SBX-HY	0.388	0.281	0.439	0.369
Baseline	0.314	0.302	0.381	0.332
svart	0.163	0.223	0.401	0.262
BykovDmitrii	0.274	0.202	0.307	0.261
fdzr	0.217	0.25143	0.065	0.178

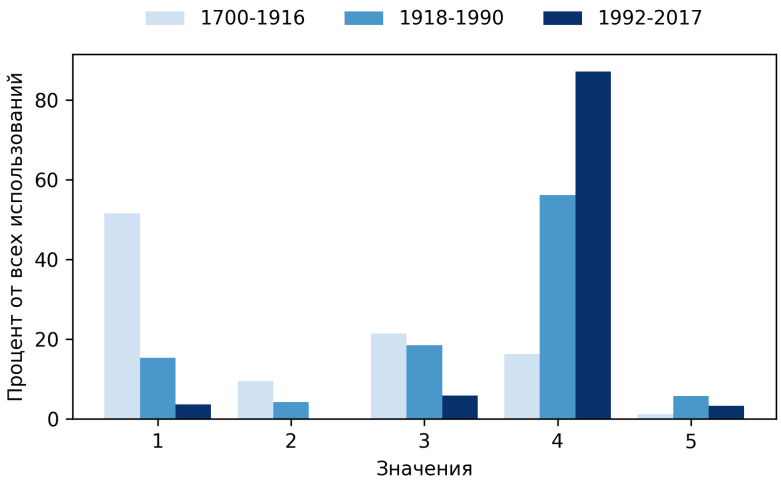
Как видно из таблицы, настоящее решение лучше по качеству большинства аналогов из соревнования Rushifteval, хоть и уступает некоторым, использующим модели XLM-R. Два решения с самым высоким качеством были описаны ранее в Главе 1.

3.4. Визуализация результатов работы модели

Для создания визуализаций семантических изменений слов используются используются библиотеки Matplotlib и Scikit-learn. Полученные с помощью модели определения векторизуются с помощью дообученного на материале Rusemshift в прошлой главе векторизатора. Так как для слов, имеющих одинаковое значение, модель склонна генерировать семантически близкие, однако не идентичные дословно определения, для группировки таких схожих определений применяется алгоритм кластеризации DBSCAN из библиотеки Scikit-learn на основе векторных представлений. Алгоритм кластеризации может настраиваться вручную через два ключевых параметра: «eps» и «min_samples». Параметр «eps» определяет максимальное расстояние между двумя точками, чтобы они считались находящимися в одном соседстве. «Min_samples» определяет минимальное количество точек, которые должны образовывать плотно связанную группу, чтобы она образовывала кластер. Затруднительно сказать заранее, какие параметры кластеризации подойдут для визуализации каждого конкретного слова. Представляется хорошим вариантом сначала выбирать небольшие значения и после повышать их, пока близкие определения, сформулированные похожим, но разным образом, не объединяться в единые кластеры. После этого, для каждого полученного кластера выбирается прототипическое определение, векторное представление которого наиболее близко к центру кластера. Данное определение выбирается для описания данного значения (кластера). Затем библиотека Matplotlib применяется для создания столбиковых диаграмм, отражающих частоту употреблений различных значений слова во времени, и для обес-

печения наглядности с помощью цветовой градации и легенд, содержащих прототипические определения каждого из значений.

Результатом анализа является график по типу иллюстрации, где представлена столбчатая диаграмма, показывающая процентное соотношение значений исследуемого слова за разные периоды времени. Каждая категория обозначена на диаграмме своим цветом и соответствующим временным интервалом: светло-синий цвет для 1700-1916, средне-синий для 1918-1990 и темно-синий для 1992-2017. Под диаграммой находится расшифровка значений, а также использованные параметры визуализации.



1: столкновение, драка.
2: беспорядочное, беспорядочное движение, толкотня.
3: беспорядочная, беспорядочная схватка.
4: место, где свалены, свалены в кучу какие-либо отходы.
5: то, что свалено, свалено в кучу.

Параметры: eps=0.12, min_samples=25

Рис. 3.1.. Изменение значений слова *свалка*

3.5. Код работы

Код, использованный во время выполнения настоящей работы, выложен в открытый доступ на сайте GitHub. ()

Проект включает следующие модули:

- config: Запуск тестов и проверок проекта с помощью CI
- latex: Написание текста работы в формате LaTeX
- model: Обучение модели и тестирование метриками
- mas_parser: Краулер и парсер словаря МАС
- vizvector: Векторизация определений и визуализация результатов
- rushifteval: Тестирование алгоритма на материале соревнования Rushifteval
- ruscorpora: Позволяет осуществить качественный анализ алгоритма

Наличие тестов и открытого доступа к коду проекта позволяет сделать исследование воспроизводимым, а также готовым к внедрению в другие проекты.

Выводы

Таким образом, была дообучена большая языковая модель FRED-T5-1.7B для задачи генерации определений с помощью датасета на основе «Малого академического словаря». Проведенные эксперименты показали, что модель способна генерировать семантически близкие, но не всегда идентичные определения. Несмотря на сравнительно низкие значения метрик BLEU и ROUGE-L, отражающих формальное сходство сгенерированных определений с эталонными, модель демонстрирует высокие результаты по метрике BERT-F1, учитывающей семантическую близость текстов. Это говорит о том, что модель способна продуцировать определения, имеющие схожий смысл с эталонными, хоть и зачастую сформулированные иными словами.

Применение модели для анализа семантических сдвигов на материале соревнования Rushifteval показало, что решение на основе FRED-T5-1.7B вместе с дообучением векторизатора способно добиться высокого результата в лидерборде. Таким образом, данная модель демонстрирует хорошие результаты в задаче выявления семантических изменений.

Разработанная визуализация на основе кластеризации векторных представлений определений позволяет наглядно представить семантические изменения слов во времени, что может быть полезно для лингвистических и исторических исследований.

Глава 4. Анализ результатов работы модели

Для дальнейшего анализа результатов алгоритма использовались 20 слов с изменившемся значением из книги «Два века в двадцати словах» (Данова [и др.], 2018). Использования данных слов брались из диахронического корпуса НКРЯ.

Для каждого рассматриваемого слова из каждого периода (досоветсткий, советский и постсоветсткий) бралась выборка из 300 вхождений, где для каждого использования слова генерировалось определение, а после строился график по аналогии с описанием визуализации выше.

Далее для каждого слова описана семантика слов на основе словарей в соответствии с рекомендациями издания И.А. Стернина (Стернин, Рудакова, 2017). В соответствии с ними, так как невозможно построить полное описание значения слова с использованием только одного словаря, требуется обобщение данных нескольких словарей. В качестве материала будут взяты три словаря современного русского языка «Большой толковый словарь» (далее *БТС*) (Кузнецов, 1998), «Толковый словарь русского языка Дмитриева» (далее *ТСД*) (Ахапкин [и др.], 2003) и «Толковый словарь русского языка» Ожегова и Шведовой (далее *ТСО*), а также книга «Два века в двадцати словах». Книга «Два века в двадцати словах» будет использована при обобщении значений, поскольку наряду со словарями содержит описания значений. После чего будет проведено сравнение выявленных при семантическом описании лексемы значений и тех, что выявлены алгоритмом.

Кроме того, произведено сравнение статистической информации по использованию слов в разные периоды для значений, соотносимых со значениями из книги «Два века в двадцати словах».

Следует учитывать то, что в книге исследуются периоды длиной меньше, чем в настоящей работе. Например, вместо досоветского выделяют 1800-

1849, 1850-1874, 1875-1899, а также 1900-1924, в связи с чем не представляется возможным выявить изменения между короткими периодами из книги.

Определения будут оцениваться по классификации, описанной в главе 1 и построенной на основе работ Huang et al. и Noraset et al.

Далее представлен разбор результата работы алгоритма для 3 слов, чьи результаты анализа значительно отличаются друг от друга. После этого представлены общие выводы по качественному анализу.

Подробный анализ результатов по каждому из остальных 17 слов представлен в Приложении.

4.1. Тройка

В результате анализа семем лексемы *тройка* в толковых словарях были выделены семь девять значений, которые можно условно сформулировать следующим образом:

1. Цифра 3. (*«Цифра 3»* в БТС, *«Цифра 3»* в ТСО, *«Тройка — это цифра 3»* в ТСД)
2. Количество три. (*«Количество три.»* в БТС, *«Тройное количество чего-либо.»* в ТСД, *«(о сходных или однородных предметах) количество три.»* в ТСО, *«Количество, сумма из трех единиц.»* в «Два века в двадцати словах»)
3. Оценка успеваемости в пятибалльной системе, означающая удовлетворительно. (*«Оценка успеваемости в пятибалльной системе, означающая удовлетворительно.»* в БТС, *«Школьная учебная отметка «удовлетворительно»».* в ТСО, *«В пятибалльной системе тройкой называют удовлетворительную, посредственную оценку чьих-либо знаний.»* в ТСД, *«Оценка в учебе.»* в «Два века в двадцати словах»)
4. Упряжка в три лошади. (*«Три лошади в одной упряжке.»* в БТС, *«Упряжка в три лошади.»* в ТСО, *«Тройкой называют упряжку из*

трёх лошадей — коренной и двух пристяжных.» в ТСД, *«Лошади.»* в «Два века в двадцати словах»)

5. Игральная карта с тремя очками. (*«Игральная карта в три очка.»* в БТС, *«В картах тройкой называют игральную карту в три очка»* в ТСД, *«Игральная карта с тремя очками»* в «Два века в двадцати словах»)
6. Транспортное средство, обозначенное цифрой 3. (*«Название автобуса, трамвая, троллейбуса третьего маршрута»* в БТС, *«Название чего-н. (обычно транспортного средства), обозначенного цифрой 3»* в ТСО, *«Маршрут автобуса, трамвая, троллейбуса, который пронумерован цифрой 3»* в ТСД)
7. Костюм, состоящий из пиджака (или жакета), брюк (или юбки) и жилета. (*«Костюм, состоящий из пиджака (или жакета), брюк (или юбки) и жилета»* в БТС, *«Костюм, состоящий из пиджака, брюк (или жакета, юбки) и жилета»* в ТСО, *«Костюм, который состоит из пиджака (или жакета), брюк (или юбки) и жилета»* в ТСД, *«Костюм»* в «Два века в двадцати словах»)
8. Три человека. (*«Тройкой называют устойчивую группу людей.»* в ТСД, *«Три человека»* в «Два века в двадцати словах»)
9. Звено боевых истребителей. (*«Тройкой называют звено боевых истребителей»* в ТСД)

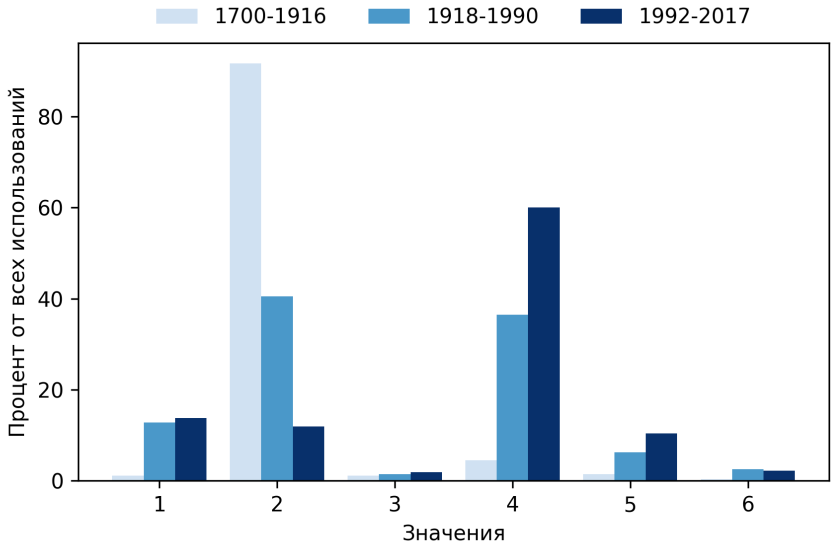


Рис. 4.1.. Изменение значений слова *тройка*

Значения для визуализации слова «Тройка» (Параметры: eps=0.28, min_samples=12).

1. Неудовлетворительная оценка по какому-либо предмету.
2. Одна из трех лошадей, запряженных в такую повозку.
3. Игральная карта с тремя одинаковыми мастями.
4. Группа лиц, состоящая из трех лиц.
5. Число 3.
6. Старинная мужская верхняя одежда из сукна или бархата, застегивавшаяся спереди на пуговицы.

Анализ значений слова *тройка*

Пятое определение корректно сформулировано. Четвёртое определение избыточно. Первое, второе, третье и шестое определения не соответствуют обобщенным значениям.

- ’Число 3.’ полностью соответствует ’Число 3’.
- ’Неудовлетворительная оценка по какому-либо предмету.’ является близким значением, так как в обобщенных значениях имеется «школьная оценка» (’Оценка успеваемости в пятибалльной систе-

ме, означающая удовлетворительно.’), но она означает «удовлетворительно», а не «неудовлетворительно».

- ’Одна из трех лошадей, запряженных в такую повозку.’ является близким обобщённому значению ’Упряжка в три лошади.’, но формулировка «одна из трех лошадей» некорректна, так как денотатом является упряжка, включающая в себя все три лошади и повозку.
- ’Игральная карта с тремя одинаковыми мастями.’ является близким значением, так как в обобщенных значениях имеется игральная карта с тремя очками, но формулировка «с тремя одинаковыми мастями» некорректна, мастью же является одна из четырёх категорий карт (В БТС: «Масть - один из четырёх разрядов на которые делится колода карт по цвету и форме очков.»).
- ’Группа лиц, состоящая из трех лиц.’ полностью соответствует ’Три человека’, так как включает те же семы «группа лиц», «три», однако содержит повторение слова *лиц*, которое делает его избыточным.
- ’Старинная мужская верхняя одежда из сукна или бархата, застегивавшаяся спереди на пуговицы.’ является некорректным определением, так как такое описание больше соответствует историческим видам одежды, таким как кафтан или сюртук, но не «костюму», представляющему собой комплект из брюк, пиджака и жилета.

Отсутствующие значения:

- ’Цифра 3.’ отсутствует среди предложенных моделью значений. Можно предположить, что информации из контекста использований недостаточно для выделения этого значения.
- ’Транспортное средство, обозначенное цифрой 3.’ также отсутствует в визуализации. Однако, модель способна на выделение данного значения.

- ’Звено боевых истребителей.’ отсутствует среди предложенных моделью значений. Модель могла не выделить это значение из-за ограниченного количества использований.

Таким образом, для лексемы *тройка* представлены:

- Корректные: 1
- Избыточность или чрезмерное использование общих фраз: 1
- Некорректные: 1
- Близкие: 3

Перейдем к частотности значений.

Судя по книге «Два века в двадцати словах», в XIX веке для слова *тройка* преимущественным является значение ’Упряжка в три лошади.’ с около 95% использований, что также отражено в нашей визуализации с около 90% использований близкого значения ’Одна из трех лошадей, запряженных в такую повозку.’ Начало советского периода, утверждается в книге, характерно увеличением использования значения ’Три человека’ в 20-ые годы, а также ’Оценка успеваемости в пятибалльной системе, означающая удовлетворительно.’ в 40-ые годы. К концу советского периода используется множество значений и ’Упряжка в три лошади.’ больше не является лидирующим. Это информация отражена в нашей визуализации, где ’Одна из трех лошадей, запряженных в такую повозку.’ падает до 40% в советское время и 15% постсоветское. Вместо него самым частым становится ’Группа лиц, состоящая из трех лиц.’ с около 60% в постсоветский период, а также частыми становятся значения школьной оценки и числа 3 с 10-15% использования.

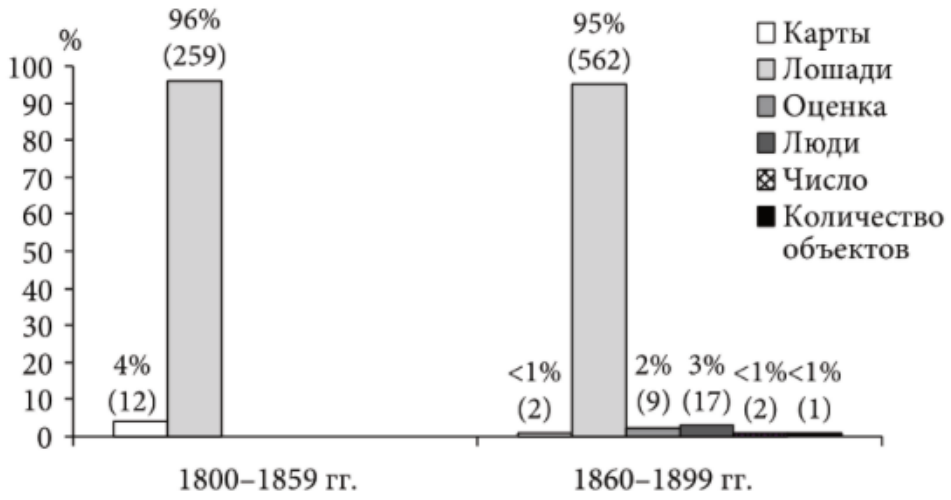


Рис. 4.2.. График для слова *тройка* для 1800-1899 из книги «Два века в двадцати словах».

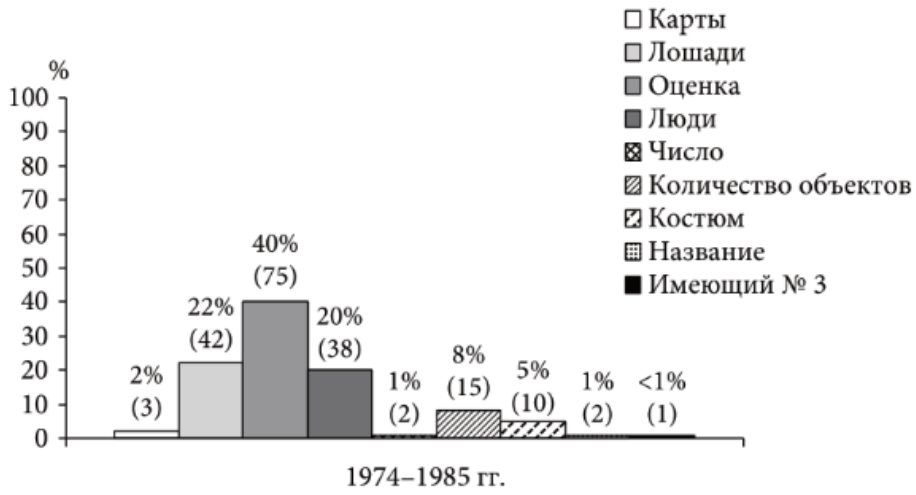


Рис. 4.3.. График для слова *тройка* для 1974-1985 из книги «Два века в двадцати словах».

Таким образом, алгоритм по большей части отражает значения, в которых использовалось слово *тройка*, не полностью согласуясь с данными из толкового словаря, но соответствуя историческим исследованием.

4.2. Выводы

В результате обобщения словарных дефиниций было составлено 120 значений для 20 слов.

Всего в результате работы алгоритма было получено 92 определения для 20 слов.

Таким образом, без учёта некорректных определений алгоритмом было успешно выявлено 72.5% значений.

Таблица 4.1.. Типы классификаций и их количество

Тип классификации	**Количество**	**Процент**
Корректные	58	$\frac{58}{92} \times 100 \approx 63.04\%$
Близкие	9	$\frac{9}{92} \times 100 \approx 9.79\%$
Некорректные	5	$\frac{5}{92} \times 100 \approx 5.43\%$
Недостаточно конкретизированные	5	$\frac{5}{92} \times 100 \approx 5.43\%$
Избыточность или чрезмерное использование общих фраз	4	$\frac{4}{92} \times 100 \approx 4.35\%$
Близкое значение, а также избыточность или чрезмерное использование общих фраз	1	$\frac{1}{92} \times 100 \approx 1.09\%$
Избыточно конкретизированные	1	$\frac{1}{92} \times 100 \approx 1.09\%$
Самореференция	0	$\frac{0}{92} \times 100 \approx 0.00\%$
Противоположное значение	0	$\frac{0}{92} \times 100 \approx 0.00\%$
Неправильная часть речи	0	$\frac{0}{92} \times 100 \approx 0.00\%$

Как видно, из результатов большинство изученных определений являются корректными без каких-либо ошибок или недочётов (63.04%).

Кроме того, проблема с самореференцией была успешно решена.

Из частых проблем можно выделить:

- Близкое значение

  Примером можно привести определение ’Насекомое, похожее на червя, а также его личинка.’ для слова *червяк*, где допущена ошибка, так как червяк не может быть взрослым насекомым. Возможно,

часть таких ошибок связана с относительно небольшим размером модели, что не позволяет ей иметь уверенные знания об окружающем мире.

- Избыточность или чрезмерное использование общих фраз

  В нашем случае эта проблема проявляется в повторении слов в определении. Например, для слова *свалка* в сгенерированном определении ’Беспорядочная, беспорядочная схватка.’ повторяется слово *беспорядочная*. На наш взгляд, это может быть связано, во-первых, с обилием синонимических рядов в определениях обучающего датасета на основе «Малого академического словаря», что является одним из способов описать значение слова в лексикологии.

К сожалению, для 2 слов из списка представляется невозможным полноценно проанализировать статистическую информацию. Этими словами являются *публика* и *сволочь*.

Так, для слова *публика* в книге «Два века в двадцати словах» не даётся графиков частотности и точных для слова *публика*. Гооврится лишь о преобладании значения ’аудитория’ и о его оттенках, которые не удается полноценно сравнить из-за того, что алгоритм предложил довольно общие значения.

Для слова *сволочь* из 4 значений, выявленных при обобщении дефиниций, алгоритмом были выявлены только следующие два значения:

1. Употребляется как бранное слово.
2. О подлом, гнусном человеке.

К сожалению, оба выделенных значения подпадают под значение ’Индивидуальное оскорбление.’ в книге «Два века в двадцати словах», поэтому анализ изменений значения сделать не представилось возможным.

Среди большинства оставшихся слов визуализации совпадают с данными из книги «Два века в двадцати словах». Исключением является слово *пока*.

Обобщёнными для него являлись значения:

1. В течение некоторого времени; до сих пор ещё; впредь до чего-л.
2. В то время как.
3. До того времени как.
4. Употребляется при прощании, до свидания.

