**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ**

**ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

**Факультет гуманитарных наук**

**Программа подготовки бакалавров по направлению**

**45.03.03 «Фундаментальная и прикладная лингвистика»**

***Буракова Надежда Вячеславовна, Татаринов Максим Дмитриевич***

**КУРСОВАЯ РАБОТА**

**Лексические и семантические изменения в текстах русскоязычных новостей с 2000 по 2022 год**

|  |  |
| --- | --- |
|  | **Научный руководитель**  **(ст. преподаватель)**  **П.Д. Петрова** |

**Нижний Новгород, 2023**

**СОДЕРЖАНИЕ**

Введение 3

2. Теоретические аспекты 6

2.1. Лексические изменения 6

2.2. Семантические изменения 8

3. Проведение анализа данных 13

3.1. Проведение анализа частот слов 13

3.2. Обучение моделей на основе исходного кода ShiftRy 24

3.3. Анализ и объяснение полученных результатов 25

Заключение 36

Список использованной литературы 38

Источники материала исследования 41

Приложение 42

**Введение**

Лексика является одним из уровней языка, где изменения проявляются довольно часто и ощущаются всеми носителями языка. В нашей предыдущей курсовой работе были изучены лексические изменения на основе лингвистических метрик удобочитаемости и лексического разнообразия, показатели которых увеличились за последние 20 лет. Однако эти данные не позволяет в полной мере оценить изменения лексики. Именно поэтому в настоящей курсовой работе на базе дополненного новостного корпуса данных, собранных нами ранее, будут рассмотрены изменения частотности лексем, а также семантические сдвиги, которые могли произойти за период с 2000 по 2022 год.

**Актуальность** данной курсовой работы состоит в том, что, несмотря на некоторые зарубежные исследования (Mitra et al., 2015; Basile et al, 2016), вопрос о лексических и семантических изменениях в русском языке недостаточно изучен. Кроме того, исследования русского языка (Хакимзянова Д.Ф., Хованская Е.С., 2015; Fomin V. et al., 2019), посвященные этим изменениям, нуждаются в более свежих данных и глубоком анализе.

**Новизна** настоящей курсовой работы состоит в подробном анализе лексики и семантики новостных ресурсов от 2000 вплоть до 2022 года с использованием программных методов, а также в установлении возможных причин изменений и их влияния на язык в целом.

**Практическая значимость** данной работы заключается в том, что результаты исследования можно применять для определения степени усвоения той или иной лексемы языком. Кроме того, результаты данной работы могут быть использованы в лексикологии, где необходимы актуальные данные построения новых словарей.

**Объектом** настоящего исследования являются лексические единицы в корпусах новостных текстов, а **предметом** – лексические и семантические изменения.

**Целью** данной работы является изучение лексических и семантических изменений новостных корпусов от 2000 до 2022 года.

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

1. Составить корпус новостных текстов от 2000 до 2022 года.

2. Написать компьютерную программу на языке программирования Python для подсчета лексических изменений.

3. Обучить векторные модели, используя исходный код ShiftRy.

4. Написать компьютерную программу на языке программирования Python для выявления значимых семантических изменений.

5. Проанализировать результаты.

В качестве **материалов исследования** использовались:

* новостной корпус газеты «Лента.ру» (Россия) (URL: <https://lenta.ru>), где материал представлен временем от 01.01.2000 до 31.12.2022;
* новостной корпус газеты «Коммерсантъ» (Россия) (URL: https://www.kommersant.ru/), где материал представлен временем от 04.10.2001 до 31.12.2022;
* новостной корпус газеты «РИА новости» (Россия) (URL: <https://ria.ru/>), материал которого представлен временем от 01.01.2004 по 31.12.2022 г.

Общий размер корпуса составил 627190528 токенов и занимал 9 ГБ.

Для настоящего исследования были применены следующие **методы**:

1. *Программный метод* для написания специального кода, позволяющего изучить собранный материал;
2. *Методы обработки естественного языка*, которые использовались при при обработке полученного материала;
3. *Анализ* и *сравнение* полученных результатов;
4. *Индукция*, так как на основе частных случаев формулируется общий вывод об изменениях.

В качестве **инструментов исследования** выступили язык программирования Python3, необходимый для разработки скриптов. Для получения материалов исследования использовались скрипты, написанные на языке Python3. Так, для газеты «Лента.ру» – скрипт, расположенный на github.com, который был модифицирован для нужд настоящей курсовой работы. Для газет «РИА Новости» и «Коммерсант» были написаны скрипты с нуля. Код, использованный для выполнения данной курсовой работы, находится по ссылке в приложении 1.

**2. Теоретические аспекты**

**2.1. Лексические изменения**

Лексические изменения – это изменения, касающиеся активного словарного состава языка.

Частота использования лексем может колебаться в зависимости от периода изучения. Исследование изменений в лексике англоязычных публикаций на сайте Reddit (Hanna Mahler, 2020) показало, что частотность лексем возрастает периодическими «вспышками» и в большинстве случаев не удерживается на одном значении на протяжении долгого времени. И хотя это и может зависеть от анализируемой структуры данных, непостоянство частотности некоторых лексем говорит об их связи с контекстом. Кроме того, исследование обнаружило, что внезапно возросшие в частотности слова вполне могут употребляться до периода увеличения их частоты, однако употребления являются единичными и специфическими.

Во многих исследованиях изменений частотности лексем частоту измеряют в «слове на миллион» (Brysbaert M. et al., 2011; McCreath G.A., 2017). Такое измерение используется для корпусов с значительно различающимися размерами данных для того, чтобы избежать некорректных подсчетов изменений частот. При таком измерении, частота лексемы выражается следующим образом: если, например, выявленная частота равна 0,39, то значит, что данная лексема появляется один раз на каждые 2,56 миллиона слов. По причине удобства данного измерения и разницы между данными корпусов каждого года, в настоящей курсовой работе также применено данное измерение частотности слов.

Говоря о самом подсчете частоты лексем, в исследовании методов и моделей автоматического извлечения ключевых слов (С.О. Шереметьева, П.Г. Осминин, 2015) приводится достаточно простой алгоритм выявления ключевых (частотных) слов. В начале из текста удаляются стоп-слова, так как они не несут никакой смысловой нагрузки (предлоги, союзы, частицы, вводные слова, и т. д.). Общая частота ключевых слов подсчитывается путем сравнения словоформ, нормализованных к основе или лемме. При статистических подходах к извлечению ключевых слов используют стемминг-алгоритмы, которые нормализуют словоформу к ее квази-основе. Наиболее известным является стемминг-алгоритм Портера (Porter M.F., 1997). Нормализованные словоформы ранжируются по частоте и те из них, чья частота выше заданного порога, считаются ключевыми.

В исследовании лексических изменений научных текстов за более чем 250 лет (Bizzoni Y. et al., 2020) использовались три вида вычислительных языковых моделей, которые раскрывают различные аспекты диахронических изменений:

1. Модели N-грамм (леммы и POS), которые указывали на общий тренд укрепления лексем в лингвистическом употреблении. Лексический уровень колебался между высокими пиками, обозначающими лексические инновации, и минимумами, указывающими на стабильное лингвистические использование.

2. Анализ на основе диахронической тематической модели, который показывал, что темы изменяются с течением времени, индексируясь за счет увеличения неопределенности дистрибуции слова/темы, что является показателем формирования регистра.

3. Анализ на основе диахронических вложений слов, который выявил, что лексическое пространство расширяется и становится более фрагментированным, что является показателем диверсификации словоупотребления.

**2.2. Семантические изменения**

В Лингвистическом энциклопедическом словаре дается следующее определение лексического значения слова: «содержание слова, отображающее в сознании и закрепляющее в нем представление о предмете, свойстве, процессе, явлении и т.д.» (Лингвистический энциклопедический словарь, 2002). Соответственно, семантические изменения – это изменения лексического значения слова. Семантические изменения могут проявляться в количественном и качественном отношении как из-за языкового, так и внеязыкового свойства.

