Санкт-Петербургский государственный университет

ТОКАРЕВА Анна Александровна

Выпускная квалификационная работа

Процедура тематической атрибуции русских текстов с использованием деривационного анализа

Уровень образования: бакалавриат

Направление 45.03.02 «Лингвистика»

Основная образовательная программа СВ.5106. «Прикладная, компьютерная и математическая лингвистика (английский язык)»
Профиль «Прикладная, компьютерная и математическая лингвистика (английский язык)»

Научный руководитель:
доцент, Кафедра математической лингвистики,
Азарова Ирина Владимировна
Рецензент:
старший научный сотрудник, Институт когнитивных исследований,
Алексеева Светлана Владимировна

Санкт-Петербург
2022

Аннотация

В данной работе предпринимается попытка разработки базового алгоритма тематической атрибуции с использованием деривационного анализа. Для решения поставленной задачи используются как статистические и лингвистические методы определения темы текста. В теоретической части данного исследования рассматриваются вопросы автоматического тематического аннотирования и описываются основополагающие принципы деривационного анализа. В практической части работы применяются гибридные подходы к выделению тематически маркированных слов, включая качественные методы, основанные на экспертных оценках, и количественные методы, опирающиеся на подсчет коэффициента тематичности. Среднее значение эффективности предложенного алгоритма показало результат в 70 %.

Ключевые слова: *тематическая атрибуция, тема, деривация, тематически маркированная лексика, терминологичность*.

The graduation qualification paper represents an attempt to develop a basic algorithm for thematic attribution using derivational analysis. Both statistical and linguistic methods of topic modelling are applied to solve the problem in question. This study deals with the theoretical issues of automatic thematic annotation and describes the fundamental principles of derivational analysis. The practical part of the work relies on hybrid approaches to the extraction of thematically labeled words. Among these methods are qualitative methods based on expert evaluations and quantitative methods drawn on the thematicality coefficient calculation. The average efficiency of the proposed algorithm showed the result of 70%.

Keywords: *thematic attribution, topic, derivation, thematically labeled lexicon, terminology*.

Оглавление

Введение

Тематическая атрибуция текста – это нетривиальная задача обработки естественного языка. Она представляет собой некоторый параметр, который находит себе применение в задачах информационного поиска, фильтрации документов, определения тематических рубрик для электронных ресурсов. Тематическая атрибуция (тематическое моделирование) используется для поиска информации по смыслу, а не по ключевым словам, для построения профилей интересов пользователя и для аннотирования изображений.

В нашем видении проблему тематической атрибуции текста можно решить путем нахождения слов-дериватов, несущих в себе основную смысловую нагрузку. Общепринятым методом выявления темы текста является поиск ключевых выражений. Как правило, они частотны и представляют собой номинативные части речи. Соответственно, ключевые слова в структуре текста занимают определенные синтаксические позиции и регулярно повторяются. На этой идее основано наше представление о подходе к нахождению темы текста. Мы считаем, что повторяются в тексте те самые слова дериваты, поэтому нахождение общей основы-представителя упростило бы задачу тематического моделирования.

В связи с этим мы сталкиваемся с еще одной проблемой – морфемной сегментацией текста. На сегодняшний день стемминг (иначе выделение основы слова) – не только необходимая часть обработки любого естественного языка, но и способ нахождения тематически маркированной лексики. Основа слова – это его семантическое ядро, но, как правило, значимые для темы слова всегда имеют в тексте множество производных с той же основой, то есть являются ключевыми элементами текста. Даже в информационном поиске ключевыми терминами для запроса являются именно стеммы (или основы слова), а не исходные слова. Таким образом, для современной лингвистики очень важны исследования в области автоматического морфемного членения, и в данной работе осуществляется попытка

разработать формальный подход к нахождению темы текста посредством деривационного анализа.

Итак, наша **гипотеза** заключается в том, что слова-дериваты, объединенные в классы эквивалентности на основании их принадлежности к одному словообразовательному гнезду, занимающие в тексте разные синтаксические позиции, но регулярно повторяющиеся, отражают тему данного текста. Включение поиска таких слов в существующие алгоритмы тематического моделирования позволит улучшить эффективность этих алгоритмов.

Актуальность исследования обусловлена малым количеством работ по тематической атрибуции, основанной на комбинированном подходе к поиску тематически важных слов.

Предметом исследования являются тематически маркированные слова в тематическом корпусе текстов.

Объектом изучения становятся слова-дериваты, которые составляют единое словообразовательное ядро и являются тематическими маркерами.

Материалом исследования является собранный тематический корпус текстов объемом 2000 словоформ.

Научная новизна данного исследования состоит в разработке подхода к определению темы текста, для которого на второй план уходит частеречная принадлежность «тематичного» слова, поскольку все синтаксические позиции в тексте заняты одной основой-представителем.

Целью исследования становится разработка алгоритма тематической атрибуции текстов определенной предметной области с опорой на правила синтаксической (транспозиционной) деривации.

Задачи:

1. Дать определение тематической атрибуции.
2. Изучить теоретические основы деривационного анализа с акцентом на словах-дериватах, для которых в процессе словообразовательного анализа не происходит изменения значения.

3. Рассмотреть существующие подходы к проблеме актуального членения, разграничить понятия темы и ремы.
4. Разработать процедуру транспозиционного деривационного анализа для получения основ (стемм), объединяющих элементы словообразовательного гнезда.
5. Оценить эффективность метода в определении темы.

Данная работа состоит из введения, трех глав, тридцати двух таблиц, десяти иллюстраций, заключения, списка литературы и двух приложений. В первой главе рассматриваются теоретические вопросы, связанные теорией словообразования, на которую мы опирались при реализации алгоритма. Вторая глава посвящена описанию тема-рематической организации текста и основным статистическим методам определения темы. В третьей главе рассматриваются алгоритмы, направленные на выделение ключевых слов и терминов, а также статистические алгоритмы выделения темы документа. В четвертой главе описан эксперимент по присвоению темы на основании синтаксической деривации, представлены результаты и их оценка.

Глава 1. Деривационный анализ

1.1 Словообразование как компонент языковой структуры

Некоторые лингвисты считают, что словообразование представляет собой раздел языкознания, который подразделяется на морфемику и деривацию. Однако, словообразование правильнее рассматривать в качестве компонента лексикологии, поскольку лексический состав русского языка в основе своей пополняется именно за счет образования новых слов, реже – за счет заимствований. Большая часть слов русского языка – это производные.

В.В. Виноградов отмечает, что именно изучение словарного фонда языка, его лексического состава и связей между ними является основой для исследований закономерностей словообразования. Лингвист акцентирует внимание на изменчивости словарного состава, отмечая, что он постоянно пополняется новыми словами, которые возникают в связи с изменением социального строя, с развитием производства, с развитием науки и культуры. [В.В. Виноградов, 1975]

Основной единицей лексикологии является слово, для словообразования – это производное слово. Тем не менее, словообразовательный анализ включает в себя не только исследование производных слов, но и слов непроизводных, которые впоследствии служат основой для образования дериватов. Производные слова, в свою очередь, подчиняются законам лексики, приобретая новые значения и вступая в разные отношения. Таким образом, лексикология и словообразование находятся в отношениях сходства и смежности.

Очевидна и связь словообразования с другими дисциплинами, в частности с синтаксисом и морфологией. Части речи и грамматические категории представляют собой отдельные системы, которые соотносятся со словообразовательной системой, так как каждая часть речи характеризуется наличием собственных способов образования новых слов, а также своим набором словообразовательных аффиксов. Кроме того, новые слова в языке

возникают в качестве представителей определенной части речи, типа словоизменения, лексико-грамматического и семантического разряда.

Важно отметить, что словообразование представляет собой динамический процесс, который реализуется посредством словообразовательных моделей. Результатом процесса словообразования являются слова-дериваты, которые с одной стороны вошли в лексикон и плотно укоренились в структуре языка, а с другой стороны – не встретились в корпусе в явном виде и представляют собой так называемые «потенциальные слова», образованные по существующей в языке словообразовательной модели. Важными для словообразования являются понятия актуальности и потенциальности. Актуальность связана с реально встречающимися словами в языке. Наряду с актуальными словами встречаются и слова, в отношении которых могут возникать сомнения по поводу их реального функционирования в языке. Такие слова называются потенциальными. Однако, встает вопрос, считать ли малочастотное слово актуальным. Допустим, в корпусе объемом несколько миллионов словоформ слово встретилось лишь один раз, реально ли это слово. Если посчитать относительную частоту (число употреблений слова на миллион слов в корпусе) данного слова в ipm (instances per million), значение будет стремиться к нулю. Таким образом, вопрос об актуальности данного слова остается открытым, ведь мы наблюдаем его существование в языке. Тем не менее, решение можно найти, определив некоторое пороговое значение, согласно которому слово будет считаться реальным или потенциальным. В данной работе мы не будем претендовать на правильность определения данного значения и лишь для примера примем его равным 0,1 ipm.

Большой вклад в становление словообразования внесли представители Казанской лингвистической школы (КЛШ) И.А. Бодуэн де Куртенэ и Н.В. Крушевский. Лингвисты задались вопросом членения слов и изучением их словообразовательной структуры.

И.А. Бодуэн де Куртенэ одним из первых поставил вопрос об основной единице морфологического уровня, а также выделил минимальную значимую часть слова, то есть морфему. Ученый разделял диахронию и синхронию и предлагал рассматривать словообразовательные связи с обеих перспектив, объясняя это тем, что в языке конкретного периода присутствует след разных эпох, поэтому вместо рассмотрения языковых явлений в одной конкретной плоскости, правильнее их рассматривать в исторической перспективе [Бодуэн де Куртенэ, 1963].

Синхроническое словообразование объясняет языковые явления, происходящие одновременно. Словообразование в синхронической перспективе описывает словообразовательные связи, воспринимаемые носителем как реальные, характерные для данного времени.

Диахроническое словообразование, наоборот, направлено на изучение словообразовательных связей и единиц, существовавших в определенный временной период. Таким образом, при диахроническом подходе важны, в первую очередь, свидетельства изменения словообразовательной структуры. Как отмечает Земская, синхронное словообразование рассматривает отношения единиц, которые существуют в одно время, а диахроническое словообразование изучает процессы превращения одних единиц в другие [Земская, 2011].

Диахронические исследования представляют собой большой интерес, однако, в рамках данного исследования мы придерживаемся строго синхронического подхода, который является отражением сознания носителей языка, и который можно рассматривать при автоматическом анализе. При деривационном анализе для нас важны слова, связанные синхроническими отношениями. Слова, связанные диахроническими отношениями, могут иметь разные значения, даже имея общую основу, поэтому объединение таких основ при работе разработанного нами алгоритма даст только отрицательный результат.

Нынешнее состояние словообразовательной системы зависит от языковых явлений, которые имели место в прошлом, хотя и в исторической перспективе словообразовательная структура описывается с помощью синхронии. Однако, представляется очевидным, что нельзя анализировать явления, происходящие в языке на данный момент, с явлениями, имевшими место в прошлом, поэтому требование И.А. Бодуэна де Куртенэ разграничить диахронию и синхронию совершенно справедливо.

Н.В. Крушевский, как один из последователей Бодуэна де Куртенэ, воспринимал словообразование как отдельную систему. Большое внимание лингвист уделял изучению словообразовательных аффиксов, разработав ряд правил идентификации морфем и нахождения морфемных границ. Среди них:

1) Высокая частотность морфемы в языке. Данный пункт требует уточнения: частоту морфемы можно рассматривать как некоторую последовательность н-грамм, то есть количество раз, когда данная н-грамма встретилась в тексте или корпусе.

2) Наличие определенного регулярно выраженного значения. Данное утверждение требует уточнения относительно понимания смысла выражения «регулярно выраженное значение». Дело в том, что говорить о наличии одного конкретного значения у морфемы весьма трудно. Применительно к корню, рассуждать о наличии общего значения еще представляется возможным, если не утверждать об устойчивости данного значения. Напротив, задача поиска общего значения у префикса или суффикса кажется еще более затруднительной, так как в русском языке такое значение ускользает и размывается. В качестве примера приведем суффикс феминитива -их, который служит как для обозначения профессии (ткачиха, повариха), так и для обозначения особи женского рода (слониха). В данной ситуации трудно ответить на вопрос, какое из этих значений является наиболее регулярным.

Рассуждать о наличии общего регулярного значения мы можем в рамках концепции В. А. Плунгяна об аддитивной модели морфологии, которая подразумевает существование некоторого каркаса языка, который

складывается из отдельных составных частей. В рамках данной теории слова складываются из морфем без дополнительных преобразований, равно как и смыслы этих морфем соединяются в одно целое значение, которое и будет общим для всех. Для аддитивной модели характерен изоморфизм семантического и формального членений, то есть при разложении некоторого смысла А на элементы х и у цепочка фонем, которая выражает этот смысл, схожим образом будет распадаться на подцепочки /х/ и /у/ [Плунгян, 2003].

Аддитивная модель составляет ядро того, что мы видим в тексте, но, тем не менее, в языке регулярно встречаются отклонения в виде кумуляции, идиоматичности, контекстной вариативности и фузии. Аддитивная морфология – это основа, на которую опирается данная работа, так как следуя принципам аддитивной морфологии, мы можем находить закономерности в сегментации морфем и проверять их на соответствие норме русского языка.

3) Системные взаимосвязи морфем. Данный критерий можно рассматривать с двух сторон: с точки зрения сосуществования морфем в контексте аддитивной морфологии и с точки зрения семантических отношений между морфемами. Под первым мы понимаем формальные закономерности сочетания морфем и отклонения морфем от принципов аддитивной морфологии. С другой стороны, к системным отношениям между морфемами также относят и отношения омонимии, синонимии и антонимии. Например, для выражения значения деятеля используются суффиксы, такие как -тель, -ник, -щик [там же].

4) Отношения деривата с производящей основой и мотивационные отношения. Если говорить об основных типах семантических отношений между производящим словом и производным, то принято пользоваться классификацией М. Докулила, который рассматривал три следующих типа:

- модификацию (семантическое включение) – отношения, при которых значение производного слова совпадает со значением производящего, но содержит дополнительный семантический компонент (дом - домик);

- мутацию (семантическое пересечение) – отношения, которые предполагают наследование производным словом части семантических компонентов производящего слова (чай как «напиток» – чайник «сосуд с ручкой и носиком, используемый для заваривания чая»);
- транспозицию (нулевая семантическая оппозиция) – отношения, при которых происходит перенос значения производящего слова на производное, но при этом меняется часть речи производного слова: белый(прил.) – белизна (сущ.) [Dokulil, 1962].

К мотивационным отношениям относятся отношения непосредственной и опосредованной мотивации, которые зависят от позиции производного слова в словообразовательном гнезде.

Особое внимание ученый уделял структуре производного слова (деривата), предложив идею его бинарности и выделив особый оттенок в значении деривата. Наблюдая за словообразовательным строем русского языка, Н.В. Крушевский изучал варианты морфем (морфологических элементов), и, по мнению Н.А. Николиной, предвосхитил «выделение в XX в. морфов одной морфемы», чье функционирование, в свою очередь, обусловлено историческими и живыми позиционными чередованиями звуков [Николина, 2015].

Большой вклад в теорию словообразования внес В.В. Виноградов. В своих работах «Вопросы современного русского словообразования (1953)» и «Словообразование в его отношении к грамматике и лексикологии» (1952)» лингвист впервые представил относительно полную для того времени классификацию способов словообразования. Туда входили:

1) морфологический (морфемный, или аффиксальный) способ словообразования, который состоит в модификации исходных лексических единиц и их превращении в новые путем прибавления аффиксальных морфем.

2) лексико-семантический способ словообразования — то есть «расщепление» многозначных слов.

3) лексико-синтаксический способ словообразования, для которого характерно сращение двух или нескольких единиц и их превращение в устойчивую лексическую единицу, функционирующую в качестве самостоятельного слова.

4) морфолого-синтаксический способ словообразования или, иными словами, переход слова из одной части речи в другую.

Данная классификация Пражского лингвистического кружка широко ценилась в кругах лингвистов, однако, в начале 70-х г.г. была пересмотрена, так как учитывала не все языковые реалии.

В течение нескольких десятков лет теория словообразования претерпевала многочисленные изменения. Многие ученые, включая О. П. Ермакову, П. А. Соболеву, А. Н. Тихонова, Е. С. Кубрякову, Е.А. Земскую, В.А. Богородицкого, В. В. Виноградова, Г. О. Винокура исследовали данный раздел языкознания, предлагали новые термины и классификации, благодаря которым на сегодняшний момент мы имеем достаточно полную и подробно описанную структуру словообразовательной системы русского языка.

Другим неотъемлемым компонентом изучения системы русского языка является морфемный анализ. Морфемный и словообразовательный анализ являются взаимосвязанными аспектами изучения слова в системе языка. Подтверждение этому можно найти в словах В.В. Лопатина о том, что подлинный морфемный анализ начинается только тогда, когда вследствие последовательного словообразовательного анализа уже установлена словообразовательная структура данного слова, а также известны значения входящих в это слово морфем [Лопатин 1977, с.28-29]. Таким образом, по мнению лингвиста, словообразовательный анализ выходит на первое место, а морфемный анализ занимает вторичное положение.

1.2. Принципы морфемного анализа и понятие морфемы

Работая с текстами, необходимо иметь в виду, что они представляют собой грамматически и синтаксически связанные словоформы, которые

состоят из линейной последовательности знаковых единиц — морфем, которые подвержены контекстному варьированию. Изучением морфов и морфем как элементарных значимых частей слов и морфем занимается морфемика.

Очевидно, что объект изучения морфемики – морфема. Ввел понятие морфемы И.А. Бодуэн де Куртенэ, который определял морфему как наименьший, далее неделимый, дальше неразложимый морфологический элемент языкового мышления. Ученый говорил о неделимости морфемы, однако, утверждал, что морфема может делиться на морфологические дроби, которые сами по себе являются морфемами: «Если морфему можно делить дальше на ее составные части, то эти составные части должны быть с нею однородны, должны также иметь значение, в них должна пульсировать психическая жизнь» [Бодуэн де Куртенэ, 1963].