Алгоритмом были сформулированы следующие значения:

1. В настоящее время, до тех пор.
2. Употребляется при обозначении времени, в течение которого совершается действие.
3. Употребляется при прощании с кем-л.

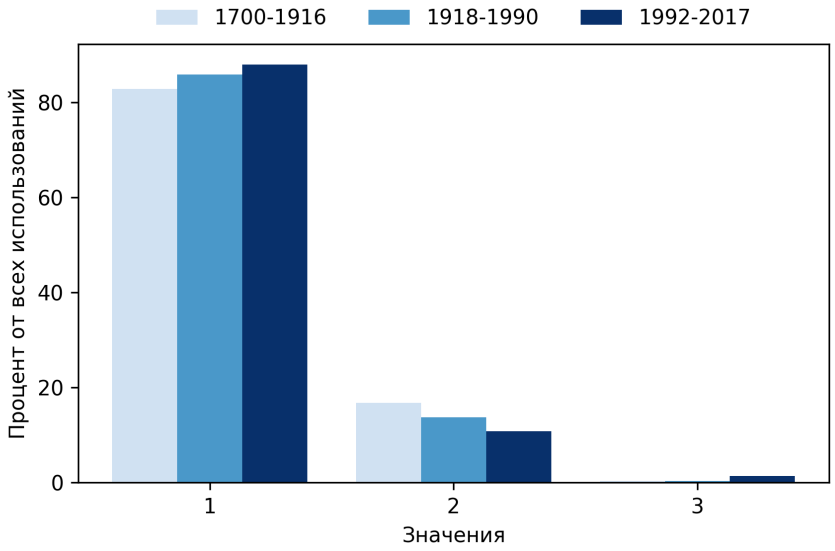


Рис. 4.4.. Изменение значений слова *пока*

В книге сообщается, что изначальным и всегда преобладающим было использование слова в качестве союза, после чего в XIX веке появилось использование как наречие, а затем в советский период – как этикетное слово. Данные из визуализации алгоритма (график снизу) поддерживают появление значения ’Употребляется при прощании с кем-л.’ поздно – несколько процентов для постсоветского периода, однако данные для наречия и союза не совпадают. Можно предположить, что модели сложно различать эти значения из-за их схожести. Например, для «Когда мы забирали щенка, нас

предупредили, что ей категорически нельзя наверх забираться, пока у нее слабые лапы.» было сгенерировано ’В настоящее время, до тех пор.’, что относит его к наречию, но из примера видно, что пока связывает части предложения и является союзом.

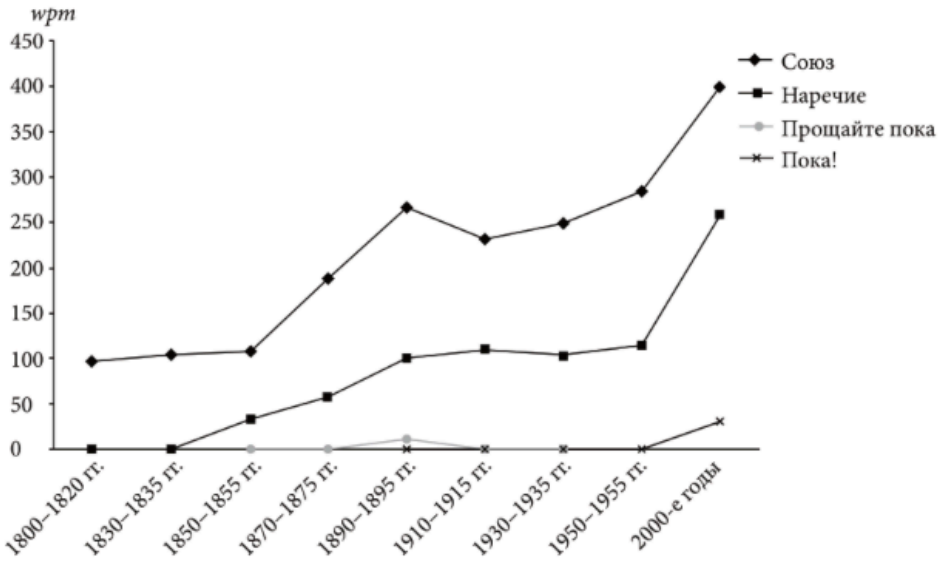


Рис. 4.5.. График для слова *пока* из книги «Два века в двадцати словах».

Заключение

В ходе выполнения выпускной квалификационной работы была обучена генеративная языковая модель на основе архитектуры Трансформер для задачи генерации определений слов на основе их контекста использования. Модель показала высокие результаты метрики сходства BERTScore для тестовой выборки, а также успешно показала себя на тестовом материале Rushifteval, имея результаты сопоставимые с лидирующими решениями. Кроме того, был создан алгоритм визуализации результатов модели, благодаря которому наше решение имеет высокую степень интерпретируемости. после чего был произведен качественный анализ работы алгоритма. В нем алгоритм показал высокие результаты, выявив большинство значений, а также верно составив визуализацию изменения использования значений для большинства слов. В целом, выполнения настоящей работы было доказано, что моделирование определений может быть успешно применено для задачи детектирования семантических изменений.

Результаты настоящей работы можно применять для определения степени семантического сдвига лексем, с наличием визуализации и определений для каждого выявленного значения, что может быть использовано в лексикологии, где необходимы актуальные данные для построения новых словарей (Giulianelli [и др.], 2023). Кроме того, модель, позволяющая автоматически генерировать качественные словарные определения, может быть полезна в таких задачах обработки естественного языка, как анализ тональности, машинный перевод и разграничение семантической неоднозначности (Gardner [и др.], 2022).

Перспективами развития настоящего исследования является:

- Использование более одного словаря в качестве материала для обучения модели.

- Использование генеративных моделей большего размера, чем используется в работе.

Ограничениями подхода можно считать необходимость в значительных вычислительных ресурсах. Несмотря на то, что FRED-T5-1.7B запускается на ЦПУ, запуск на большом количестве вхождений займет значительное число времени. Для запуска на ГПУ же необходима видеокарта с 8 ГБ видеопамяти.

Код, использованный во время выполнения настоящей работы, выложен в открытый доступ на сайте GitHub и может быть воспроизведен. ()

Список литературы

1. *Kutuzov A.*, *Øvrelid L.*, *Szymanski T.*, *Velldal E.* Diachronic word embeddings and semantic shifts: a survey // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA : Association for Computational Linguistics, 08.2018. — С. 1384—1397.
2. *Rodina J.*, *Trofimova Y.*, *Kutuzov A.*, *Artemova E.* ELMo and BERT in semantic change detection for Russian. — 2020.
3. *Giulianelli M.*, *Luden I.*, *Fernández R.*, *Kutuzov A.* Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. — 2023.
4. *Gardner N.*, *Khan H.*, *Hung C.-C.* Definition modeling: literature review and dataset analysis // Applied Computing and Intelligence. — 2022. — Т. 2. — С. 83—98.
5. *Kutuzov A.*, *Pivovarova L.* RuShiftEval: a shared task on semantic shift detection for Russian // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021. — С. 533—545.
6. *Tahmasebi N.*, *Borin L.*, *Jatowt A.* Survey of computational approaches to lexical semantic change detection // Computational approaches to semantic change. — Berlin : Language Science Press, 2021. — С. 1—91.
7. *Periti F.*, *Cassotti P.*, *Dubossarsky H.*, *Tahmasebi N.* Analyzing Semantic Change through Lexical Replacements. — 2024.
8. *Майсак Т. А.* Грамматикализация. — 2016 ; — Accessed: 2024-05-21. *Большая российская энциклопедия. Электронная версия*.
9. *Bloomfield L.* Language. — New York : Holt, Rinehart, Winston, 1933.

10. *Harris T. M.* Semantic Shift in the English Language //. — 2014.

11. *Стернин И. А.*, *Рудакова А. В.* Словарные дефиниции и семантический анализ. — Воронеж, 2017. — С. 34.

12. *Виноградов В.*, *Шведова Н.* История слов: около 1500 слов и выражений и более 5000 слов, с ними связанных. — Институт русского языка им. В.В. Виноградова РАН, 1999.

13. *Данова М. К.*, *Добрушина Н. Р.*, *Опачанова А. С.* [и др.]. Два века в двадцати словах. — Москва : Издательский дом Высшей школы экономики, 2018. — 455 с.

14. *Jatnika D.*, *Bijaksana M.*, *Ardiyanti A.* Word2Vec Model Analysis for Semantic Similarities in English Words // Procedia Computer Science. — 2019. — Янв. — Т. 157. — С. 160—167.

15. *Mikolov T.*, *Chen K.*, *Corrado G.*, *Dean J.* Efficient Estimation of Word Representations in Vector Space. — 2013.

16. *Kutuzov A.*, *Fomin V.*, *Mikhailov V.*, *Rodina J.* SHIFTRY: WEB SERVICE FOR DIACHRONIC ANALYSIS OF RUSSIAN NEWS //. — 01.2020. — С. 500—516.

17. *Schlechtweg D.*, *McGillivray B.*, *Hengchen S.*, *Dubossarsky H.*, *Tahmasebi N.* SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection // CoRR. — 2020. — Т. abs/2007.11464.

18. *Kutuzov A.* Distributional Word Embeddings in Modeling Diachronic Semantic Change : Doctoral Thesis / Kutuzov Andrey. — University of Oslo, 2020. — Accessed: 2020-11-16T12:34:15Z.

19. *Rachinskiy M.*, *Arefyev N.* Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection //. — 06.2021. — С. 578—586.

20. *Conneau A.*, *Khandelwal K.*, *Goyal N.*, *Chaudhary V.*, *Wenzek G.*, *Guzmán F.*, *Grave E.*, *Ott M.*, *Zettlemoyer L.*, *Stoyanov V.* Unsupervised Cross-lingual Representation Learning at Scale // CoRR. — 2019. — T. abs/1911.02116.

21. GlossReader. — URL: https://github.com/myrachins/RuShiftEval (дата обр. 18.01.2024).

22. *Arefyev N.*, *Fedoseev M.*, *Protasov V.*, *Panchenko A.*, *Homskiy D.*, *Davletov A.* DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model //. — 06.2021. — С. 16—30.

23. DeepMistake. — URL: https://github.com/Daniil153/DeepMistake (дата обр. 18.01.2024).

24. *Noraset T.*, *Liang C.*, *Birnbaum L.*, *Downey D.* Definition Modeling: Learning to define word embeddings in natural language. — 2016.

25. *Gadetsky A.*, *Yakubovskiy I.*, *Vetrov D.* Conditional Generators of Words Definitions // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — С. 266—271.

26. *Huang H.*, *Kajiwara T.*, *Arase Y.* Definition Modelling for Appropriate Specificity // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. — Online, Punta Cana, Dominican Republic : Association for Computational Linguistics, 11.2021. — С. 2499—2509.

27. *Papineni K.*, *Roukos S.*, *Ward T.*, *Zhu W. J.* BLEU: a Method for Automatic Evaluation of Machine Translation. — 2002. — Окт.

28. *Lin C.-Y.* ROUGE: A Package for Automatic Evaluation of summaries //. — 01.2004. — С. 10.

29. *Zhang T.*, *Kishore V.*, *Wu F.*, *Weinberger K. Q.*, *Artzi Y.* BERTScore: Evaluating Text Generation with BERT. — 2020.

30. *Кузнецов С. А.* Большой толковый словарь русского языка: А-Я. — СПб. : Норинт, 1998. — С. 1534. — РАН. Ин-т лингв. исслед. Сост., гл. ред. канд. филол. наук С. А. Кузнецов.

31. *Zmitrovich D.* [и др.]. A Family of Pretrained Transformer Language Models for Russian. — 2023.

32. *Shavrina T.*, *Fenogenova A.*, *Anton E.*, *Shevelev D.*, *Artemova E.*, *Malykh V.*, *Mikhailov V.*, *Tikhonova M.*, *Chertok A.*, *Evlampiev A.* RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Association for Computational Linguistics, 2020.

33. *Евгеньева А. П.* Словарь русского языка: В 4-х т. — Москва : Русский язык, 1981-1984. — В 4-х томах.

34. Evaluate. — URL: https://github.com/huggingface/evaluate (дата обр. 15.11.2023).

35. paraphrase-multilingual-mpnet-base-v2. — URL: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (дата обр. 19.04.2024).

36. encodechka. — URL: https://github.com/avidale/encodechka (дата обр. 19.04.2024).

37. *Tatarinov M. D.* Work Definition Modeling. — 2024 ; — Accessed: 2024-04-24. https://github.com/tatarinovst2/work-definition-modeling.

38. *Ахапкин Д.* [и др.]. Толковый словарь русского языка : Ок. 2000 слов-вар. ст., свыше 12000 значений. — Москва : Астрель [и др.], 2003. — С. 989. — ГУП ИПК Ульян. Дом печати.

Приложение А. Качественный анализ

Знатный

В результате анализа семем лексемы *знатный* в толковых словарях были выделены пять групп значений, которые можно условно сформулировать следующим образом:

1. Известный, знаменитый, прославленный своей деятельностью. (*«Известный, знаменитый, прославленный.»* в БТС, *«Прославившийся своей деятельностью, такой, к-рого знают все.»* в ТСО, *«Выдающийся в труде.»* в «Два века в двадцати словах»)
2. Принадлежащий к знати, к аристократии, к верхушке привилегированного класса. (*«Принадлежащий к знати, к верхушке привилегированного класса.»* в БТС, *«Принадлежащий к аристократии, к знати.»* в ТСО, *«Принадлежащий к знати, высокий по чину.»* в «Два века в двадцати словах»)
3. Отличный, высокий по качеству. (*«Отличный, высокий по качеству; сильный.»* в БТС, *«Отличный, высокий по качеству; сильный (прост.).»* в ТСО, *«Хороший.»* в «Два века в двадцати словах»)
4. Существенный, серьезный (усилитель). (*«Существенный, серьезный (усилитель).»* в «Два века в двадцати словах»)
5. Знаемый, известный, видимый. (*«Знаемый, известный, видимый.»* в «Два века в двадцати словах»)

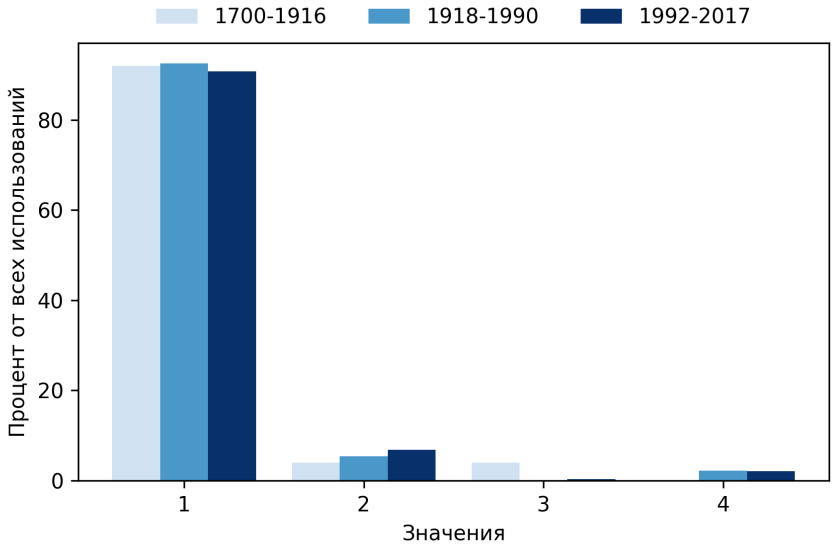


Рис. А.1.. Изменение значений слова *знатный*

Значения для визуализации слова «Знатный» (Параметры: eps=0.2, min_samples=10).

1. Принадлежащий к знати, имеющий высокое общественное положение.
2. Очень хороший, превосходный.
3. Значительный по значению, важный, значительный.
4. Знающий свое дело, искусный, опытный.

Анализ значений слова *знатный*

Первое и второе определения корректно сформулированы. Третье и четвертое определения соответствуют обобщенным значениям. Пятое определение не соответствует обобщенным значениям.

- ’Принадлежащий к знати, имеющий высокое общественное положение.’ имеет общий смысловой элемент с ’Принадлежащий к знати, к аристократии, к верхушке привилегированного класса.’, а именно семы «принадлежность к знати», «высокое общественное положение».

- ’Очень хороший, превосходный.’ полностью соответствует ’Отличный, высокий по качеству.’, так как включает те же семы «отличный», «высокий по качеству», «превосходный».
- ’Значительный по значению, важный, значительный.’ соответствует ’Существенный, серьезный (усилитель).’, так как включает те же семы «важность», «серьезность».
- ’Знающий свое дело, искусный, опытный.’ частично соответствует ’Известный, знаменитый, прославленный своей деятельностью.’, так как включает семы «искусный», «опытный», которые подразумевают известность и признание в своей деятельности.

Отсутствующие значения:

- ’Знаемый, известный, видимый’ также отсутствует в визуализации. Это значение могло быть не выделено из-за недостаточной частотности.

Ошибок в написании определений (орфографических, синтаксических, повторения слова) не обнаружено.

Таким образом, для лексемы *знатный* представлены:

- Корректные: 3
- Близкие: 1

Перейдем к частотности значений.

В книге «Два века в двадцати значениях» можно выделить три основных момента. Во-первых, наличие до конца XVIII века значения ’Знаемый, известный, видимый’, однако оно не было выделено алгоритмом. Во-вторых, преимущественное использование значения ’Принадлежащий к знати, имеющий высокое общественное положение.’ на протяжении всего исследуемого времени. Такой же результат наблюдается и в визуализации, с 90% использования на протяжении трех эпох. В-третьих, появление в советский период значения, связанного с трудом, что так же отражается на графике.

Рис. А.2.. График для слова *Знатный* для 1891-1920 из книги «Два века в двадцати словах».

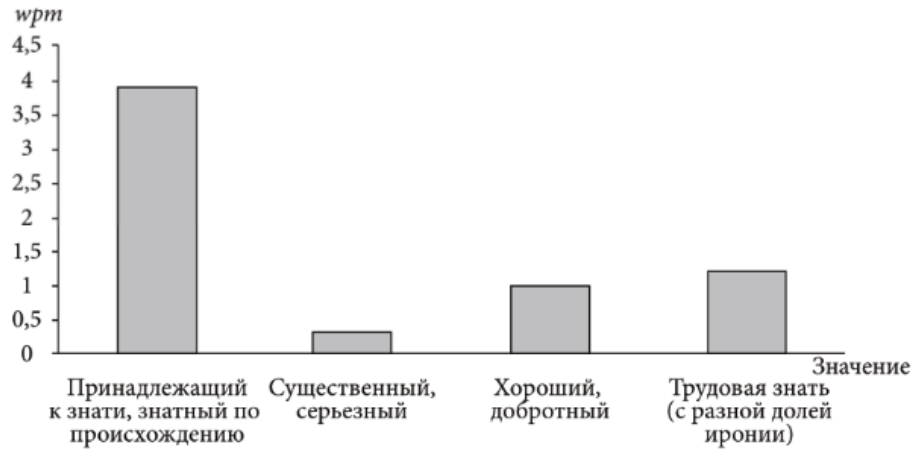


Рис. А.3.. График для слова *Знатный* для 1990-2010 из книги «Два века в двадцати словах».

Таким образом, алгоритм в целом отражает значения, в которых использовалось слово *знатный*, согласуясь с данными из толкового словаря и историческим исследованием. Алгоритм выделяет как основное значение, связанное с высоким положением, так и большинство менее частых.

Кануть

В результате анализа семем лексемы *кануть* в толковых словарях были выделены четыре группы значений, которые можно условно сформулировать следующим образом:

1. Упасть каплей; капнуть. (*«Упасть каплей; капнуть.»* в БТС, *«Капнуть, упасть каплей (устар.).»* в ТСО, *«Капнуть.»* в «Два века в двадцати словах»)
2. Погрузиться, утонуть. (*«Упав куда-л., во что-л., погрузиться.»* в БТС, *«Утонуть, упасть на дно.»* в «Два века в двадцати словах»)
3. Бесследно исчезнуть, пропасть. (*«Пропасть, исчезнуть, скрыться.»* в БТС, *«Бесследно пропасть, исчезнуть.»* в ТСО, *«Пройти, минуть, исчезнуть.»* в «Два века в двадцати словах»)
4. Исчезнуть из виду. (*«Исчезнуть из виду.»* в «Два века в двадцати словах»)

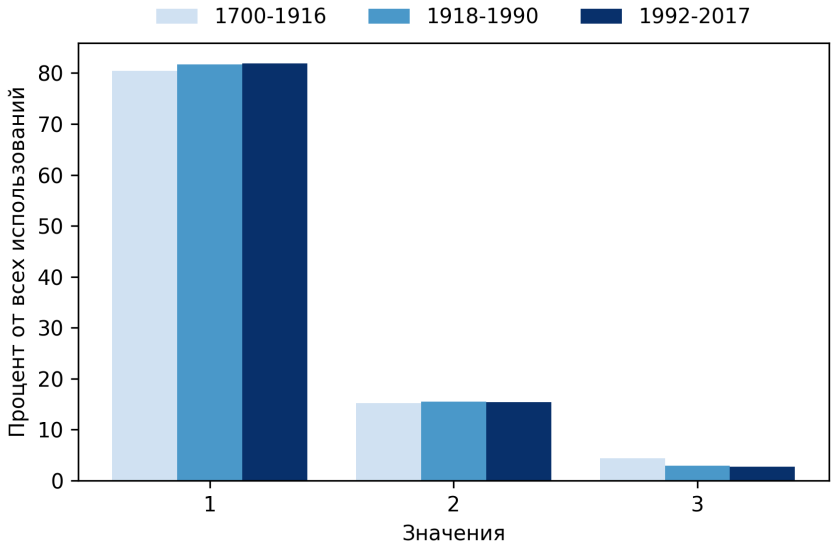


Рис. А.4.. Изменение значений слова *кануть*

Значения для визуализации слова «Кануть» (Параметры: eps=0.18, min_samples=10).

1. Исчезнуть, пропасть.

2. Пройти, миновать, исчезнуть.
3. Упасть, погрузиться.

Анализ значений слова *кануть*

Первые три определения корректно сформулированы.