На практике существуют различные теории того, как понимать понятие «семантического сдвига» и семантики слов. Традиционное понимание семантического сдвига связано с изменениями, которые, как правило, происходят на большой временной шкале, такой как век или более. Такие сдвиги могут привести к полному изменению значения слова. В то же время, дистрибутивная теория подразумевает, что семантические сдвиги в основном связаны с изменением типичных контекстов и ассоциаций слова. Это означает, что значение слова меняется в зависимости от того, как оно используется в различных контекстах. В статье, посвященной ShiftRy, авторы придерживаются дистрибутивной теории, поскольку она наиболее подходит для работы с корпусами текстов и автоматического обнаружения семантических сдвигов.

В настоящей курсовой работе мы будем понимать семантические сдвиги на основе дистрибутивной теории. Однако, так как изменение вектора часто может быть вызвано изменением контекста, в котором использовалось слово, без изменения семантики самого слова, мы будем также подразделять следующие изменения: 1. изменение семантики, при котором меняется значение слова или слово начинает чаще употребляться в ранее не главном значении; 2. изменение контекста слова, как правило, вызванное появлением новостей на какую-либо определенную тему.

Одну из самых полных классификаций типов изменения семантики слова разработал немецкий лингвист Г. Пауль (Пауль Г., 1960). Он выделил следующие типы:

1. Сужение значения, или специализация, при котором слово начинает употребляться в более узких сферах общения за счет конкретизации своего значения. Например, слово *платье* раньше означало «любую одежду, которую носят поверх белья», однако сейчас *платье* – «женская одежда особого покроя».
2. Расширение значения, при котором значение слова становится более общим, способным применяться во многих сферах общения. Например, *мошенник* имел значение «вор-карманник», но теперь описывает «любого нечестного на руку человека».
3. Перенос значения, или метонимия, на основе смежности или близости ассоциативных связей. Например, слово *вздор* изначально означало «то, что содрано при очистке дерева от коры; стружки, хлам, отбросы», однако позже приобрело значение «бессмыслица, чепуха» из-за ассоциативных связей.
4. Метафора, заключающаяся в изменении значения на основе скрытого сравнения. Например, глагол *boil* имел значение «пузыриться», позже по сходству характера действия получил значение «кипятить».
5. Возвышение значения, при котором значение слова улучшается в течение времени. Например, слово *chiffon* обозначающее нежную деликатную воздушную ткань, ранее имело значение «тряпка».
6. Деградация значения, когда значение слова ухудшается в течение времени. Например, слово *gossip* первоначально обозначало «крестного или крестную», а в современном языке приобрело значение «сплетник, человек, рассказывающий скандальные истории другим людям».
7. Гипербола, при которой значение слова нарочно преувеличивается.
8. Литота, когда значение слова нарочно подвергается преуменьшению.

Семантика может быть подвержена изменению не только из-за каких-либо языковых причин, но и из-за экстралингвистических факторов, то есть факторов, не связанных напрямую с языком. М.С. Атрощено выделила следующие связанные с культурным аспектом языка экстралингвистические факторы, влияющие на значения слов (Атрощенко М.С., 2018):

1. Историко-политические, экономические и социальные изменения. Различные изменения в социуме отражаются в языке путем изменения значений некоторых слов, относящихся так или иначе к этим изменениям, или путем появления абсолютно новых слов, актуальных для периода их появления.
2. Гетерогенность языкового сообщества, когда одни языки влияют на языковой состав других языков в результате их пересечения (например, сочетание *черный ящик,* существующее во многих языках и приобретшее образное значение).
3. Развитие науки и техники, в результате которого ранее существовавшие лексемы приобретают совершенно новую семантику, относящуюся к научным и техническим инновациям (например, у слова *мышь* появилась новая семантика, описывающая устройство для работы на компьютере).

При анализе изменений в значении слов по частям речи в исследовании английского языка (Волкова Е.А., 2016) было установлено, что процесс изменения значения чаще наблюдается у существительных – 57,5%, глаголов – 24%, и прилагательных –16%.

Говоря о способах автоматического нахождения семантических сдвигов, существуют следующие методы:

1. Метод, основанный на контекстных векторах и использующий семантическую плотность (Sagi et al., 2009) или взаимную информацию (Gulordava et al., 2013) для выявления семантических сдвигов. Однако, данный метод не может располагать вектора на протяжении времени или точно определить, что изменилось в значениях слов.
2. Метод, использующий тематические модели (англ. topic-based models) и принимающие «topics» в качестве значений слов. Новые значения слов определялись сравнением более новых моделей со старыми (Cook et al., 2014; Lau et al., 2012) или кластеризацией топиков на протяжении времени (Wijaya et al., 2011). Однако это обычно не позволяло полноценно проследить изменение значения слова на протяжении времени.
3. Модель, основанная на графах и призванная раскрыть сложные отношения в семантике слов через моделирование их значений при помощи извлечения лексических смыслов неоднозначных слов (англ. word sense induction) и их расположения на протяжении времени, то есть такие модели позволяли различать отдельные значения слов (Mitra et al., 2014; Tahmasebi et al., 2017).

Проект Shiftry является одним из последних проектов, связанным с поиском семантических сдвигов на новостном датасете. На основе кода этого проекта были обучены векторные модели для данной курсовой работы. Проект анализирует диахронические изменения в употреблении слов и использует диахронические дистрибутивно-семантические векторные модели, обученные на русскоязычном новостном корпусе из текстов за временной период с 2010 по 2020 годы с ежегодными обновлениями (Kutuzov et al., 2020).

Для определения, произошел ли семантический сдвиг между словом из разных моделей, используется косинусное расстояние. Косинусное расстояние – это метрика, используемая для измерения сходства между двумя векторами. Расстояние вычисляется как косинус угла между двумя векторами. Чем ближе угол между двумя векторами, тем больше косинус расстояния между ними.

В моделях Word2Vec, используемых в проекте Shiftry, косинусное расстояние используется для нахождения семантических сдвигов. Это позволяет при работе с моделью находить семантические отношения между словами. Например, можно использовать косинусное расстояние для определения противоположности слов «холодный» и «горячий».

**3. Проведение анализа данных**

**3.1. Проведение анализа частот слов**

Все полученные в ходе работы скриптов статьи газет были разделены по годам их написания и обработаны с помощью UDPipe, программного продукта для морфологического анализа русского языка. В результате работы UDPipe получены лемматизированные и приведенные к нижнему регистру тексты, где слова отмечены соответствующей частью речи, например, *крым\_PROPN*, и очищены от стоп-слов. В настоящей курсовой работе мы принимаем во внимание следующие части речи: имена существительные, имена прилагательные, глаголы, наречия. Поскольку UDPipe выделяет имена собственные отдельно, в данной курсовой работе мы придерживаемся такой же классификации и также выделяем и их. Также важно уточнить, что некоторые части речи (например, причастия) UDPipe может тегировать неправильно из-за того, что у него нет отдельных для них тегов, однако в данной работе мы не считаем это за ошибку.

С помощью написанного скрипта высчитывалась частота лексем за каждый год, которая измеряется в слове на миллион. Далее высчитывалась частота N-грамм на миллион за год. После получения частотности N-грамм за каждый год сравнивались частотные изменения лексем, выраженные в частном между частотой лексемы предыдущего года и следующего. Также вычислялись частотные биграммы. Мы изучили 100 лексем и 100 биграмм с самыми большими изменениями частотности в каждом году с 2000 по 2022.