И.А. Бодуэн де Куртенэ рассматривал морфему как двустороннюю единицу, имеющую как план содержания, так и план выражения. Еще одним важным термином в языкознании стала нулевая морфема, которая лишена всякого произносительно-слухового состава, но ассоциируется с известными семасиологическими и морфологическими представлениями [Там же].

Для В.В. Виноградова слово – это система морфем, организованная по законам грамматики данного языка. Очень важным для теории словообразования является положение Виноградова о слитности корневых морфем с суффиксами словообразования, о продуктивности и непродуктивности служебных морфем. Лингвиста волнует вопрос соотношения морфемного состава слова и его морфологической структуры.

Существует несколько трактовок понятия «морфема». Для Блумфилда очень важна двусторонность данной единицы. По его мнению, морфема – это языковая форма, которая лишена частичного фонетико-семантического сходства с какой-либо другой формой [Блумфилд, 1968].

Представители Пражского лингвистического кружка морфемой называли морфологическую единицу, неразложимую на более мелкие

морфологические единицы, в целом ряде слов имеющую одну и ту же формальную функцию, то есть такую единицу, которую невозможно разложить на более дробные части, обладающие этим свойством.

Изучив работы лингвистов, можем выделить несколько основных трактовок морфемы:

- Морфема – это единица плана выражения; значение не имеет никакого отношения к ее определению; морфема существует только как повторяющаяся часть отдельных высказываний.
- Морфема – это чисто функциональная единица языка, и ее форма несущественна для выполнения ее функций.
- Морфема – это знаковая, двусторонняя единица языка, соотносящая план выражения с планом содержания.
- Морфема – минимальная значащая единица, связанная со словом и неспособная перемещаться в пределах этого слова.

Е.С. Кубрякова считает, что трактовка морфемы должна согласовываться со следующими теоретическими предпосылками:

1) морфема понимается как элементарная, но не единственная единица описания морфологического уровня, которая представляет собой минимальную и предельную единицу данного уровня;

2) морфема – знаковая единица, которая имеет отношение к передаче разных типов лингвистического значения;

3) морфема — двусторонняя единица, поэтому для ее характеристики важно учитывать и ее форму, и ее значение [Кубрякова, 1974].

Что касается классификации морфем, то их можно разделить, во-первых, по их роли и функции в составе слова, то есть на служебные и неслужебные морфемы. Неслужебные морфемы несут в себе основную смысловую нагрузку слова, а служебные морфемы задают форму слова. Традиционно по данному критерию различают корневые и аффиксальные морфемы. К корневым морфемам относится корень, к аффиксальным морфемам – приставка, суффикс, постфикс, флексия (окончание). Корневые морфемы – обязательная,

центральная часть слова. Важно отметить, что не все корни можно отнести к неслужебным морфемам, и не все аффиксы являются исключительно служебными.

Корни могут выполнять как служебную, так и не служебную функцию. В большинстве случаев корневые морфемы в действительности являются неслужебными и выступают в роли ядра образования, составляют основную, обычно неизменяемую часть слова, часто совпадая с основой слова. Аффиксальные морфемы, в отличие от корневых, имеют, как правило, закрепленную позицию в составе слова и могут употребляться только в словах с одинаковыми структурами изменения. При этом в большинстве случаев аффиксальные морфемы не имеют реального вещественного значения и не соотносятся напрямую с объектами окружающей действительности, как это делает корень [Кубрякова, 1974: с. 113-116].

Следующий аспект классификации морфем – это их способность употребляться независимо и изолированно. Корни, которые могут употребляться самостоятельно, называются свободными. В русском языке их большинство. Связанными называются корни, которые могут употребляться только в сочетании с аффиксами.

Однако, согласно точке зрения Кубряковой, при определении статуса морфемы, при дифференциации морфем на свободные и связанные необходимо рассматривать данную особенность на разных уровнях языка: на уровне морфологиии и синтаксиса. Аффиксальные морфемы, к примеру, обычно существуют как части более сложных конструкций. Напротив, корневые морфемы выступают и как составные части более сложных конструкций, и как самостоятельные слова [Кубрякова, 1974: с. 118-125].

Корневые морфы и морфемы в составе словоформ и слов обладают следующими характеристиками:

1) они всегда материально выражены и обязательны

2) корень заключает в себе лексическое значение.

3) существенный признак морфов и морфем – их способность употребляться в свободном и связанном виде. Аффиксальные морфы в составе словоформы всегда связаны с другими морфами. Свободными они не бывают. Лишь корневые морфы и морфемы могут характеризоваться как свободные и связанные.

Еще один признак дифференциации морфем – их участие в формо- или словообразовании. По данному критерию морфемы делятся, соответственно, на формообразующие (служат для образования формы слова) и словообразующие (служат для образования новых слов). К первым относятся флексии и словоизменительные суффиксы, а ко вторым – префиксы и суффиксы.

Стоит также разграничить понятия морфа и морфемы. Морф – линейная синтагматическая единица, предстает как вариант конкретной морфемы в словоформе. Морфема – совокупность алломорфов. Алломорфы имеют одинаковое семантическое значение и находятся в отношении дополнительной дистрибуции.

Немаловажным является вопрос о морфемной членимости, так как она является одним из важнейших признаков слова. Актуальность проблемы разбиения слов на морфемы становится очевидной при подробном рассмотрении подходов, предложенных лингвистами.

Все многообразие подходов к морфемному членению слов можно свести к двум подходам. Широко признанным является подход, при котором морфемное членение слова осуществляется с акцентом на словообразовательные отношения слова. Метод опирается на установление словообразовательных связей данного слова с другими словами. Второй подход ориентирован на структурную соотносительность слов, на наличие в других словах частей, формально и семантически тождественных тем, которые выделяются в данном слове. Тем не менее, стоит внимательнее рассмотреть несколько точек зрения.

И.А. Кузьмина считает, что деление слова на составные части, описание их морфологической структуры может быть осуществлено с двух точек зрения: 1) когда морфема рассматривается как минимальная значимая часть цельнооформленного слова; 2) когда морфема рассматривается как средство выражения грамматических значений слова. [Кузьмина, 2010]

По мнению Ю.С. Маслова, вычленение морфем — частей слов — основывается на параллелизме между частичными различиями, наблюдаемыми во внешнем облике (звучании) слов и их форм, и частичными различиями в значениях (лексических и грамматических), передаваемых этими словами и формами. Путем сравнения форм, частично различных (и тем самым частично сходных) по звучанию и по значению, выявляются различия (и сходства) в звучании, параллельные различиям (и, соответственно, сходствам) в значении, и таким образом устанавливаются единицы, в которых за определенным экспонентом (отрезком звучания, иногда нулем звучания и т. д.) закреплено определенное содержание (значение). Если эти единицы окажутся минимальными, т. е. не поддающимися дальнейшему членению на основе того же принципа, то это и будут морфемы. [Маслов, 1975]

Для М.А. Михайлова членимость слова – это его способность выделять в своем составе сегменты (основы, флексии, префиксы, корни, суффиксы, а также различные сочетания), каждый из которых связан с определенной семантикой. Для ученого важна связь смысловой и звуковой сторон морфемы. Он пишет, что формы членимы благодаря связям, которые устанавливаются между этими формами, но и эти связи сами укрепляются благодаря выделимости в них тождественных сегментов. [Михайлов, 1974]

Кубрякова Е.С. говорит о степени членимости, предлагая в качестве критерия лингвистической формы использовать свойства отдельных ее частей, связанные с их дистрибуцией и семантикой. Лингвист выделяет: 1) живое морфологическое членение слов; 2) условное членение слов; 3) дефектное членение слов. [Кубрякова, 2008]

В соответствии с тем типом членения, которому подлежит языковая форма (в силу ее объективных свойств, выделяются четыре группы последовательностей:

1) характеризующихся полной нечленимостью;

2) обладающих признаками живой морфологической членимости;

3) обладающих признаками условной членимости;

4) обладающих признаками дефектной членимости [Там же].

Опираясь на выделенные свойства, Кубрякова предлагает следующую классификацию:

1) Во-первых, способность/неспособность слова выделять отдельные значащие части, которые повторяются в другом лингвистическом окружении. Согласно этому признаку, все слова делятся на членимые и нечленимые.
2) Во-вторых, способность выделенных отрезков в членимых образованиях встречаться в уникальном (неуникальном) окружении. По данному признаку членимые образования делятся на формально - членимые, каждая из выделенных частей которых повторяется в разном окружении, и дефектно - членимые, у которых одна из выделенных частей встречается в уникальном или почти уникальном окружении.
3) В-третьих, способность выделенных отрезков у формально-членимых слов повторяться в разных окружениях с тем же или другим значением. По данному признаку слова делятся на обладающие свойством живой морфологической или же условной членимости. [Там же]

Как можно заметить, мнений по поводу членимости слова довольно много, однако, когда мы имеем дело с морфемной сегментацией, все постулаты о членимости слов сводятся к нескольким общим принципам морфемного анализа.

1) Принцип системности (принцип двойного сравнения), который предполагает наличие большого количества однокоренных слов для исходного слова, а также определенной системности аффиксов.

Одним из первых на данный принцип обратил внимание А.М.

Пешковский: «В одном из ... рядов должна являться та же основа с другими формальными частями (вертикальный ряд), в другом— та же формальная часть с другими основами (горизонтальный ряд...)». [Пешковский, 1933]

2) Принцип производности. Здесь на помощь приходит критерий Г.О. Винокура о том, что «значение слов с производной основой всегда определимо посредством ссылки на значение соответствующей первичной основы». Сделать вывод о морфемных границах и характере аффиксов можно только основываясь на соотношении производного и производящего слов. [Винокур, 1959]

3) Принцип синхронизма, при котором морфемная сегментация должна проводиться исключительно с точки зрения синхронии.

Важным для нашего исследования является понятие морфотактики - сочетаемостных характеристик (валентностей) морфем. В каждом языке существует некоторая комбинаторика присоединения аффиксальных морфем к основе слова, а также просто сочетаний корней, суффиксов и флексий. В процессе выполнения практической части данной исследовательской работы мы прибегали к рассмотрению таких комбинаторных групп.

Подробно разобрав основные определения морфемики, изучив подходы лингвистов, можем перейти ко второму аспекту словообразования – деривации.

1.3. Основные понятия деривации

Впервые термин деривация ввел в обиход польский ученый Ежи Курилович. Лингвист понимает деривацию как процесс, который направлен на трансформацию первичной функции исходной единицы и который

приводит либо к изменению этой функции, либо к изменению значения исходной единицы [Курилович, 1962].

Е.С. Кубрякова под деривацией понимает процесс или результат образования в языке любого вторичного знака, который можно объяснить с помощью единицы, принятой за исходную, или выделить из нее, применив определенные правила. [Кубрякова, 1974]

Е.А. Земская обращает внимание на функциональный аспект деривации. Для нее «деривационный механизм» неоднороден по выполняемым функциям, которые в общем можно разделить на пять:

1) собственно номинативная, то есть создание некоторого наименования процесса, объекта, явления;

2) конструктивная – позволяет изменить синтаксической пространство речи;

3) компрессионная функция связана с номинативной и помогает в создании более краткой номинации;

4) экспрессивная функция связана с созданием экспрессивной формы выражения;

5) стилистическая, позволяющая использовать средство выражения, которое соответствует той или иной сфере речи. Иными словами, функция согласования выбранного способа выражения с определенной сферой речи [Земская, 2005].

Деривацию можно подразделить на синтаксическую и лексическую. Такое разграничение предложил Е.Курилович.

Под синтаксической деривацией понимается специфический тип семантических отношений между производным и производящим словами, основанный на функциональной транспозиции, под которой мы «понимаем использование одной языковой формы в функции другой формы, т.е. ее противочлена в парадигматическом ряду». [ЛЭС, 1990]

Синтаксическая деривация — наиболее продуктивный тип словообразовательной транспозиции, так как в ней отражается противоречие

между категориальным значением производного и производящего слов. Вообще говоря, при синтаксической деривации производное слово главным образом отличается от производящего исключительно своими синтаксическими свойствами, при этом лексическое значение сохраняется.

Синтаксическим дериватом же, как писал Е. Курилович, называется форма с тем же лексическим значением, что и у исходной формы, но с другой синтаксической функцией. [Курилович, 1962]

О.П. Ермакова выделяет основные типы синтаксических дериватов:

1) имена действия (наказание, рисование, пение, тление, умение, издание, задание, уныние, спасенье, борьба, прибытие, забытье, ковка, рубка, грабеж и др.);

2) имена качества (новизна, глубина, открытость, резвость, непредсказуемость, ширь, высь и др.);

3) относительные прилагательные (зерновой, цифровой, стенной, металлический, завтрашний, вчерашний и др.);

4) наречия на -о, образованные от качественных имен прилагательных (быстро, громко, откровенно, терпеливо, суетливо, радостно и др.)

[Ермакова, 1984]

Рассуждая о роли синтаксических дериватов, Е.А. Земская отмечает, что синтаксические дериваты облегчают говорящему построение высказывания, и что «это “облегчение” касается употребления более кратких конструкций, включающих синтаксические дериваты, синонимических более длинным, включающим более сложные синтаксические построения» [Земская, 1981]

Е. Курилович рассматривает деривацию с точки зрения синтаксических функций производных форм, формулируя правило, касающееся отношения первичных синтаксических функций к вторичным, которое звучит следующим образом: «если изменение синтаксической функции некоторой формы (некоторого слова) А влечет за собой формальное изменение А в В (при той же лексической функции), первичной синтаксической функцией является та,

что соответствует исходной форме, а вторичной — та, что соответствует производной форме». [Курилович, 1962]

В данной работе мы имеем дело с синтаксической деривацией, или транспозицией по Докулилу. Согласно нашей гипотезе, именно синтаксические дериваты, занимающие разные позиции в тексте, способны выразить семантику текста.

Что касается лексической (семантической) деривации, то лингвист считает, что лексическая деривация предполагает, что исходные и производные слова идентичны друг другу по первичной синтаксической функции. Лексическая деривация направлена на преобразование лексического значения исходной единицы. Интересно сопоставление семантической деривации у Куриловича и таких явлений, как модификация и мутация у Докулила. Данные понятия соотносимы.

О. Н. Трубачев о семантической деривации говорит как о процессе появления у слова семантически производных значений, семантических коннотаций, дополнительных значений, то есть о процессе расширения семантического объёма слова, приводящего к возникновению семантического синкретизма, результатом которого является появление так называемой полисемии. [Трубачев, 1988]

Одними из самых распространенных типов семантической деривации являются метонимический и метафорический переносы.

Метонимизация есть процесс, аналогичный компрессии. В результате компрессии устраняются компоненты, семантика которых является чисто грамматической или целиком выражается другим предложением текста. В основе метафоры лежит контаминация двух предложений. Особенность метафоризации определяется типом оператора: используется оператор сравнения, получающий в результирующем предложении семантическое выражение. [Хасанов, 2018]

Таким образом, лексическая и синтаксическая деривация являются важными источниками пополнения словарного состава русского языка. Стоит

отметить, что вопрос функционирования дериватов обоих типов заслуживает большего внимания и более пристального изучения.

К определению деривации необходимо добавить несколько важных терминов, которые составляют теорию словообразования. В первую очередь, следует сказать о производном и производящем словах. Можно догадаться, что производящим называется слово, от которого непосредственно образовано производное слово (дериват).

Процессы деривации завершаются не только получением вторичной, результативной единицы, но также установлением особых деривационных отношений между производным и производящим. Эти отношения имеют место как между единицами одного и того же уровня, так и между единицами разных уровней. В этом отношении понятие «деривации» отражает два подхода: межуровневый подход, позволяющий выявить словообразовательные механизмы получения более сложных единиц «верхнего» уровня из менее сложных единиц «нижнего» уровня, и «внутриуровневый» подход, позволяющий объяснить механизмы синтагматической сочетаемости единиц. Отсюда самостоятельный статус получает деривационная морфология, которая описывает с способы организации морфологических структур слова в языках разного типа.

Одним из типов особых отношений являются формально-смысловые отношения. Е.А. Земская перечисляет следующие виды формально-смысловых отношений между производной и производящей основой:

1) производная основа семантически мотивируется производящей, то есть по смыслу сложнее, чем производящая, и по форме является более сложной, чем производящая;

2) производная основа по форме сложнее, чем производящая, а по смыслу они имеют равную сложность, различаясь лишь принадлежностью к разным частям речи;

3) производная и производящая основы имеют равную сложность по форме, а по смыслу производная мотивируется производящей и сложнее последней;

4) отношения производности могут быть у однокоренных основ, которые имеют равную сложность и по форме, и по смыслу;

5) если из двух однокоренных основ одна стилистически нейтральна, а другая - стилистически окрашена, производной основой является последняя. [Земская, 2011]

Еще одним важным понятием является понятие мотивации. Словообразовательная мотивация – это отношение между двумя однокоренными словами, значение одного из которых либо а) определяется через значение другого, либо б) тождественно значению другого во всех своих компонентах, кроме грамматического значения части речи. Одно из слов, связанных отношениями мотивации, является мотивирующим, а другое – мотивированным.

Отношения мотивации связывают между собой производящее и производное слово. В соответствие с типом мотивации С.Б. Козинец выделяет три класса производных слов по соотношению с производящим:

1 класс. Производные, связанные с производящими отношениями прямой мотивации. При прямой мотивации производное мотивируется прямым значением или частью значения производящего.

2 класс. Производные, связанные с производящими отношениями образной (метафорической) мотивации. К ней относятся такие случаи, при которых производное слово получает образно-метафорическое значение. Этот класс слов интересен тем, что во многих случаях прямое значение производящего слова получает метафорическую интерпретацию в слове. Все остальные семы нейтрализуются.

3 класс. Производные, связанные с производящими отношениями условной мотивации. [Козинец, 2014]

Различают несколько типов мотивации: непосредственная мотивация – это мотивационное отношение двух слов, одно из которых отличается от другого только одним формантом; опосредованная мотивация – то мотивационное отношение двух слов, одно из которых отличается от другого совокупностью формантов.