- ’Исчезнуть, пропасть.’ корректное определение, так как имеет общий смысловой элемент с ’Бесследно исчезнуть, пропасть.’, а именно семы «исчезнуть» и «пропасть».
- ’Пройти, миновать, исчезнуть.’ частично соответствует ’Бесследно исчезнуть, пропасть.’ и ’Исчезнуть из виду.’, так как включает те же семы «пройти», «исчезнуть». Визуализация добавляет семы «миновать», что расширяет значение, делая его более широким.
- ’Упасть, погрузиться.’ частично соответствует ’Погрузиться, утонуть.’, так как включает семы «упасть» и «погрузиться».
- ’Упасть каплей; капнуть.’ не выделяется алгоритмом. Можно предположить, что это значение недостаточно часто встречается. В «Двух веках в двадцати словах» указано, что на протяжении XIX века оно заменяется другими значениями.

Ошибок в написании определений не обнаружено.

Таким образом, для лексемы *кануть* представлены:

- Корректные: 1
- Близкие значения: 2

Перейдем к частотности значений.

Затруднительно провести анализ рассматриваемого слова, так как, судя по «Двум векам в двадцати словах» выделенные алгоритмом значения появляются в досоветский период и продолжают использоваться дальше. Подтверждается информация о том, что с 1900 года ’исчезнуть, сгинуть, пропасть’ является основным значением слова, однако книга не предоставляет визуализаций частоты использования значений слова по периодам.

Классный

В результате анализа семем лексемы *классный* в толковых словарях были выделены пять групп значений, которые можно условно сформулировать следующим образом:

1. Имеющий отношение к школьному обучению. (*«к Класс (3 зн.)»* в БТС, *«Классным называют то, что имеет отношению к классу.»* в ТСД, *«Имеющий отношение к школьному обучению.»* в «Два века в двадцати словах»)
2. Имеющий определённый класс, разряд, соответствующий требованиям такого класса, разряда. (*«Имеющий определённый класс, разряд, соответствующий требованиям такого класса, разряда.»* в БТС, *«Имеющий класс (разряд).»* в «Два века в двадцати словах»)
3. Имеющий определённый ранг, чин. (*«Имеющий определённый ранг, чин.»* в БТС, *«Классный чиновник.»* в «Два века в двадцати словах»)
4. Специалист, обладающий высоким мастерством в своей области. (*«Принадлежащий к высшему классу, разряду по квалификации, по мастерству в чём-л.»* в БТС, *«Классным называют специалиста, который обладает высоким мастерством в своей области.»* в ТСД)
5. Отличный, высокого качества. (*«Отличный.»* в БТС, *«Принадлежащий к высшему классу (в 4 знач.), высокого качества (разг.).»* в ТСО, *«Классным называют человека, предмет, событие и т. п., которые имеют замечательные свойства, обладают высоким качеством; разговорный стиль.»* в ТСД, *«Хороший, отличный.»* в «Два века в двадцати словах»)

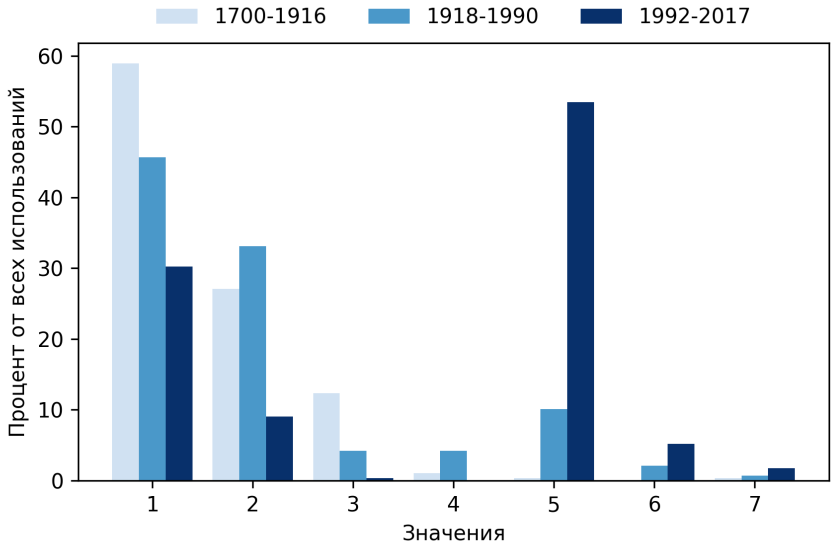


Рис. А.5.. Изменение значений слова *классный*

Значения для визуализации слова «Классный» (Параметры: eps=0.22, min_samples=5).

1. Служащий в классе, занимающийся в классе.
2. Предназначенный для класса.
3. Помещение для занятий в школе.
4. Предназначенный для пассажиров первого класса.
5. Очень хороший, замечательный.
6. Отличающийся высоким мастерством в каком-л. деле.
7. Связанный с присвоением какого-либо звания, чина.

Анализ значений слова *классный*

Третье и четвертое определения не соответствуют обобщенным значениям. Остальные определения являются корректными.

- ’Служащий в классе, занимающийся в классе.’ соответствует определению ’Имеющий отношение к школьному обучению.’, объединяя семы «служащий», «занимающийся», «класс».
- ’Предназначенный для класса.’ также относится к определению ’Имеющий отношение к школьному обучению.’ с теми же семами.

- ’Очень хороший, замечательный.’ соответствует значению ’Отличный, высокого качества.’, объединяя семы «хороший», «отличный», «высокого качества».
- ’Отличающийся высоким мастерством в каком-л. деле.’ соответствует значению ’Специалист, обладающий высоким мастерством в своей области.’, объединяя семы «высокий мастерство», «специалист», «область».
- ’Связанный с присвоением какого-либо звания, чина.’ аналогично определению ’Имеющий определённый ранг, чин.’, объединяя семы «звание», «чин», «ранг».

Слишком специфичные значения:

- ’Помещение для занятий в школе.’ является слишком специфичным и может быть включено в ’Имеющий отношение к школьному обучению.’
- ’Предназначенный для пассажиров первого класса.’ также является слишком специфичным, может быть включено в ’Имеющий определённый класс, разряд, соответствующий требованиям такого класса, разряда.’.

Таким образом, для лексемы *классный* представлены:

- Корректные: 5
- Слишком специфичные: 2

Перейдем к частотности значений.

Основными моментами из книги «Два века в двадцати словах» является появление в советское время значения ’Очень хороший, замечательный.’, становление его основным в постсоветский период и преобладание значений, связанных со школой, до этого. Всё вышеперечисленное выводится из нашей визуализации, где значение ’Очень хороший, замечательный.’ набирает около 10% от всех использований в советский период и около 55% в постсоветский. Значения, связанные со школьным обучением (1, 2, 3) вместе

набирают более 90% от всех использований в досоветский период. Информация из книги «Два века в двадцати словах» приведена в графике ниже.

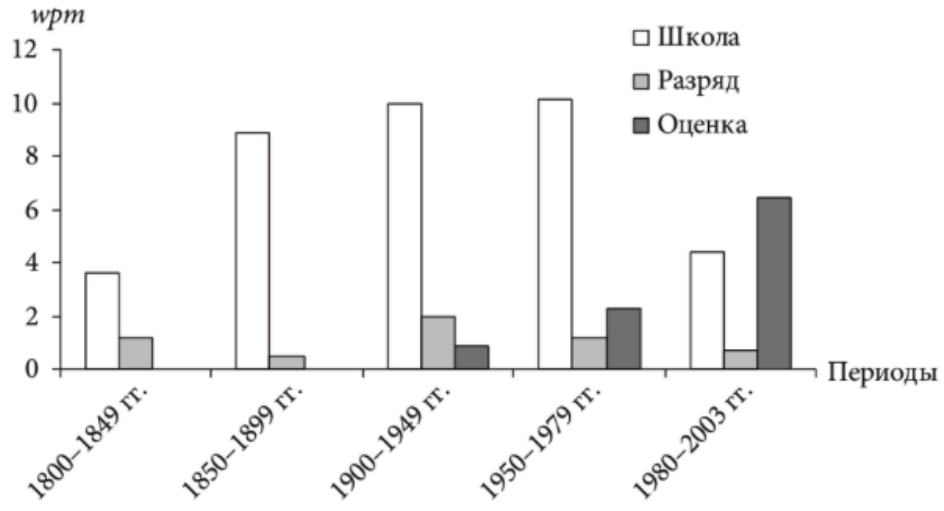


Рис. А.6.. График для слова *Классный* из книги «Два века в двадцати словах».

Таким образом, алгоритм отражает значения, в которых использовалось слово *классный*, согласуясь с данными из толкового словаря и историческим исследованием, однако некоторые определения являются слишком специфичными.

Мама

В результате анализа семем лексемы *мама* в толковых словарях были выделены пять групп значений, которые можно условно сформулировать следующим образом:

1. Женщина, являющаяся родительницей ребёнка. (*«Женщина по отношению к своим детям»* в ТСО, *«Женщина по отношению к рождённым ею детям»* в БТС и ТСД, *«Генетическая мать»* в «Два века в двадцати словах»)
2. Обращение ребёнка к своей матери. (*«Мамой ребёнок называет свою мать»* в ТСД, *«Мама — это обращение ребёнка к своей матери»* в ТСД)

3. Обращение к тёще или свекрови. (*«Тёща или свекровь (обычно в семейном обращении)»* в БТС, *«Мамой в разговорной речи иногда называют тёщу или свекровь»* в ТСД)
4. Обращение к опекуну или кормильцу без родственных связей. (*«Комилица, няня»* в «Два века в двадцати словах»)
5. Наименование компонента компьютера – материнской платы. (*«Материнская плата»* в «Два века в двадцати словах»)

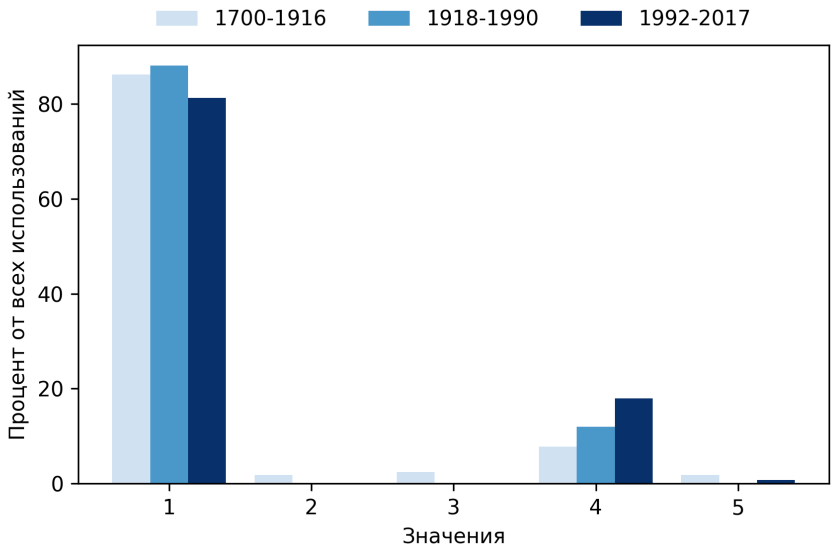


Рис. А.7.. Изменение значений слова *мама*

Значения для визуализации слова «Мама» (Параметры: eps=0.08, min_samples=5).

1. Женщина, мать.
2. Фамильярное обращение к пожилому мужчине.
3. В дореволюционной России: женщина, занимавшаяся воспитанием детей.
4. Женщина по отношению к своим детям.
5. Ласковое обращение к женщине.

Анализ значений слова *мама*

Первое, третье и четвертое определения корректно сформулированы. Второе и пятое определения не соответствуют обобщенным значениям.

- ’Женщина, мать.’ имеет общий смысловой элемент с ’Женщина, являющаяся родительницей ребёнка.’, а именно семы «женщина» и «родитель».
- ’Женщина по отношению к своим детям.’ полностью соответствует ’Женщина, являющаяся родительницей ребёнка.’, так как включает те же семы «женщина», «родитель».
- ’В дореволюционной России: женщина, занимавшаяся воспитанием детей.’ имеет соответствие с ’Обращение к опекуну или кормильцу без родственных связей.’, так как семы «женщина», «занимающаяся воспитанием детей» отражают смысл «няня».
- ’Фамильярное обращение к пожилому мужчине.’ является некорректным значением, в обобщенных значениях нет упоминаний о мужчине.
- ’Ласковое обращение к женщине.’ близко к ’Обращение ребёнка к своей матери.’, но имеет более широкий смысл, включающий всех женщин, а не только матерей или нянь, что делает его недостаточно специфичным.

Отсутствующие значения:

- ’Обращение к тёще или свекрови’ отсутствует среди предложенных моделью значений. Можно предположить, что информации из контекста использований недостаточно для отделения этого значения от ’женщина, мать.’.
- ’Наименование компонента компьютера – материнской платы’ также отсутствует в визуализации. Однако, модель способна на выделение данного значения.

Таким образом, для лексемы *мама* представлены:

- Корректные: 3
- Некорректные: 1
- Недостаточно специфичные: 1

Перейдем к частотности значений.

Основным моментом в Двух веках в двадцати словах для слова *мама* является появление значения ’Женщина, являющаяся родительницей ребёнка.’, сменившего ’Обращение к опекуну или кормильцу без родственных связей.’ в середине XIX века. Это отражено в визуализации алгоритма, где ’В дореволюционной России: женщина, занимавшаяся воспитанием детей.’ присутствует только для досоветского периода. Основным значением для остальных периодов является ’Женщина, являющаяся родительницей ребёнка.’, что также отражено в визуализации алгоритма. Информация из книги приведена в графике ниже. Также в книге упоминается появление значения ’Наименование компонента компьютера – материнской платы’ в конце XX века, однако он не был выделен алгоритмом.

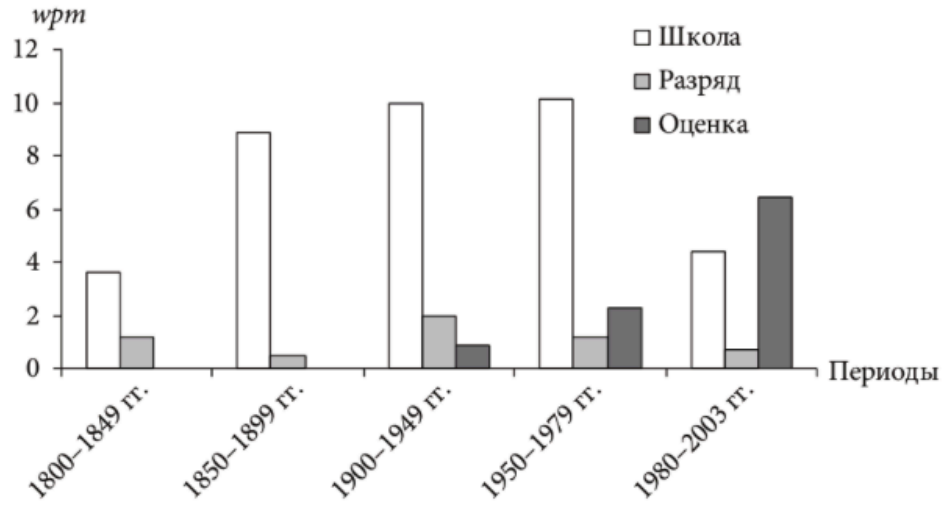


Рис. А.8.. График для слова *Мама* для 1760-1889 из книги «Два века в двадцати словах».

Таким образом, алгоритм отражает часть значений, в которых использовалось слово *мама*, согласуясь с данными из толкового словаря и истори-

ческим исследованием. Часть значений предложена некорректно, так как в них упоминается употребление слова по отношению к мужчинам, а также по отношению к абсолютно всем женщинам.

Машина

В результате анализа семем лексемы *машина* в толковых словарях были выделены девять групп значений, которые можно условно сформулировать следующим образом:

1. Механическое устройство, совершающее полезную работу с преобразованием энергии, материалов или информации. (*«Механизм или совокупность механизмов, совершающие какую-л. полезную работу путём преобразования одного вида энергии в другой.»* в БТС, *«Механическое устройство, совершающее полезную работу с преобразованием энергии, материалов или информации.»* в ТСО, *«Машина — это механизм, который совершает какую-либо полезную работу.»* в ТСД, *«Бытовой прибор»* в «Два века в двадцати словах»)
2. Автомобиль, средство передвижения. (*«Автомобиль, автомашина.»* в БТС, *«То же, что автомобиль.»* в ТСО, *«Машина — это средство передвижения, автомобиль.»* в ТСД, *«Автомобиль»* в «Два века в двадцати словах»)
3. Поезд, паровоз. (*«Поезд, паровоз.»* в «Два века в двадцати словах»)
4. Количество груза, вмещающееся в кузов грузового автомобиля. (*«О количестве груза, вмещающегося в кузов грузового автомобиля (обычно от 3 до 5 тонн).»* в БТС, *«Машиной чего-либо в разговорной речи называют количество груза, которое помещается в одну машину.»* в ТСД)
5. Движущийся или летающий механизм. (*«О самодвижущихся механизмах различного значения (комбайне, тракторе, мотоцикле и*

т.п.).» в БТС, *«Машиной называют любой движущийся, летающий механизм.»* в ТСД)

6. Мотоцикл, велосипед. (*«У спортсменов: мотоцикл, велосипед.»* в ТСО)
7. Организация, действующая подобно механизму, налаженно и чётко. (*«О какой-л. организации, ведомстве и т.п., действующих, подобно механизму, бесперебойно, точно, ритмично.»* в БТС, *«Об организации, действующей подобно механизму, налаженно и чётко.»* в ТСО, *«Машиной называют политическую, военную и т. п. организации, которые действуют точно и бесперебойно, как механизм.»* в ТСД)
8. Человек, лишённый каких-л. эмоций, действующий машинально, автоматически. (*«О человеке, лишённом каких-л. эмоций, действующем машинально, автоматически.»* в БТС, *«Машиной в разговорной речи называют человека, который никак не проявляет своих чувств, эмоций и совершает поступки машинально, автоматически.»* в ТСД)
9. Устройство для работы с информацией, компьютер. (textit«Комплекс технических, аппаратных и программных средств, предназначенных для автоматического сбора, хранения, обработки, передачи информации и её использования.» в БТС, textit«Компьютер.» в «Два века в двадцати словах»)

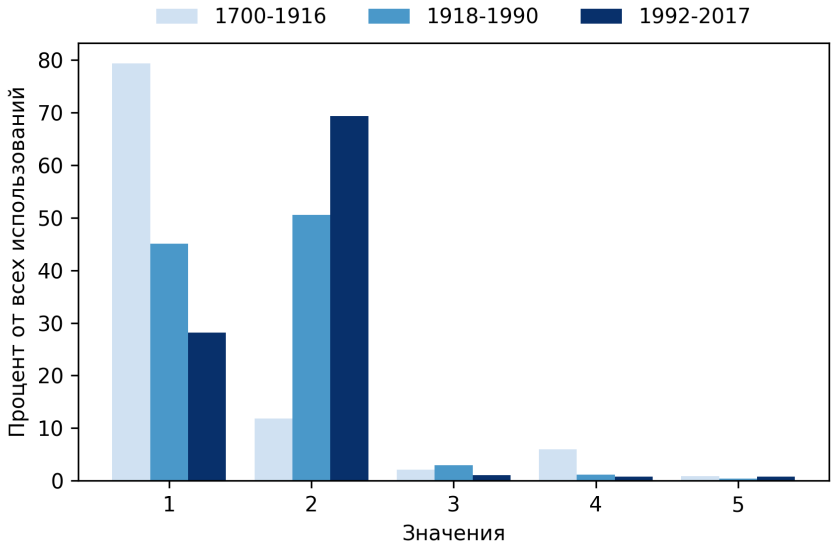


Рис. А.9.. Изменение значений слова *машина*

Значения для визуализации слова «Машина» (Параметры: eps=0.15, min_samples=5).

1. Приспособление, устройство, служащее для выполнения какой-л. работы.
2. Автомобиль, транспортное средство.
3. Самолет, вертолет и т. п.
4. О человеке, действующем механически, бездумно.
5. Система, совокупность каких-либо учреждений, организаций, предприятий и т. п.

Анализ значений слова *машина*

Первое, второе и четвертое определения корректно сформулированы. Третье и шестое определения не соответствуют обобщенным значениям.

- ’Приспособление, устройство, служащее для выполнения какой-л. работы.’ имеет общий смысловой элемент с ’Механическое устройство, совершающее полезную работу с преобразованием энергии, материалов или информации.’, а именно семы «устройство» и «выполнение работы».

- ’Автомобиль, транспортное средство.’ полностью соответствует ’Автомобиль, средство передвижения.’, так как включает те же семы «автомобиль» и «средство передвижения».
- ’О человеке, действующем механически, бездумно.’ имеет общий смысловой элемент с ’Человек, лишённый каких-л. эмоций, действующий машинально, автоматически.’, а именно семы «человек» и «действующий механически».
- ’Система, совокупность каких-либо учреждений, организаций, предприятий и т. п.’ близко к ’Организация, действующая подобно механизму, налаженно и чётко.’, так как включает семы «организация» и «действующая налаженно и чётко».
- Несмотря на то, что как ’Самолет, вертолет и т. п.’ не выделяется среди обобщённых значений, являясь более узким, чем ’Движущийся или летающий механизм.’, под обобщённое определение также попадают и другие, как ’Автомобиль, средство передвижения.’ или ’Поезд, паровоз.’, где оба являются типами движущихся механизмов. В связи с этим на равне с ними представляется разумным признать данное определение как корректное.

Отсутствующие значения:

- ’Количество груза, вмещающееся в кузов грузового автомобиля’ отсутствует среди предложенных моделью значений. Можно предположить, что информации из контекста использований недостаточно для выделения этого значения.
- ’Устройство для работы с информацией, компьютер’ также отсутствует в визуализации. Однако, модель способна на выделение данного значения.
- Другими отсутствующими значениями являются различные типы транспорта: ’Поезд, паровоз.’ и ’Мотоцикл, велосипед.’.

Ошибки в написании определений: отсутствуют.

Таким образом, для лексемы *машина* представлены:

• Корректные: 5

Перейдем к частотности значений.

’Механическое устройство, совершающее полезную работу с преобразованием энергии, материалов или информации.’ и схожие с ним значения, судя по «Двум векам в двадцати словах», присутствовало в течении всего рассматриваемого периода, что отражается в значении ’Приспособление, устройство, служащее для выполнения какой-л. работы.’, предложенное нашим алгоритмом. Однако в течение советского времени ’Автомобиль, средство передвижения.’ становится основным. Так, на нашей визуализации значение 2 набирает с около 15% за досоветский период до около 70% за постсоветский, данное увеличение произошло в ущерб значению 1 – ’Приспособление, устройство, служащее для выполнения какой-л. работы.’. Похожая ситуация наблюдается на графиках из книги, приведенных ниже.