Для начала, мы сравнили частоты лексем разных периодов: 2004 года и 2022, а также 2000 и 2022 года. В 2022 году, по сравнению в 2004, стали чаще упоминаться такие лексемы, как: *диетолог\_NOUN* – в 285,1 раз, *размещать\_ADJ* – в 207,5 раз, *локация\_NOUN* – в 183 раза, *рэпер\_NOUN* – в 120,8 раз. Стали употребляться следующие лексемы: *канал\_NOUN*, *сбер\_PROPN*, *мобилизовать\_ADJ*, *ник\_NOUN*, *гибридный\_ADJ*, *русофобия\_NOUN*, *деэскалация\_NOUN*, *коронавирус\_NOUN*, *онлайн\_ADV*. Лексемы *размещать\_ADJ* и *локация\_NOUN* стали употреблять по отношению к интернету и смартфонам, по той же причине появились такие слова, как *-канал\_NOUN*, *ник\_NOUN*, *онлайн\_ADV*, имеющие отношение к интернету.

Участившиеся биграммы: *северный\_ADJ поток\_NOUN* (в 3126 раз), *потолок\_NOUN* *цена\_NOUN* (в 902,4 раза), *оспа\_NOUN обезьяна\_NOUN* (в 563,8 раз), *видео\_NOUN опубликовывать\_VERB* (в 242 раза).

По сравнению с 2000, помимо указанных раннее, возросло употребление следующих лексем: *бренд\_NOUN* – в 133,2 раза, *собеседница\_NOUN* – в 38,8 раз, *логистика\_NOUN* – в 37,7 раз, *опция\_NOUN* – в 34,7 раз, *аксессуар\_NOUN* – в 31,4 раза. Кроме того, появились новые лексемы: *смартфон\_NOUN*, *аккаунт\_NOUN*, *видеообращение\_NOUN*, что опять же связано с развитием интернета и технологий.

В 2022 году, по сравнению с 2000, увеличилась частота у таких биграмм, как: *народный\_ADJ республика\_NOUN* (в 769 раз), *начало\_NOUN* *спецоперация\_NOUN* (в 126 раз), *эфир\_NOUN радио\_NOUN* (в 89,5 раз), *иностранный\_ADJ агент\_NOUN* (в 50,5 раз).

Переходя к частотам лексем соседних лет (например, 2000 и 2001, затем 2001 и 2002 и т. д.), при сравнении частотности лексем новостных статей 2000 и 2001 годов было обнаружено увеличение частоты лексем, связанных с событиями в Афганистане. Участились упоминания государственных лиц: *bush\_PROPN* – в 65 раз, *маквей\_PROPN* – в 36,3 раз; террористических организаций: *аль-каида\_PROPN* – в 37,7 раз, населённых пунктов: *мазари-шариф\_PROPN* – в 90,6 раз, *кундуз\_PROPN* – в 59,9 раз, *кабул\_PROPN* – в 15,7 раз. Кроме того, увеличилась частота лексем, связанных с войной:

* *антиталибский\_ADJ* – в 38,7 раз,
* *пво\_PROPN* – в 11,9 раз,
* *бомбовой\_ADJ* – в 9,9 раз.

Участившиеся биграммы также были связаны с войной в Афганистане:

* *осама\_PROPN ладен\_PROPN* – в 164,3 раза,
* *антиталибский\_ADJ коалиция\_NOUN* – в 70,1 раза,
* *сша\_PROPN афганистан\_PROPN* – в 53,8 раз,
* *режим\_NOUN талибан\_PROPN* – в 47 раз,
* *террористический\_ADJ атака\_NOUN* – в 46 раз,
* *удар\_NOUN афганистан\_PROPN* – в 41,4 раза.

Другие лексемы, частота которых увеличилась: легочный\_ADJ – в 27,6 раз, возбудитель\_NOUN – в 16,3 раз. Кроме того, одной из биграмм с наибольшей увеличенной частотой является *сибирский\_ADJ язва\_NOUN* (в 50 раз), что вызвано вспышкой сибирской язвы в США, о которой сообщали СМИ.

В 2002 году в первую очередь заметно участились лексемы, относящиеся к отрицательным событиям, произошедшим в этот период, и местам, где они произошли:

* *дубровка\_PROPN* – в 133,4 раза,
* *галашка\_NOUN* – в 74,6 раз,
* *норд-ост\_PROPN* – в 27,5 раз,
* *бали\_PROPN* – в 37 раз,
* *каспийск\_PROPN* – в 24 раза,
* *заложница\_NOUN* – в 6,3 раза.

Биграммы:

* *центр\_NOUN дубровка\_PROPN* – в 195,7 раз,
* *здание\_NOUN театральный\_ADJ* – в 119 раз,
* *заложник\_NOUN москва\_PROPN* – в 78,9 раз,
* *мюзикл\_NOUN норд-ост\_PROPN* – в 35,2 раза.

Однако, помимо этого, участились слова, связанные со спортом, в частности, с футбольным матчами:

* *уралан\_PROPN* – в 22,5 раза,
* *премьер-лига\_NOUN* – в 14,8 раз,
* *чемпионат\_PROPN* – в 7,4 раза.

Биграммы:

* *советский\_ADJ спорт\_NOUN* – в 59,8 раз,
* *футбол\_NOUN сборная\_NOUN* – в 11,3 раза,
* *игра\_NOUN солт-лейк-сити\_PROPN* – в 11 раз.

В 2003 году заметно участились лексемы, связанные с событиями в Ираке. Упоминалось множество различных населённых пунктов: *неджеф\_PROPN, фаллудж\_PROPN, тикрит\_PROPN, басра\_PROPN*. Также, участились слова, относящиеся непосредственно к этим событиям:

* *смертница\_NOUN* – в 71 раз,
* *антииракский\_ADJ* – в 21,2 раза,
* *иракец\_NOUN* – в 14,5 раз,
* *шахида\_NOUN* – в 11,8 раз,
* *пехотный\_ADJ* – в 9,5 раз,
* *партизан\_NOUN* – в 8,7 раз.

Участившиеся биграммы также относятся к данной тематике:

* *войска\_NOUN ирак\_PROPN* – в 190,2 раза,
* *центр\_NOUN багдад\_PROPN* – в 112,9 раз,
* *террористка\_NOUN смертница\_NOUN* – в 99 раз,
* *антииракский\_ADJ коалиция\_NOUN* – в 41,7 раз.

Помимо этого, лексема *матрица\_PROPN* участилась в 35 раз, что связано с выходом одноимённого фильма.

В 2005 году заметно возросла частота упоминания лексемы *корякия\_PROPN* – в 215,5 раз, из-за объединения с Камчатским краем. Также, часто стали встречаться имена политических деятелей, вступивших в этот год на высокопоставленную должность: *бакиев\_PROPN* (в 847 раз), *меркель\_PROPN* (в 87,7 раз), *юрий::ехануров\_PROPN* (в 54 раза).

Такие биграммы, как: *выборы\_NOUN мосгордума\_PROPN* (в 204,1 раза), *совпадать\_VERB позиция\_NOUN* (в 64,3 раза), *пункт\_NOUN процент\_NOUN* (в 60,6 раз), *международный\_ADJ комиссия\_NOUN* (в 11 раз), также свидетельствуют о распространенности темы выборов в 2005 году.

В 2006 году выросла частота упоминания лексемы *кондопога\_PROPN* (в 551 раз чаще) из-за массовых беспорядков в городе. Данные лексемы также стали употребляться чаще:

* *карикатура\_NOUN* – в 49 раз,
* *суррогатный\_ADJ* – в 36,4 раза,
* *соцпартия\_PROPN* – в 15,9 раз,
* *нацпроект\_NOUN* – в 11,5 раз,
* *пророк\_NOUN* – в 11,5 раз,
* *антикризисный\_ADJ* – в 11,3 раза.

Частотные биграммы связаны с преобразованиями в сфере политики, а также с экономикой:

* *проект\_NOUN образование\_PROPN* – в 72,4 раза,
* *подготавливать\_VERB интернет\_NOUN* – в 72,7 раз,
* *русский\_ADJ марш\_NOUN* – в 65,8 раз,
* *торги\_NOUN классический\_ADJ* – в 36,4 раза,
* *мировой\_ADJ биржа\_NOUN* – в 19,9 раз,
* *бумага\_NOUN российский\_ADJ* – в 17,6 раз.