Важнейшей единицей описания словообразовательной системы является понятие словообразовательного гнезда, которое можно определить как упорядоченную совокупность однокоренных дериватов, связанных отношениями непосредственной или опосредованной мотивации с одним непроизводным (базовым) словом, которое называется вершиной данного гнезда. Элементами словообразовательного гнезда, кроме его вершины, являются также словообразовательная цепочка и словообразовательная парадигма.

Словообразовательная цепь – это ряд однокоренных слов, находящихся в отношениях последовательной производности. Словообразовательная парадигма – совокупность производных слов, связанных отношениями словообразовательной производности с одним и тем же производящим словом и находящиеся на одной ступени словопроизводства.

Производные слова восходят к одному непроизводному, возглавляющему словообразовательное гнездо. Дериваты удалены от исходного слова на одну или несколько словообразовательных ступеней. Все слова гнезда связаны корневой морфемой - носителем лексического значения.

В структуре словообразовательного гнезда тесно переплетаются синтагматические и парадигматические отношения производных слов, Любое словообразовательное гнездо одновременно может быть представлено и как совокупность словообразовательных цепочек (по горизонтали), и как совокупность словообразовательных парадигм (по вертикали). Разные парадигмы одного гнезда сводятся к исходному слову - вершине гнезда - только при помощи словообразовательных цепочек. А члены цепочек, как правило, являются вершинами гнездовых парадигм. [Ивсеева, Лузгина, 2007]

Целью словообразовательного является установление отношений словообразовательной производности между производным словом и производящим. Производное слово в своем составе содержит мотивирующую основу и словообразовательный формант.

Регулярно образуясь по одному и тому же принципу, используя одни и те же словообразовательные средства и получая слова определенных частей речи, единицы абстрактно выстраиваются в словообразовательные типы – такие схемы построения слов определенной части речи, которые абстрагированы от конкретных лексических единиц. Иными словами, к словообразовательному типу относятся слова, образованные от одной части речи с помощью одного словообразовательного форманта, имеющим одно и то же словообразовательное значение.

Выводы по главе 1

В первой главе мы ознакомились с различными взглядами лингвистов на словообразование как компонент языковой структуры. Словообразование – динамический процесс, результатом которого являются слова-дериваты. Словообразование можно рассматривать с точки зрения диахронии и синхронии, но данная работа опирается на синхроническое описание, поскольку при рассмотрении словообразовательных процессов в реальном времени есть возможность отследить значение того или иного слова.

Важным аспектом функционирования языка является предложенная Плунгяном аддитивная модель морфологии, которая является каркасом для всех явлений, происходящих в языке. Именно аддитивная модель морфологии позволяет проводить аналогии и находить соответствия при морфемном членении. Принципам этой модели мы следуем в данной работе.

Еще одним ключевым понятием, рассмотренным в первой главе, является понятие морфемы, которая понимается как минимальная двусторонняя значащая и при этом функциональная единица языка, соотносящая план выражения с планом содержания. Морфемы способны

вступать в комбинаторные сочетания друг с другом. Изучением таких сочетаний занимается морфотактика.

Последний аспект, рассмотренный в данной главе, касается понятия деривации, введенного в обиход Е. Куриловичем. Лингвист предлагает подразделять деривацию на лексическую и синтаксическую. В данной работе нас интересует синтаксическая деривация (транспозиция), поскольку является способом образования новых слов, выступающих в новой синтаксической функции без изменения их лексического значения.

На основе изученного теоретического материала можно сделать вывод о том, что деривация представляет собой усложнение структуры слова, в которой есть некоторый прототип (мотивирующая основа или целое слово), к которому в процессе словообразовательного анализа присоединяются некоторые аффиксальные элементы, которые могут иметь свое лексическое значение, а могут совпадать со значением корня. С точки зрения лексического значения, усложненные лексемы могут передавать то же лексическое значение на синтаксическом уровне, что и простые лексемы, или могут это значение модифицировать. В нашем случае на уровне смысла лексемы с разными синтаксическими функциями не приобретают новое значение, а несут в себе значение мотивирующего слова, что служит обоснованием для стемминга.

Глава 2. Тема-рематическая организация текста

2.1 Взгляды на коммуникативное членение текста

2.1.1 Традиционное членение высказывания на тему и рему

Процедура тематической атрибуции непосредственно связана с пониманием коммуникативной организации как одного предложения, так и всего текста в целом. Говоря о коммуникативной организации предложения, лингвисты, как правило, обращаются к проблеме актуального (коммуникативного) членения, которая связана с разделением предложения на несколько составляющих: так называемую отправную точку высказывания (тему), которая является менее важной в плане передачи информации, и новую часть высказывания (рему), информативно более содержательную и выражающую то, что сообщается об отправной точке высказывания.

Актуальное членение принято рассматривать только на уровне предложения, однако, в данной работе мы будем исходить из коммуникативной организации цельного текста, так как любой связный текст состоит из высказываний, которые в совокупности придают тексту смысловую целостность, реализуя тем самым коммуникативную (функциональную) перспективу сообщения.

Существует несколько взглядов на проблему актуального членения. Первая точка зрения связана с делением высказывания на тему и рему. Основным приверженцем данной теории является чешский лингвист Вилем Матезиус. Матезиус противопоставляет актуальное членение предложения формальному (синтаксическому). Он считает, что в отличие от формального членения, которое состоит в разложении предложения на некоторые грамматические элементы, актуальное членение «выясняет способ включения предложения в предметный контекст, на базе которого оно возникает». К основным элементам актуального членения лингвист относит основу высказывания, известную или понятную часть высказывания, и ядро высказывания, новую информацию об исходной точке высказывания. При

этом, по мнению лингвиста, тема не несет в себе новой информации, но является необходимым элементом связи предложения с контекстом [Матезиус, 1967, с 239-245].

Хотя В. Матезиус рассматривал проблему актуального членения скорее с семантической перспективы, ученый уделил внимание и синтаксической структуре предложения, связав актуальное членение с главным, по его мнению, средством его выражения на синтаксическом уровне – порядком слов. Он считал, что тема и рема в предложении сочетаются по-разному. Лингвист предложил «объективный порядок», согласно которому начальная часть предложения, как правило, принимается в качестве исходного пункта высказывания, а конец предложения – в качестве ядра высказывания. Данную последовательность ученый объясняет движением от известного к неизвестному, что упрощает восприятие информации получателем сообщения. Однако, Матезиус обращает внимание и на обратный порядок расположения компонентов высказывания, который он называет «субъективным» в связи с отсутствием при таком порядке естественного перехода от известного к неизвестному и акцентированием значимости ядра высказывания, то есть ремы. [Матезиус]

Позиция К.Г. Крушельницкой являет собой продолжение идей чешского лингвиста. Она определяет актуальное членение как процесс выделения в предложении или тексте тех компонентов сообщения, в отношении которых определяется коммуникативное значение предложения. Лингвист рассматривает актуальное членение, опираясь на смысловые веса членов предложения. Она отвергает логико-психологическую интерпретацию актуального членения, при которой логическим, психологическим, или смысловым субъектом называют член предложения, который называет предмет сообщения, а член предложения, содержащий основную информацию в сообщении, — логическим, психологическим, или смысловым предикатом. [Крушельницкая К.Г., 1956, с 55-67].

По мнению лингвиста, явление актуального членения «и по своей сущности, и по языковому выражению органически входит в синтаксис языка». Вместо исходной точки высказывания и ядра она использует термины «данное», нечто известное слушателю, и «новое», неизвестная информация. Как отмечает Крушельницкая, коммуникативная структура предложения заключается в сочетании известной и неизвестной информации и «выступает в таких функционально-стилевых видах речи, как разговор, деловая проза». При этом значения данного и нового, которые получают члены предложения в коммуникативном акте не лишают основных значений членов предложения, которые выражаются соответствующими грамматическими формами, а, наоборот, накладываются на грамматические значения и несут в себе основную коммуникативную нагрузку [там же].

Исследованиями проверено, что любой член предложения может выступать и как данное, и как новое. По словам Крушельницкой, чаще всего данным в предложении является подлежащее. Дополнение, обстоятельства времени, места, причины и цели могут выступать в качестве обоих компонентов высказывания, выражая и тему, и рему. Сказуемое и обстоятельство образа действия преимущественно выступают как новое. Что касается определения, то его смысловая нагрузка тесно связана со смысловым весом определяемого слова. Таким образом, члены предложения в каждом конкретном высказывании представляют собой единство двух значений: синтаксического значения и «коммуникативной нагрузки»— как «данного» или как «нового». [там же].

Значительное внимание актуальному членению уделил Пржемысл Адамец, обратив внимание не только на порядок слов, но и на фразовое ударение. Так, лингвист считает недостаточной мысль о том, что основа высказывания преимущественно находится в начале, а ядро в конце. Он добавляет, что данное предположение приобретает смысл лишь в том случае, если фразовое ударение стоит на последнем слове. Для лингвиста фразовое ударение является важным элементом для выражения актуального членения.

Он считает, что порядок слов и фразовое ударение находятся в отношении функциональной заменяемости и могут поочередно брать на себя главную функцию. Тем не менее, для Адамца очевидна главная роль порядка слов — выражение актуального членения. Важным для нас является то, что по словам ученого актуальное членение можно отнести к одному из верхних уровней широко понимаемого синтаксиса, и, следовательно, вопросы порядка слов относить к вопросам синтаксическим. [Адамец, 1966: 20]

Взгляд Адамца на компоненты актуального членения находит много общего с теорией В. Матезиуса. В частности, П. Адамец в предложении выделяет основу и ядро. При этом основа может состоять только из одного компонента или из нескольких. П. Адамец говорит о ступенчатой структуре предложения при наличии в нем нескольких компонентов основы. Суть состоит в присоединении к примарной основе оставшейся части предложения, которая принимается в качестве ядра, а далее – в разложении остатка предложения на новые основу и ядро. Стоит отметить, что в позиции Адамца также прослеживается идея о разном характере основы и ядра, то есть темы и ремы. Когда в основу входит только один компонент, то в ее роли выступает субстантивный компонент, то есть темой чаще всего является подлежащее, дополнение или обстоятельственная характеристика. Если темой является субстантивный компонент, она имеет выразительно тематический характер и являет собой субстанцию, о которой что-то сообщается. В случае, когда тема представлена обстоятельственной характеристикой, она принимает характер простой, по Фирбасу, ситуационной кулисы". В отношении ремы для Адамца неточным является понимание ремы (ядра) как новой, важной в коммуникативном отношении информации. В частности, важным лингвисту кажется не само ядро, а его соотнесение с основой. Ядро в равной степени с основой может состоять из одного или нескольких компонентов. Что касается простого ядра, чаще всего в его роли выступает глагол, иначе — подлежащее, дополнение или обстоятельство [там же].

Рассуждая о явлении актуального членения в целом, автор утверждает, что «актуальное членение представляет собой некоторую надстройку над синтаксической структурой и лексическим наполнением предложения, находясь с ними в постоянных взаимоотношениях». Лингвист делает вывод о существовании взаимообусловленности между лексико-синтаксической структурой предложения и его актуальным членением, разделяя отношение актуального членения к синтаксической структуре, к лексическому наполнению и к степени индивидуализованности отдельных компонентов, а также зависимость актуального членения от конкретной внешней ситуации [там же: 39].

Большой вклад в изучение средств актуализации коммуникативного центра предложения внесла И.И. Ковтунова, которая наглядно показала, что выразителями актуального членения могут быть не только интонация и порядок слов, но и, например, частицы, а также подробно разобрала роль каждой части речи в актуальном членении высказывания. Дополнительным средством выражения актуального членения, по мнению лингвиста, служат частицы. Так, отмечается, что тема, как правило, выделяется частицами ***а*** и ***же***. Частицами, выражающими рему, выступают частицы ***только***, ***лишь*** и некоторые другие. [Ковтунова]

Ковтунова обращает внимание на синтагматику высказывания и выражает мысль о том, что высказывание существует в синтагматических и парадигматических отношениях с рядом других высказываний, и приходит к выводу о том, что актуальное членение одного предложения зависит от предшествующего или ряда предшествующих ему предложений. Она утверждает, что тема и рема определяются конкретным коммуникативным заданием данного высказывания, которое зависит от контекста. Таким образом, в зависимости от предыдущего контекста темой или ремой могут стать разные компоненты синтаксической структуры предложения [там же, стр. 18-19]. Что касается функций темы и ремы, здесь лингвист согласна с многими лингвистами в том, что функцию темы преимущественно берут на

себя слова с предметным значением, а функцию ремы – слова со значением признака. Иначе говоря, для слов со значением предметности функция темы обусловлена самим их значением, а для слов со значением признака функция темы может быть обусловлена лишь контекстом. То же правило работает и в сторону ремы: функция ремы для слов со значением признака обусловлена их значением, а для слов со значением предметности – контекстом [стр. 92-93].

Следует также отметить ценную монографию И. П. Распопова «Актуальное членение предложения». По мнению Распопова, актуальное членение напрямую зависит от определенного коммуникативного задания, так как роль речевой ситуации при определении актуального членения можно объяснить тем, что, как и при прямом взаимодействии в диалоге, так и в монологической речи, говорящий учитывает степень осведомленности собеседника в том, что является материалом для данного сообщения. При этом, важно понимать, что актуальное членение – не только фактор семантики, но и фактор структуры предложения [Распопов, 2009: 27-34].

И. П. Распопов рассматривает актуальное членение как структурный фактор, тесно связанный с категорией предикативности, чья роль состоит в отнесении содержания предложения к действительности. Как отмечает автор, именно указание на объект действительности выражается одним из компонентов актуального членения – исходной информацией или основой (темой), а содержание более информативного компонента, некоторым образом, предицитируется. Иными словами, лингвист избегает использования терминов «тема» и «рема», но вводит понятия «основы высказывания» и «предицируемой части», причем предицирование не может быть отделимо от актуального членения, и наоборот [там же: 37-43]. Важным моментом в монографии является указание на связь актуального членения с грамматическим. Распопов утверждает, что оформление актуального членения зависит от его взаимодействия с членением грамматическим. Более того, в актуальном членении используются средства языка, не обязательно выражающие грамматическое членение. Взаимодействие этих средств

обеспечивает выражение в структуре предложения как актуального, так и грамматического членения и позволяет перераспределить функциональную нагрузку [там же: 49-50].

Лингвист выделил шесть важнейших типов соотношения между основой высказывания и предицируемой частью:

1) Предложение не содержит словесно выраженной основы и весь его лексико-грамматический состав представлен предицируемой частью. Такой тип соотношения между частями актуального членения характерен для однословных односоставных предложений. Основой такого предицирования является действительность в целом. Иными словами, такие предложения содержат нулевую основу.

2) При данном типе основа может быть выражена словами, которые ограничивают действительность и определяют условия, при которых имеет место какой-либо факт, в то время как предицируемая часть раскрывают содержание данного факта при данных условиях. Содержание основы вводит в определенную речевую ситуацию и объясняет ее, таким образом предвещая содержание предицируемой части.

3) Здесь основа высказывания содержит указание на какое-то лицо или предмет, на котором сосредоточено внимание говорящего и слушающего. Предицируемая часть, в свою очередь, содержит развернутое высказывание о лице или предмете. Такой тип соотношения компонентов актуального членения наиболее характерен для двусоставных предложений. В состав основы, по мнению Распопова, входит как грамматическое подлежащее, так и грамматическое дополнение, а состав предицируемой части часто формирует группа сказуемого.

4) Для данного типа характерно назначение в качестве основы какого-либо лица или предмета с указанием конкретных условий, при которых данное лицо или предмет представляют реальную действительность. Предицируемая часть содержит в себе развернутое высказывание об этом лице или предмете, которое уточняется и ограничивается указанными условиями

или обстановкой. В качестве основы выступают обстоятельственный второстепенный член и подлежащее (дополнение), остальные члены предложения входят в состав предицируемой части. Как отмечает Распопов, в данном типе основа и предицируемая часть совмещают себе функции, которые они выполняют раздельно при втором и третьем типах.

5) Основа содержит указание на какой-то факт, имеющий место в действительности, а в предицируемой части этот факт в том или ином отношении уточняется. При этом в состав основы обязательно входит грамматическое сказуемое или один главный член, к которым присоединяются и другие члены предложения. В состав предицируемой части может входить любой член предложения, не включенный в основу высказывания.

6) Для последнего типа соотношения основы и предицируемой части имеет место содержание в составе основы названия лиц и предметов, находящихся в определенных отношениях друг с другом. В предицируемой части происходит раскрытие реального содержания данного отношения. Распопов включает сюда двусоставные предложения субъектно-обьектного строя, в основу представляет грамматическое подлежащее и дополнение, называющие субъект и объект действия, а предицируемая часть включает простое глагольное сказуемое с иногда примыкающим к нему обстоятельственным второстепенным членом меры и степени или качества действия) [Распопов: 58-73].

Рассмотрев шесть типов соотношения элементов актуального членения, автор приходит к мысли, что состав данных компонентов необходимо рассматривать не только в отношении включения в него грамматических членов, но и в плане коммуникативном, проводя дополнительное членение этих компонентов на коммуникативно значимые элементы.

По мнению Панфилова, актуальное членение предложения отражает субъектно-предикатную структуру выражаемой мысли. Из этого утверждения вытекает справедливый вопрос: если актуальное членение носит формально-грамматический характер, стоит ли относить его к синтаксическому

членению, как, например, считает Крушельницкая, или все-таки вынести его за пределы синтаксиса, как предполагал Матезиус, противопоставляя актуальное членение грамматическому. Панфилов предлагает вынести коммуникативное членение на «надсинтаксический» уровень, объясняя свою позицию тем, что коммуникативную нагрузку, проявляющуюся в членении предложения на тему и рему, можно привязать к любому члену предложения без изменения его на синтаксическом уровне, и, следовательно, необходим особый, «логико-грамматический» уровень предложения, обусловленный его актуальным членением, то есть, в терминах Панфилова, членением предложения на логико-грамматические субъект и предикат. Исходя из представлений о надсинтаксическом уровне предложения, Панфилов делает вывод о том, что предложение само по себе также является принадлежностью логико-грамматического уровня языка и не может получить определения на синтаксическом уровне [Панфилов, 1963].