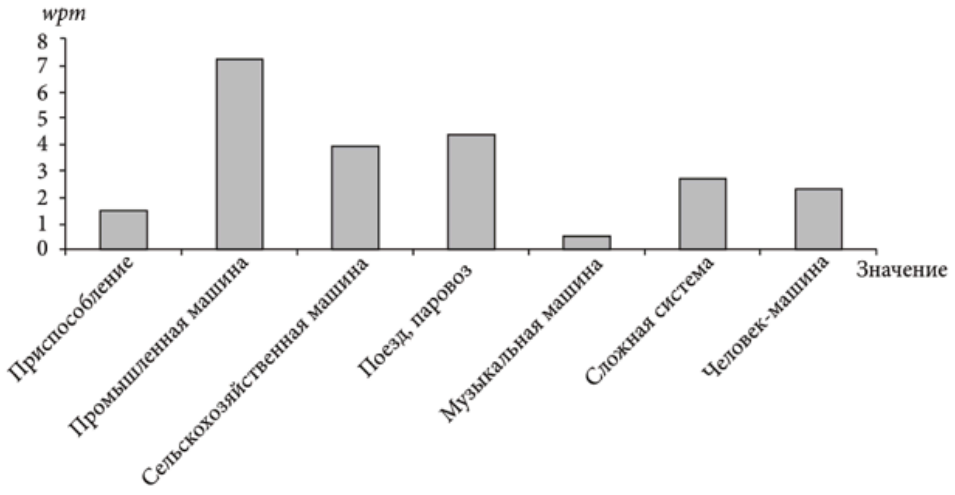


Рис. А.10.. График для слова *машина* для 1861-1890 из книги «Два века в двадцати словах».

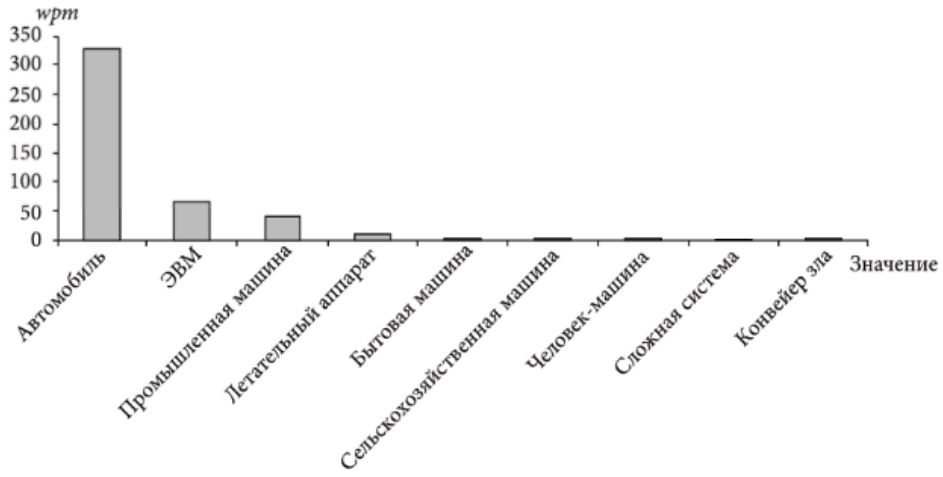


Рис. А.11.. График для слова *машина* для 1960-1970 из книги «Два века в двадцати словах».

Тем не менее, в книге обсуждается использование слова в значении ’Поезд, паровоз.’ в досоветский период, а также в значении ’Устройство для работы с информацией, компьютер’, которые были не были выделены алгоритмом.

Таким образом, алгоритм в целом отражает значения, в которых использовалось слово *машина*, согласуясь с данными из толкового словаря и историческим исследованием. Алгоритм выделяет изменение в частотности двух основных значения, но не вывел информацию по менее важным.

Молодец

В результате анализа семем лексемы *молодец* в толковых словарях были выделены пять групп значений, которые можно условно сформулировать следующим образом:

1. Молодой, крепкий, статный мужчина. (*«Молодой человек, достигший расцвета лет, крепкий и статный.»* в БТС, *«Молодой человек, сильный, крепкого сложения.»* в ТСО, *«обычно мн. Человек, обычно сильный, смелый, бесшабашный.»* в ТСО *«Молодцом называют молодого, здорового, привлекательного парня.»* в ТСД)

2. Удалец, храбрец, герой (в народной словесности). (*«Сильный и смелый герой; удалец, храбрец.»* в БТС, *«В народной словесности: удалец, храбрец.»* в ТСО)
3. Похвала, одобрение чьих-либо действий. (*«О том, чьи действия вызвали одобрение, удовлетворение у кого-л.»* в БТС, *«Выражение похвалы тому, кто делает что-н. хорошо, ловко, умело.»* в ТСО, *«Молодцом вы называете того, чьи действия вы одобряете, хвалите.»* в ТСД, *«Похвала»* в «Два века в двадцати словах»)
4. Слуга, помощник, служащий. (*«Служащий»* в «Два века в двадцати словах»)
5. Бандит или приспешник вражеских групп. (*«Пренебр. =Молодчик (3 зн.: Пособник, приспешник или участник каких-л. реакционных, вражеских или преступных групп, организаций.).»* в БТС)

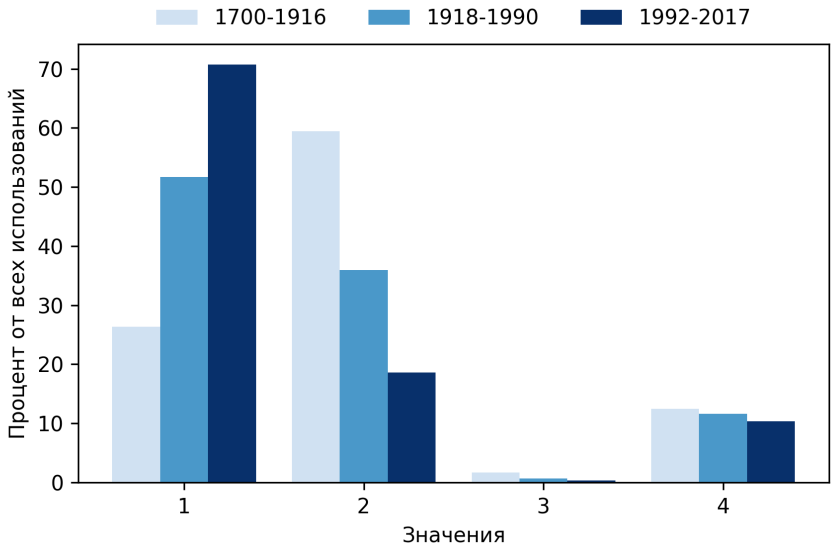


Рис. А.12.. Изменение значений слова *молодец*

Значения для визуализации слова «Молодец» (Параметры: eps=0.18, min_samples=10).

1. Употребляется как похвала, одобрение.
2. Молодой человек, юноша.
3. Употребляется как бранное слово.

4. О молодом человеке, отличающемся храбростью, удалью и т. п.

Анализ значений слова *молодец*

Первые четыре определения в визуализации имеют следующие соответствия с обобщенными значениями из словарей:

- ’Употребляется как похвала, одобрение.’ Это определение соответствует третьему обобщенному значению: *«Похвала, одобрение чьих-либо действий.»* Общие семы: «похвала», «одобрение». Данный пример является корректным.
- ’Молодой человек, юноша.’ Это определение соответствует первому обобщенному значению: *«Молодой, крепкий, статный мужчина.»* Общие семы: «молодой», «человек». Хотя в определении визуализации отсутствуют семы «крепкий» и «статный», это определение можно считать корректным, но недостаточно специфичным.
- ’Употребляется как бранное слово.’ Это определение можно считать близким к пятому обобщенному значению: *«Бандит или приспешник вражеских групп.»* Общие семы: «бранное слово» (имеется в виду негативная коннотация). Данное определение можно считать имеющим близкое значение, но не полностью соответствующим, так как не уточняется денотат.
- ’О молодом человеке, отличающемся храбростью, удалью и т. п.’ Это определение соответствует второму обобщенному значению: *«Удалец, храбрец, герой (в народной словесности).»* Общие семы: «молодой человек», «храбрость», «удаль». Это определение можно считать корректным.

Отсутствующие значения:

- **Слуга, помощник, служащий.**

Это значение отсутствует в визуализации. Возможно, это значение не достаточно распространено или редко упоминается в используемых данных.

Статистика по лексеме *молодец*

- Корректные: 2 (определения 1 и 4)
- Недостаточно специфичные: 1 (определение 2)
- Имеющие близкое значение: 1 (определение 3)

Таким образом, модель в целом корректно определяет основные значения слова «молодец», но не охватывает все возможные значения, представленные в словарях.

Перейдем к частотности значений.

Основным моментом, выделяемым книгой «Два века в двадцати словах», является выход значения ’Похвала, одобрение чьих-либо действий.’ на лидирующие позиции вместо изначального значения ’Молодой, крепкий, статный мужчина.’, что вы можете увидеть на графике снизу. В визуализации, сделанной алгоритмом, это также отображено. Значение ’Употребляется как похвала, одобрение.’ растёт с 30% в досоветский период растёт до 70% в постсоветский за счёт уменьшения использования ’Молодой человек, юноша.’. Также утверждается уменьшение использлвания значения ’Слуга, помощник, служащий.’, который не был выделен алгоритмом.

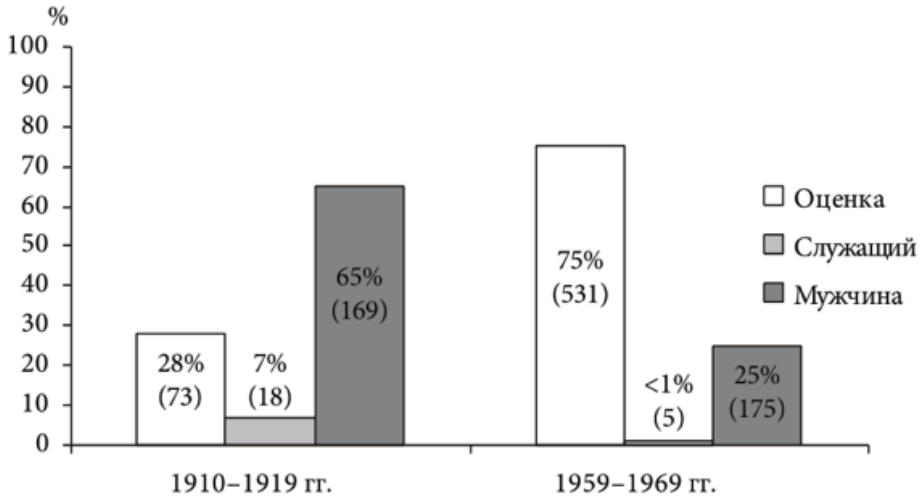


Рис. А.13.. График для слова *молодец* для 1910-1969 из книги «Два века в двадцати словах».

Таким образом, алгоритм в целом отражает значения, в которых использовалось слово *молодец*, согласуясь с данными из толкового словаря и историческим исследованием. Алгоритм выделяет основных значения, но не вывел ’Слуга, помощник, служащий.’.

Пакет

1. Упакованный в бумажную или иную обёртку какой-л. предмет (предметы); свёрток. (*«Упакованный в бумажную или иную обёртку какой-л. предмет (предметы); свёрток.»* в БТС, *«Бумажный сверток, упаковка с чем-н.»* в ТСРЯ, *«Предмет, который завёрнут в бумажную или другую упаковку.»* в ТСД, *«Упаковка, сверток»* в «Два века в двадцати словах»)
2. Бумажный или полиэтиленовый мешок для упаковки каких-л. предметов, продуктов и т.п. (*«Бумажный кулёк для упаковки каких-л. предметов, продуктов и т.п.»* в БТС, *«Бумажный мешок для продуктов, кулек.»* в ТСРЯ, *«Бумажный или полиэтиленовый кулёк с ручками или без для упаковки каких-либо предметов, продуктов и т. п.»* в ТСД, *«Ёмкость, тара»* в «Два века в двадцати словах»)

3. Конверт с письмом официально-делового содержания. (*«Конверт с письмом официально-делового содержания.»* в БТС, *«Конверт с письмом официального назначения.»* в ТСРЯ, *«Конверт с письмом официально-делового содержания.»* в ТСД, *«Письмо, конверт, почтовое отправление»* в «Два века в двадцати словах»)
4. Комплект документов, официальных бумаг. (*«Комплект документов, официальных бумаг.»* в БТС, *«В нек-рых сочетаниях: комплект документов, официальных бумаг.»* в ТСРЯ, *«Комплект документов или официальных бумаг.»* в ТСД)
5. Стопка ящиков или одинаковых деталей, строительных материалов и т.п., уложенных на специальный поддон для погрузки, перевозки и т.п. (*«Стопка ящиков или одинаковых деталей, строительных материалов и т.п., уложенных на специальный поддон для погрузки, перевозки и т.п.»* в БТС, *«Стопка грузов, уложенная на поддон (спец.).»* в ТСРЯ, *«Комплект одинаковых деталей, строительных материалов и т. п.»* в ТСД)
6. Совокупность информации, собранной для разовой передачи по компьютерной сети. (*«Совокупность информации, собранной для разовой передачи по компьютерной сети.»* в БТС)
7. Набор взаимосвязанных элементов, объединённых общей целью. (*«Наборы» ('нематериальная совокупность')* в «Два века в двадцати словах»)
8. Некоторое число акций какого-либо предприятия или компании, которым владеет человек или какая-либо организация, предприятие. (*«Пакетом акций является некоторое число акций какого-либо предприятия или компании, которым владеет человек или какая-либо организация, предприятие.»* в ТСД, *«Пакет акций»* в «Два века в двадцати словах»)

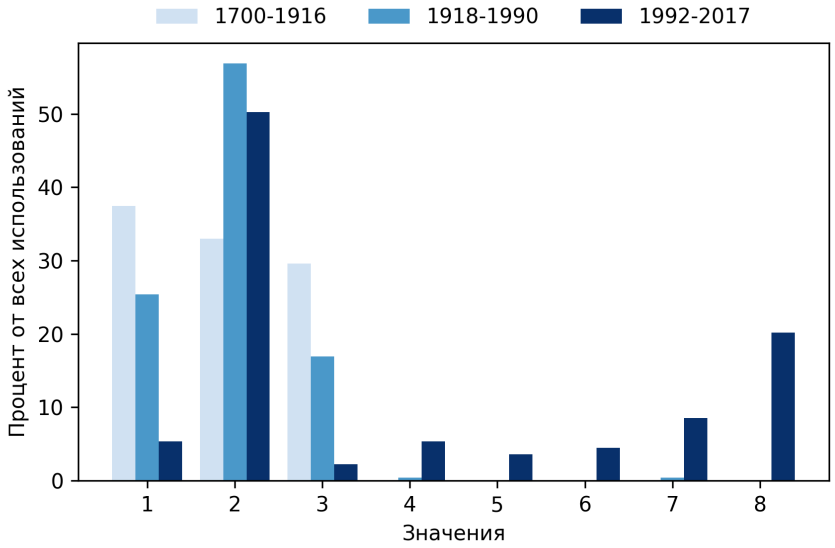


Рис. А.14.. Изменение значений слова *пакет*

Значения для визуализации слова «Пакет» (Параметры: eps=0.12, min_samples=8).

1. Письмо, посылка и т. п. в таком виде.
2. Бумажный или матерчатый мешочек с чем-либо для хранения, перевозки и т. п.
3. Письмо, посылка и т. п., запечатанные в такой конверт.
4. Совокупность каких-либо однородных, связанных между собой предметов, явлений и т. п.
5. Совокупность программных средств, объединенных по какому-либо признаку.
6. Часть чего-либо, принадлежащая кому-либо на определенных условиях.
7. Совокупность каких-либо однородных предметов, документов и т. п.
8. Совокупность акций какого-либо акционерного общества.

Анализ значений слова *пакет*

Все определения, кроме шестого, корректно сформулированы.

- ’Письмо, посылка и т. п. в таком виде.’, а также ’Письмо, посылка и т. п., запечатанные в такой конверт.’ имеет общий смысловой элемент с ’Конверт с письмом официально-делового содержания.’, а именно семы «письмо», «посылка», «конверт».
- ’Бумажный или матерчатый мешочек с чем-либо для хранения, перевозки и т. п.’ соответствует ’Бумажный или полиэтиленовый мешок для упаковки каких-л. предметов, продуктов и т.п.’, общие семы «мешок», «бумажный».
- ’Совокупность программных средств, объединенных по какому-либо признаку.’ частично соответствует ’Набор взаимосвязанных элементов, объединённых общей целью.’, так как включает те же семы «совокупность», «объединенных объектов», но является более узким, так как касается только программных средств. Среди примеров, которые были выделены алгоритмом, находятся такие, как «Для обработки же растровых изображений и конкретно цифровых фотографий у компании ”Corel” существует пакет Corel Paint Shop Pro Photo.», где значение слова действительно может быть описано как ’Совокупность программных средств, объединенных по какому-либо признаку.’, поэтому мы будем считать это определение корректным.
- ’Совокупность акций какого-либо акционерного общества.’ полностью соответствует ’Некоторое число акций какого-либо предприятия или компании, которым владеет человек или какая-либо организация, предприятие.’, так как включает те же семы «совокупность», «акции».
- ’Совокупность каких-либо однородных, связанных между собой предметов, явлений и т. п.’, , а также ’Совокупность каких-либо однородных предметов, документов и т. п.’ соответствует ’Набор взаимосвязанных элементов, объединённых общей целью.’.

- ’Часть чего-либо, принадлежащая кому-либо на определенных условиях.’ не имеет схожих определений среди обобщенных и является некорректным. Анализ примеров, для которых алгоритм дал такое определение, показывает, что большинство примеров связано с акциями, например, «Но контрольный пакет акций был размыт.». В данном случае логичным является определение, акцентирующее «совокупность».

Отсутствующие значения:

- ’Упакованный в бумажную или иную обёртку какой-л. предмет (предметы); свёрток.’ отсутствует среди предложенных моделью значений.
- ’Стопка ящиков или одинаковых деталей, строительных материалов и т.п., уложенных на специальный поддон для погрузки, перевозки и т.п.’ также отсутствует в визуализации. Это значение акцентируется на физической стопке предметов (ящиков, деталей) на поддоне, что могло быть причиной отсутствия в визуализации.

Ошибок в написании определений (орфографических, синтаксических) не обнаружено.

Таким образом, для лексемы *пакет* представлены:

- Корректные: 7
- Некорректные: 1

Перейдем к частотности значений.

Основные изменения в значениях слова *пакет*, указанные в книге «Два века в двадцати словах», расположены на графике снизу. В досоветский период значения, обозначающие ’почтовое отправление’ и ’упаковка, свёрток’ уступили место ’мешку, в том числе из полиэтилена’, ’набору’ и ’картонной упаковке’. Информация из визуализации алгоритма частично соответствует этим данным. ’Письмо, посылка и т. п. в таком виде.’ и ’Письмо, посылка и т. п., запечатанные в такой конверт.’ вместе преобладают в досовет-

ский период, уступая место значению ’Бумажный или матерчатый мешочек с чем-либо для хранения, перевозки и т. п.’ в советское время. Кроме того, согласно алгоритму, в постсоветское время появляются такие значения, как ’Совокупность каких-либо однородных, связанных между собой предметов, явлений и т. п.’, ’Совокупность акций какого-либо акционерного общества.’, что соответствует информации из книги.

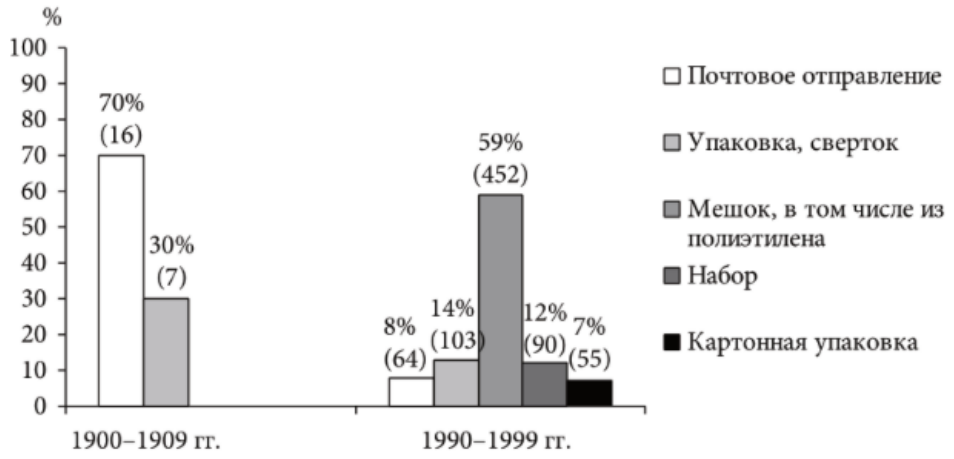


Рис. А.15.. График для слова *пакет* для 1900-1999 из книги «Два века в двадцати словах».

Таким образом, алгоритм в целом отражает значения, в которых использовалось слово *пакет*, согласуясь с данными из толкового словаря и историческим исследованием.