В 2007 году лексема *самообороный\_NOUN* имеет один из самых высоких значений изменения частоты – она возросла в 191,4 раза. Участившиеся биграммы связаны с проходившим саммитом России и Евросоюза: *указ\_NOUN роспуск\_NOUN* (в 43,5 раз), *марш\_NOUN несогласный\_ADJ* (в 21,7 раз), *федеральный\_ADJ список\_NOUN* (в 17,8 раз), *пятый\_ADJ созыв\_NOUN* (в 17,3 раза).

В 2008 году прилагательное *неустрашимый\_ADJ* возросло в частотности в 65,9 раз. Участились такие биграммы, как: *отвод\_NOUN войска\_NOUN* (в 344 раза), *осетия\_PROPN обстрелять\_VERB* (в 131,8 раз), *жертва\_NOUN вооружать\_ADJ* (в 106,3 раза), *независимость\_NOUN южный\_ADJ* (в 96,8 раз), *грузия\_PROPN мир\_NOUN* (в 84 раза), что связано с войной в Грузии, происходившей в тот период.

В 2009 году из-за вспышки свиного гриппа участились лексемы, связанные с болезнью и лечением:

* *свиной\_NOUN* – в 354,7 раз
* *контактный\_PROPN* – в 61,8 раз,
* *пандемия\_NOUN* – в 16,6 раз,
* *грипп\_NOUN* – в 12,6 раз,
* *воз\_PROPN* – в 7,5 раз,
* *орви\_NOUN* – в 6 раз,
* *заболеть\_VERB* – в 6 раз,
* *летальный\_ADJ* – в 5,3 раза,
* *карантин\_NOUN* – в 4,3 раза,
* *пневмония\_NOUN* – в 4 раза.

Биграммы, частота которых возросла, также связаны с болезнью: *угроза\_NOUN пандемия\_NOUN* (в 409,6 раз), *человек\_NOUN грипп\_NOUN* (в 134,7 раз), *заболевание\_NOUN грипп\_NOUN* (в 64,5 раз), *вирус\_NOUN грипп\_NOUN* (в 21,8 раз).

В 2010 году возросла частотность лексики, связанной с пожарами из-за лесных и торфяных пожаров в России. Данные слова стали употребляться намного чаще:

* *смог\_NOUN* – в 124,1 раза,
* *торфяник\_NOUN* – в 35,4 раза,
* *гарь\_NOUN* – в 34 раза,
* *погорелец\_NOUN* – в 25 раз,
* *торфяной\_ADJ* – в 16,9 раз,
* *аномальный\_ADJ* – в 11,5 раз,
* *жара\_NOUN* – в 8,6 раз,
* *жар\_NOUN* – в 7,3 раза,
* *пепел\_NOUN* – в 7,3 раза,
* *пожароопасный\_ADJ* – в 6,7 раз,
* *засуха\_NOUN* – в 5,5 раз.

Данная тема также отражается в участившихся биграммах:

* *аномальный\_ADJ жара\_NOUN* – в 197,1 раза,
* *природный\_ADJ пожар\_NOUN* – в 100 раз,
* *торфяной\_ADJ пожар\_NOUN* – в 21,9 раз,
* *температурный\_ADJ рекорд\_NOUN* – в 9,2 раза.

В 2011 году часто употреблялись лексемы, связанные с нестабильностью в Ливии и Сирии: *мятежника\_NOUN* (в 17,5 раз), *повстанец\_NOUN* (в 8,5 раз), *антиправительственный\_ADJ* (в 6,6 раз). Эта же тема прослеживается в частоте биграмм: *выступление\_NOUN\_требование\_NOUN* (в 134 раза), *противостояние\_NOUN правительственный\_ADJ* (в 129 раз).

В 2012 году возросла частота упоминаний таких лексем, как *pussy\_PROPN* (в 1440,3 раз), *riot\_PROPN* (в 207 раз), *панк-группа\_NOUN* (в 121,1 раз), *надежда::толоконников\_PROPN* (в 104 раза), *участница\_NOUN* (в 3,3 раза), *христос::спаситель\_PROPN* (в 3,3 раза) из-за уголовного дела Pussy Riot. Кроме того, лексема паралимпиада\_PROPN также стала одной из частой лексем 2012 года.

По биграммах также заметно, что в этот год появляются такие понятия как *иностранный\_ADJ\_агент\_NOUN* (в 120 раз) и *пропаганда\_NOUN\_гомосексуализм\_NOUN* (в 12,25 раз), связанные с законами, принятыми в этот год.

В 2014 году участились множество лексем, связанных с событиями в Украине:

* *славянск\_PROPN* – в 1769,2 раз,
* *коктейль\_PROPN* – в 722,5 раза,
* *молотов\_PROPN* – в 36,7 раз,
* *порошенко\_PROPN* – в 274,3 раза,
* *луганский\_ADJ* – в 148 раз,
* *летчица\_NOUN* – в 145 раз,
* *ополченец\_NOUN* – в 128 раз,
* *госпереворот\_NOUN* – в 119,8 раз,
* *крым\_PROPN* – в 64,5 раз,
* *самопровозглашенный\_ADJ* – в 44,9 раза,
* *донбасс\_PROPN* – в 22,5 раза,
* *радикал\_NOUN* – в 21,3 раз,
* *санкционный\_ADJ* – в 20,2 раз,
* *легитимность\_NOUN* – в 20,2 раз,
* *пленный\_ADJ* – в 18,9 раз.

Также, стали распространены следующие лексемы: *пророссийский\_NOUN, деэскалация\_NOUN, евромайдан\_NOUN*.

Среди биграмм чаще стали использоваться *крым\_PROPN россия\_PROPN* (в 2145 раз), *украина\_PROPN опасаться\_VERB* (в 2121 раз), *присоединение\_NOUN крым\_PROPN* (в 1883 раза), *восток\_NOUN украина\_PROPN* (в 1369 раз).

С 2016-2017 годов появились упоминания telegram-каналов. Также, стали употребляться такие лексемы, как *криптовалюта\_NOUN, блокчейн\_NOUN, биткоин\_NOUN* в связи с возросшей популярностью криптовалют.

Самые возросшие биграммы в 2017 год связаны с прошедшими выборами в США и новым президентом: *сша\_PROPN\_дональд::трамп\_PROPN* (в 464 раза), *трамп\_PROPN\_подписывать\_VERB* (в 214 раз), а также с соцсетями: *слово\_NOUN\_twitter\_PROPN* (в 605 раз), *паблик\_NOUN\_фейсбук\_PROPN* (в 315 раз), *паблик\_NOUN\_вконтакт\_PROPN* (в 209 раз).

В 2018 году частотность следующих лексем увеличилась из-за дела об отравлении Сергея Скрипаля:

* *сергей::скрипаль\_PROPN* – в 1349,1 раз,
* *новичок\_PROPN* – в 207 раз,
* *причастно\_ADV* – в 98,2 раз,
* *нервный\_ADV* – в 25,8 раз,
* *паралитический\_ADJ* – в 18,2 раз,
* *отравить\_VERB* – в 18,2 раз,
* *бессознательный\_ADJ* – в 8,2 раз,
* *отравление\_NOUN* – в 6,4 раз.

Биграммы:

* *причастно\_ADV российский\_ADJ* – в 871 раз,
* *офицер\_NOUN гру\_PROPN* – в 726 раз,
* *экс-полковник\_NOUN гру\_PROPN* – 317 раз.

Кроме того, такие лексемы как *самовыдвижений\_NOUN, общенародный\_ADJ, грудинин\_PROPN* стали упоминаться чаще во время выборов президента РФ.