2.1.2 Нетривиальные взгляды на функциональную перспективу сообщения

Среди различных трактовок проблемы актуального членения интерес вызывает позиция Я. Фирбаса, который предложил понятие коммуникативного динамизма (КД). Согласно теории КД, элементы высказывания, линейно следующие друг за другом, постепенно передают информацию. Само понятие коммуникативного динамизма можно определить как вклад, которой тот или иной элемент вносит в функциональную перспективу предложения, тем самым развивая процесс коммуникации. Так, разные элементы в составе предложения передают разные степени коммуникативного динамизма. Лингвист различает тему, которая передает самую низкую степень коммуникативного динамизма, рема, наоборот, передает самую высокую степень КД, а элементы, находящиеся между темой и ремой лингвист переходными [Фирбас, 1972]. Таким образом, шкала КД Фирбаса выглядит так: собственно тема, остаток темы, собственно переход, остаток перехода, рема и собственно рема, то есть степень КД возрастает к

концу высказывания. По утверждению Фирбаса, тема-рематическая организация текста, а точнее факторы, формирующие ее, такие как порядок слов, фразовое ударение, интонация, семантика и контекст, оказывают непосредственное влияние на степень коммуникативного динамизма. Из этого следует, что компоненты актуального членения выделяются в соответствии с информативной нагрузкой, которую они вносят в коммуникацию.

Концепция Ф. Данеша о тематической прогрессии также играет немаловажную роль в становлении теории о коммуникативном членении высказывания. Важным является место актуального членения в теории синтаксиса. Ученый подразделяет синтаксис на три подуровня: 1) уровень грамматической структуры предложения, 2) уровень семантической структуры предложения, 3) уровень организации высказывания. Актуальное членение, по мнению лингвиста, относится к третьему, то есть к уровню организации высказывания. Сущность самого явления заключается в том, что основная информация заключена в тематическом компоненте высказывания. При этом каждое последующее предложение опирается на предыдущее, тем самым «продвигая» высказывание от предшествующего к данному и развивая тему всего текста. Таким образом, текст выстраивается в виде определенной иерархии тем высказываний и их связи между собой, то есть представляет собой так называемую «тематическую прогрессию», которую можно подразделить на несколько типов. Среди них:

1. Линейная прогрессия – самый распространенный тип прогрессии. Данный тип можно описать как процесс постепенного развертывания информации, при котором рема предыдущего предложения становится темой последующего, то есть происходит развертывание текста от темы к реме (от данного к новому).

2. Прогрессия со сквозной темой (константная). Для данного типа характерно наличие одной темы, которая сохраняется в каждом последующем высказывании, связывая их в цельный текст.

3. Прогрессия с производными темами. Здесь «гипертема» текста, то есть основная тема, либо названа говорящим, либо текст содержит фрагменты, на нее указывающие. Частные темы являются производными от гипертемы, поэтому в тексте отсутствует последовательная тематизация.

4. Прогрессия с расщепленной темой. Данный тип характеризуется наличием сложной ремы, которая при тематизации расщепляется на отдельные тематические прогрессии, которые могут представлять собой описанные выше типы и последовательно присутствовать в тексте.

5. Прогрессия с тематическим прыжком. Главный признак данного типа – наличие разрыва в тема-рематической цепочке, который, как правило, встречается в текстах с последовательной тематизацией. Однако, не составляет никакого труда восстановить разрыв из контекста [Данеш, 1974: 117].

Несмотря на то, что любой из типов тематический прогрессий не может быть однозначно применен к тому или иному тексту, тематические прогрессии Данеша определяют принципы построения коммуникативной организации текста.

Еще одной нетривиальной трактовкой явления актуального членения является «принцип напряжения» (das Prinzip der Spannung) К. Бооста. В основу его теории входит разделение предложения на две противопоставленных части: содержащую вопрос тему, с которой и начинается напряжение, и содержащую ответ рему, заключающую в себя само высказывание, в котором напряжение разрешается. Важно отметить, что, хотя эти части и противопоставлены, они составляют единое целое. Как и В. Матезиус, Боост рассматривает членение на тему и рему (die Thema-Rhema-GIiedertmg) как аспект семантики, противопоставляя его грамматическому членению. Тема может быть образована не только подлежащим, но и другими членами предложения, а рема, в свою очередь, включает в себя не только сказуемое. В целом, позиция Бооста находит много общего с концепцией Матезиуса,

поскольку в обеих работах актуальное членение рассматривается с опорой на коммуникативную функцию высказывания [Боост, 1955].

Проблеме актуального членения посвящено множество работ. Мы рассмотрели интересующие нас концепции, и теперь в рамках данной работы считаем необходимым более углубленно изучить один из компонентов актуального членения – тему, поскольку данная научная работа связана с нахождением, прежде всего, темы текста.

Выводы по главе 2

В данной главе мы рассмотрели основные взгляды на проблему коммуникативного членения предложения. Существует несколько основополагающих позиций, среди которых деление высказывания на тему и рему, на «данное» и «новое», «исходный пункт сообщения» и ядро. Согласно данной концепции, выразителями актуального членения могут быть интонация, порядок слов, фразовое ударение, частицы.

К более новаторксим подходов к функциональной перспективе сообщения можно «принцип напряжения» К. Бооста, идеей о тематических прогрессиях Ф. Данеша, теория «куммуникативного динамизама» Фирбаса.

Относительно проблемы тематической атрибуции нас интересует тема, представленная в тексте. Однако, в тексте может присутствовать несколько тем, то есть текст имеет многоуровневую структуру, которая состоит из более мелких текстов. Между этими темами устанавливаются когезивные отношения, которые реализуются с помощью анафорической замены и синонимии. Тем не менее, в тексте, содержащем только одну тему, регулярно встречаются повторения этой самой темы в роли синтаксических дериватов или синонимов. Именно на основе такой лексической повторяемости строится наша практическая работа.

Однако, результативность метода, основанного исключительно на лексической повторяемости, не будет гарантированно высокой в силу сущности тематической прогрессии. Дело в том, что достаточно регулярно

транспозиция (синтаксическая деривация) используется для перехода рематических элементов в тематические.

Глава 3 Подходы к автоматическому определению темы

3.1 Статистические методы определения темы

В рамках данного исследования тематический компонент выходит на первый план, поэтому далее будет рассматриваться исключительно тема, понимаемая как номинативно выраженное содержательное ядро целого текста. Однако, наличие корпуса текстов большого объема значительно затрудняет задачу присвоения темы тому или иному тексту, поэтому в настоящее время принято пользоваться автоматическими методами определения тематики текстов, которые будут описаны ниже.

Прежде всего, одним из статистических и автоматизированных методов «тематического» анализа является тематическое моделирование. Тематическая модель (topic model) — модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. В основном, используется вероятностное тематическое моделирование, представляющее собой набор алгоритмов, направленных на анализ слов в большой коллекции документов и извлечение тем в этих документах, а также связей между темами.

В целом, работу тематической модели можно охарактеризовать как задачу «мягкой» кластеризации слов и документов по одному и тому же множеству кластеров – тем на основании их семантической близости. Мягкая кластеризация в данном контексте означает, что что один и тот же документ можно отнести к нескольким темам, поэтому кластеры могут быть нечетко обозначены. При этом каждая тема представляется как дискретное распределение на множестве слов, а документ – как вектор в пространстве скрытых тем [Воронцов, 2013].

Перейдем к формальной постановке задачи. Предположим, что $\boldsymbol{D}$ — коллекция текстовых документов, а $\boldsymbol{W}$ — множество терминов, употребляемых в этих документах, то есть словарь. Каждый документ $\boldsymbol{d \in D}$ состоит из последовательности $\boldsymbol{n_d}$ терминов $\boldsymbol{(w1, \ldots, w_{nd})}$, лежащих в словаре $\boldsymbol{W}$, причем слово (термин) может встретиться в документе сколько угодно раз.

Предположим также, что множество тем T конечно, и любое употребление слова w связано с неизвестной темой $t \in T$. Тогда коллекцию документов мы будем рассматривать как множество случайным образом выбранных из дискретного распределения $p(d, w, t)$ троек (d, w, t). Документы и термины, следовательно, представляют собой наблюдаемые переменные, а тема —скрытую переменную. Тема t, в свою очередь, представлена как вероятностное распределение $p(w|t)$ на множестве термов, то есть, иными словами, вероятность отнесения термина w к теме t. Данное утверждение принято называть ***гипотезой независимости*** [Воронцов, 2014].

Кроме того, существует так называемая ***гипотеза «мешка слов» (bag of words)***, в основе которой лежит предположение о несущественности порядка слов в документах для определения темы текста. Следствием из этой гипотезы является ***гипотеза «мешка документов»***, согласно которой порядок документов в коллекции также может быть произвольным.

Отсюда мы приходим к противоречию в понимании темы в лингвистическом и статистическом плане. Из вышеприведенного анализа подходов к проблеме актуального членения становится ясной первоочередная роль порядка слов в членении высказывания на тему и рему. Напротив, при статистическом анализе текста порядок слов теряет свою значимость, поэтому даже при произвольной перестановке терминов в документе тема текста будет определена.

Исходя из определений полной и условной вероятностей и гипотезе условной независимости получаем следующую функцию правдоподобия для вероятностной модели порождения данных:

$$p(w|d) = \sum_{t \in T} p(t|d) \times p(w|t),$$

где $p(t|d)$ – порождение документа некоторым распределение на темах, а $p(w|t)$ – порождение темы некоторым распределением тем.

Таким образом, задача построения тематической коллекции документов сводится к нахождению для данной коллекции документов D множества всех

использующихся в ней тем T, а также к восстановлению для каждого документа из коллекции по распределению слов по документам распределений тем в документе $\boldsymbol{p(t|d)}$ и слов по темам $\boldsymbol{p(w|t)}$, то есть к выявлению скрытых тем и оценке распределений.

Тематическим моделированием, следовательно, называется процедура восстановления распределений вероятностей тем в тексте, который рассматривается как случайная независимая выборка слов (мешок слов), которая была порождена некоторым набором тем.

Что касается определения темы, то в терминах вероятностных моделей данное понятие можно трактовать как набор терминов, совместно встречающихся в документе (коллекции документов), или как вероятностное распределение $\boldsymbol{p(w|t)}$ на терминах.

Основные алгоритмы в тематическом моделировании

К одному из наиболее популярных алгоритмов тематического моделирования можно отнести метод ***латентного размещения Дирихле*** (Latent Dirichlet Allocation, LDA). Алгоритм принадлежит семейству порождающих вероятностных моделей, где каждая тема представлена как вероятность появления данного слова в заданном наборе. Данная модель в явном виде задает распределение слов по темам и априорное распределение по документам. Одним из преимуществ данного метода является способность к обнаружению неявных связей между словами с учетом полисемии. Более того, еще одна особенность данной модели заключается в нечеткости тем, то есть в возможности попадания одних и тех же слов в разные темы. При этом, выделенные темы будут считаться независимыми [Бенгфорт, 2019].

Латентно-семантический анализ (Latent Semantic Analysis, LSA) — еще один популярный алгоритм на основе векторов, который впервые был представлен в 1990 году Дирвестером (Deerwester) и его коллегами [S. Deerwester, 1990]. Метод также известен как латентно-семантическое индексирование (Latent Semantic Indexing) и используется для анализа

отношений между документами и терминами в коллекции, а также для поиска латентных (скрытых) ассоциативно-семантических связей между словами и сочетаниями слов с помощью сокращения разреженного вероятностного пространства (матрицы) «термы на документы». Разреженное пространство представлено матрицей, строки которой занимают слова (термы), а столбцы – документы. Элементы матрицы содержат, как правило, веса, которые учитывают частоту встречаемости каждого терма в каждом документе, которая нормализуется с помощью меры TF–IDF. Далее к матрице применяется метод сингулярного разложения матриц (Singular Value Decomposition, SVD), в результате чего получают матрицы, представляющие собой матрицы «термы на темы», важность тем и «темы на документы» [Воронцов, 2013].

Не вдаваясь в подробности, можно сказать, что LSA переносит документы коллекции и термины, содержащиеся в них, в латентное пространство свойств, в котором размерности в идеале соответствуют высокоуровневым компонентам. Следовательно, каждый документ представляется как взвешенная комбинация таких компонентов, а каждый термин в то же время может быть в той или иной степени связан с иными понятиями.

Основная идея латентно-семантического анализа состоит в следующем: если в исходном вероятностном пространстве, которое состоит из векторов слов, между двумя любыми векторными представлениями слов может не наблюдаться никакой зависимости, то после некоторых алгебраических преобразований векторного пространства такая зависимость может появиться, а величина ее будет определять силу ассоциативно-семантической связи между этими двумя словами.

Вероятностный латентно-семантический анализ **(pLSA)** — метод, использующий для анализа представления вероятности вхождения терм-документ в виде многомерного распределения. В данном методе выделяются наборы совместно встречающихся терминов, объединённых общей тематикой.

В данном алгоритме каждый документ представляется в виде числового вектора, числа которого отображают одну из тем. При увеличении числа документов размер вектора также растет. Отсюда вытекает один из минусов метода – количество параметров модели линейно зависит от размера обучающего корпуса. Также недостатком PLSA является склонность к переобучению из-за большого числа параметров, которые находятся в прямой зависимости от числа документов; это затрудняет работу модели на больших корпусах данных [Кольцов, Кольцова, Митрофанова, Шиморина 2014].

На сегодняшний момент создано и успешно применяется большое количество различных тематических моделей. Их значительная часть основана на двух базовых алгоритмах – LDA и pLSA. Большинство моделей построены на использовании модели мешка слов, которая, тем не менее, не позволяет учитывать связь слов в контексте, что, несомненно, является недостатком.

3.2 Термин и термин-кандидат в качестве маркеров темы

Алгоритмы тематического моделирования определяют специфичность текста, выявляя так называемые «специальные» слова и словосочетания, отражающие конкретную предметную область. Когда речь идет о тексте или о корпусе текстов, представляющем определенную тему, о тематически маркированных словах принято говорить не просто как о ключевых словах и кандидатах в них, а о терминах и терминоэлементах.

В науке существует множество подходов к определению термина. В Большом энциклопедическом термин трактуется как слово или словосочетание, обозначающее понятие специальной области знания или деятельности. Иными словами, о термине можно говорить как о единице естественного языка, в результате некоторой коллективной договорённости обладающей специальным терминологическим значением, отражающим основные черты области, к которой данный термин принадлежит.

Б.Н. Головин считает, что термин – это отдельное слово или образованное на базе имени существительного подчинительное

словосочетание, которое обозначает некое профессиональное понятие и служит для удовлетворения специфических нужд общения в определенной сфере (научной, технической, производственной, управленческой) [Головин, 1987].

В основном, взгляды на определение данного слова похожи, но все же для каждой области данное понятие приобретает свой оттенок значения. Такое многообразие трактовок данного понятия можно объяснить тем, что термин – объект нескольких наук одновременно. Включение термина в разные терминологические системы создает межнаучную терминологическую омонимию. А. В. Суперанская объясняет многообразие определений понятия «термин» тем, что «…у представителей разных дисциплин оно связывается со своими особыми понятиями и представлениями, имеет неравный объём содержания и определятся по-своему». [Суперанская, 2012]

Важным для данного исследования является взгляд Г. О. Винокура, который считает, что термины — это слова в особой функции, и что в роли термина может выступать любое слово, даже самое тривиальное [Винокур, 1939]. Именно этого видения понятия «термин» мы придерживаемся в настоящем исследовании. По нашему мнению, в качестве термина может выступать любое слово или словосочетание, отражающее специфику данного текста (корпуса).

В корпусной лингвистике принято использовать понятие «термин-кандидат», поскольку термин не может существовать вне терминологической системы. Решение относительно того, является ли данное слово или словосочетание термином в конкретном тексте зависит от того, отражает ли это слово конкретное научное понятие, которое уже существует в логико-понятийной системе. Помимо этого, оценка степени «терминологичности» термина является довольно субъективной и требует согласования со специалистом конкретной научной области.

Чтобы правильно выделить в тексте термины и терминоэлементы, необходимо понимать, какие признаки отличают термин от обычных слов. В

[Захаров, Азарова, 2019] о свойствах термина написано, что те задаются связью самого термина с некоторым понятием специальной области знаний (терминологичность) и лексическими единицами языка, которые используются для его оформления. Лингвисты считают, что несмотря на фиксацию терминов в терминологических словарях, существуют проблемы в отношении полноты устойчивости и непротиворечивости терминосистемы, так как в связи со свободным характером лексикализации понятий появляются новые синонимичные термины, что нарушает «стабильность» понятия.

С.В. Гриневым-Гриневичем был разработан ряд свойств, характеризующих термин:

1) специфичность употребления – отнесенность термина к специальной области употребления обусловлена тем, что он используется для называния понятий;

2) содержательная точность – четкость, ограниченность значения термина;

3) дефинированность – специальное понятие имеет точные границы, устанавливаемые с помощью научного определения – дефиниции, которая одновременно является и определением значения термина;

4) независимость от контекста;

5) однозначность;

6) стилистическая нейтральность – термин не должен порождать какихлибо добавочных ассоциаций;

7) конвенциональность – целенаправленный характер появления термина, когда необходимость удобных названий для новых понятий требует создания или тщательного выбора из существующих лексических средств выражения понятий;

8) номинативный характер – в качестве терминов по большей части используются имена существительные [Гринев-Гриневич, 2008].

Таким образом, заключим, что термин – многофункциональная единица, ограниченная в некоторых аспектах употребления. Тем не менее, эти

ограничения никак ни умаляют важность употребления терминов в научных текстах и необходимость их дальнейшего изучения.

3.3 Подходы к выделению терминов и терминоэлементов

Среди подходов к выделению терминов и сочетаний терминов можно выделить: статистический метод, который основан на подсчёте частоты совместной встречаемости компонентов словосочетания и мерах ассоциации (учитывают как частоту совместной встречаемости); лингвистический метод, в основе которого лежит использование определённых морфосинтаксических шаблонов и сочетание статистического и лингвистического методов в виде смешанного подхода.