Передовой

1. Идущий, движущийся впереди остальных; ведущий. (*«Идущий, движущийся впереди остальных; ведущий.»* в БТС, *«Движущийся или находящийся впереди.»* в ТСО, *«Идущий впереди»* в «Два века в двадцати словах»)
2. Находящийся, действующий впереди, в авангарде (о военных силах). (*«Находящийся, действующий впереди, в авангарде (о военных силах).»* в БТС, *«В военном деле передовыми называют от-*

ряды, позиции и т. д., находящиеся ближе всего к месту боевых действий.» в ТСД, *«Расположенный в авангарде» (о военных действиях)* в «Два века в двадцати словах»)

3. Превосходящий других по уровню своего технического развития. (*«Превосходящий других по уровню своего развития; прогрессивный.»* в БТС, *«Не останавливающийся в развитии, прогрессивный.»* в ТСО, *«Передовыми называют очень современные, сложные и интересные методы, технологии и т. д.»* в ТСД)
4. Содержащий, излагающий прогресивные идеи, часто свободолюбивые, либеральные или демократические мысли. (*«Содержащий, излагающий свободолюбивые, либеральные мысли; демократический.»* в БТС, *«Передовыми называют новые современные идеи, книги, способствующие какому-то развитию общества, науки, литературы и т. п.»* в ТСД, *«Прогрессивный»* в «Два века в двадцати словах»)
5. Руководящая статья в газете, журнале, печатаемая на первом месте. (*«Передовая статья (руководящая редакционная статья в газете, журнале, печатаемая на первом месте).»* в ТСО, *«Статья»* в «Два века в двадцати словах»)
6. Превосходивший других по своим успехам в работе или опережавший других по производственным показателям. (*«В СССР: превосходивший других по своим успехам в работе или опережавший других по производственным показателям.»* в БТС, *«Передовым называли (во времена СССР) человека, коллектив и т. д., который достиг наибольших успехов в работе.»* в ТСД)
7. Человек, отправленный для передачи информации или выполнения определенной миссии; посланник, гонец. (*«Посланник, гонец»* в «Два века в двадцати словах»)

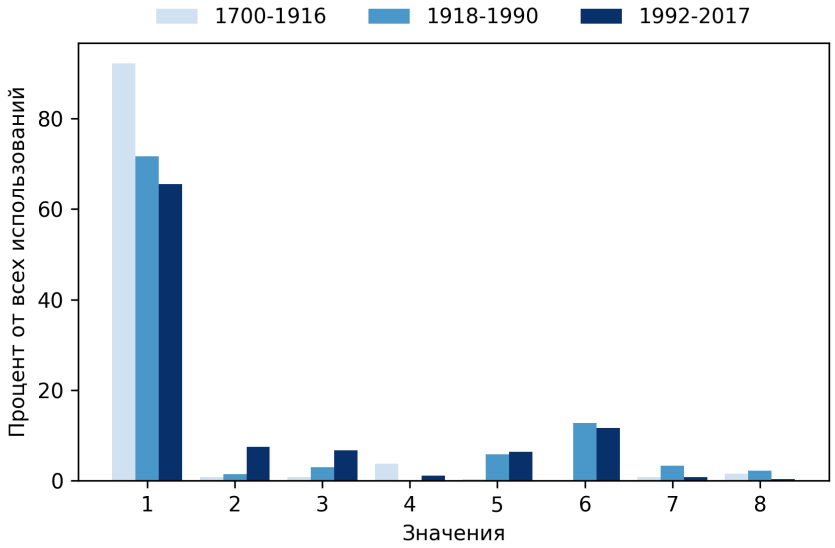


Рис. А.16.. Изменение значений слова *передовой*

Значения для визуализации слова «Передовой» (Параметры: eps=0.18, min_samples=12).

1. Находящийся в авангарде, в первых рядах чего-либо.
2. Основанный на новейших достижениях науки, техники и т. п.
3. Содержащий в себе новое, прогрессивное.
4. Являющийся передовицей.
5. Передний край обороны, боевых действий.
6. Передний край, линия фронта.
7. Главная, основная часть газеты, журнала.
8. Находящийся на более высокой ступени общественного развития по сравнению с другими.

Анализ значений слова *передовой*

Первое и четвертое определения корректно сформулированы. Второе и пятое определения не соответствуют обобщенным значениям.

- ’Находящийся в авангарде, в первых рядах чего-либо.’ имеет общий смысловой элемент с ’Идущий, движущийся впереди остальных; ве-

дущий.’, а именно семы «находящийся», «в авангарде/в первых рядах».

- ’Основанный на новейших достижениях науки, техники и т. п.’ полностью соответствует ’Превосходящий других по уровню своего технического развития.’, так как включает те же семы «новейшие достижения», «наука, техника».
- ’Содержащий в себе новое, прогрессивное.’ имеет общий смысловой элемент с ’Содержащий, излагающий прогрессивные идеи, часто свободолюбивые, либеральные или демократические мысли.’, а именно семы «содержащий», «новое/прогрессивное», однако представляет собой более широкое определение, так как не ограничивается только идеями.
- ’Являющийся передовицей.’ и ’Главная, основная часть газеты, журнала.’ полностью соответствуют ’Руководящая статья в газете, журнале, печатаемая на первом месте.’, так как семы «главная/основная», «часть газеты, журнала/статья» полностью отражают смысл «руководящая статья».
- ’Передний край обороны, боевых действий.’ и ’Передний край, линия фронта.’ вместе имеют частичное соответствие с ’Находящийся, действующий впереди, в авангарде (о военных силах).’, так как семы «передний край/в авангарде», «боевых действий/военных сил» частично отражают смысл «впереди, в авангарде». Однако, данное определение ближе подходит к значению субстантивированного прилагательного «передовая» – ’Участок оборонительной линии, соприкасающейся с неприятельским фронтом; передовая линия.’. Например, для контекста «Такое случилось еще раз, потому что отказать во встрече уходящему на передовую, когда к Москве подступают немцы, было невозможно.» было сгенерировано ’Передний

край обороны, боевых действий.’. Исследуемые определения будут считаться нами как корректные.

Отсутствующие значения:

- ’Превосходивший других по своим успехам в работе или опережавший других по производственным показателям.’ отсутствует среди предложенных моделью значений. Можно предположить, что информации из контекста использований недостаточно для отделения этого значения от ’Превосходящий других по уровню своего развития; прогрессивный.’.
- ’Человек, отправленный для передачи информации или выполнения определенной миссии; посланник, гонец.’ также отсутствует в визуализации. Однако, модель способна на выделение данного значения.

Таким образом, для лексемы *передовой* представлены:

- Корректные: 7
- Недостаточно специфичные: 1

Перейдем к частотности значений.

В книге лишь значение ’Основанный на новейших достижениях науки, техники и т. п.’ указывается как появившееся после 1917 года, что подтверждается в визуализации алгоритма, где оно представлено только в советский и постсоветский период. Кроме того, указано, что военное значение ’Находящийся в авангарде, в первых рядах чего-либо.’ уступает переносным значениям, акцентирующимся на прогрессивности, чего не наблюдается в визуализации, сделанной алгоритмом.

Снизу вы можете увидеть графики из книги «Два века в двадцати словах».

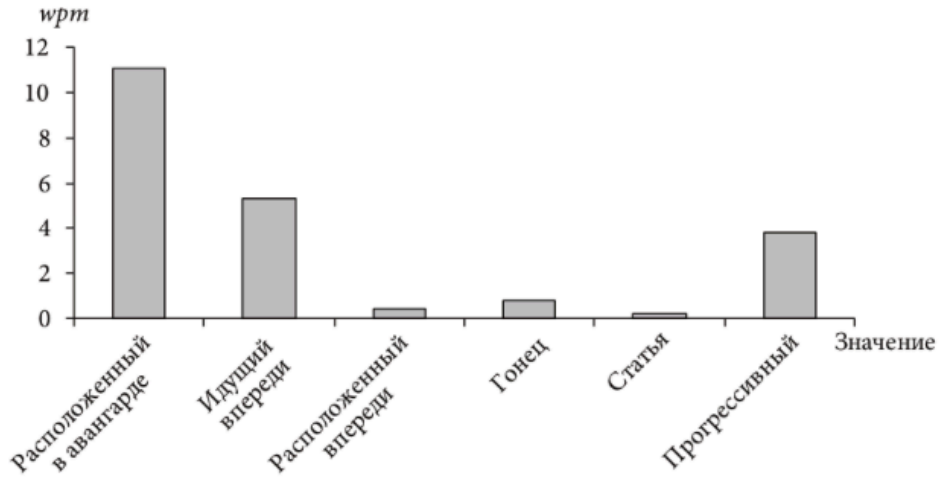


Рис. А.17.. График для слова *передовой* для 1830-1859 из книги «Два века в двадцати словах».

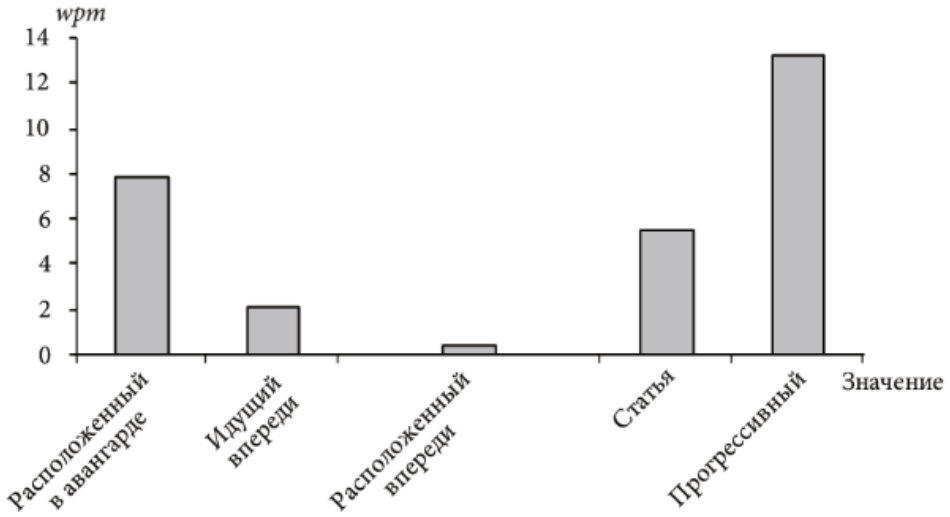


Рис. А.18.. График для слова *передовой* для 1930-1939 из книги «Два века в двадцати словах».

Таким образом, нельзя сказать, что алгоритм отражает значения, в которых использовалось слово *передовой*, так как они не полностью согласуются с данными из толкового словаря и историческим исследованием.

Пионер

1. Человек, впервые проникший в неисследованную страну, область и поселившийся в ней. (*«Человек, впервые проникший в неисследо-*

ванную страну, область и поселившийся в ней.» в БТС, *«Человек, к-рый одним из первых пришел и поселился в новой неисследованной стране, местности.»* в ТСО, *«Первый поселенец на какой-либо территории»* в «Два века в двадцати словах»)

2. Тот, кто положил начало чему-либо в какой-либо сфере деятельности, в науке, культуре; новатор, зачинатель. (*«Тот, кто прокладывает новые пути в какой-л. сфере деятельности, в науке, культуре; новатор, зачинатель.»* в БТС, *«Человек, к-рый положил начало чему-н. новому в области науки, культуры »* в ТСО, *«Первооткрыватель»* в «Два века в двадцати словах»)
3. Член добровольной самодеятельной детской организации. (*«В СССР: член добровольной самодеятельной детской организации, объединявшей детей и подростков от 10 до 15 лет.»* в БТС, *«Член детской организации в СССР и ряда детских организаций в нек-рых других странах.»* в ТСО, *«Член детской организации»* в «Два века в двадцати словах»)
4. Профессия, связанная со строительством мостов и укреплений. (*«Сапер.»* в «Два века в двадцати словах»)

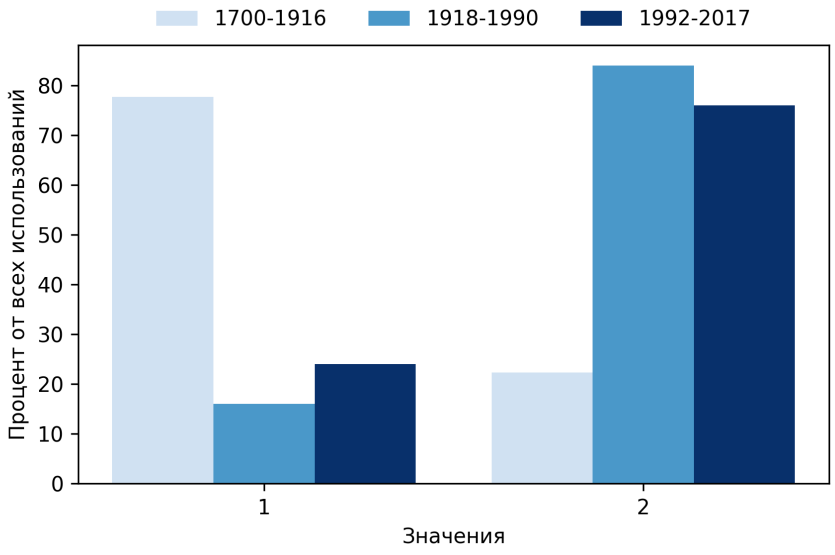


Рис. А.19.. Изменение значений слова *пионер*

Значения для визуализации слова «Пионер» (Параметры: eps=0.18, min_samples=15).

1. Тот, кто первым начал что-либо делать, кто является основоположником чего-либо.
2. Юный член пионерской организации.

Анализ значений слова *пионер*

Оба определения корректно сформулированы.

- ’Тот, кто первым начал что-либо делать, кто является основоположником чего-либо.’ имеет общий смысловой элемент с ’Тот, кто положил начало чему-либо в какой-либо сфере деятельности, в науке, культуре; новатор, зачинатель.’, а именно семы «начало», «деятельность», «новаторство».
- ’Юный член пионерской организации.’ полностью соответствует ’Член добровольной самодеятельной детской организации.’, так как включает те же семы «член», «детская организация».

Отсутствующие значения:

- ’Человек, впервые проникший в неисследованную страну, область и поселившийся в ней.’ отсутствует среди предложенных моделью значений. В книге «Два века в двадцати словах» указывается, что это в русском имело единичные использования. Можно предположить, что оно не вошло в исследуемую выборку. Однако, модель способна на выделение данного значения.
- ’Профессия, связанная со строительством мостов и укреплений.’ также отсутствует в визуализации. В книге «Два века в двадцати словах» указывается, что это значение редкое. Можно предположить, что оно не вошло в исследуемую выборку. Однако, модель способна на выделение данного значения.

Ошибок в написании определений не обнаружено.

Таким образом, для лексемы *пионер* представлены:

- Корректные: 2

Перейдем к частотности значений.

Значения, связанные с первооткрывательством, а также с сапёром, появились до 1917 года. Новым значением является ’Член добровольной самодеятельной детской организации.’, появившееся в советское время. В визуализации алгоритма указано, около 20% использований слова в таком значении в досоветский период, что объясняется двусмысленностью части примеров, например, «Пионеры слушают это и восхищаются.», где необходим дополнительный контекст для установления значения.

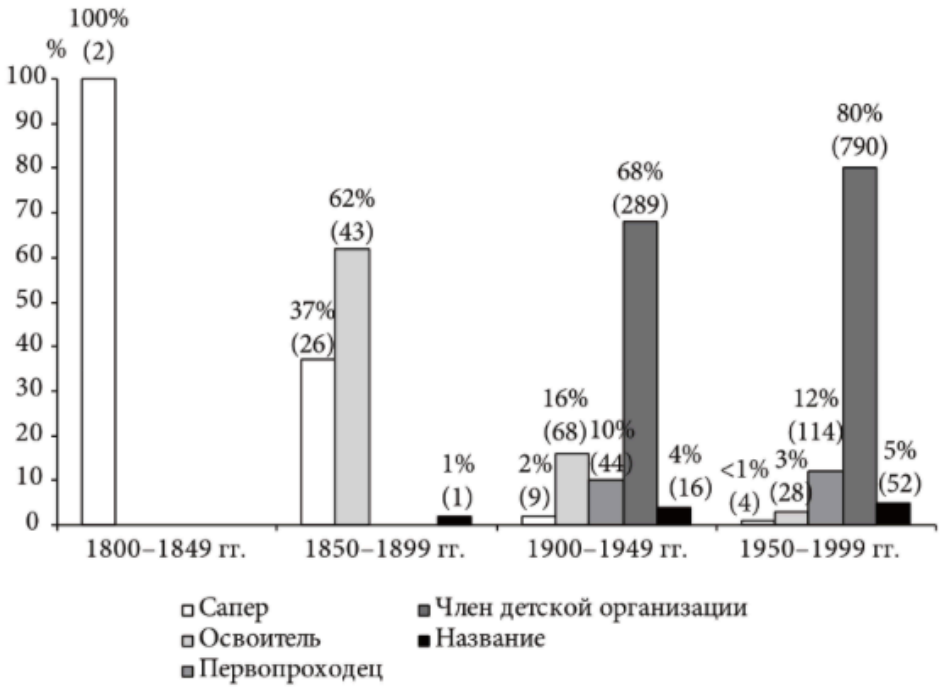


Рис. А.20.. График для слова *пионер* из книги «Два века в двадцати словах».

Таким образом, алгоритм лишь частично отражает значения, в которых использовалось слово *пионер*, так как из 4 значений выделено только 2, из которых около 20% размечено неверно.

Пожалуй

1. Вежливое обращение или просьба. (*«Повелительный наклон, при вежливом обращении.»* в БТС, *«Будь добр.»* в «Два века в двадцати словах»)
2. Выражение допущения или вероятности. (*«Вводное слово, выражающее допущение возможного, склонность согласиться.»* в ТСО, *«Словом пожалуй обозначают вероятность чего-либо.»* в ТСД, *«Возможно.»* в «Два века в двадцати словах»)
3. Выражение намерения совершить действие. (*«Слово пожалуй употребляется в том случае, если кто-либо сообщает о своём намерении совершить какое-либо действие, которое кажется этому человеку наиболее приемлемым в какой-либо ситуации.»* в ТСД, *«Склоняюсь к тому, что...»* в «Два века в двадцати словах»)
4. Выражение нерешительного, неопределённого согласия. (*«Частица, выражающая не уверенное согласие.»* в ТСО, *«Словом пожалуй обозначают нерешительное, неопределённое согласие что-либо сделать.»* в ТСД, *«Ладно, согласен.»* в «Два века в двадцати словах»)

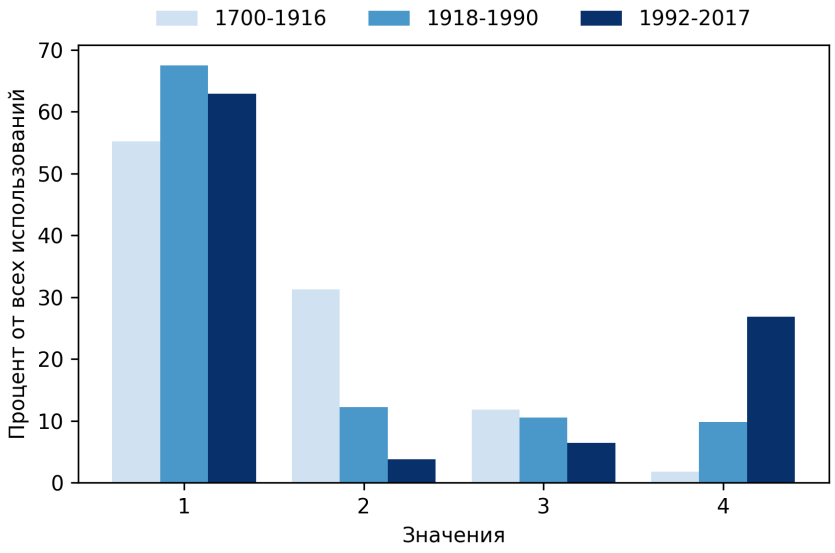


Рис. А.21.. Изменение значений слова *пожалуй*

Значения для визуализации слова «Пожалуй» (Параметры: eps=0.15, min_samples=10).

1. Употребляется для выражения сомнения, неуверенности в чем-либо.
2. Употребляется для выражения просьбы, приглашения к чему-либо.
3. Вполне возможно, вероятно.
4. Употребляется для присоединения предложений или отдельных членов предложений, усиливающих или уточняющих высказанную мысль.

Анализ значений слова *пожалуй*

Первое, второе и третье определения корректно сформулированы. Четвертое определения не соответствуют обобщенным значениям.

- ’Употребляется для выражения сомнения, неуверенности в чем-либо.’ не имеет похожих обобщенных определений, а примеры, которые имеют данное определение, например, «От отца я, пожалуй, кроме книг ничего в подарок и не получал.» было бы корректно отнести к ’Вполне возможно, вероятно.’
- ’Употребляется для выражения просьбы, приглашения к чему-либо.’ соответствует ’Вежливое обращение или просьба’ с общими семами «просьбы».
- ’Вполне возможно, вероятно.’ соответствует ’Выражение допущения или вероятности.’, так как включает те же семы «возможности», «вероятности».
- ’Употребляется для присоединения предложений или отдельных членов предложений, усиливающих или уточняющих высказанную мысль.’ является некорректным значением, так как описывает такое функциональное использование в синтаксисе, что отсутствует в словарях.

Отсутствующие значения:

- ’Выражение намерения совершить действие’ также отсутствует в визуализации. Однако, модель способна на выделение данного значения.

Ошибок в написании определений (орфографических, синтаксических) не обнаружено.

Таким образом, для лексемы *пожалуй* представлены:

- Корректные: 2
- Некорректные: 2

Перейдем к частотности значений.

В книге сообщается, что изначальные значения ’Вежливое обращение или просьба’ и ’Выражение нерешительного, неопределённого согласия.’ были вытсенены в течение XIX века преобладающим на сегодняшний момент значением ’Вполне возможно, вероятно.’ В визуализации данная информация подтверждается для значения ’Употребляется для выражения просьбы, приглашения к чему-либо.’, имевшее в постсовесткий период более 30% использований и около 5% в постсовесткий. Тем не менее, затруднительно установить корректность статистики далее из-за некорректных определений.

Снизу вы можете увидеть графики из книги «Два века в двадцати словах».

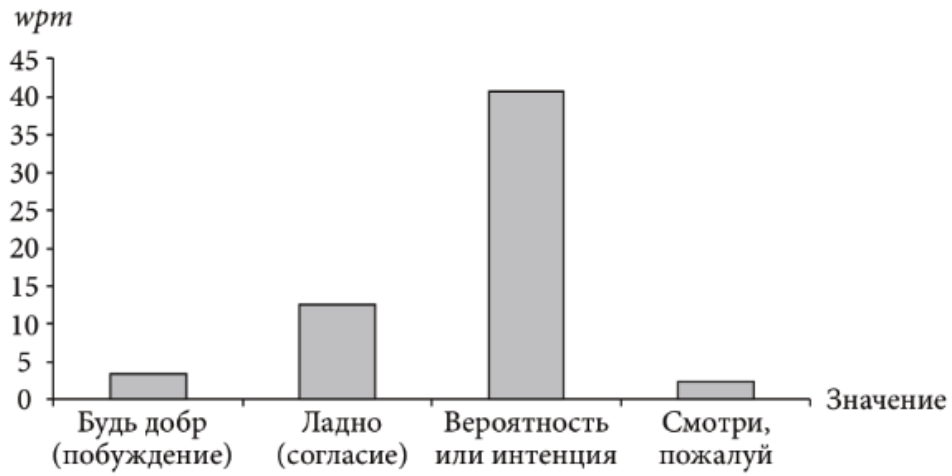


Рис. А.22.. График для слова *пожалуй* для 1831-1860 из книги «Два века в двадцати словах».