В 2019 года во время президентских выборов в Украине лексемы *зеленский\_ADJ, слуга\_PROPN, шоумен\_NOUN* также возросли в частоте.  
Самые участившиеся в 2019 биграммы:

* *слуга\_PROPN народ\_NOUN* – в 768 раз,
* *владимир::зеленский\_PROPN заявлять\_VERB* – в 335 раз,
* *временный\_ADJ президент\_NOUN* – 238 раз.

Из-за пандемии коронавируса в 2020 году начали употребляться такие лексемы, как: *самоизоляция\_NOUN, дистанцирование\_NOUN, самоизолироваться\_VERB*. Также возросла частотность следующих лексем:

* *пандемия\_NOUN* – в 1488,2 раза,
* *оперштаб\_NOUN* – в 824,7 раз,
* *заражений\_NOUN* – в 313,3 раза,
* *бессимптомный\_ADJ* – в 283,8 раз,
* *удаленка\_NOUN* – в 238,6 раз,
* *ивл\_NOUN* – в 208,2 раза,
* *выздороветь\_VERB* – в 151,5 раз,
* *излечиваться\_VERB* – в 147 раз,
* *вирусолог\_NOUN* – в 131,6 раз,
* *переболеть\_VERB* – в 103,7 раз,
* *антисептик\_NOUN* – в 96,3 раза,
* *смертельный\_NOUN* – в 90,3 раза,
* *вакцина\_NOUN* – в 87,1 раза,
* *антитело\_VERB* – в 78,3 раза,
* *карантин\_NOUN* – в 69,9 раз.

Самые участившиеся в 2020 биграммы:

* *вспышка\_NOUN новый\_ADJ* – в 4035 раз,
* *время\_NOUN пандемия\_NOUN* – в 2700 раз,
* *объявлять\_VERB вспышка\_NOUN* – в 2002 раза,
* *число\_NOUN заражать\_VERB* – в 1472 раза,
* *российский\_ADJ вакцина\_NOUN* – в 1030 раз.

Кроме того, в 2021 году также появились лексемы, связанные с коронавирусом: *ковивак\_PROPN, ревакцинация\_NOUN, постковидный\_ADJ*. Участились следующие лексемы:

* *непривитый\_ADJ* – в 65,4 раза,
* *qr-кодов\_NOUN* – в 22,3 раза,
* *дельта\_NOUN* – в 50 раз,
* *однокомпонентный\_ADJ* – в 47,6 раз.

Самые участившиеся в 2021 биграммы:

* *сертификат\_NOUN вакцинация\_NOUN* – в 1145 раз,
* *входить\_VERB кабул\_PROPN* – в 656 раз,
* *сша\_PROPN джо::байден\_PROPN* – в 614 раз,
* *дело\_NOUN ив\_PROPN* – в 528 раз,
* *путин\_PROPN байден\_PROPN* – в 459 раз,
* *темп\_NOUN вакцинация\_NOUN* – 120 раз.

Как видно из результатов биграмм, также темами стали президентство Байдена, уход США из Афганистана и дело Ив Роше.

В 2022 из-за событий в Украине появились следующие лексемы: *денацификация\_NOUN*, *демилитаризаций\_NOUN*, *сво\_PROPN*.

Участились также такие лексемы, как:

* *энергодар\_PROPN* – в 424,6 раз,
* *мобилизовать\_VERB* – в 266,7 раз,
* *приазовский\_ADJ* – в 164 раза,
* *северодонецк\_PROPN* – в 150,3 раза,
* *каховка\_PROPN* – в 148,7 раз,
* *запорожский\_ADJ* – в 139,7 раз,
* *град\_PROPN* – в 34 раза,
* *биолаборатория\_NOUN* – в 32 раза,
* *мобилизация\_NOUN* – в 31,5 раз,
* *дальнобойный\_ADJ* – в 30,4 раза,
* *залповый\_ADJ* – в 18,1 раз,
* *кровавый\_ADJ* – в 17,6 раз,
* *бронированный\_ADJ* – в 17,1 раз,
* *снаряд\_NOUN* – в 15,6 раз,
* *артиллерия\_NOUN* – в 15,2 раза.

Самые участившиеся в 2022 биграммы:

* *инфраструктура\_NOUN украинский\_ADJ* – в 4740 раз,
* *проводить\_VERB демилитаризаций\_NOUN* – в 2291 раз,
* *начало\_NOUN спецоперация\_NOUN* – в 2223 раза,
* *наступление\_NOUN группировка\_NOUN* – в 1663 раз.

В результате анализа изменения частотностей как отдельных лексем, так и биграмм, можно сделать вывод, что их частота в большинстве случаев зависит от экстралингвистических факторов, таких как политические, социальные, экономические изменения. Также, можно заметить, что частота многих лексем не остается постоянной после резкого возрастания: многие лексемы являются частоупотребляемыми только в определенный период времени (например, лексема *пандемия\_NOUN*). Однако, при сравнении далеких друг от друга промежутков времени можно выявить новые лексемы, которые со временем закрепились в языке и в дальнейшем не имели больших изменений в частоте.

**3.2. Обучение моделей на основе исходного кода ShiftRy**

Для обучения моделей использовались те же лемматизированные, приведенные к нижнему регистру и очищенные от стоп-слов тексты, что и для подсчета частот.

Все модели были обучены с размерностью вектора 300, окном контекста 5 слов слева и 5 слов справа, в течение 10 эпох, без уменьшения выборки. Слова с частотой в корпусе менее 10 были проигнорированы.

Для выравнивания моделей использовалось выравнивание Прокруста. Выравнивание Прокруста – это ортогональное преобразование, основанное на методе главных компонент, используемое для того, чтобы как можно ближе выровнять одно пространство эмбеддингов с другим (например, модель, обученная на корпусе 2021 года, и модель, обученная на корпусе 2022 года). Пространство эмбеддингов всех моделей было выравнено относительно модели 2017 года. После стало возможно рассчитать косинусное расстояние между эмбеддингами слов из разных моделей, как будто они были обучены вместе.

В итоге были обучены 23 модели за период времени с 2000 по 2022 год с количеством векторов от 11944 до 71515, каждая относящаяся к определенному году.

После чего, был написан скрипт, подсчитывающий семантические сдвиги между моделями и сохраняющий их в JSON.

Он был использован для всех моделей, следующих друг за другом по времени статей.

**3.3. Анализ и объяснение полученных результатов**

Для начала мы приведем общие результаты.

Результаты показывают, что доля слов, имеющих семантический сдвиг, варьируется от 2 до 9 процентов и увеличивается с течением времени. Наиболее высокий показатель был зафиксирован в 2013-2014 годах (0,09), а наименьший – в 2000-2001 годах (0,02).

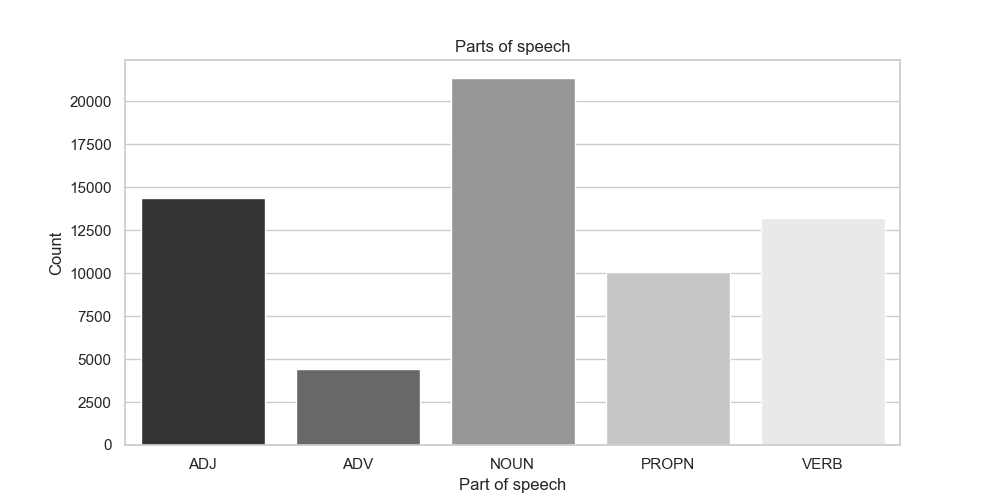


Рисунок 1: Абсолютные данные распределения сдвигов по частям речи

Как видно из абсолютных результатов, больше всех слов с семантическими сдвигами имеют существительные, за ними следуют прилагательные, глаголы, имена собственные и наречия. Те же результаты представлены на абсолютном большинстве изменений между годами.