В [Захаров, Азарова и др., 2019] предлагаются следующие параметры автоматического выявления прототипов терминов:

1) статистические параметры прототипов терминов - связаны с тем, что они выражают важные аспекты в рамках специального содержания текстов предметной области, поэтому имеют высокую частоту встречаемости в них, так что можно ввести: а) эмпирически подобранное пороговое значение частотности, которое может быть использовано для «отсечения» наиболее вероятных прототипов терминов; б) коэффициенты превышения частоты встречаемости, среди которых наиболее известны MI-score и T-score. Таким образом могут быть выделены и подтверждены не только ядерные, но и периферийные прототипы терминов; в) специальные стоп-списки строевых (служебных) слов.

2) статистико-комбинаторные параметры прототипов терминов при выявлении неоднословных единиц – их необходимо учитывать при выявлении неоднословных единиц. Обычно большая часть терминов – это сочетания слов, так называемые многокомпонентные термины.

3) синтаксические параметры направлены на выявление синтаксических конструкций, в которых термины встречаются в текстах.

4) морфологические параметры прототипов терминов, характеризующие главные компоненты описанных выше синтаксических

конструкций. Они характеризуют главные компоненты описанных выше синтаксических конструкций.

Статистический метод основан на использовании мер ассоциации, к которым относятся mutual information (MI), log-likelihood, chisquare ($\chi 2$), t-score, LL и многие другие. Меры ассоциации, как правило, используются для выделения статистически устойчивых и терминологических словосочетаний.

Мера MI (или коэффициента взаимной информации) позволяет найти в корпусе редкие словосочетания. Таким образом, вес каждой отдельной коллокации тем больше, чем реже она встречается. [Хохлова, 2017] Основной задача данной меры ассоциации – присвоить большее значение сочетаниям с редкими словами.

Мера t-score является скорректированным ранжированием словосочетаний по частоте встречаемости. В отличие от MI, она не завышает значение низкочастотных словосочетаний терминов. Полезна при выявлении устойчивых конструкций. [Крутченко, 2016]

Для выделения терминов является мера log-likelihood (логарифмическая функция правдоподобия) является наиболее эффективной, так как успешно справляется как с высокочастотными, так и с низкочастотными терминосочетаниями.

Еще одним статистическим показателем является частота совместной встречаемости, которая в некоторых случаях работает даже лучше, чем меры ассоциации. Минус данного показателя состоит в том, что он малоэффективен при выделении редких терминов. Иногда даже низкочастотные терминологические сочетания могут быть ключевыми в отражении тематической области документа или корпуса текстов.

Статистический подход для выделения терминологических сочетаний может быть реализован по-разному. Первый вариант заключается в нахождении n-словных сочетаний (n-грамм) по заданным частотным характеристикам. Это могут быть значения абсолютных или относительных частот для данных сочетаний слов в корпусе или значение некоторой

статистической меры, согласно которой данная конструкция была найдена и выдана среди результатов. Далее возможно использование порога отсечения по заданному значению. [Захаров, Азарова и др., 2019]

В основе *лингвистического метода* лежит предположение о том, что терминологические словосочетания строятся по более или менее закреплённым морфосинтаксическим шаблонам – образцам построения терминологических слов и словосочетаний [Ефремова, 2010]. Нельзя не отметить важность определения грамматических характеристик компонентов словосочетаний и установления между ними синтаксических отношений.

Лингвистический подход для выделения терминологических сочетаний заключается в предварительном описании моделей построения терминов. Далее корпус анализируется с целью нахождения словосочетаний, соответствующих этим моделям. Фактически этот подход является комбинированным, так как объединяет и лингвистический, и статистический методы [Там же].

Смешанный (гибридный) подход заключается в совместном использовании статистического и лингвистического методов. То есть в одном случае термины и терминосочетания изанчально выделяются на основании их языковых характеристик, и только потом происходит оценка их терминологического характера с помощью мер ассоциации, а в другом случае только после отбора статистически значимых словосочетаний производится морфологический и синтаксический анализы. Среди гибридных подходов наиболее всего известны алгоритмы KEA, RAKE, TextRank и ряд других.

Можно сказать, что при гибридном подходе осуществляется трехступенчатый анализ языковых данных. На первом этапе отбираются данные по заданным лингвистическим моделям. На втором этапе происходит статистическая оценка полученных языковых выражений. На третьем формируется список терминов (ключевых слов) для исследуемого тематического корпуса по результатам сравнения с эталонным корпусом [Захаров, Азарова, 2019].

В эксперименте по автоматическому выявлению терминологических сочетаний авторы применяли статистические методы, используя сравнительный корпус. Эксперимент проводился на основе системы Sketch Engine[1], в которой возможно получить частоту слов и словосочетаний.

По мнению лингвистов, существует несколько способов реализовать статистический подход. Во-первых, задав частотные характеристики, возможно выявить n-граммы, то есть сочетания слов. Во-вторых, можно предварительно описать модели построения терминов, а далее искать словосочетания на основе описания модели. В своем эксперименте ученые используют метод выявления устойчивых сочетаний с использованием грамматики лексико-синтаксических шаблонов для терминологических сочетаний для русского языка. [Там же]

Лингвисты в своем эксперименте используют еще один способ извлечения терминов – программу Sketch Engine, которая представляет собой программное обеспечение для управления корпусами и анализом текста, разработанное компанией Lexical Computing Limited.

Sketch Engine сравнивает кандидатов в термины и терминологические сочетания по относительной частоте с такими же единицами фонового корпуса и выдает первые N единиц, отсортированные по убыванию «коэффициента терминологичности». Далее применяется гибридный подход, при котором сравниваются показатели мер ассоциации для слов или их сочетаний в тематическом корпусе с соответствующими значениями в эталонном корпусе. Гибридный подход в системе Sketch Engine реализован в двух режимах: выявление ключевых слов для данного тематического корпуса, который заключается в сравнении частот и мер ассоциации для слов и словосочетаний в тематическом и эталонном корпусах, и выявление терминов, при котором выбор терминов (кандидатов в термины) осуществляется на основе так называемой «терминологических правил», в которых описаны

[1] https://www.sketchengine.eu/

типичные модели, по которым строятся многословные термины. Затем для всех сочетаний каждой из данных моделей аналогичным образом подсчитывается значение меры в тематическом и эталонном корпусах. [Там же]

Проведя исследование, основанное на нескольких подходах, лингвисты выяснили, что более эффективным является гибридный подход, так как бóльшая часть терминов, найденных при использовании статистического подхода, была включена в списки результатов гибридного подхода. Также гибридный подход выдает меньше «шума» за счет заранее описанных моделей.

Выводы по главе 3

В третьей главе мы рассмотрели основные алгоритмы тематического моделирования. Алгоритм LDA использовался в данной работе для наблюдения за распределением тем по корпусу.

Обосновав использование понятия «термин-кандидат» и его использование в рамках тематической атрибуции тематической направленностью и некоторой предметной спецификой собранного нами корпуса, мы рассмотрели возможные подходы к автоматическому извлечению кандидатов в термины.

В нашем эксперименте мы пользовались несколькими методами, включая Sketch Engine и Rutermextract. Стоит отметить, что эффективность работы Skectch Engine довольно мала, так как результаты регулярно не соответствуют правде.

Что касается классификации самих подходов, то в данном исследовании мы используем гибридный подход, то есть подход с применением как лингвистического, так и статистического способов выделения терминов.

Глава 4. Процедура тематической атрибуции с использованием деривационного анализа

На первом этапе работы необходимо было собрать небольшой корпус текстов по музыкальной тематике. Решение о малом объеме связано с тем, что корпус меньшего объема легче анализировать и, тем самым, корректировать ошибки в работе будущего алгоритма. Материалом для проведения исследования стали десять текстов, представленных на сайте yugZone[2]. Объем итогового корпуса составил чуть больше двух тысяч словоформ.

Далее корпус был токенизирован и очищен от знаков препинания. Стоп-слова также были удалены. Список стоп-слов нам любезно предоставила Митрофанова Ольга Александровна. Морфологическая разметка была проведена с помощью морфоанализатора Pymorphy2[3]. Каждому токену были присвоены соответствующие грамматические характеристики и приписаны леммы.

К сожалению, Pymorphy2 не всегда правильно присваивает граммемы, поэтому ввиду малого объема корпуса все ошибки были исправлены вручную. В частности, большинству причастий морфоанализатор ставил в соответствие тег ADJF, поэтому для упрощения и прилагательные, и причастия были сведены к данной граммеме.

4.1. Выявление тематически маркированных слов

4.1.1 Статистический способ тематического моделирования

Многие лингвисты сходятся во мнении, что основную смысловую нагрузку текста несут в себе ключевые слова (КС). Они составляют семантическую доминанту текста, отражают его тему, то есть служат тематическими маркерами, или предметную область. Тем не менее, в зависимости от стиля текста и его структуры ключевыми могут быть как наиболее частотные слова, так и редкие слова, встречающиеся в тексте всего

[2] https://www.yugzone.ru/brainmusic.htm

[3] https://pymorphy2.readthedocs.io/en/stable/

несколько раз. Поэтому считаем необходимым учитывать как низкочастотные, так и высокочастотные КС.

В нашем понимании тематически маркированными словами являются слова с наибольшим коэффициентом «тематичности». Для подсчета данного показателя мы обратились к частотному словарю русской лексики О.Н.Ляшевской и С.А.Шарова[4] для сравнения значений частот слов нашего корпуса и частот в словаре. Посчитав относительную частоту токена по исходному корпусу и получив для каждого токена значение относительной частоты в ipm по фоновому корпусу (Ляшевской и Шарова), мы разделили одно значение на другое, таким образом, получив коэффициент «thematicality», который, во-первых, учитывает в том числе и низкочастотные слова, а, во-вторых, позволяет выделить слова, действительно отражающие тему текста. Иначе говоря, данный показатель отражает специфичную для данного корпуса лексику.

Применив описанный выше метод сначала ко всему корпусу, а затем и к каждому тексту по отдельности мы получили следующщю картину:

Таблица 1 *- коэффиент "тематичности" для всего корпуса*

Word	corpus_freq	vocab_freq(ipm)	thematicality
испытуемый	0,003842459	0,5	76,84918348
живой	0,005283381	1,8	29,35211869
гаджет	0,000960615	0,4	24,01536984
томография	0,000960615	0,4	24,01536984
озвучание	0,00192123	0,9	21,34699541
звукоизоляция	0,001440922	0,8	18,01152738
прослушивание	0,007684918	4,3	17,87190313
веселие	0,001440922	0,9	16,01024656
сенсор	0,001440922	0,9	16,01024656

В столбце «Word» представлены леммы, в столбцах «corpus_freq» и «vocab_freq(ipm)» относительные частоты по данному корпусу и по словарю Ляшевской и Шарова соответственно, а в столбце «thematicality» - значения коэффициента тематичности по убыванию.

[4] http://dict.ruslang.ru/freq.php

Приведем также таблицу для текста 1.

Таблица 2 *- коэффициент "тематичности" для текста 1*

Word	corpus_freq	vocab_freq(ipm)	thematicality
трек	0,035830619	4,2	85,31099736
испытуемый	0,003257329	0,5	65,1465798
засыпание	0,003257329	0,6	54,2888165
метроном	0,003257329	0,6	54,2888165
расслаблять	0,013029316	2,5	52,11726384
стрессовый	0,009771987	2,4	40,71661238
симпатический	0,003257329	0,9	36,19254434
артериальный	0,009771987	2,8	34,89995347
усыплять	0,003257329	1,2	27,14440825
головоломка	0,003257329	1,6	20,35830619
минимум	0,003257329	1,7	19,16075877
замедляться	0,003257329	1,9	17,14383679
композиция	0,045602606	26,8	17,01589771
бессонница	0,013029316	7,9	16,49280501
прослушивание	0,006514658	4,3	15,1503674

Далее мы обратились к еще одному простому методу извлечения КС – TF-IDF[5] (англ. TF — term frequency, IDF — inverse document frequency), который сравнивает частоту встречаемости токена в конкретном документе с частотой встречаемости токена в целом.

Формула данной метрики состоит из двух составляющих:

$$\mathrm{tf}(t,d) = \frac{ni}{\sum_k n_k} \text{ и } \mathrm{idf}(t, D) = \frac{\log |D|}{|di \supset ti|},$$

где показатель tf выражается отношением количества вхождений слова в документе к общему количеству слов в документе. При этом n_i – это число вхождений слова t в документ, а выражение в знаменателе — общее число слов в данном документе. Второй показатель idf или обратная частота документа измеряет, насколько часто встречается данное слово во всех документах. Здесь $|D|$ – число документов в корпусе, $|(di \supset ti)|$ – количество документов коллекции D, содержащих слово t_i (при $n_i \neq 0$).

В целом, формула TF-IDF выглядит как произведение этих двух показателей.

[5] https://ru.wikipedia.org/wiki/TF-IDF

$$\mathbf{TF\text{-}IDF(t, d, D) = tf(td) \times idf(t, D)}$$

Данная метрика не очень хорошо работает для небольших корпусов, как наш, но позволяет выявлять наиболее «тематичные» слова среди слов с общей семантикой. Полученные в результате работы программы слова далее будут сравниваться с оценками экспертов.

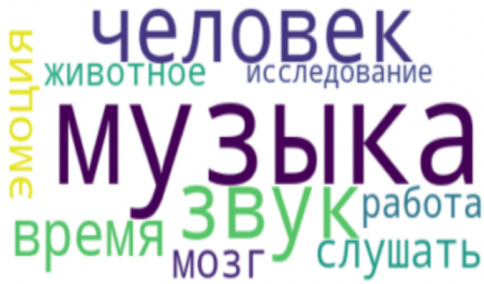
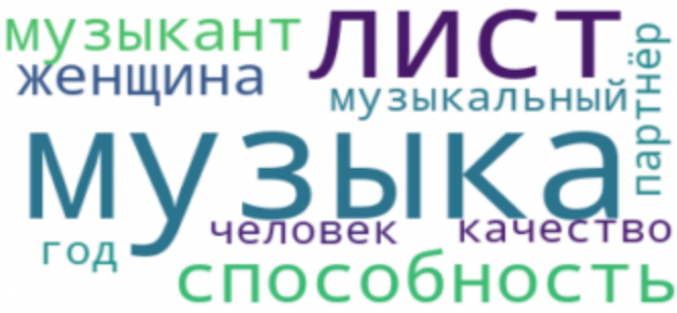
Новым для нас алгоритмом выделения ключевых выражений стала библиотека под названием **rutermextract**[6], созданная Игорем Шевченко. Алгоритм показал достоверные результаты, поэтому было решено включить его в работу.

Воспользовавшись встроенной в SketcthEngine функцией работы с корпусом, для каждого текста мы также получили списки КС, с которыми будем работать позднее.

Если внимательно взглянуть на корпус, станет понятно, что главная его особенность заключается в неоднородности тем. Общей темой для каждого текста является влияние музыки на организм человека, на процессы жизнедеятельности. Но в то же время каждый текст содержит в себе подтемы, поэтому корпус можно подразделить на несколько подкорпусов с пересекающимися темами. Для того чтобы посмотреть на разброс подтем, мы воспользовались алгоритмом тематического моделирования LDA, описанного в главе 2. Обучив модель на десяти текстах, мы пришли к пяти основным темам, представленным ниже.

[6] https://github.com/igor-shevchenko/rutermextract
https://pypi.org/project/rutermextract/

Таблица 3 *- распределение тем в корпусе*

Тема 1	**Тема 2**
	
Тема 3	**Тема 4**
	
Тема 5	
	

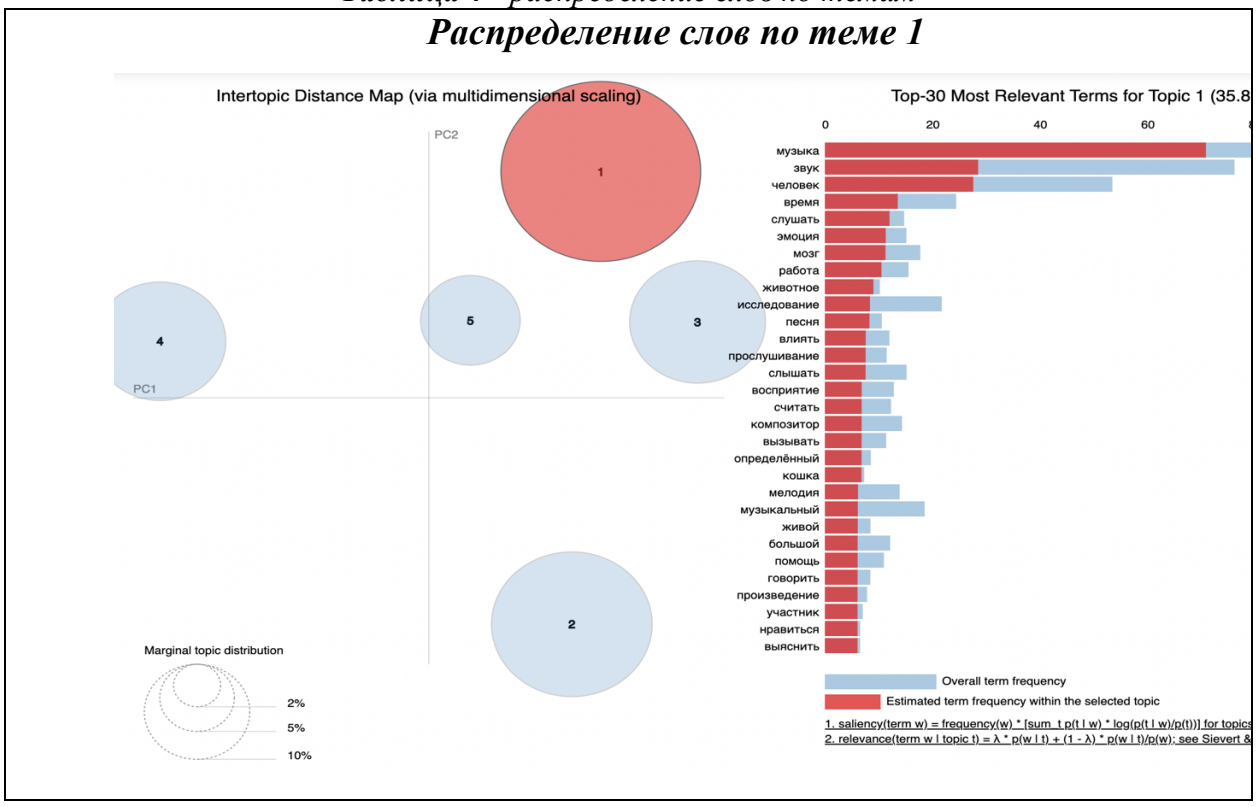
Если обратиться к таблице 4, то, во-первых, можно заметить, что в каждом из текстов присутствует основная тема – музыка. Доказательством тому служат слова, покрывающие каждую тему: *музыка, звук, трек, композиция*. При этом каждая из пяти групп содержит более узкие темы или подтемы.