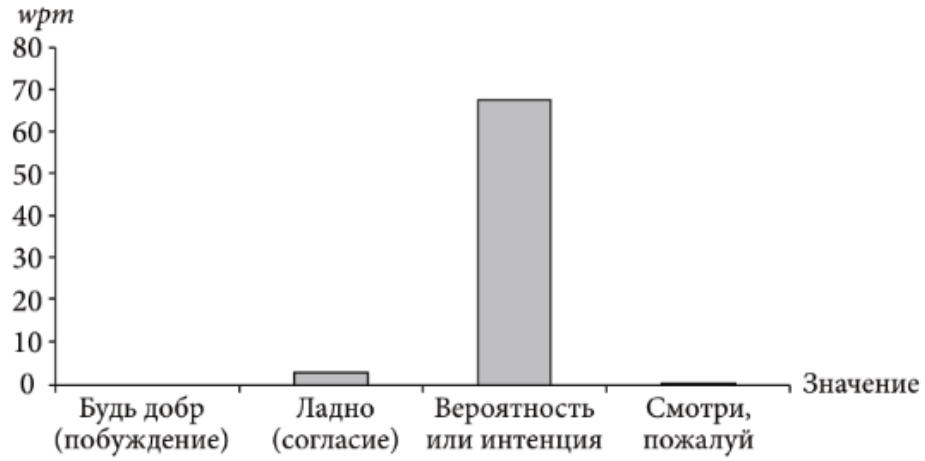


Рис. А.23.. График для слова *пожалуй* для 1900-1910 из книги «Два века в двадцати словах».

Таким образом, алгоритм лишь частично отражает значения, в которых использовалось слово *пожалуй*, так как из 4 значений корректны только 2.

Пока

1. В течение некоторого времени; до сих пор ещё; впредь до чего-л. (*«В течение некоторого времени; до сих пор ещё; впредь до чего-л.»* в БТС, *«В течение нек-рого времени, впредь до чего-н.; до сих*

пор еще.» в ТСО, *«Наречие – в течение некоторого времени, до сих пор еще.»* в «Два века в двадцати словах»)

2. В то время как. (*«В то время как; до того времени как.»* в БТС, *«В течение того времени как.»* в ТСО, *«Союз с фоновым значением (’в то время как").»* в «Два века в двадцати словах»)
3. До того времени как. (*«В то время как; до того времени как.»* в БТС, *«Союз с предельным значением ("вплоть до того как’).»* в «Два века в двадцати словах»)
4. Употребляется при прощании, до свидания. (*«Приветствие при прощании, до свидания.!»* в ТСО, *«Элемент формулы прощания.»* и *«Этикетное слово — до свидания.»* в «Два века в двадцати словах»)

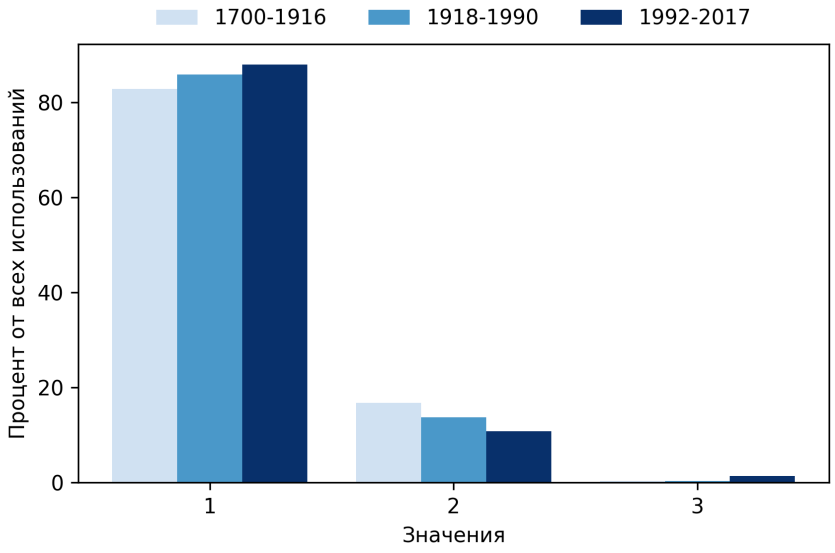


Рис. А.24.. Изменение значений слова *пока*

Значения для визуализации слова «Пока» (Параметры: eps=0.23, min_samples=5).

1. В настоящее время, до тех пор.
2. Употребляется при обозначении времени, в течение которого совершается действие.
3. Употребляется при прощании с кем-л.

Первое, второе и третье определения корректно сформулированы.

- ’В настоящее время, до тех пор.’ имеет общий смысловой элемент с ’В течение некоторого времени; до сих пор ещё; впредь до чего-л.’, а именно семы «время», «до сих пор», «в течение».
- ’Употребляется при обозначении времени, в течение которого совершается действие.’ полностью соответствует ’В то время как.’, так как включает те же семы «время», «совершение действия».
- ’Употребляется при прощании с кем-л.’ полностью соответствует ’Употребляется при прощании, до свидания.’, так как включает те же семы «прощание», «до свидания».

Отсутствующие значения:

- ’До того времени как.’ отсутствует среди предложенных моделью значений. Это значение близко к ’В то время как’, но с акцентом на предельность времени, что могло привести к отсутствию этого значения в визуализации.

Ошибок в написании определений (орфографических, синтаксических) не обнаружено.

Таким образом, для лексемы *пока* представлены:

- Корректные: 3

Перейдем к частотности значений.

В книге сообщается, что изначальным и всегда преобладающим было использование слова в качестве союза, после чего в XIX веке появилось использование как наречие, а затем в советский период – как этикетное слово. Данные из визуализации алгоритма (график снизу) поддерживают появление значения ’Употребляется при прощании с кем-л.’ поздно – несколько процентов для постсоветского периода, однако данные для наречия и союза не совпадают. Можно предположить, что модели сложно различать эти значения из-за их схожести. Например, для «Когда мы забирали щенка, нас

предупредили, что ей категорически нельзя наверх забираться, пока у нее слабые лапы.» было сгенерировано ’В настоящее время, до тех пор.’, что относит его к наречию, но из примера видно, что пока связывает части предложения и является союзом.

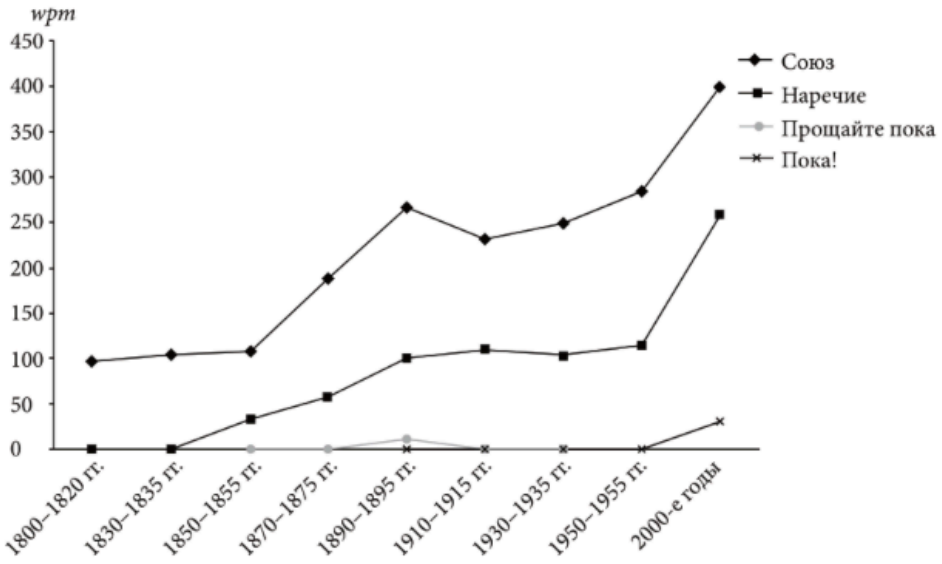


Рис. А.25.. График для слова *пока* из книги «Два века в двадцати словах».

Таким образом, алгоритм лишь по большей части не отражает значения, в которых использовалось слово *пожалуй*, так как из 3 выделенных значений, хоть и правильно сформулированных, статистика использования не согласуется с данными книги.

Привет

1. Обращение к кому-либо с выражением дружеского расположения, дружеских чувств, доброжелательства. (*«Обращённое к кому-л. выражение дружеского расположения, дружеских чувств, доброжелательства.»* в БТС, *«Обращенное к кому-н. выражение чувства личной приязни, доброго пожелания, солидарности.»* в ТСРЯ, *«Словесное или несловесное выражение внимания к собеседнику.»* в «Два века в двадцати словах»)

2. Дружелюбное, ласковое обращение с кем-либо. (*Дружелюбное, ласковое обращение с кем-либо.* в «Два века в двадцати словах»)
3. Вежливо-фамильярная форма приветствия при встрече или расставании. (*«Дружеское или фамильярное приветствие, обращённое к кому-л. при встрече или расставании.»* в БТС, *«Приветствие при встрече или расставании.»* в ТСРЯ, *«Если кто-либо говорит Привет! при встрече с каким-либо человеком или группой людей, значит, он просто употребляет вежливо-фамильярную форму приветствия.»* в ТСД, *«Здравствуйте».* в «Два века в двадцати словах»)
4. Выражение удивления, несогласия, иронии. (*«Выражение удивления, несогласия, иронии.»* в БТС, *«Выражение недоумения, удивленного несогласия.»* в ТСРЯ, *«Если один человек говорит Привет! другому в ответ на какие-либо не понравившиеся ему слова или действия, значит, он тем самым выражает удивление, несогласие, иронию.»* в ТСД)
5. Формула заключения письма с выражением внимания к собеседнику. (*«С приветом, друзья! Ну я ухожу, п.! С дружеским, сердечным, большим и т.п. приветом; с приветом (заключительная формула письма).»* в БТС, *«Иногда слова С (большим, пламенным и т. п.) приветом используются в качестве формально-вежливой заключительной фразы в письме.»* в ТСД, *«Формула конца письма с выражением внимания к адресату письма.»* в «Два века в двадцати словах»)
6. Формула выражения внимания к третьему лицу. (*«Формула конца письма с выражением внимания к третьему лицу.»* в «Два века в двадцати словах»)
7. Отсутствие ответа или реакции на обращение. (*«Ни ответа ни привета. Об отсутствии ответа, отзыва на чьё-л. обращение, пись-*

мо.» в БТС, *«Ни ответа ни привета - нет никакого ответа от кого-н., никаких известий о ком-н.»* в ТСРЯ, *«Говоря, что от кого-либо не слышно ни ответа, ни привета, вы подразумеваете под этим долгое отсутствие какого-либо отклика со стороны этого человека на ваше обращение, письмо.»* в ТСД)

8. Описание состояния человека, ведущего себя странно или глуповато. (*«С приветом, в зн. прил. Разг. Со странностями, глуповатый или не совсем нормальный (о человеке).»* в БТС, *«С приветом кто (прост.) - со странностями, не совсем нормален.»* в ТСРЯ, *«Если вы говорите, что кто-либо (совсем) с приветом!, вы в грубоватой или ироничной форме выражаете своё мнение о том, что этот человек — со странностями, глуповат или не совсем нормален, или ведёт себя таким образом в данной ситуации.»* в ТСД)

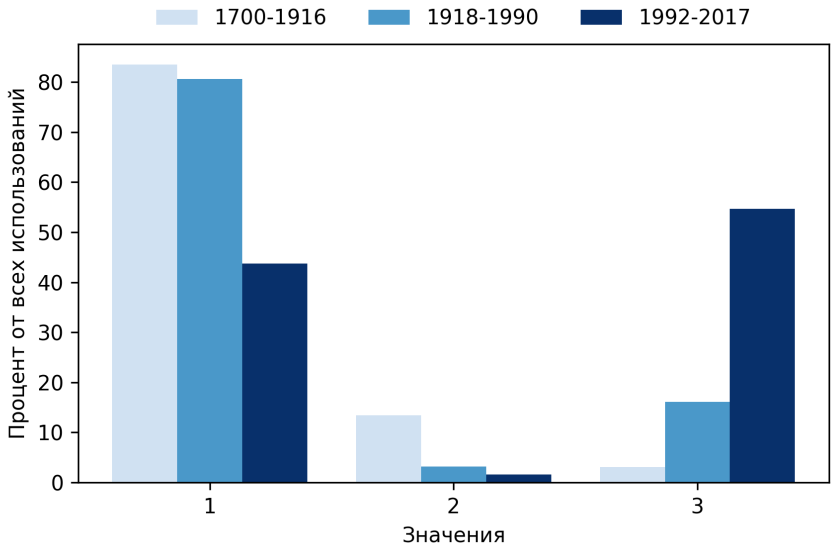


Рис. А.26.. Изменение значений слова *Привет*

Значения для визуализации слова «Привет» (Параметры: eps=0.1, min_samples=20).

1. Устное или письменное обращение к кому-либо с пожеланием доброго здоровья, счастья, успехов и т. п.
2. Доброжелательное отношение к кому-, чему-либо.

3. Добрый день, доброе утро.

Анализ значений слова *привет*

Первое и третье определения корректно сформулированы. Второе определение имеет обобщенное значение и соответствует конкретным значениям из словарей.

- ’Устное или письменное обращение к кому-либо с пожеланием доброго здоровья, счастья, успехов и т. п.’ имеет общий смысловой элемент с ’Обращение к кому-либо с выражением дружеского расположения, дружеских чувств, доброжелательства.’, а именно семы «обращение» и «доброжелательность».
- ’Доброжелательное отношение к кому-, чему-либо.’ соответствует значению ’Дружелюбное, ласковое обращение с кем-либо.’, так как включает те же семы «доброжелательность/дружелюбность» и «отношение/обращение».
- ’Добрый день, доброе утро.’ соответствует значению ’Вежливо-фамильярная форма приветствия при встрече или расставании.’, так как представляет из себя синонимичный ряд форм приветствия при встрече.

Отсутствующие значения:

- ’Выражение удивления, несогласия, иронии.’ отсутствует среди предложенных моделью значений. Можно предположить, что информации из контекста использований недостаточно для отделения этого значения от общего «доброжелательного обращения».
- ’Формула заключения письма с выражением внимания к собеседнику.’ и ’Формула выражения внимания к третьему лицу.’ также отсутствуют в визуализации. Однако, модель способна на выделение данных значений при более детализированном контексте.

- ’Отсутствие ответа или реакции на обращение.’ отсутствует среди предложенных моделью значений. Это может быть связано с недостаточной частотностью данного значения в корпусе данных.
- ’Описание состояния человека, ведущего себя странно или глуповато.’ также отсутствует в визуализации. Причиной может быть ограниченность контекста или недостаточная выраженность данной семы в корпусе.

Ошибок в написании определений (орфографических, синтаксических, повторений слова и так далее) не обнаружено.

Таким образом, для лексемы *привет* представлены:

- Корректные: 3

Перейдем к частотности значений.

Судя по книге «Два века в двадцати словах» (график снизу), изначальные использования слова *привет* имели значения ’Обращение к кому-либо с выражением дружеского расположения, дружеских чувств, доброжелательства.’ и ’Дружелюбное, ласковое обращение с кем-либо.’, где первое значение было более распространено. Ближе к завершению советского периода и после на первое место выходит использование слова *привет* как аналог *здравствуйте*. Все эти данные подкрепляются в нашей визуализации, где значение ’Добрый день, доброе утро.’ растёт с меньше 5% в досоветский период до около 55% в постсоветский, вытесняя ’Устное или письменное обращение к кому-либо с пожеланием доброго здоровья, счастья, успехов и т. п.’ с около 80% до 45% и ’Доброжелательное отношение к кому-, чему-либо.’ с 15% до 2-3%. Более того, на графики из книги заметно, что использование ’Доброжелательное отношение к кому-, чему-либо.’ снижается уже в начале советского периода, что так же отражено в нашей визуализации.

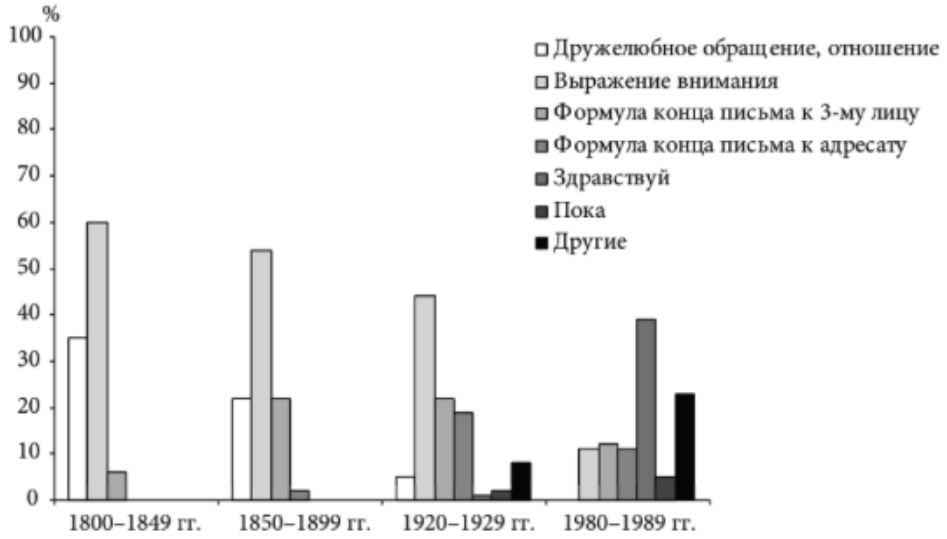


Рис. А.27.. График для слова *привет* из книги «Два века в двадцати словах».

Таким образом, алгоритм отражает основные значения, в которых использовалось слово *привет*, согласуясь с данными из толкового словаря и историческим исследованием.

Пружина

1. Упругая узкая металлическая пластина или нить, согнутая преимущественно в форме спирали. (*«Узкая упругая металлическая пластина или закрученная спиралью металлическая нить (служащая обычно для приведения в действие механизмов, амортизации ударов и т.п.)»* в БТС, *«Упругая узкая металлическая пластина или нить, согнутая преимущ. спиралью.»* в ТСО, *«Он носил маску с железною пружиною, которая не мешала ему есть.»* в «Два века в двадцати словах»)
2. Переносно, движущая сила в каком-то деле. (*«Движущая сила в каком-н. деле.»* в ТСО, *«Движущая сила чего-л.»* в БТС *«Природа есть первоначальная всему причина и самодвижущаяся пружина.»* в «Два века в двадцати словах»)

3. Метафора сжатости, а именно, упругость как свойство объекта или субъекта. (*«Метафора сжатости (движения).»* в «Два века в двадцати словах»)

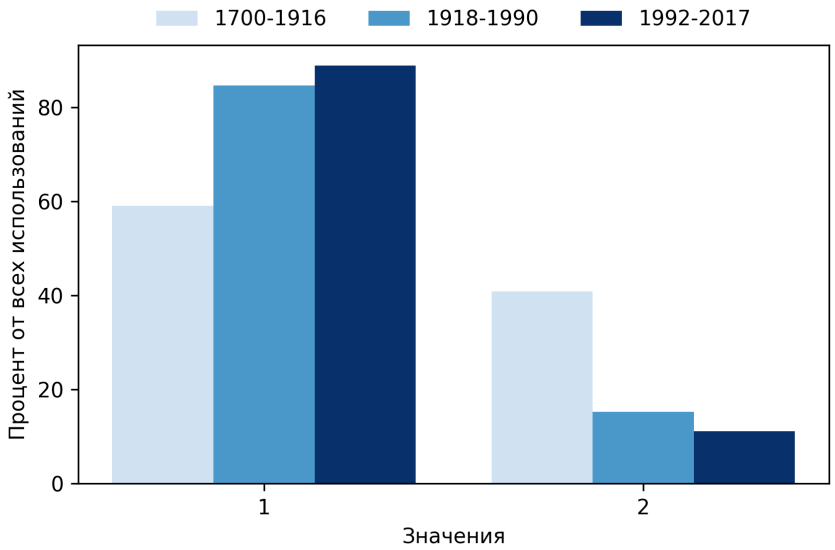


Рис. А.28.. Изменение значений слова *Пружина*

Значения для визуализации слова «Пружина» (Параметры: eps=0.25, min_samples=50).

1. Механизм, приводимый в действие сжатием и разжатием упругого стержня.
2. То, что является движущей силой, источником чего-либо.

Анализ значений слова *машина*

Первое и второе определения корректно сформулированы.

- ’Механизм, приводимый в действие сжатием и разжатием упругого стержня.’ имеет общий смысловой элемент с ’Упругая узкая металлическая пластина или нить, согнутая преимущественно в форме спирали.’, а именно семы «механизм», «упругость», «сжатие/разжатие», «стержень» (подразумевается металлическая пластина или нить).

- ’То, что является движущей силой, источником чего-либо.’ полностью соответствует ’Переносно, движущая сила в каком-то деле.’, так как включает те же семы «движущая сила», «источник».
- Метафорическое значение ’Метафора сжатости, а именно, упругость как свойство объекта или субъекта.’ не представлено в визуализации. Это значение, возможно, недостаточно распространено в исследуемом материале, поэтому не вошло в визуализацию.

Ошибок в написании определений (орфографических, синтаксических, повторений слов и т.д.) не обнаружено.

Таким образом, для лексемы *пружина* представлены:

- Корректные: 2

Перейдем к частотности значений.

В книге «Два века в двадцати словах» (графики ниже) сообщается о постепенной замене преобладающего метафорического значения слова (’Переносно, движущая сила в каком-то деле.’) на прямое (’Упругая узкая металлическая пластина или нить, согнутая преимущественно в форме спирали.’) в конце XIX века и о нынешнем преобладании прямого значения. Такие же данные представлены в нашей визуализации, где использования ’То, что является движущей силой, источником чего-либо.’ падают с 40% до 10%.