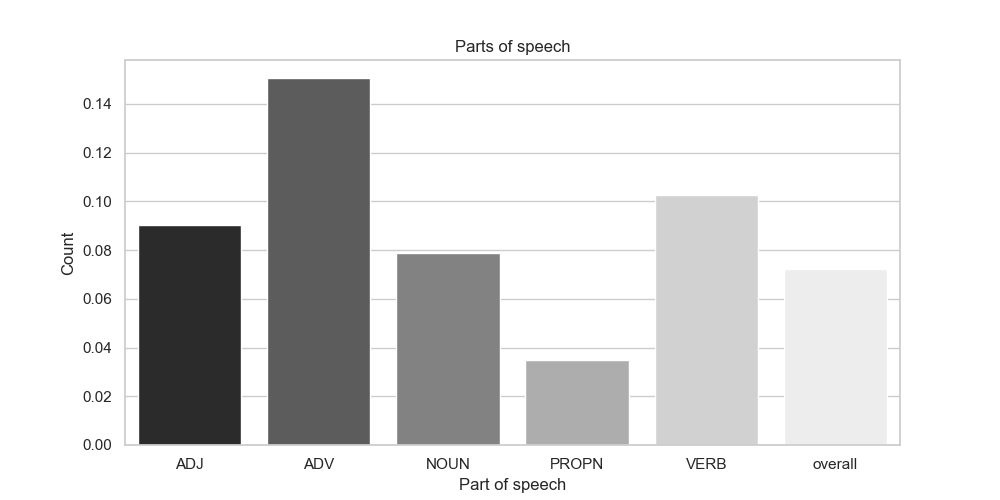


Рисунок 2: Относительные данные распределения сдвигов по частям речи

Из относительных результатов, представленных на графике выше, можно увидеть, что частота использования различных частей речи в течение годов существенно не изменялась. Самый высокая встречаемость семантического сдвига была зафиксирована у наречий, далее шли глаголы, прилагательные и существительные. Самый низкий процент наличия семантического сдвига в большинстве моделей был у имен собственных.

Также было проанализировано, какие слова наиболее часто имеют измененные вектора. Из полученных данных был сделан вывод, что словами с наиболее часто меняющимися векторами, являются слова: *целесообразно\_ADJ, обещать\_ADJ, необязательный\_ADV, предполагаемый\_VERB, срочность\_NOUN, размен\_NOUN, предусматривать\_ADJ, одномоментно\_ADV*. Все они входили 5 или более раз в топ-100 слов с измененными векторами.

Самыми частыми частями речи слов с наиболее часто изменяющимися векторами были существительные и прилагательные. Они вместе составили более половины всех слов. Другие части речи, такие как наречия, глаголы и собственные имена, были менее представлены в результатах. Также были анализированы изменения между моделями 2000 и 2022 года, а также 2004 и 2022 года. Выбор данных моделей вызван тем, что 2022 является моделью с самым новым материалом, когда модель 2000 года – с самым старым, а 2004 – с самым старым, но обученная на достаточном количестве данных из всех источников.

Приведем результаты сравнения моделей 2004 и 2022 года. 22 слова было отброшено из-за ошибок парсинга. В результате анализа 100 слов с наибольшим изменением вектора было обнаружено 68 изменений контекста слов, чаще всего в связи с определенными событиями, а также 10 слов с непосредственными изменениями в значении. Было выявлено 4 частые темы: политика (26), военная тематика (6), названия мест или объектов (24) и фамилии (6). Слова с изменениями в значении:

* *мобилизированный\_VERB*,
* *азов\_PROPN*,
* *удаленный\_VERB*,
* *кейс\_NOUN*,
* *эко\_PROPN*,
* *заморозка\_NOUN*,
* *преднамеренный\_VERB*,
* *призывать\_ADJ*,
* *замещать\_VERB*,
* *пропорциональной\_ADJ*.

Разберем главные слова с изменениями в значении.

Для 2004 года характерно следующее значение лексемы *мобилизованный\_VERB*: направить (направлять) на выполнение какой-л. работы. Например: «*Налоговых и неналоговых доходов, контролируемых ФНС России, в федеральный бюджет страны мобилизовано 1 триллион 386,7 миллиарда рублей, что на 15,8% больше, чем за соответствующий период 2003 года*.» В 2022 наиболее частотным является данное значение лексемы: «тот, кто призван в армию по мобилизации». Например: «*Мобилизованный военнослужащий ЛНР с позывным* «*Береза*» *обратился к своему сыну, воюющему в Донбассе на стороне Киева, с призывом сложить оружие.*» Таким образом, можно сделать вывод о том, что по сравнению с 2004 годом в связи с произошедшими общественно-политическими событиями, такими как военная операция на Украине, произошло сужение значения лексемы мобилизованный. Основное значение сменилось с направленности в целом на какую-либо работу, до конкретной направленности «в армию», характерной для военного дискурса.

В новостях 2004 года слово *азов\_PROPN* используется по отношению к Азовскому морю, а также к городу Азов в Ростовской области. Например: «*При этом Азов разделяется линией государственной границы, а использование Керченского пролива осуществляется в соответствии с соглашением сторон. В Азове Ростовской области открыта первая в истории Дона женская колония общего режима.*» В 2022 году чаще обозначает украинское вооруженное формирование, признанное в РФ террористической организацией. Например, «*На Украине решили возродить полк Азов (признан террористическим и запрещен в России).*» Следовательно, в связи с событиями на Украине в 2022 году по сравнению с 2004 годом основное значение лексемы азов\_PROPN изменилось с «Азовское море» на «украинское вооруженное формирование». Данное значение не присутствует в датасете 2004 года, что свидетельствует о »расширении» значения.

*Удаленный\_VERB* употребляется в 2004 году в своих двух значениях причастия от глагола «удалить» и «находящегося на расстоянии». Например, «*При этом индийские ученые считают, что могли быть полностью уничтожены 6 племен, проживавшие в полной изоляции на самых удаленных островах архипелага. В сборной Англии вместо защитника Ледли Кинга на поле вышел центральный защитник* «*Челси*» *Джон Терри, а в сборной Швейцарии вместо удаленного в матче с хорватами Йохана Фогеля вышел Фабио Челестини.*» а 2022 году намного чаще во втором значении «находящийся, работающий или осуществляемый на расстоянии», причем, как правило, подразумевается удаленность работы, часто из-за рубежа, и имеет в синонимах по векторам такие слова как «дистанционный\_ADJ». Пример из корпуса: «*Помимо этого, психиатр отметил, что с массовым переходом на удаленную работу существенно увеличилось число разводов.*» В связи с этим, удаленный\_VERB претерпело «сужение» основного значения.

В 2004 году *кейс\_NOUN* используется исключительно в значении «чемодан». Пример: «*После этого подозрительным объектом занялись взрывотехники. Как сообщается на сайте телеканала ORF, специалисты установили, что никакой опасности кейс не представляет.*» В 2022 году значение из профессиональной области «случай, описание случая» уже совершенно превалирует над прежним значением: «*Пока что все пробуют и подглядывают друг за другом, пытаясь понять, какой из кейсов уже становится неким правилом и неким алгоритмом для дальнейших действий всех остальных», — сказал он.*» Использование лексемы во втором значении произошло из-за заимствования из английского языка. Важно сделать оговорку, что следует рассматривать данный случай не как изменение значения, а как случай омонимии, где неологизм идентичен уже существующему слову, так как два значения не имеют общих сем.