Так, тема 1 включает в себя несколько подтем: влияние музыки и звуков на работу и эмоции, а также описание строения слухового аппарата у животных. Тема 2 связана с созданием атмосферы в играх и фильмах с

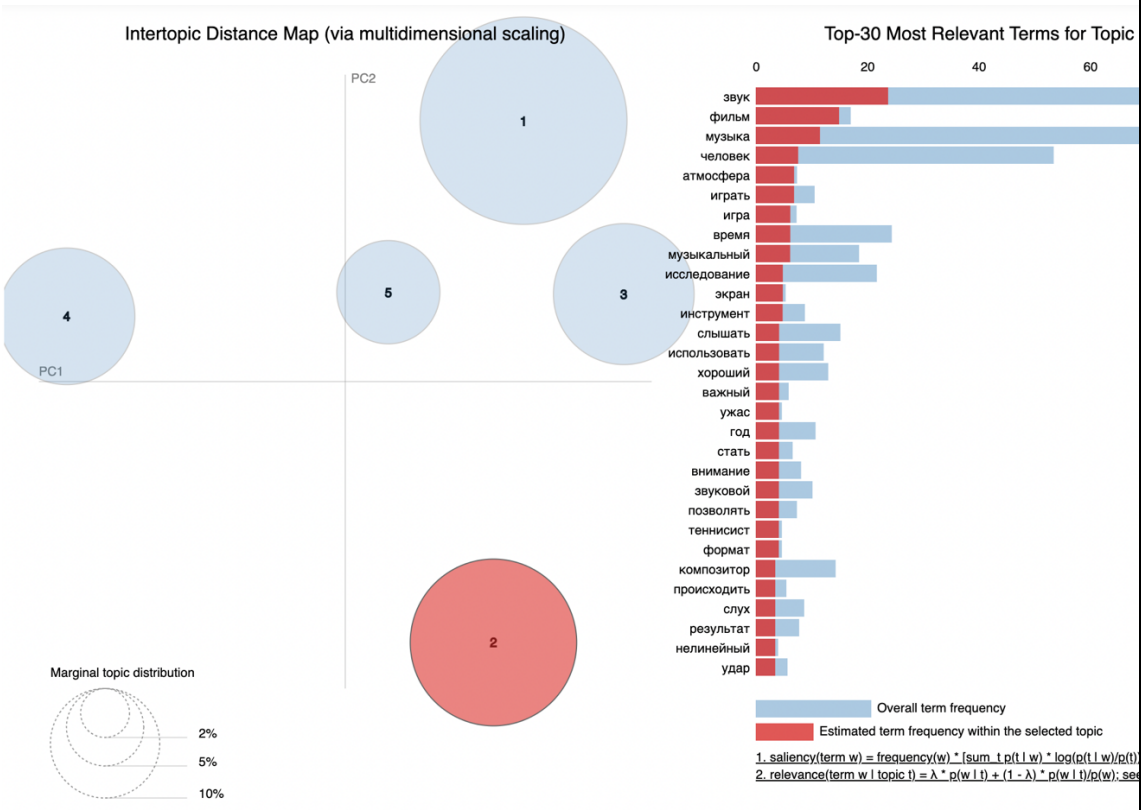
помощью музыки. Тема 3 охватывает такие аспекты, как эксперименты и исследования, касаемые влияния музыки и треков на здоровье и уровень стресса человека. В теме 4 представлены следующие подтемы: влияние музыки на восприятие и синестезия. Тему 5 отражают такие подтемы, как влияние музыки Ф. Листа на интеллектуальные способности и построение человеческих отношений.

Следует также отметить, что во всех рассмотренных выше темах основную долю составляют слова-дериваты, такие как: *музыка – музыкальный – музыкант, игра – играть, звук – звуковой.* Данное служит еще одним аргументом в пользу высказанной нами гипотезы.

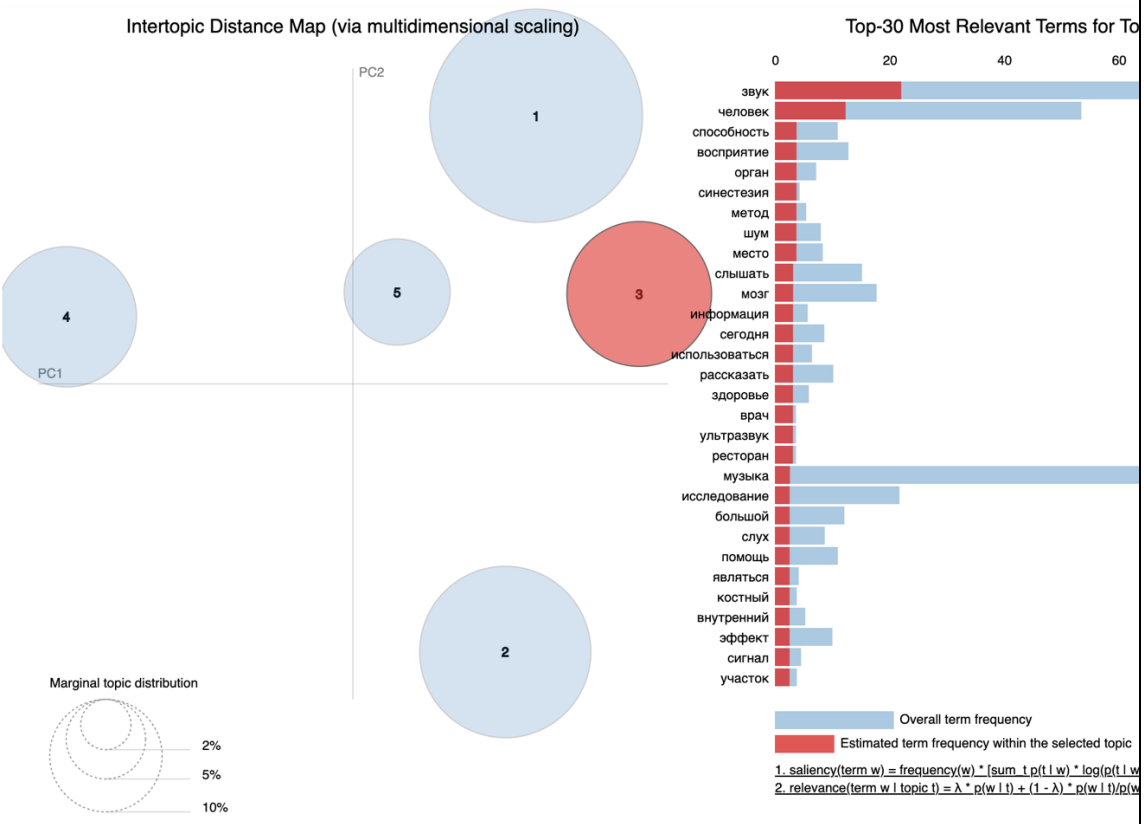
Таблица 4 *- распределение слов по темам*



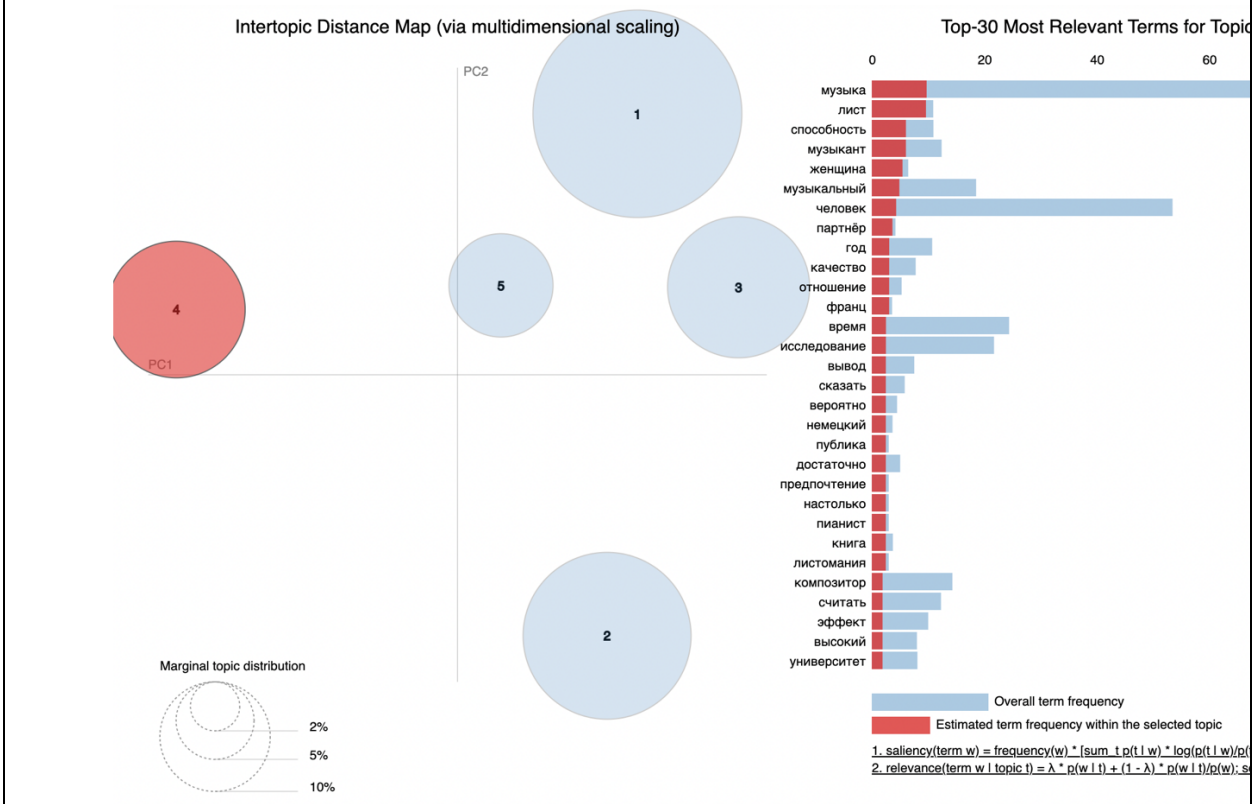
Распределение слов по теме 2



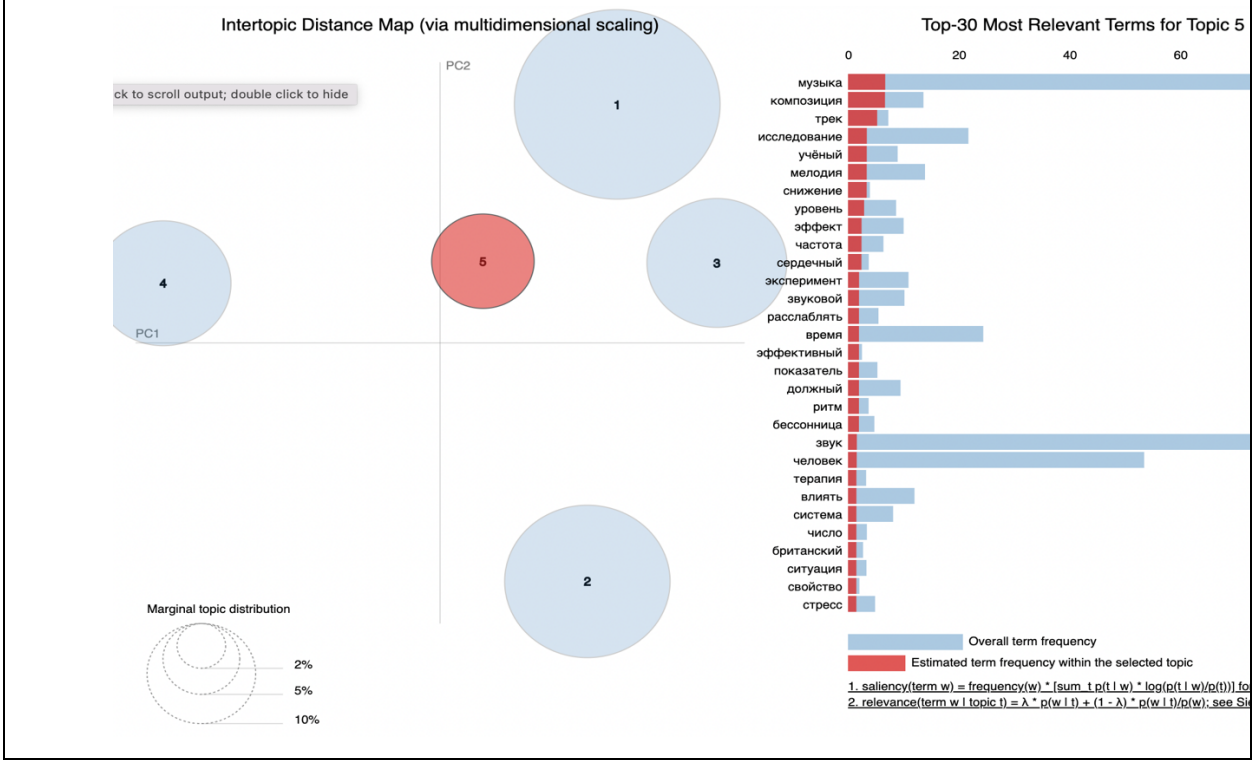
Распределение слов по теме 3



Распределение слов по теме 4



Распределение слов по теме 5



4.2.2 Лингвистический способ тематического моделирования

В качестве лингвистической основы для предстоящего эксперимента необходимо было составить некоторый набор ключевых слов, который послужил бы «золотым стандартом» для последующего сравнения с результатами статистического метода. Для создания эталонного списка мы прибегли к помощи экспертов.

Итак, экспертам были предложены исходные десять текстов, и перед ними была поставлена задача выделить те слова и словосочетания, которые, по их мнению, отражают тему текста. В итоге на каждый текст приходилось по четыре эксперта. Никаких ограничений, связанных с количеством КС, не предполагалось.

Обработка выделенных КС проводилась следующим образом: слова приводились к начальной форме, однокоренные слова объединялись в одно слово (музыкальный, музыкант → музыка), словосочетания подразделялись на составные слова (живая музыка → живой, музыка), английские слова не включались.

Таблица 5 *- список КС для текста 1*

эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Программа TF-IDF**	**Sketch Engine**	**Rutermextract**
бессоница	артериальное давление	Бессоница	артериальное давление	бессонница	академия звуковой терапии	артериальное давление
влияние	бессонница	Британские ученые	влияние	давление	артериальное давление	бессонница
исследование	замедление дыхания	Влияние музыки на организм	исследвание	композиция	артериальный	звуковая терапия
музыка	исследование	Жанр	композиция	мелодия	бессонница	исследование
пробуждение	классическая музыка	Исследование	мозговая активность	ритм	британская академия	композиция
расслабление	комфорт	Мелодия	ощущение комфорта	свойства	британская академия звуковой терапии	музыка
снижение	мозговая активность	Музыка	пробуждение	сердечный	звуковая терапия	сердечное сокращение
стресс	музыка	Музыка для расслабления	расслабляющая кмпозиция	снижение	расслабляющий эффект	сердечный ритм
ученый	расслабление	Нервная система	расслабляющий эффект	сокращения	сердечный ритм	снижение
	сердечный ритм	Пробуждение	свойство трека	стресс	снижение уровня	стрессовая ситуация
	снижение пульса	Снижение сердечного ритма	сердечный ритм	трек	стрессовый ситуация	трек
	темп	Стресс	стрессовая ситуация	уровень	уровень стресса	уровень
	тревога	Темп	темп	ученые	утренняя музыка	учёный
	уровень стресса	Трек	уровень стресса	частота	участница	частота
	частота		частота дыхания		участница эксперимента	
	эффект		эксперимент		частота дыхания	

Разбив слова и выражения на отдельные элементы текста, мы сравнили оценки экспертов между собой, а также с КС, предложенными статистическими алгоритмами. Каждому слову ставилась в соответствие единица, если слово встретилось у данного эксперта, и ноль, если слово не было выделено. Далее считался общий процент совпадений.

Таблица 6 *- оценки экспертов для текста 1*

текст 1					
	музыка	**исследование**	**ученый**	**бессонница**	**расслабление**
эксперт 1	1	1	1	1	1
эксперт 2	1	1	0	1	1
эксперт 3	1	1	1	1	0
эксперт 4	0	1	0	0	1
Sketch Engine	1	0	0	1	1
Программа TF_IDF	0	0	1	1	0
Rutermextract	1	1	1	1	0
Процент совпадений	**71,4**	**71,4**	57,1	**85,7**	57,1

Просмотрев все десять текстов, мы пришли к соглашению, что слова, для которых процент совпадений больше 70% составляют средний набор КС, а слова, процент совпадения которых превышает 85%, составляют малый набор КС.

Таким образом, для первого текста мы получили следующий предварительный список слов-тематизаторов: *музыка, исследование, бессонница, снижение, стресс, сердце, ритм, давление.*

Тем не менее, для большей точности мы также проверили корреляцию экспертных оценок друг с другом и с программами. Корреляция высчитывалась в Excel с помощью функции =СУММПРОИЗВ(--(интервал 1 = интервал 2))/количество слов*100.