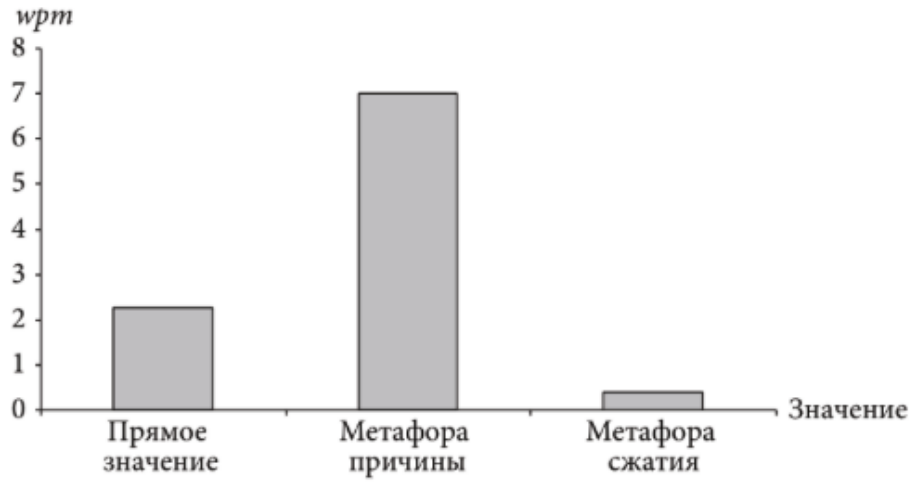


Рис. А.29.. График для слова *пружина* для 1830-1859 из книги «Два века в двадцати словах».

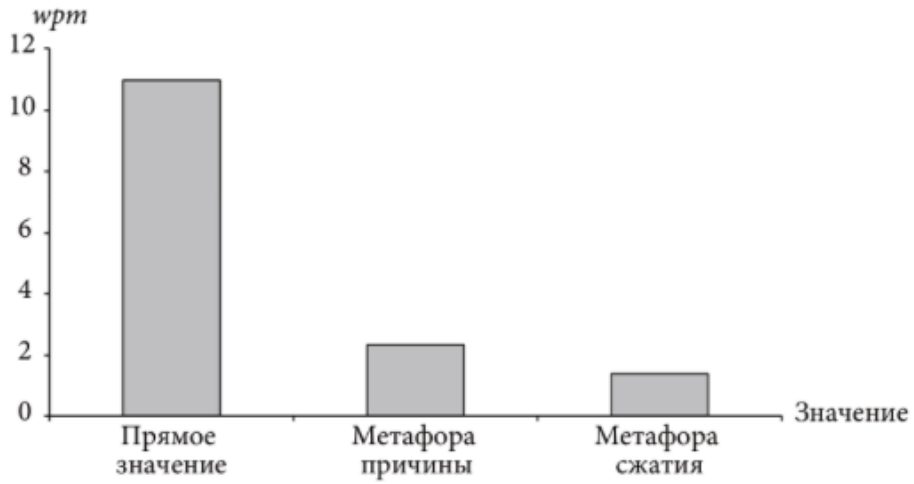


Рис. А.30.. График для слова *пружина* для 1960-2008 из книги «Два века в двадцати словах».

Таким образом, алгоритм отражает основные значения, в которых использовалось слово *пружина*, согласуясь с данными из толкового словаря и историческим исследованием.

Публика

В результате анализа семем лексемы *публика* в толковых словарях были выделены следующие группы значений:

1. Люди, присутствующие в качестве зрителей, слушателей, посетителей. (*«Лица, находящиеся где-либо в качестве посетителей, зрителей, слушателей.»* в БТС, *«Люди, находящиеся где-нибудь в качестве зрителей, слушателей, пассажиров.»* в ТСО, *«Публикой называют людей, которые собираются, присутствуют где-либо в качестве зрителей, слушателей.»* в ТСД, *«Публикой называют людей, которые собираются, присутствуют где-либо в качестве посетителей.»* в ТСД, *«Аудитория, зрители, слушатели»* в «Два века в двадцати словах»)
2. Люди и общество вообще. (*«Люди, общество.»* в БТС, *«Вообще люди, общество.»* в ТСО, *«Публикой иронично называют категорию людей, которым свойственны какие-либо общие признаки.»* в ТСД)
3. Светское общество, привилегированный класс населения. (*«Светское общество.»* в «Два века в двадцати словах»)
4. Группа лиц, объединённые по общим признакам, часто с негативной коннотацией. (*«Неодобрительно о лицах, объединённых по каким-либо признакам.»* в БТС, *«Общество или отдельные лица, объединённые по каким-н. общим признакам.»* в ТСО, *«Публикой иронично называют категорию людей, которым свойственны какие-либо общие признаки.»* в ТСД)
5. Пассажиры, люди в общественном транспорте. (*«Пассажиры»* в «Два века в двадцати словах», *«Люди, находящиеся где-нибудь в качестве зрителей, слушателей, пассажиров.»* в ТСО)
6. Читатели, аудитория, воспринимающая литературные или иные творческие произведения. (*«Читатели»* в «Два века в двадцати словах»)
7. Обозримая группа людей. (*«Народец (обозримая группа людей).»* в «Два века в двадцати словах»)

8. Скопление народа, уличная толпа, масса. (*«Уличная толпа, масса.»* в «Два века в двадцати словах»)

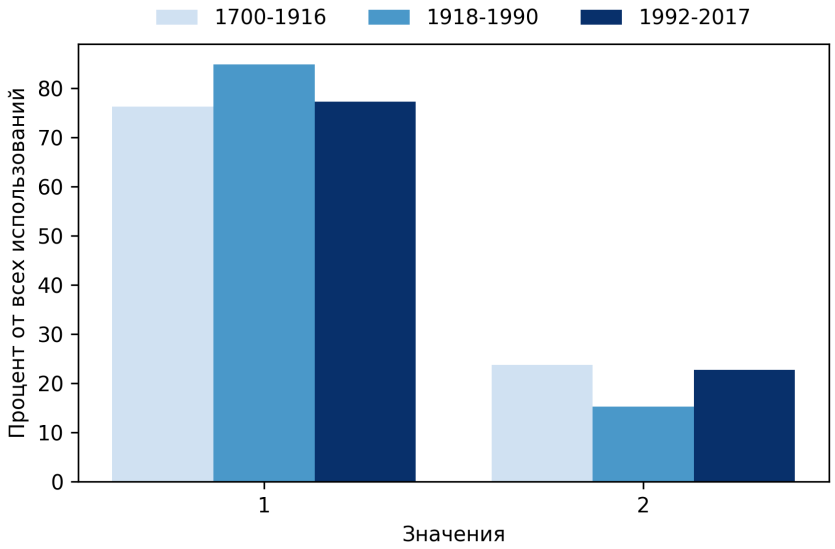


Рис. А.31.. Изменение значений слова *Публика*

Значения для визуализации слова «Публика» (Параметры: eps=0.11, min_samples=10).

1. Люди, присутствующие на каком-л. собрании, спектакле, концерте и т. п.
2. Общество, народ.

Анализ значений слова *публика*

Оба определения корректно сформулированы.

Оба определения имеют явные аналоги в составленном ранее описании значений слова.

- ’Люди, присутствующие на каком-л. собрании, спектакле, концерте и т. п.’ имеет соответствие с ’Люди, присутствующие в качестве зрителей, слушателей, посетителей.’. Общими смысловыми элементами являются «люди», «присутствующие», «мероприятие».
- ’Общество, народ.’ соответствует значению ’Люди и общество вообще.’. Общими семами являются «люди», «общество/народ».

Не предлагаются следующие значения:

- ’Светское общество, привилегированный класс населения.’.
- ’Группа лиц, объединённые по общим признакам, часто с негативной коннотацией.’.
- ’Пассажиры, люди в общественном транспорте.’.
- ’Читатели, аудитория, воспринимающая литературные или иные творческие произведения.’.
- ’Обозримая группа людей.’.
- ’Скопление народа, уличная толпа, масса.’.

Ошибок в написании определений (орфографических, синтаксических и так далее) не обнаружено.

Статистика по лексеме «публика»:

- Корректные: 2

Перейдем к частотности значений.

К сожалению, в книге «Два века в двадцати словах» не даётся графиков частотности для слова *публика*. Гооврится лишь о преобладании значения ’аудитория’ и о его оттенках, которые не удается полноценно сравнить из-за того, что алгоритм предложил довольно общие значения.

Свалка

В результате анализа семем лексемы *свалка* в толковых словарях были выделены семь групп значений, которые можно условно сформулировать следующим образом:

1. Место для сбора мусора, нечистот. (*«Место, куда свозят, выбрасывают мусор, нечистоты, негодные вещи.»* в БТС, *«Место, куда вывозят, выбрасывают мусор, нечистоты, негодные вещи.»* в ТСД, *«Место для сбора мусора, нечистот»* в «Два века в двадцати словах»)

2. Процесс сваливания. (*«к Свалить»* в БТС, *«Процесс сваливания»* в «Два века в двадцати словах»)
3. Всеобщая драка. (*«Свалкой называют всеобщую драку, в которой участвует много людей»* в ТСД, *«Драка»* в «Два века в двадцати словах»)
4. Скопление людей, толпа. (*«Скопление людей, толпа.»* в «Два века в двадцати словах» и в БТС)
5. Груда, куча, нагромождение чего-либо. (*«Беспорядочно накиданная груда, куча чего-л.»* и *«Если кто-либо превращает квартиру в свалку, то это означает, что там в беспорядке нагромождаются предметы, мебель и пр.»* в БТС, *«Свалкой называют беспорядочно накиданную груду каких-либо предметов.»* в ТСД, *«Груда»* в «Два века в двадцати словах»)
6. Вооруженное столкновение войск, битва. (*«Битва»* в «Два века в двадцати словах»)

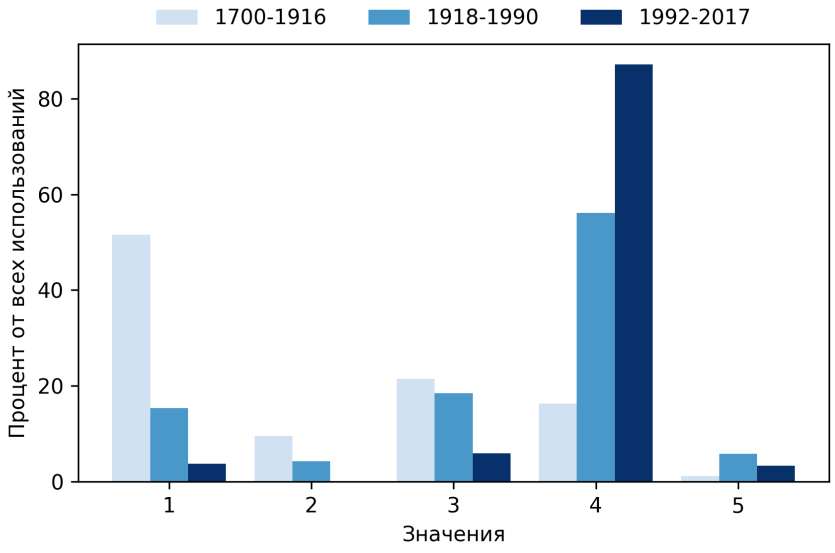


Рис. А.32.. Изменение значений слова *Свалка* (Параметры: eps=0.12, min_samples=25)

Значения для визуализации слова *Свалка*:

1. Столкновение, драка.

2. Беспорядочное, беспорядочное движение, толкотня.
3. Беспорядочная, беспорядочная схватка.
4. Место, где свалены, свалены в кучу какие-либо отходы.
5. То, что свалено, свалено в кучу.

Перейдем к анализу определений.

Первое определение корректно сформулированы. Остальные определения имеют разного рода ошибки.

- ’Столкновение, драка.’ соответствует ’Всеобщая драка.’, так как имеет общий смысловой элемент, а именно семы «столкновение» и «драка».
- ’Место, где свалены, свалены в кучу какие-либо отходы.’ соответствует ’Место, куда свозят, выбрасывают мусор, нечистоты, негодные вещи.’, так как включает те же семы «место», «свалены», «отходы». Однако повторение слова «свалены» является ошибкой в генерации.
- ’Беспорядочное, беспорядочное движение, толкотня.’ частично соответствует ’Скопление людей, толпа.’, которое также указано в «Двух веках в двадцати словах» как ’Толпа, давка.’, так как подразумевает собрание большого количества людей. Однако повторение слова «беспорядочное» является ошибкой в генерации, что также относит определение к избыточным.
- ’Беспорядочная, беспорядочная схватка.’ соответствует ’Всеобщая драка.’, включает семы «потасовки», «с участием большого количества людей». Повторение слова «беспорядочная» является ошибкой в генерации, соответственно данное определение будет определено как избыточное.
- ’То, что свалено, свалено в кучу.’ соответствует ‘Груда, куча, нагромождение чего-либо.’, так как оба определения акцентируют внимание на неорганизованном скоплении чего-либо. Повторение слова

«свалено» является ошибкой в генерации, поэтому мы классифицируем это определение как имеющее избыточность.

Отсутствующие значения:

- ’Процесс сваливания’ также отсутствует в визуализации. Это значение указывает на процесс, а не на результат, что могло быть причиной его отсутствия в предсказаниях модели.
- ’Вооруженное столкновение войск, битва’ отсутствует среди предложенных моделью значений. Это значение является редким, что могло быть причиной его невключения в результат.

Определения, предложенные алгоритмом с повторением слов, далее будут написаны без повторения.

Таким образом, для лексемы *свалка* представлены:

- Корректные: 1
- Избыточность или чрезмерное использование общих фраз: 3
- Близкое значение, а также избыточность или чрезмерное использование общих фраз: 1

Перейдем к частотности значений.

В книге «Два века в двадцати значениях» как появившееся в 1900-ых годах указано значение *«Место для сбора мусора, помойка.»*, соответствующее четвёртому значению, предложенному алгоритмом ’Место, где свалены, в кучу какие-либо отходы.’. Как видно из графика результатов алгоритма, оно почти не используется в досоветский период, но становится главным с 60% использования в советский период и доминирует в постсоветский с около 85%. Эти данные совпадают с тем, что говорится в книге, где утверждается 87% использования значения *«Помойка.»* в 1998-1997 годы, 32% для 1925-1949 годов.

Уменьшается же судя по графику преимущественно значение 1 (’Столкновение, драка.’), которое падает с 50% использований в досоветский период до 5% в постсоветский. В книге результаты схожи. Так, утвер-

ждается, что в 1875-1899 году слово имело значение ’Драка.’ в 71% использований, а к 1998-1997 значение упало до 12%.

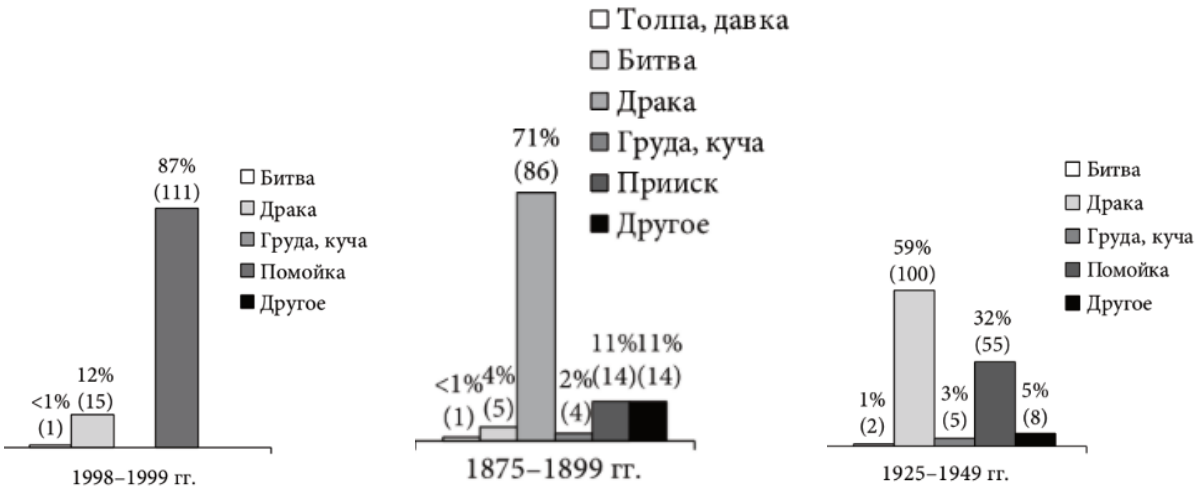


Рис. А.33.. Визуализации для слова *свалка* из книги «Два века в двадцати словах».

Таким образом, модель довольно точно отражает реальное изменение значений слова «свалка» во времени, согласуясь с данными из толкового словаря и историческим исследованием. Она адекватно выделяет как наиболее широко используемое сегодня значение, связанное с местом сбора мусора, так и менее очевидные значения, включая драку, однако предложенные моделью определения имеют излишние повторения слов.

Сволочь

В результате анализа семем лексемы *сволочь* в толковых словарях были выделены шесть групп значений, которые можно условно сформулировать следующим образом:

1. Подлый, скверный человек; негодяй. (*«Грубо. Скверный, подлый человек; негодяй.»* в БТС, *«Негодяй, мерзавец.»* в ТСО, *«Подлец»* в «Два века в двадцати словах»)
2. Собирательное наименование для дрянных, подлых людей; сброд, подонки. (*«собир. Дрянные, подлые люди; сброд, подонки.»* в БТС,

«собир. Сброд, подлые люди.» в ТСО, *«Сброд»* в «Два века в двадцати словах»)

3. Военный сброд, разброд войска. (*«Вольница, военный сброд»* в «Два века в двадцати словах»)
4. Малые люди, чернь, мелкая канцелярская чернь. (*«Маленькие люди, чернь»* в «Два века в двадцати словах»)
5. Сборище, компания. (*«Сборище»* в «Два века в двадцати словах»)
6. Экспрессивное восклицание, выражающее негативные эмоции. (*«Экспрессивное восклицание»* в «Два века в двадцати словах»)

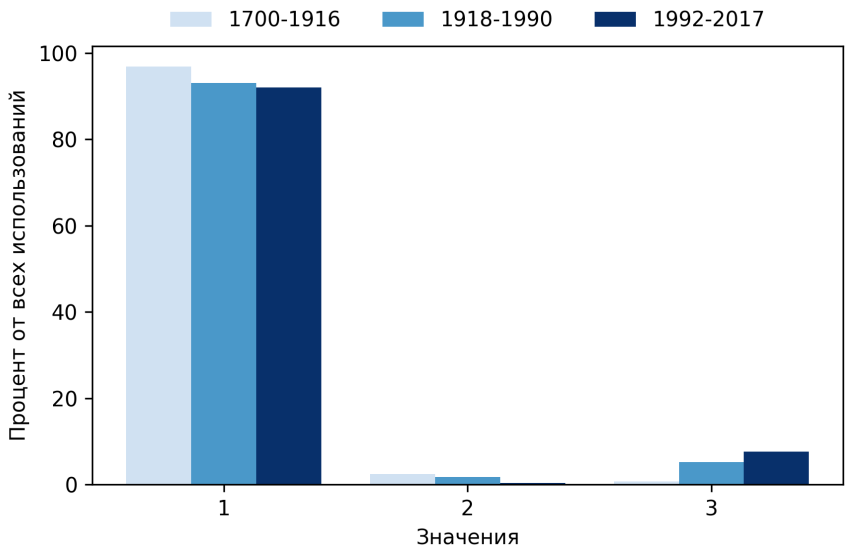


Рис. А.34.. Изменение значений слова *сволочь*

Значения для визуализации слова «Сволочь» (Параметры: eps=0.1, min_samples=10).

1. Употребляется как бранное слово.
2. О подлом, гнусном человеке.

Анализ значений слова *сволочь*

Первое и третье определения корректно сформулированы. Второе определение не соответствует обобщенным значениям.

- ’Употребляется как бранное слово.’ имеет общий смысловой элемент с ’Экспрессивное восклицание, выражающее негативные эмоции.’, а именно семы «бранное слово» и «экспрессивное восклицание».
- ’О подлом, гнусном человеке.’ полностью соответствует ’Подлый, скверный человек; негодяй.’, так как включает те же семы «подлый», «гнусный».

Отсутствующие значения:

- ’Собирательное наименование для дрянных, подлых людей; сброд, подонки’ отсутствует среди предложенных моделью значений. Возможно, информации из контекста использований недостаточно для выявления этого значения.
- ’Военный сброд, разброд войска’ также отсутствует в визуализации. Это может быть связано с редкостью использования данного значения в современных контекстах.
- ’Малые люди, чернь, мелкая канцелярская чернь’ также не представлено в визуализации, что может указывать на недостаточное количество примеров с этим значением в датасете.
- ’Сборище, компания’ также не было явно выделены.

Таким образом, для лексемы *сволочь* представлены:

- Корректные: 2

Перейдем к частотности значений.

К сожалению, оба выделенных значения подпадают под значение ’Индивидуальное оскорбление.’ в книге «Два века в двадцати словах», поэтому анализ изменений значения сделать не представляется возможным.