*Эко\_PROPN* в 2004 году чаще всего обозначал «Организацию экономического сотрудничества». Пример: «*Его участниками станут представители всех антинаркотических структур государств-членов ЭКО – Азербайджана, Афганистана, Ирана, Казахстана, Киргизии, Пакистана, Таджикистана, Турции, Туркменистана, Узбекистана, а также секретариата Организации.»* когда в 2022 году чаще всего относился к «технологии экстракорпорального оплодотворения». Пример: «*Но женщина, желающая стать матерью генетически своего ребенка, должна понимать, что возможности ЭКО в этом плане ограничены.*» В обоих случаях лексемы представляют собой аббревиатуру. Они обозначают разные вещи и являются омонимами. Эко в значении 2022 года является неологизмом.

*Заморозка\_NOUN* в 2022 году используется чаще в переносном значении вместо основного. Пример для 2004 года: «*Последний был изготовлен методом заморозки смеси жидкого азота и воды.*» Пример для 2022 года: «*Введение мер в отношении НРД, который уже находился под ограничениями со стороны международных депозитариев Euroclear и Clearstream, также повлекло за собой заморозку полагающихся инвесторам выплат — дивидендов и купонов.*» Таким образом, можно сделать вывод об изменениях в значении, которое является метафорой по используемой нами классификации.

*Преднамеренный\_VERB* в 2004 употребляется преимущественно в экономической сфере по отношению к банкротству. Пример: «*Однако, из-за того, что преднамеренные банкротства совершаются хорошо подготовленным специалистами по крайне запутанной схеме, а обращения от потерпевших не поступают вообще или поступают несвоевременно, за время работы отдела было расследовано и направлено в суд всего одно уголовное дело*»*, – сообщил следователь.*» В 2022 году лексема приобретает более негативную коннотацию и употребляется со словами провокация, дезинформация, эскалация, теракт. Например: «*Все эти факты позволяют с полной уверенностью утверждать, что противник на данном участке осуществляет преднамеренные провокации и намеревается обострить и без того сложную ситуацию*»*, – сказал представитель ведомства.*» Таким образом, коннотации данной лексемы стала более негативной.

Приведем результаты сравнения моделей 2000 и 2022 года. Результаты были во многом аналогичны предыдущим. 15 слов было отброшено из-за ошибок парсинга. В результате анализа 100 слов с наибольшим изменением вектора было обнаружено 80 изменений контекста слов, чаще всего в связи с определенными событиями, а также 5 слов с непосредственными изменениями в значении. Слова с изменениями в значении: *кейс\_NOUN, тишина\_NOUN, вылиться\_VERB, потолок\_NOUN, мобилизация\_NOUN*.

Случай с *кейс\_NOUN* аналогичен предыдущим результатам.

Слово *тишина\_NOUN* на момент 2022 года чаще употреблялось в значении «отсутствие вражды, ссор, столкновений, боевых действий» по сравнению со своим основным значением «отсутствие звуков, шума», а также как часть названия тюрьмы «Матросская тишина» в 2000. Примеры для 2000 года: «*Ковалев пробыл в СИЗО «Матросская тишина» год и два месяца. «Меня спасло тиканье американки, которое я услышал в утренней тишине кабинета», – сказал глава администрации.*» Пример для 2022 года: «*У нас сегодня тишина. Возможно, тишина перед бурей, но мы готовы стоять до конца и наш русский город Херсон сдавать никто никому не собирается», — подчеркнул замглавы администрации.*» С учетом этого, представляется возможным сделать вывод, что произошло изменение значения, которое можно охарактеризовать как метафора.

В новостях 2022 года *вылиться\_VERB* чаще употребляется в переносном значениях, «принять какую-н. форму, выразиться в чем-н.», чем в основном «вытечь из чего-либо или откуда-либо (о жидкости)», как в 2000. Пример для 2000 года: «*На асфальт вылилось несколько тонн солярки.*» Пример для 2022 года: «*Юрий Аванесов предупредил, что разведение россиянами цветов на балконе может вылиться в денежный штраф.*» Данное изменение можно классифицировать как «метафору».

Аналогичная ситуация с *потолок\_NOUN*, где прямым значением является «верхнее покрытие внутри помещения», а переносным – «предельное значение какой-либо величины, верхняя граница; предел, предельная степень чего-либо». Пример для 2000 года: «*Предположительно, потолок, сделанный из обожженной на солнце глины, не выдержал веса гостей, собравшихся на втором этаже.*» Пример для 2022 года: «*Лидеры «семерки» и так договорились ввести потолок цен на нефть.*» Таким образом, есть основания говорить об изменении значения, которое является «метафорой».

*Мобилизация\_NOUN* в датасете 2022 года чаще используется как «призыв на военную службу», чем «направление на выполнение какой-либо работы, каких-либо задач». Пример для 2000 года: »*Он не стал раскрывать суть тенденций, однако заявил, что от правительства, администрации и регионов* «*потребуется мобилизация усилий*»*, передает РИА* «*Новости*». Пример для 2022 года: «*Всеобщая мобилизация — это очень сложный сценарий, она не объявляется по всяким изменениям ситуации.*» Данное изменение можно определить как «сужение» значения.

Говоря про результаты изменения значений между моделями следующих друг за другом лет, было проанализировано топ-50 слов с измененными векторами для каждой последующей пары моделей. Результаты показали, что абсолютное большинство изменений векторов были вызваны изменением контекста, но не изменением семантики слов.

Как правило, изменение контекстов связано с определенными событиями, о которых было написано в датасете. Так, для результатов по 2000-2001 году чаще встречаются слова, связанные с конфликтом в Чечне (*пленный\_ADJ, возмездие\_NOUN*) и вспышкой сибирской язвы в США (*сибирская\_ADJ, язва\_NOUN*). Для результатов 2002-2003 войной в Ираке и нестабильности, вызванной ей, (*капитуляция\_NOUN, багдада\_PROPN, эр-рияд\_PROPN, корпус\_PROPN*)

Слово *идиотизм\_NOUN* встречается как в топ-50 словах с максимальным изменением вектора в 2008-2009, так и 2009-2010.

Для 2013-2014 годов характерны темы конфликта на Донбассе и присоединения Крыма (*украинский\_NOUN, крымчанин\_NOUN, донбасс\_PROPN, крымский\_ADJ, русскоязычной\_ADJ, малайзийский\_ADJ*). Для 2014-2015 – конфликта в Сирии (*сирия\_NOUN, асад\_NOUN*). Для 2019-2020 – пандемии (*карантина\_NOUN, пандемия\_NOUN, выздороветь\_VERB*).  
  
Исключениями выступили слово *призванный\_VERB* в 2001-2002 годах, а также *релокация\_NOUN* в 2021-2022 годах, изменение векторного представления которых невозможно объяснить изменением исключительно контекста.

Слово *призванный\_VERB* претерпело изменения, так как в статьях чаще стал использоваться в общем значении «направленный на выполнение какой-либо работы, каких-либо задач». Пример: «*Документ, призванный содействовать сохранению этнокультурного многообразия и укреплению единства российской нации, был разработан по поручению главы Мордовии*» В датасете 2000 года чаще наблюдается значение «призванный на военную службу». Пример: «*Всем призванным на военную службу резервистам Генштаб гарантирует выплату необходимых компенсаций и денежного содержания.*» Можно сделать вывод об изменении значения, которое классифицируется как «расширение».

В 2021 году под *релокация\_NOUN* подразумевался переезд в целом, обычно из региона в регион. Пример: «*Раньше релокацию сотрудника было провести крайне сложно: директор одного магазина должен был в ручном режиме помочь сотруднику пройти 10 кругов согласований, договориться с директором магазина в другом городе и так далее.*» В 2022 году в абсолютном большинстве случаев имелся в виду переезд заграницу. Пример: «*Он также отметил, что процентное соотношение по сравнению с маем изменилось - количество тех, кто рассматривает вариант возвращения, выросло на 2,3%. Эксперт сообщил, что релокация сохраняет стихийный характер.*» Таким образом, произошло «сужение» значения лексемы.