Таблица 7 *- корреляция согласия экспертных оценок для текста 1*

	эксперт 1	эксперт 2	эксперт 3	эксперт 4	Sketch Engine	Программа TF-IDF	Rutermextract
эксперт 1		39,3	**71,4**	25,0	53,6	53,6	**60,7**
эксперт 2			35,7	**60,7**	**67,9**	39,3	53,5
эксперт 3				35,7	42,9	**64,3**	57,1
эксперт 4					46,4	39,3	39,2
Sketch Engine						**60,7**	**67,9**
Программа TF-IDF							**75**
Rutermextract							

Положительной корреляцией будем считать значения, составляющие шестьдесят пять процентов и более. Анализируя получившиеся данные, мы пришли к неоднозначному результату. С одной стороны, для некоторых текстов наблюдается сильное расхождение между оценками экспертов, что, на наш взгляд, можно объяснить несколькими причинами: во-первых, неоднородностью текстов, поскольку часть из них «монотематичны», то есть статичны в развитии темы, а остальные тексты содержат в себе несколько подтем, последовательно переходящих друг в друга, что усложняет задачу нахождения потенциально тематичных слов. Во-вторых, такие низкие значения корреляции могут свидетельствовать о недостаточной строгости и полноте инструкций, полученной экспертами при выделении прототипов в тематические слова.

С другой стороны, для некоторых текстов наблюдается довольно высокий уровень корреляции оценок экспертов и статистических алгоритмов. Это, в свою очередь, говорит о произвольном выборе слов экспертами, что возможно связано с недостаточным пониманием содержания текста, либо о выборе слов, наиболее часто встречающихся в тексте.

Выдвинув гипотезу о возможных причинах подобных расхождений, при создании правдоподобного набора слов-тематизаторов мы рассматривали исключительно корреляции экспертных оценок, не учитывая результаты

машинных алгоритмов. Так, для текста 1 в ядерную группу вошли слова, выделенные экспертом 1 и экспертом 3, так как корреляция между ними равна 71,3. В итоге для первого текста «эталонными» стали такие тематические слова: *музыка, исследование, ученый, расслабление, влияние, пробуждение бессонница, снижение, стресс, ритм, темп, сердце*.

Необходимо добавить, что итоговый «эталонный» набор тематичных слов может разниться с предварительным в силу того, что при рассмотрении корреляции в согласии экспертов конечное решение по включению/исключению данного слова в качестве ключевого выносится на основании наибольшей частоты индексирования, то есть наибольшего процента совпадений между экспертами без учета статистических методов. Таким образом, слово, изначально не попавшее в предварительный список КС из-за недостаточного процента совпадения между экспертами и программами, может быть включено в итоговый список при условии, что среди экспертов с высоким коэффициентом согласия данное слово имеет не менее 70% совпадений. Рассмотрим подробно каждый из текстов.

Таблица 8 *- оценки экспертов для текста 2*

текст 2					
	орган	**звук**	**синестезия**	**цвет**	**мозг**
эксперт 1	1	1	1	1	1
эксперт 2	1	0	1	1	1
эксперт 3	1	0	1	0	0
эксперт 4	1	0	1	1	1
Sketch Engine	1	0	1	0	1
Программа	1	1	1	0	1
Rutermextract	1	1	1	0	1
Процент совпадений	**100**	**42,9**	**100**	**42,9**	**85,7**

Предварительный список тематически значимых слов для текста 2: *орган, синестезия, мозг, кость, проводимость, информация, способность, чувство*.

Таблица 9 *- корреляция согласия экспертных оценок для текста 2*

	эксперт 1	эксперт 2	эксперт 3	эксперт 4	Sketch Engine	Программа TF-IDF	Rutermextract
эксперт 1		58,3	58,3	62,5	45,8	54,2	62,5
эксперт 2			50,0	**79,2**	54,2	45,8	45,8
эксперт 3				45,8	**70,8**	62,5	54,2
эксперт 4					**66,7**	50,0	58,3
Sketch Engine						**83,3**	**83,3**
Программа TF-IDF							**83,3**
Rutermextract							

В данной таблице присутствует довольно сильная корреляция в оценках экспертов 2 и 4, а также между экспертом 4 и экспертом 1. При этом отчетливо виден высокий уровень согласия между статистическими методами. Важно уточнить, что текст 2 мы относим к «политематичным».

Итак, для текста 2 список прототипов в тематические слова выглядит следующим образом: *орган, синестезия, цвет, мозг, кость, проводимость, сигнал, музыка, хромостезия, информация, способность, чувство.*

В таблице ниже представлены оценки экспертов для текста 3.

Таблица 10 *- оценки экспертов для текста 3*

текст 3					
	Звук	**ухо**	**диапазон**	**колебание**	**слух**
эксперт 1	1	1	1	1	1
эксперт 2	1	1	1	1	1
эксперт 3	1	1	1	0	1
эксперт 4	1	1	0	1	1
Sketch Engine	1	0	1	1	1
Программа	1	1	1	0	1
Rutermextract	1	1	0	0	1
Процент совпадений	**100**	**85,7**	**71,4**	**57,1**	**100**

Список слов-тематизаторов, полученный в результате экспертных оценок: *звук, ухо, слух, восприятие, музыка, животное, диапазон, кузнечик, кошка*.

Таблица 11 *- корреляция согласия экспертных оценок для текста 3*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		**72**	60	52	52	60	44
эксперт 2			48	48	40	48	48
эксперт 3				60	60	68	**68**
эксперт 4					60	44	60
Sketch Engine						60	44
Программа							**68**
Rutermextract							

В таблице прослеживается довольно высокая корреляция между экспертными оценками и автоматическими программами, но наибольший показатель был получен между оценками экспертов 1 и 2, поэтому будем опираться на них. Текст 3 также содержит только одну основную тему – восприятие музыки животными. Однако, столь небольшие показатели в согласии можно объяснить наличием в тексте перечисления разных видов животных, поэтому неудивительным кажется расхождение экспертов во мнении, какие именования стоит перечислять.

Итак, проанализировав слова, имевшие высокий процент совпадения у экспертов 1 и 2, мы получили следующие КС: *звук, ухо, канал, эмоция, диапазон, слух, восприятие, музыка*.

Таблица 12 *- оценки экспертов для текста 4*

текст 4					
	воздействие	**шум**	**тишина**	**звук**	**звукозапись**
эксперт 1	1	1	1	1	1
эксперт 2	1	1	1	1	1
эксперт 3	1	1	1	1	1
эксперт 4	1	1	1	1	0
Sketch Engine	1	1	0	1	0
Программа TF-IDF	1	1	1	1	0
Rutermextract	0	1	0	1	0
Процент совпадений	**85,7**	**100**	**71,4**	**100**	**42,9**

Для текста 4 в качестве слов, несущих в себе основную смысловую нагрузку, экспертами были отмечены: *воздействие, шум, тишина, звук, здоровье, звукоизоляция, ультразвук*. У обозревателя данного исследования может возникнуть вопрос касательно повторения корней в приведенных словах, так как при описании процедуры анализа выделенных людьми сочетаний мы утверждали, что все однокоренные слова сводились к одному слову-представителю. В действительности так и есть, но, если взять в качестве примера слова звук и ультразвук, то станет понятно, что несмотря на один корень, приставка ультра- придает слову дополнительное лексическое значение, поэтому мы склонны считать эти два слова разными. Еще одним важным уточнением мы считаем «политематичность» данного текста, так как данный критерий безусловно влияет на общий уровень согласия.

Таблица 13 *- корреляция согласия экспертных оценок для текста 4*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		61,1	55,6	55,6	27,8	**66,7**	50,0
эксперт 2			61,1	50,0	55,6	61,1	55,6
эксперт 3				**77,8**	61,1	44,4	50,0
эксперт 4					50,0	44,4	38,9
Sketch Engine						61,1	55,6
Программа TF-IDF							**72,2**
Rutermextract							

Что касается согласия экспертных оценок, то в таблице 14 наивысшее значение достигло 77,8 %, следовательно, при подборе КС для данного текста мы ориентировались на эксперта 3 и эксперта 4. Высокий показатель корреляции между алгоритмами TF-IDF и Rutermextract говорит о том, что в данном тексте по большей части в качестве ключевых выделялись менее частотные, редкие слова.

В результате сравнения предложенных двумя экспертами слов-тематизаторов был составлен список, в который входили слова: *воздействие, ультразвук, шум, тишина, звук, здоровье, звукоизоляция, эффект, вред*.

Таблица 14 *- оценки экспертов для текста 5*

текст 5					
	музыка	**живой**	**эмоции**	**слышать/слушать**	**эксперимент**
эксперт 1	1	1	1	1	0
эксперт 2	1	1	1	0	1
эксперт 3	1	1	0	1	1
эксперт 4	1	1	1	1	1
Sketch Engine	1	1	0	1	1
Программа TF-IDF	1	1	0	1	1
Rutermextract	1	1	1	1	0
Процент совпадений	**100**	**100**	**57,1**	**85,7**	**71,4**

Предварительный список прототипов в слова-тематизаторы для текста 5 включает: *музыка, живой, слышать/слушать, эксперимент, восприятие , исполнитель*.

Таблица 15 *- корреляция согласия экспертных оценок для текста 5*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		**84,6**	53,8	**69,2**	61,5	46,2	**76,9**
эксперт 2			53,8	**69,2**	61,5	46,2	61,5
эксперт 3				38,5	**92,3**	**76,9**	**76,9**
эксперт 4					46,2	61,5	46,1
Sketch Engine						**69,2**	**84,6**
Программа TF-IDF							53,8
Rutermextract							

В таблице 16 представлены значения корреляции согласованности экспертов, причем наблюдается большой процент совпадений мнений экспертов друг с другом. Наивысшие показатели представлены у экспертов 1 и 2, а также у экспертов 1 и 4, 4 и 2. Текст 5 мы отнесли к «политематичным», отсюда противоречие – согласно нашему первичному предположению монотематичные тексты более просты для подобного разбора, однако, результаты анализа экспертных оценок для рассматриваемого текста демонстрируют обратное. Вероятно, такие высокие значения согласия связаны с малым объемом самого текста, следовательно, есть выделение ключевых

слов не занимало много времени и не требовало длительной концентрации внимания.

Таким образом, для текста 5 получился следующий истинный набор КС: *музыка, живой, эмоции, влияние, эксперимент, восприятие, исполнитель*.

Таблица 16 - оценки экспертов для текста 6

текст 6						
	бессонница	**слушать**	**музыка**	**стресс**	**приложение**	**сон**
эксперт 1	0	1	1	0	1	1
эксперт 2	1	0	1	1	0	1
эксперт 3	1	0	1	1	1	1
эксперт 4	1	1	1	1	1	1
Sketch Engine	1	0	1	0	0	0
Программа	0	1	1	1	1	1
Rutermextract	1	0	1	0	1	1
Процент совпадений	**71,4**	**42,9**	**100,0**	**57,1**	**71,4**	**85,7**

Кандидатами в ключевые слова стали: *бессонница, музыка, приложение, сон, звук, работа, шум, продуктивность*.

***Таблица 17** - корреляция согласия экспертных оценок для текста 6*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		50,0	44,4	**66,7**	38,9	50,0	50
эксперт 2			50,0	61,1	**66,7**	55,6	**66,7**
эксперт 3				**66,7**	50,0	**72,2**	**72,2**
эксперт 4					50,0	**72,2**	61,1
Sketch Engine						55,6	**77,8**
Программа TF-IDF							**66,7**
Rutermextract							

В данном случае наблюдается высокая корреляция между статистическими алгоритмами и решениями респондентов, что говорит о довольно произвольном выборе тематичных слов в данном тексте. Тем не менее, согласованность между экспертами есть. Сошлись в своих решениях эксперты 1 и 4, а также эксперты 1 и 3. В итоге основной набор КС для текста 6 составили слова: *бессонница, музыка, стресс, наушник, приложение, сон,*

звук, работа, шум, продуктивность, здоровье. Заметим, что текст 6 включает в себя несколько тем.

***Таблица 18** - оценки экспертов для текста 7*

текст 7					
	ужас	**атмосфера**	**звук**	**музыка**	**композитор**
эксперт 1	1	1	1	1	1
эксперт 2	1	1	1	1	0
эксперт 3	1	1	1	1	0
эксперт 4	1	1	1	1	0
Sketch Engine	1	1	1	1	1
Программа	1	1	1	1	0
Rutermextract	1	1	1	1	0
Процент совпадений	**100,0**	**100,0**	**100,0**	**100,0**	**28,6**

В первичный список КС для текста 7 попали: *ужас, атмосфера, звук, музыка, слух, хоррор, фильм , страх, игра, озвучание*.

***Таблица 19** - корреляция согласия экспертных оценок для текста 7*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		42,9	57,1	60,7	50,00	42,9	53,6
эксперт 2			57,1	**67,9**	**71,4**	64,3	53,6
эксперт 3				**75**	42,9	57,1	46,4
эксперт 4					53,6	60,7	42,9
Sketch Engine						**78,5**	53,6
Программа TF-IDF							**68**
Rutermextract							

Текст 7 относим к монотематичным, поскольку тема создания атмосферы страха и ужаса прослеживается на протяжении всего текста. Здесь наблюдается сильная положительная корреляция в согласии эксперта 4 с экспртом 2 и эксперта 4 с экспертом 3. Однако, высокий уровень согласованности эксперта 2 с ресурсом Sketch Engine наталкивает на некоторые сомнения по поводу релевантности выбора КС, поэтому основное предпочтение в выборе слов-тематизаторов отдаем экспертам 3 и 4. В итоге для текста 7 ключевыми словами являются: *ужас, атмосфера, звук, музыка,*

слух, хоррор, фильм, страх, нелинейный, частота, озвучание, нервный, окончания, эмоции, игра.

***Таблица 20** - оценки экспертов для текста 8*

текст 8					
	музыка	**область**	**мозг**	**влиять**	**социальный**
эксперт 1	1	1	1	1	1
эксперт 2	1	0	1	1	1
эксперт 3	0	1	1	1	1
эксперт 4	1	1	1	1	1
Sketch Engine	1	1	1	0	0
Программа	0	0	1	0	0
Rutermextract	1	1	1	0	0
Процент совпадений	**71,4**	**71,4**	**100,0**	**57,1**	**57,1**

Для текста 8 были выбраны следующие слова: *музыка, область, мозг, удовольствие, память, слушать, эмоции, центр.*

***Таблица 21** - корреляция согласия экспертных оценок для текста 8*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		**63,64**	**81,82**	**81,82**	36,36	45,45	45,5
эксперт 2			63,64	45,45	36,36	**63,64**	36,4
эксперт 3				**72,73**	36,36	54,55	36,4
эксперт 4					54,55	36,36	54,5
Sketch Engine						**63,64**	72,7
Программа TF-IDF							63,6
Rutermextract							

Текст 8 представлен несколькими пересекающимися темами, объединенными одной большой, тем не менее, мы склонны относить данный текст к монотематичным. По таблице 22 заметны высокие показатели согласованности экспертов 1, 3 и 4, что говорит как о простоте текста, так и о подтверждении нашей гипотезы о соотношении высоко экспертного согласия и монотематичности текста.

Уделив внимание решениям трех указанных выше экспертов, мы получили эталонный набор КС, в который вошли: *музыка, область, мозг,*

влиять, социальный, связь, удовольствие, память, слушать, эмоции, настроение, физический, форма, центр, песня.

***Таблица 22** - оценки экспертов для текста 9*

текст 9					
	звук	**база**	**данные**	**раздражитель**	**теннис**
эксперт 1	1	1	1	1	1
эксперт 2	1	1	1	1	1
эксперт 3	1	1	1	0	1
эксперт 4	1	1	1	1	1
Sketch Engine	1	0	0	1	1
Программа	1	1	1	1	1
Rutermextract	1	0	0	0	1
Процент совпадений	**100**	**71,4**	**71,4**	**71,4**	**100**

Текст 9, по мнению, экспертов, содержит такие КС: *звук, база, данные, раздражитель, теннис, мелодия, феномен, музыка, звучать*.

***Таблица 23** - корреляция согласия экспертных оценок для текста 9*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		**70,6**	58,8	**70,6**	52,9	**94,1**	52,9
эксперт 2			**64,7**	**76,5**	58,8	**76,5**	47,1
эксперт 3				**76,5**	58,8	64,7	**70,6**
эксперт 4					47,1	**76,5**	58,8
Sketch Engine						58,8	**76,5**
Программа TF-IDF							58,8
Rutermextract							

Текст 9 без всяких сомнений относится к политематичным, и высокий уровень согласованности экспертов снова подтверждает недостоверность первой нашей гипотезы. Практически полное совпадение в ответах эксперта 1 и алгоритма TF-IDF, а также высокие показатели корреляции между статистическими алгоритмами позволяют сделать вывод, что для политематичных текстов редкие и низкочастотные слова с большей вероятностью будут выделены в качестве ключевых.

Стоит к тому же обратить внимание на то, что в большинстве случаев наибольшая корреляция наблюдается между экспертом 4 и любым другим. На основании данного наблюдения можно заключить, что ответы данного эксперта можно назвать универсальными для конкретного корпуса, и что истинный набор кандидатов в тематизаторы может быть составлен на основании его ответов.

В результате анализа экспертных оценок мы получили истинный набор КС: *звук, база, данные, раздражитель, теннис, эмоция, мелодия, феномен, музыка, слушать, звучать, Иоганн Себастьян Бах*.

Таблица 24 *- оценки экспертов для текста 10*

текст 10					
	Франц Лист	**листомания**	**феномен**	**чарующий**	**эффект**
эксперт 1	1	1	1	0	0
эксперт 2	1	1	1	0	0
эксперт 3	1	1	1	1	1
эксперт 4	1	1	1	0	1
Sketch Engine	1	1	0	0	1
Программа	1	1	0	0	0
Rutermextract	1	0	0	0	0
Процент совпадений	**100,0**	**85,7**	**57,1**	**14,3**	**42,9**

Текст 10 мы отнесли к политематичным. Предварительный список ключевых слов выглядит так: *Франц Лист, листомания, музыкант, женщина, пианист, способность, вкус*.

Таблица 25 *- корреляция согласия экспертных оценок для текста 10*

	эксперт 1	**эксперт 2**	**эксперт 3**	**эксперт 4**	**Sketch Engine**	**Программа TF-IDF**	**Rutermextract**
эксперт 1		52,0	44,0	40,0	**72,0**	**80,0**	60
эксперт 2			52,0	**72,0**	48,0	56,0	44
эксперт 3				**64,0**	40,0	48,0	44
эксперт 4					44,0	44,0	40
Sketch Engine						**68,0**	**72**
Программа TF-IDF							**72**
Rutermextract							

В таблице 26 виден высокий уровень согласия между экспертом 2 и экспертом 4. В результате анализа экспертных оценок мы получили следующий истинный набор КС: *Франц Лист, листомания, феномен, музыкант, влиять, предпочтение, женщина, пианист, способность, экстаз, привлекательность, вкус*.