Стиль

В результате анализа семем лексемы *стиль* в толковых словарях были выделены следующие группы значений, которые можно условно сформулировать следующим образом:

1. Совокупность признаков, черт, приёмов, создающих целостный образ искусства определённого времени, направления, индивидуальной манеры художника. (*«Совокупность признаков, черт, создающих целостный образ искусства определённого времени, направления, индивидуальной манеры художника в отношении идейного содержания и художественной формы.»* в БТС, *«Совокупность черт, близость выразительных художественных приёмов и средств, обусловливающие собой единство какого-н. направления в творчестве.»* в ТСО, *«Стилем называют жанровую и тематическую направленность художественного произведения.»* и *«Стилем называют совокупность литературных приёмов, характерных для какого-либо направления, жанра, произведения.»* в ТСД, *«Особенности направления архитектуры»* в «Два века в двадцати словах»)
2. Индивидуальная манера художника, писателя. (*«Стилем называют индивидуальную авторскую манеру, которая ощущается читателем, зрителем в нескольких произведениях одного автора.»* в ТСД, *«Черты, свойственные конкретному человеку (например, деятелю искусства)»* в «Два века в двадцати словах»)
3. Совокупность наиболее характерных черт в искусстве какого-либо народа, страны, региона. (*«Восточный, латиноамериканский, китайский, русский с. (совокупность наиболее общих черт в искусстве какого-л. народа, страны, региона, отличающихся от искусства соседних народов и т.п.).»* в БТС, *«Стилем называют сово-*

купность наиболее характерных черт в искусстве какого-либо народа, страны и т. п.» в ТСД)

4. Способ, метод, совокупность приёмов осуществления какой-либо деятельности, работы. (*«Способ осуществления чего-л., характер деятельности, работы в их отличительных признаках.»* в БТС, *«Метод, совокупность приёмов какой-н. работы, деятельности, поведения.»* в ТСО, *«Стилем называется способ осуществления чего-либо.»* в ТСД)
5. Совокупность приёмов использования языковых средств. (*«Совокупность приёмов использования средств языка, характерная для какого-л. писателя или литературного произведения, направления, жанра.»* в БТС, *«Совокупность приёмов использования языковых средств для выражения тех или иных идей, мыслей в различных условиях речевой практики, слог2.»* и *«Совокупность приёмов использования языковых средств, а также вообще средства художественной выразительности, определяющие своеобразие творчества писателя, отдельного произведения.»* в ТСО, *«Характеристика языковых средств»* в «Два века в двадцати словах»)
6. Манера словесного изложения. (*«Построение речи в соответствии с нормами литературного языка, манера словесного изложения.»* в БТС, *«Стилем называют чью-либо манеру словесного изложения какой-либо информации.»* в ТСД)
7. Функциональная разновидность литературного языка. (*«Функциональная разновидность литературного языка.»* в БТС, *«Стилем называют функциональную разновидность литературного языка.»* в ТСД)
8. Характерная манера совершения движения, в т.ч. в спорте. (*«Совокупность признаков, черт, приёмов, выделяющих какую-л. вещь, предмет на фоне аналогичных и образующих их суть (в спорте).»*

в БТС, *«Стилем называют характерную манеру совершения движения.»* в ТСД)

9. Модное веяние в одежде. (*«Совокупность признаков, черт, отличающих направление, вещь от других (в моде, в одежде).»* в БТС, *«Стилем называют модное веяние, которое воспринято многими.»* в ТСД)
10. Изменяющаяся социальная форма жизни, деятельности. (*«Совокупность признаков общественной жизни, активности в тот или иной период.»* в БТС, *«Стилем называют изменяющуюся социальную форму жизни, деятельности.»* в ТСД)
11. Индивидуальная манера поведения, общения, одежды и т.п. (*«Индивидуальная манера осуществления какой-л. деятельности, работы, проявления личных качеств в разговоре, поведении, одежде и т.п.»* в БТС, *«Стилем называют изменяющуюся индивидуальную форму жизни, деятельности.»* и *«Манера, совокупность особенностей по отношению к широкому кругу явлений (поведение, одежда, взгляды, внешность, интерьер)»* в «Два века в двадцати словах»)
12. Способ летоисчисления. (*«Способ летосчисления.»* в БТС, ТСО и ТСД, *«Способ летоисчисления»* в «Два века в двадцати словах»)

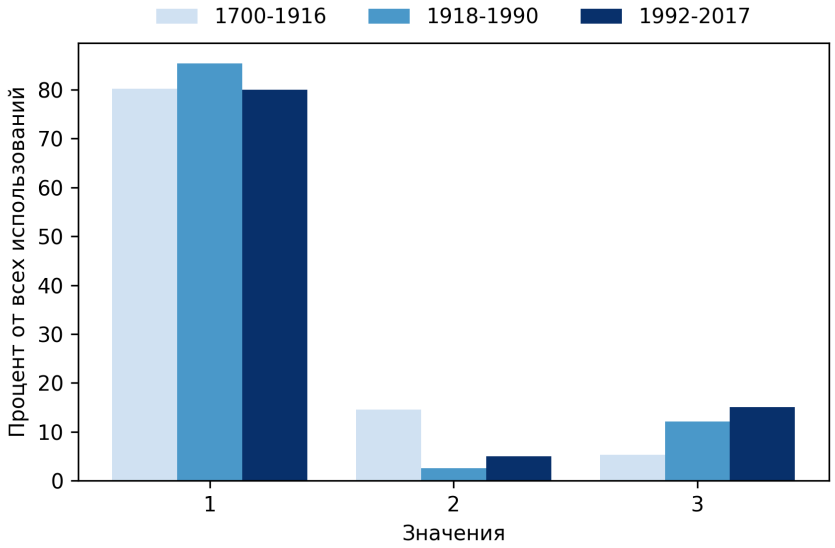


Рис. А.35.. Изменение значений слова *стиль*

Значения для визуализации слова «Стиль» (Параметры: eps=0.15, min_samples=25).

1. Совокупность художественных приемов, характерных для какого-либо искусства, литературы и т. п.
2. Система летосчисления, принятая в какой-л. стране, а также время по этой системе.
3. Характер, манера, образ действий, поведения кого-либо.

Перейдем к анализу определений.

Определения корректно сформулированы.

- ’Совокупность художественных приемов, характерных для какого-либо искусства, литературы и т. п.’ имеет общий смысловой элемент с ’Совокупность признаков, черт, приёмов, создающих целостный образ искусства определённого времени, направления, индивидуальной манеры художника.’ ’Индивидуальная манера художника, писателя.’ ’Совокупность наиболее характерных черт в искусстве какого-либо народа, страны, региона.’ ’Совокупность приёмов использования языковых средств.’ а именно семы «совокупность», «приемы», «искусство», «литература».

- ’Система летосчисления, принятая в какой-л. стране, а также время по этой системе’ полностью соответствует ’Способ летоисчисления.’, так как включает те же семы «система», «летосчисление», «страна».
- ’Характер, манера поведения кого-либо.’ наиболее близко к ’Индивидуальная манера поведения, общения, одежды и т.п.’

Отсутствующие значения:

- ’Функциональная разновидность литературного языка.’ отсутствует среди предложенных моделью значений. Можно предположить, что информации из контекста использований недостаточно для отделения этого значения от ’Совокупность художественных приемов, характерных для какого-либо искусства, литературы и т. п.’.
- ’Модное веяние в одежде.’ также отсутствует в визуализации. Однако, модель способна на выделение данного значения.
- ’Изменяющаяся социальная форма жизни, деятельности.’ и ’Характерная манера совершения движения, в т.ч. в спорте.’ также отсутствуют. Возможно, эти значения не были включены из-за их меньшей частоты в исследуемом материале.

Таким образом, для лексемы *стиль* представлены:

- Корректные: 3

Перейдем к частотности значений.

К сожалению, в книге «Два века в двадцати словах» указано, что все значения слова «стиль» появились в досоветский период, а также все графики даны только для этого периода. Однако в книге говорится, что частота употребления значения ’Способ летоисчисления.’ снижается значительно с досоветского периода, что согласуется с данными нашего исследования. В остальном, представляется затруднительным сравнение результатов модели с данными из книги.

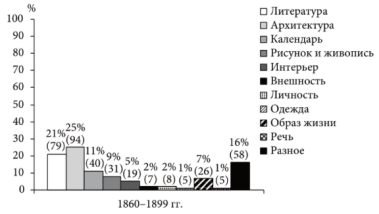


Рис. А.36.. Визуализации для слова *Стиль* для 1860-1899 из книги «Два века в двадцати словах».

Тётка

В результате анализа семем лексемы *тётка* в толковых словарях были выделены следующие группы значений:

1. Сестра отца или матери, а также жена дяди. (*«Сестра отца или матери, а также жена дяди.»* в ТСО, *«Тёткой называют сестру матери или отца. Родная, двоюродная тётка. | Тётка по материнской, по отцовской линии.»* в ТСД)
2. Обращение к незнакомой женщине. (*«Называние незнатной женщины в форме «тётка + имя»»* в «Два века в двадцати словах», *«Обращение к незнакомой женщине (ед. или мн. ч.)»* в «Два века в двадцати словах»)
3. Вообще женщина. (*«Обо всякой взрослой женщине.»* в БТС, *«Тёткой грубо называют женщину или девушку.»* в ТСД, *«Вообще женщина (чаще пожилая).»* в ТСО, *«Женщина (вне контекста обращения)»* в «Два века в двадцати словах»)
4. Карточная игра. (*«Карточная игра»* в «Два века в двадцати словах»)

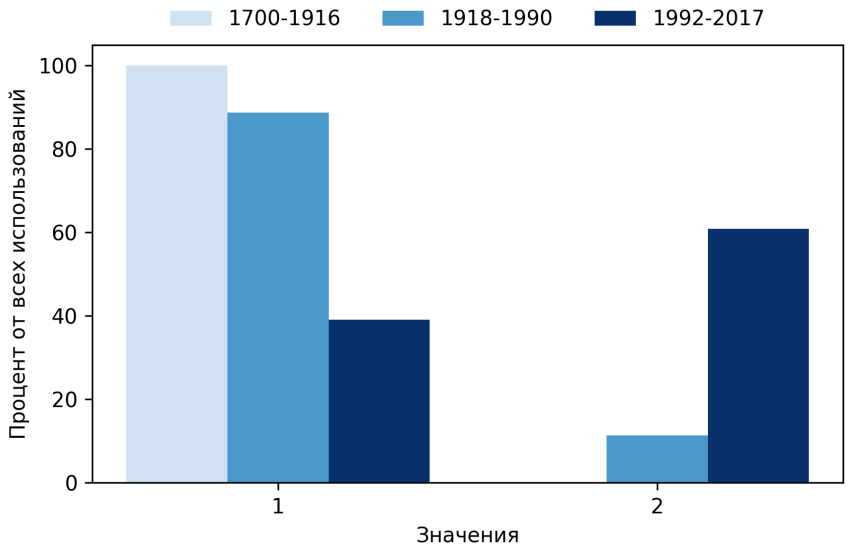


Рис. А.37.. Изменение значений слова *тётка*

Значения для визуализации слова «Тётка» (Параметры: eps=0.1, min_samples=15).

1. Родная сестра отца или матери.
2. Женщина средних лет.

Анализ значений слова *тётка*

Первое определение корректно сформулировано. Второе определение имеет слишком узкое значение, так как в словарях значение охватывает всех взрослых женщин, а не только женщин средних лет.

- ’Родная сестра отца или матери.’ имеет общий смысловой элемент с ’Сестра отца или матери, а также жена дяди.’, а именно семы «сестра» и «отец или мать».
- ’Женщина средних лет.’ соответствует ’Вообще женщина.’, хоть и более узкое, так как обобщенное определения включает всех женщин, а не только женщин средних лет. Сема «средних лет» делает его слишком узким.

Отсутствующие значения:

- ’Обращение к незнакомой женщине.’ моделью отдельно от ’Женщина средних лет.’ не выделяется.
- ’Карточная игра’ также отсутствует в визуализации. Его отсутствие обусловлено тем, что это значение не так часто встречается в исследуемом корпусе. Данное определение указано только в «Двух веках в двадцати словах», где приводятся только единичные использования слова в этом значении, не влияющие в целом на статистику значений.

Таким образом, для лексемы *тётка* представлены:

- Корректные: 1
- Избыточно конкретизированные: 1

Перейдем к частотности значений.

Главным изменением для слова *тётка* является появление в конце советского периода его использования по отношению ко всем женщинам, а не только родственницам (30% для 1980-1985 гг. и около 5% для 1910-1920 и 1940-1954 гг.), что приводится в графиках снизу. К сожалению, в книге не приводится информация об использовании слова в постсоветский период. Данные из нашей визуализации согласуются с вышеописанными изменениями. Так, значение ’Женщина средних лет.’ появляется в советский период с около 10% использований и выходит на лидирующие позиции в постсоветский период с больше 60%.

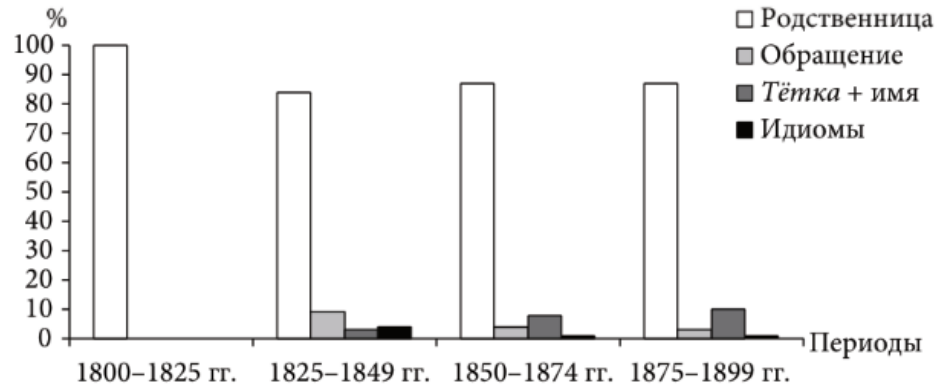


Рис. А.38.. График для слова *тётка* для 1800-1899 из книги «Два века в двадцати словах».

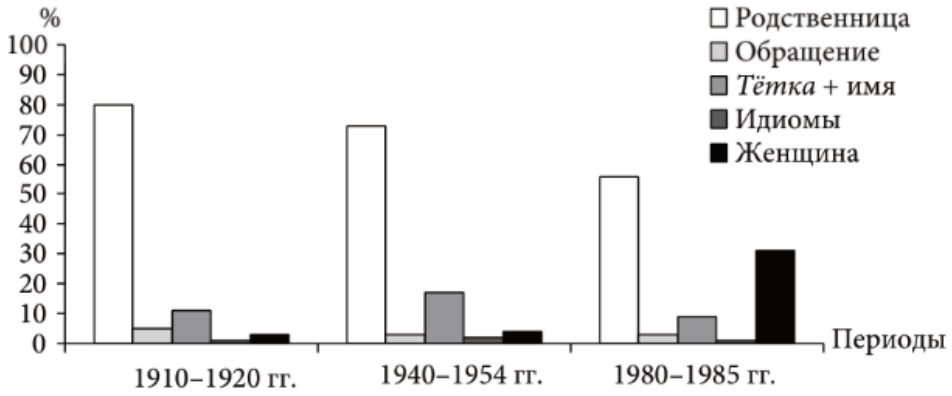


Рис. А.39.. График для слова *тётка* для 1910-1985 из книги «Два века в двадцати словах».

Таким образом, алгоритм полностью отражает значения, в которых использовалось слово *тётка*, согласуясь с данными из толкового словаря и историческим исследованием.

Червяк

В результате анализа семем лексемы *червяк* в толковых словарях были выделены четыре группы значений, которые можно условно сформулировать следующим образом:

1. Маленькое беспозвоночное животное. (*«=Червь (1.Ч.; 1-2 зн.).»* в БТС, *«То же, что червь.»* в ТСО, *«Маленькое беспозвоночное животное»* в «Два века в двадцати словах»)
2. Ничтожное, жалкое создание. (*«О жалком, ничтожном человеке (презр)»* в ТСО, *«Ничтожное создание»* в «Два века в двадцати словах»)
3. Тревожное, мучительное чувство. (*«О постоянном наличии какого-л. чувства, состояния, плохо воздействующего на кого-л.»* в БТС, *«Внутренний паразит»* в «Два века в двадцати словах»)
4. Техническое приспособление в виде винта для передачи движения. (*«Проф. Червячная передача; деталь механизма для такой передачи.»* в БТС, *«Зубчатое колесо в форме винта для передачи движения в нек-рых механизмах (спец.)»* в ТСО, *«Техническое приспособление»* в «Два века в двадцати словах»)

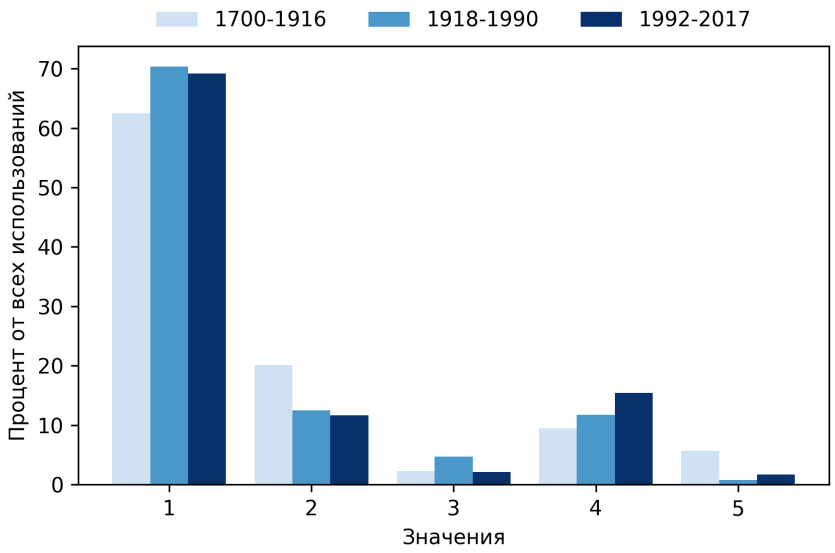


Рис. А.40.. Изменение значений слова *червяк*

Значения для визуализации слова «Червяк» (Параметры: eps=0.16, min_samples=10).

1. Насекомое, похожее на червя, а также его личинка.
2. О мелком, ничтожном человеке.

3. Употребляется как бранное слово.
4. О ком-, чем-либо маленьком, тонком, извивающемся.
5. О каком-л. неприятном, мучительном чувстве, испытываемом кем-л.

Анализ значений слова *червяк*

Первое, второе и пятое определения корректно сформулированы. Третье и четвертое определения не соответствуют обобщенным значениям.

- ’Насекомое, похожее на червя, а также его личинка.’ имеет общий смысловой элемент с ’Маленькое беспозвоночное животное.’, а именно семы «маленькое», «животное». Тем не менее, оно содержит значительные ошибки, червяк не является насекомым, а также содержит ссылку на самого себя (червя).
- ’О мелком, ничтожном человеке.’ соответствует ’Ничтожное, жалкое создание.’, так как включает те же семы «ничтожность», «жалкость».
- ’О каком-л. неприятном, мучительном чувстве, испытываемом кем-л.’ соответствует ’Тревожное, мучительное чувство.’, так как включает семы «неприятное», «мучительное» и «чувство».
- ’Употребляется как бранное слово.’ является близким значением к ’Ничтожное, жалкое создание.’, так как бранное слово может указывать на презрение, но не полностью соответствует исходному значению.
- ’О ком-, чем-либо маленьком, тонком, извивающемся.’ является близким значением к ’Маленькое беспозвоночное животное.’, которое не раскрывает отличительные стороны денотата.

Отсутствующие значения:

- ’Техническое приспособление в виде винта для передачи движения.’ отсутствует среди предложенных моделью значений. Модель спо-

собна на выделение этого значения, сгенерировав ’металлический стержень, служащий для передачи вращательного движения’ для вхождения «При установке киноаппарата в боксе надо следить за тем, чтобы червячное колесо и червяк вошли в зацепление.» Вероятно, модель не распознала это значение из-за его редкости в общем корпусе текстов.

Таким образом, для лексемы *червяк* представлены:

- Корректные: 2
- Близкие значения: 3

Перейдем к частотности значений.

В книге «Два века в двадцати словах» не даётся графиков изменения частоты использования значений для слова *червяк*, однако упоминается о том, что значения ’Ничтожное, жалкое создание.’ и ’Тревожное, мучительное чувство.’ реже используются со временем, особенно во второй половине XX века. Эти данные подтверждаются в нашей визуализации, где использование ’О мелком, ничтожном человеке.’ падает с 20% в досоветский период до 10% в советский и постсоветский, а также использование ’О каком-л. неприятном, мучительном чувстве, испытываемом кем-л.’ уменьшается с около 8% в досоветский период до 1-2% в советский и 3% в постсоветский.

Таким образом, алгоритм в целом отражает значения, в которых использовалось слово *червяк*, согласуясь историческим исследованием, но допуская различного рода ошибки в 3 из 5 сгенерированных определениях.

Приложение Б. Обучение модели

Таблица Б.1.. LoRa параметры

Параметр	Значение
r	32
lora_alpha	64
lora_dropout	0.1

Таблица Б.2.. Trainer параметры

Параметр	Значение
learning_rate	1e-3
lr_scheduler_type	linear
batch_size	16
gradient_checkpointing	true
gradient_accumulation_steps	1
weight_decay	0.1
optimizer	adafactor
num_train_epochs	6

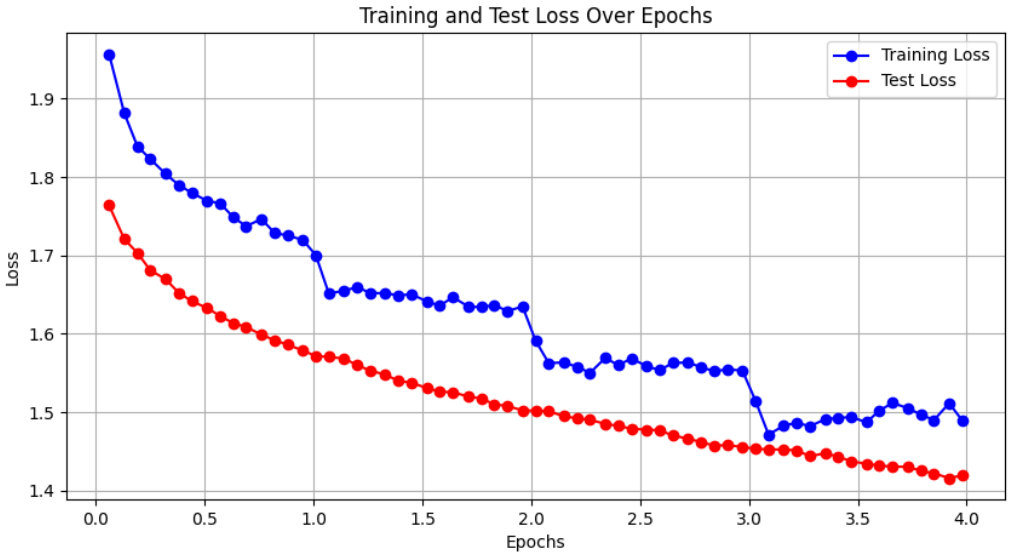


Рис. Б.1.. Лосс при обучении модели

Приложение В. Дообучение векторизатора

Hyperparameter	Value
Batch size	32
Loss function	CosineSimilarityLoss
Epochs	2
Warmup steps	100
Evaluation steps	50
Training data split ratio	80/20
Random seed	42

Таблица В.1.. Hyperparameters for model training