Таким образом, было обнаружено 5 сужений значения, 4 расширения, 4 метафоры, 2 появления неологизма, который являлся омонимом уже существующей лексемы, 1 изменение коннотации в негативную сторону. Кроме изменений значения, обнаружено множество изменений контекста слов, которое вызывало изменение их векторных представлений.

**Заключение**

Таким образом, при написании данной курсовой работы было проведен сбор и обработка датасета, а также обучение моделей за период времени с 2000 по 2022 год, которые были выровнены с использованием метода Прокруста.

В данной курсовой работе был проведен частотный анализ лексем, употребленных с 2000 по 2022 год. Результаты показали, что частота лексем напрямую зависит от ситуации в мире и контекстов. Новые лексемы, ранее не употреблявшиеся, становятся одними из самых частотных в период их появления (например: *криптовалюта, денацификация*). Однако, помимо лексем, зависящих от ситуаций в мире, за период с 2000 по 2022 год стали чаще употребляться лексемы, которые ассоциируются с интернетом и технологиями. Сравнение частот 2000 и 2022 годов и 2004 и 2022 годов показали, что в постоянном лексиконе появились новые лексемы (*смартфон, ник, аккаунт* и т.д), а также некоторые ранее существовавшие лексемы стали употребляться чаще из-за их добавочного значения, связанного с развитием технологий.

 Также была анализирована семантика лексем, используя косинусное расстояние между векторным представлением слов.   
В результате анализа данных о частоте использования слов с семантическим сдвигом, было установлено, что доля таких слов варьируется от 2 до 9 процентов. Наибольшее количество слов с семантическим сдвигом было зафиксировано среди имен существительных, за ними следовали прилагательные, глаголы, имена собственные и наречия.   
Анализ изменений между моделями 2000 и 2022 года, а также 2004 и 2022 года показал, что большинство изменений векторов были вызваны изменением контекста, а не семантики слов. Однако были выявлены слова с непосредственными изменениями в значении, такие как *мобилизированный\_VERB, азов\_PROPN, удаленный\_VERB, кейс\_NOUN, эко\_PROPN, уклонист\_NOUN, заморозка\_NOUN, преднамеренный\_VERB, призывать\_ADJ, детализировать\_VERB, отключить\_VERB, замещать\_VERB, пропорциональной\_ADJ, умалчивать\_VERB*.   
Также были выявлены частые темы, связанные с изменениями контекста слов, такие как политика, военная тематика, названия мест или объектов и фамилии.

Результаты данной курсовой работы могут быть использованы в лексикологии, где необходимы данные для построения актуальных словарей. Например, неологизм *кейс\_NOUN* почти не отражен на сегодняшний день в словарях, хотя активно употребляется в СМИ.

В ходе работы было обнаружено, что самые эффективные результаты получаются при сравнении моделей, обученных на материале, значительно отдаленном по временным рамкам, а также при моделях, обученных на как можно большем количестве данных, поэтому перспективами курсовой работы является изучения большего количества сдвигов на большем количестве материала.

**Список использованной литературы**

1. Атрощенко М.С. Экстралингвистические факторы изменения значения слова // Актуальные проблемы культуры речи : Сборник научных статей Всероссийской научно-практической конференции / Ответственный редактор И.С. Папуша. – М.: Московский государственный областной университет, 2018. – С. 93-98.
2. Волкова Е.А. Семантические изменения в английском языке // Гуманитарные и социальные науки. 2016. №1.
3. Лингвистический энциклопедический словарь / Гл. ред. В.Н. Ярцева. 2-е изд., доп. М., 2002.
4. Пауль Г. Принципы истории языка / Пер. с нем. М.: Изд-во иностранной литературы, 1960. – 500 с.
5. Сервис ShiftRy [Электронный ресурс] URL: [https://shiftry.rusvectores.org](https://shiftry.rusvectores.org/) (дата обращения: 02.03.2023)
6. Хакимзянова Д.Ф., Хованская Е.С. Семантическое развитие слова на примере глагольной лексики // Современные проблемы науки и образования. – 2015. – № 2-2.
7. Шереметьева С.О., Осминин П.Г. Методы и модели автоматического извлечения ключевых слов // Вестник ЮУрГУ. Серия: Лингвистика. 2015. №1.
8. Bizzoni Y., Degaetano-Ortlieb S. et al. Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach // Sec. Language and Computation. 2020. Vol. 3.
9. Brysbaert M., Buchmeier M., Conrad M., Jacobs A.M., Bölte J., Böhl A. The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. // Exp Psychol. — 2011. — № 58(5). — С. 412-424.
10. Cook P., Lau R., Mccarthy D., Newman D., Balwin T. Novel Word-sense Identification // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. – 2014. – P. 1624-1635.
11. Fomin V., Bakshandaeva D., Rodina Ju.,Kutuzov A. Tracing cultural diachronic semantic shifts in Russian using word embeddings // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2019». – 2019.
12. Gulordava K., Baroni M. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus // Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics. – 2013. – P. 67-71.
13. Kutuzov A., Fomin V., Mikhailov V., Rodina J. ShiftRy: Web service for diachronic analysis of russian news // Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference «Dialogue». – 2020. – P. 500-516.
14. Lau R., Cook P., Mccarthy D., Newman D., Balwin T. Word sense induction for novel sense detection // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. – 2012. – P. 591-601.
15. Mahler H. Lexical Emergence on Reddit: An Analysis of Lexical Change on the «Front Page of the Internet» / Mahler H. [Электронный ресурс] // journals.openedition : [сайт]. — URL: https://journals.openedition.org/lexis/4917#authors (дата обращения: 17.03.2023).
16. McCreath G.M., Linehan C.M.J., Mar R.A. Can Differences in Word Frequency Explain Why Narrative Fiction Is a Better Predictor of Verbal Ability than Nonfiction? // Discourse Processes. — 2017. — № 54:5-6. — С. 373-381.
17. Mitra S., Mitra R., Maity K., Riedl M., Biemann C., Goyal P., Mukherjee A. An automatic approach to identify word sense changes in text media across timescales // Natural Language Engineering – 2015. – P. 1-26.
18. Mitra S., Mitra R., Riedl M., Biemann C., Mukherjee A., Goyal P. That’s sick dude!: Automatic identification of word sense change across different timescales // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2014. – P. 1020-1029.
19. Porter M.F. An Algorithm for Suffix Stripping // Readings in Information Retrieval. Morgan Kaufmann Publishers Inc. – 1997. – P. 313–316.
20. Sagi E., Kaufmann S., Clark B. Semantic density analysis: Comparing word meaning across time and phonetic space // Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics. – 2009. – P. 104-111.
21. Tahmasebi N., Risse T. Finding individual word sense changes and their delay in appearance // Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. – 2017. – P. 741–749.
22. Wijaya D., Yeniterzi R. Understanding semantic change of words over centuries // DETECT '11: Proceedings of the 2011 international workshop on detecting and exploiting cultural diversity on the social web. – 2011. – P. 35-40.

**Источники материала исследования**

1. «Коммерсантъ». [Электронный ресурс]. URL: [https://www.kommersant.ru](https://vk.com/away.php?to=https%3A%2F%2Fwww.kommersant.ru&cc_key=) (дата обращения: 02.03.2023)

2. «Лента». [Электронный ресурс]. URL: [https://lenta.ru](https://vk.com/away.php?to=https%3A%2F%2Flenta.ru&cc_key=) (дата обращения: 02.03.2023).

3. «РИА Новости». [Электронный ресурс]. URL:<https://ria.ru/> (дата обращения: 02.03.2023).

**Приложение 1**

Код, использованный при написании данной курсовой работы: https://github.com/tatarinovst2/course-work-3