4.2. Получение основ тематически маркированных слов с помощью деривационного анализа

4.2.1 Метод отсечения суффиксов и флексий

Основная часть данного исследования – разработка программы отсечения аффиксов для получения основы – ядра слова, объединяющего все производные слова общей семантикой. На первом этапе обработки происходило отсечение флексий, постфиксов и суффиксов. Сначала отсекались окончания и формообразующие суффиксы – морфемы, служащие для образования формы слова. Формообразующие суффиксы, в отличие от окончаний, не выражают грамматических значений падежа, числа, рода и лица, но выражают значения времени, степеней сравнения, наклонения и так далее. По большей части формообразующие суффиксы представлены в глагольных формах. К ним относятся суффиксы инфинитива, формы прошедшего времени, повелительного наклонения, причастий и деепричастий, а также степеней сравнения прилагательных и наречий.

Считаем необходимым напомнить, что все слова были приведены к словарной форме (лемматизированы), полные причастия были сведены к прилагательным. Для каждой из знаменательных частей речи был составлен список формообразующих аффиксов. Так, для глагола такими были суффиксы инфинитива: ть/-ти, -чь, а также сочетания суффикс+постфикс типа –ться/тись. Для прилагательных и наречий в качестве формообразующих морфемам мы выделили флексии “ий”, ”ый”, ”ой”, ”ей”, “ое”, ”ая” (для прилагательных) и суффиксы сравнительной степени -ее(ей), -е , -ше, -же, -ейше, -айше, -ейш/-айш (для прилагательных и наречий). У существительных на данном этапе убирались только окончания.

Далее круг аффиксов, выражающих грамматическое значение, расширялся. Такое последовательное включение большего количества морфем позволило нам тщательно отслеживать полученные основы и корректировать программу в случае ошибок. Итак, для прилагательных и причастий мы получили объединенный список следующих суффиксов: -ащ-(-ящ-), -ущ-(-ющ-), -ш-, -вш-, -им-, -ом-/-ем-, -нн-, -онн-/-енн-, -анн/-янн, -т-, -ан/-ян. Деепричастий (GRND) в нашем корпусе практически не встретилось, тем не менее, в случае увеличения объема корпуса были учтены и добавлены суффиксы а/-я, -в, -вши, -ши. Чтобы избежать ошибок в отсечении лишних морфем, полученные основы проверялись вручную, но в дальнейшем, при работе с большими объемами токенов, планируется создание проверочного списка слов, на который алгоритм будет ориентироваться при отсечении той или иной морфемы. Такой список возможно получить с помощью Морфемно-орфографического словаря А.Н. Тихонова[7], содержащего полные морфемные разборы слов. Итак, мы получили неполные основы слов.

[7] https://gufo.me/dict/tikhonov

Таблица 26 *- результаты алгоритма (отсечение формообразующих суффиксов)*

слово	абс. частота	Часть речи	основа
жить	3	INFN	жи
дать	3	INFN	да
ранний	6	ADJF	ран
рано	1	ADVB	рано
петь	1	INFN	пе
явно	1	ADVB	явно
целый	5	ADJF	цел
склонный	1	ADJF	склон
кафе	1	NOUN	кафе
тихий	2	ADJF	тих
кора	3	NOUN	кор
уметь	1	INFN	уме
игра	10	NOUN	игр
умный	1	ADJF	умн

Следующим не менее важным этапом работы алгоритма был процесс отсечения транспозиционных морфем. Транспозиционные суффиксы передают словообразовательное значение, свойственное производным словам, причем семантика дериватов ничем не отличается от семантики производящих слов, кроме общечастеречного значения.

Тщательно изучив теоретический материал, мы пришли к выводу, что основными продуктивными транспозиционными суффиксами являются суффиксы синтаксической деривации. Для имен существительных характерными являются отглагольные существительные со значением «отвлеченный процесс по глаголу», для выражения которого наиболее распространены суффиксы: -к, -ниј/тиј, -ациј, - -ениј, -аниј, -циј, -ств, -ø, -от, -б, -ня, -аж, -ёж, -ок. Для отадъективных существительных со значением отвлеченного признака характерны суффиксы -ость, -от, -изн, -ств, -в, -об, -ø. Класс продуктивных транспозиционных суффиксов для имен прилагательных в основном составляют суффиксы отсубстантивных прилагательных -н, -ов-, -ск-, -льн (и его варианты), имеющие самые широкие связи с производящими основами. Такие прилагательные выражают неконкретизированное отношение к тому, что названо производящей основой. К транспозиционным

суффиксам наречия относятся суффиксы отадъективных наречий со значением «отвлеченное свойство по прилагательному» -о, -е, -у, -и, -ому/-ему. К транспозиционным суффиксам глагола мы отнесли суффиксы -и/-а/-я/-е (удар → ударить), -ну, -ова/-ева/-ыва/-ва, -ировать/-изировать (демонстрация → демонстрировать), -евывать/-евствова.

Чтобы получить минимально возможные основы мы повторили процедуру отсечения грамматических суффиксов для всех частей речи. Так, например, для наречия *активно* после работы программы была получена основа *активн,* но так как данное слово образовано от прилагательного *активный*, которое в свою очередь образовано с помощью суффикса -н, этот суффикс также отсекался, в результате оставив основу *актив.* Что касается глаголов, то при появлении субморфов мы оставляли только один вариант. Если для слова *задумать* программа предлагала разбор *задума,* а для слова *задумка – задум,* то для двух слов конечным разбором считалась основа *задум*.

После процедуры отсечения суффиксов для всего корпуса полученные основы были отсортированы по алфавиту, что позволило четко увидеть «вкладывающиеся» стеммы. В таблице 28 представлены отсортированные по алфавиту основы, которые могут быть объединены в одну основу-представителя. Чередования морфем также учитывались и объединялись.

***Таблица 27** - вероятные вложения основ*

слово	частота	Часть речи	Финальная основа
активация	1	NOUN	актив
активность	5	NOUN	актив
активизироваться	1	INFN	актив
активный	1	ADJF	актив
активироваться	2	INFN	актив
активно	1	ADVB	актив
акустико	2	NOUN	акустик
акустик	1	NOUN	акустик
акустический	1	ADJF	акустич
альтернатива	2	NOUN	альтернатив
альтернативный	3	ADJF	альтернатив

После описанных выше действий мы основа обращались к словарю Ляшевской и Шарова и в соответствие каждому слову из полученного набора данных находили относительную частоту встречаемости.

Вслед за этим перед нами стояла задача объединения всех одинаковых основ в одну для получения так называемых классов эквивалентности. Для этого была написана программа, на выходе которой мы получили набор данных, в котором для каждой из основ было выбрано слово-представитель с указанием общей относительной частоты по нашему корпусу, по частотному корпусу Ляшевской–Шарова, а также коэффициента тематичности. Выбор слова-представителя был определен наличием среди слов с данной основой представителя основного слова в словообразовательном гнезде. Допустим, с конкретной основой в нашем корпусе встретились существтельное и мотивированное им прилагательное, следовательно, словом-представителем для данной основы станет имя существительное, поскольку в данном случае оно является центром словообразовательного гнезда. Соответственно, если выбор стоял между прилагательным и глаголом, в качестве представителя выступал глагол и так далее. Итоговый объем корпуса после объединения составил 1634 слова и был представлен в следующем виде.

Таблица 28 *- корпус после работы программы*

слово	Частота_ЛШ	Частота_корпус	thematicality
живой	1,8	0,005280845	29,33802742
гаджет	0,4	0,000960154	24,00384061
томография	0,4	0,000960154	24,00384061
звукоизоляция	0,8	0,00144023	18,00288046
прослушивание	4,3	0,007681229	17,86332325
тактильный	0,6	0,000960154	16,00256041
веселие	0,9	0,00144023	16,00256041
трек	4,2	0,006240999	14,85952038
саундтрек	0,7	0,000960154	13,71648035

Данный алгоритм был применен ко всему корпусу, равно как и ко всем текстам, в результате чего мы получили одиннадцать наборов данных, которые будут описаны ниже.

4.2.2 Анализ результатов тематической атрибуции

Итак, перейдем к более детальному описанию полученных в результате процедуры отсечения суффиксов таблиц. Каждый из датасетов мы отфильтровали по убыванию коэффициента «тематичности» для того, чтобы проверить, насколько изменились результаты тематической атрибуции после проведенного эксперимента.

При анализе получившихся разборов мы руководствовались подразделением текстов корпуса на политематичные и монотематичные, поэтому далее придерживаемся этого деления. Для каждого текста было построено несколько видов графиков: графики обеих частот, график по коэффициенту тематичности, а также интервальные графики.

Количество интервалов находилось по формуле Стерджесса:

$$\mathbf{k = 1 + 3{,}322 \times \lg(n)}$$

где n - общее число единиц наблюдения (общее количество элементов в совокупности, то есть количество словоформ в нашем тексте), lg(n) – десятичный логарифм от n.

Шаг интервала вычислялся как значение интервального размаха, деленного на значение k, то есть требуемое число интервалов. Формула вычисления интервального шага представлена ниже:

$$\boldsymbol{i = \frac{x_{max} - x_{min}}{k}}$$

где x_{max} и x_{min} - максимальное и минимальное значение в выборке.

Политематичные тексты

Текст 1

Проследить изменения в данных до и после обработки удобнее всего при построении интервальных графиков. Посчитав по каждому тексту длину интервалов и интервальный шаг, мы разделили каждый набор данных на 9 неравномерных интервалов по столбцу thematicality, поскольку нас интересует именно распределение тем. Остальные графики можно найти в приложении.

Таким образом, для текста 1 мы получили два интервальных графика: до преобразований (Thematicality_1) и после (Thematicality_2).

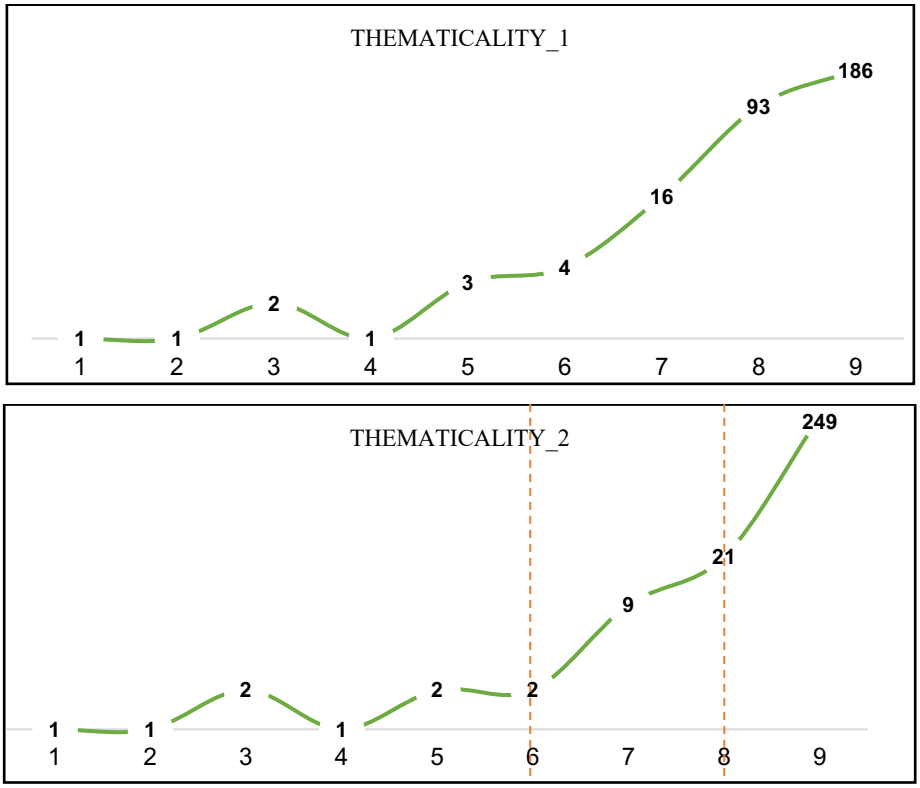


Рис. 1 *- Интервальные графики для текста 1*

На рисунке 1 изображены два графика, показывающие распределение слов по девяти интервалам. На оси абсцисс расположены номера интервалов, цифры на линии обозначают количество слов, попавших в конкретный интервал. Как можно увидеть, в обоих случаях наибольшее количество слов сосредоточено на хвосте. Это слова, имеющие самые низкие значения коэффициента «тематичности» и представляющие собой общие, неспецифичные единицы лексики. Напротив, слева сосредоточены самые специфичные для данного текста слова, значение thematicality которых преобладает.

Заметим, что оба графика мультимодальны, то есть имеют отдельно выраженные подргуппы тем, причем в наиболее выраженную группу в обоих графиках попадают одни и те же элементы (слова). Иными словами, слова, олицетворяющие специфичную лексику, неизменны вне зависимости от каких-либо деривационных преобразований. Таким образом, графики неоднородны, а отражение темы в тексте 1 динамично.

Начиная с пятого интервала значения колеблются от класса к классу, происходит наложение мод друг на друга. Это говорит о том, что в тексте присутствует несколько «подтекстов» со своими темами, что оказывает влияние на вид графика. Стоит заметить, что до преобразований моды отдельных интервалов более сглажены и слабо выражены, в то время как после процедуры отсечения аффиксов модальное распределение стало более выраженным, что говорит о возможности четко разбить графики на несколько частей, проведя разграничения по минимальным точкам.

Таким образом, разделив график Thematicality_2 на несколько промежутков, мы получили распределения низкочастотных, среднечастотных и высокочастотных слов.

Наибольший интерес для нас представляют слова, попавшие в вершины, а именно в интервалы под номером три, пять и шесть. В третьем интервале сосредоточены два слова – *засыпание, метроном*. Пятый интервал представлен словами *симпатический, артериальный*, а шестой – словами *кембриджский, усыплять*.

К самым специфичным словам с наибольшим значением коэффициента тематичности относятся слова из интервалов один, два, три, четыре, пять и шесть: *трек, испытуемый, засыпание, метроном, симпатический, кембриджмкий, усыплять*. Среднечастотную группу составляют слова из седьмого и восьмого интервалов, в которые входят слова *головоломка, минимум, замедляться, приветствоваться, композиция, бессонница, свистящий, прослушивание, подстраиваться, участница, сосудистый, вегетативный, расслабление, терапия, замедление, мелодия, предугадать, клип, сердечный, частота, сердечно, настрой, ритм*. В самую частотную группу вошли остальные слова, не будем указывать их в силу большого количества.

Текст 2

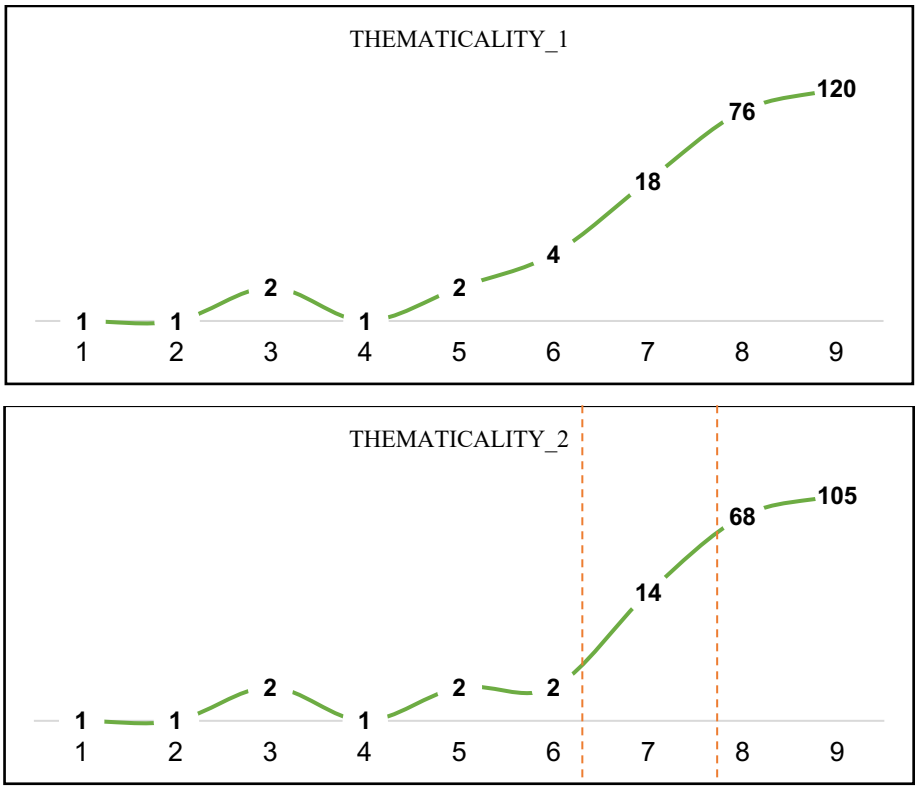


Рис. 2 *- Интервальные графики для текста 2*

На рисунке 2 изображены графики, отражающие распределение тем по тексту 2. Если внимательно рассмотреть верхний из них, можно увидеть одну ярко выраженную моду, что в принципе может говорить о данном распределении как об однородном, так как в остальном кривая равномерно возрастает. Данное наблюдение может навести на мысль о том, что данный текст вовсе не является политематичным и представлен только одной темой. Однако, при рассмотрении графика Thematicality_2, на котором изображена кривая после обработки подкорпуса для текста 2, график приобрел более выразительные моды, тем самым отразив неоднородность распределения.

Что касается трех частотных групп, в низкочастотную попали интервалы с первого по шестой с тремя ядерными группами в интервалах три, пять и шесть, куда вошли слова *испытуемый, карканье, живой, сенсор, заставка, заслать*. В интервалы один, два и четыре попали слова *тактильный, гаджет, проводимость*.

В среднечастотную группу вошли слова из интервала семь: *физиолог, фортепьяно, отрывистый, обрабатываться, переключаться, гарнитур, расшифровываться, калифорнийский, патология, разновидность, наименее, прослушивание, наушник, трость*.

Текст 4

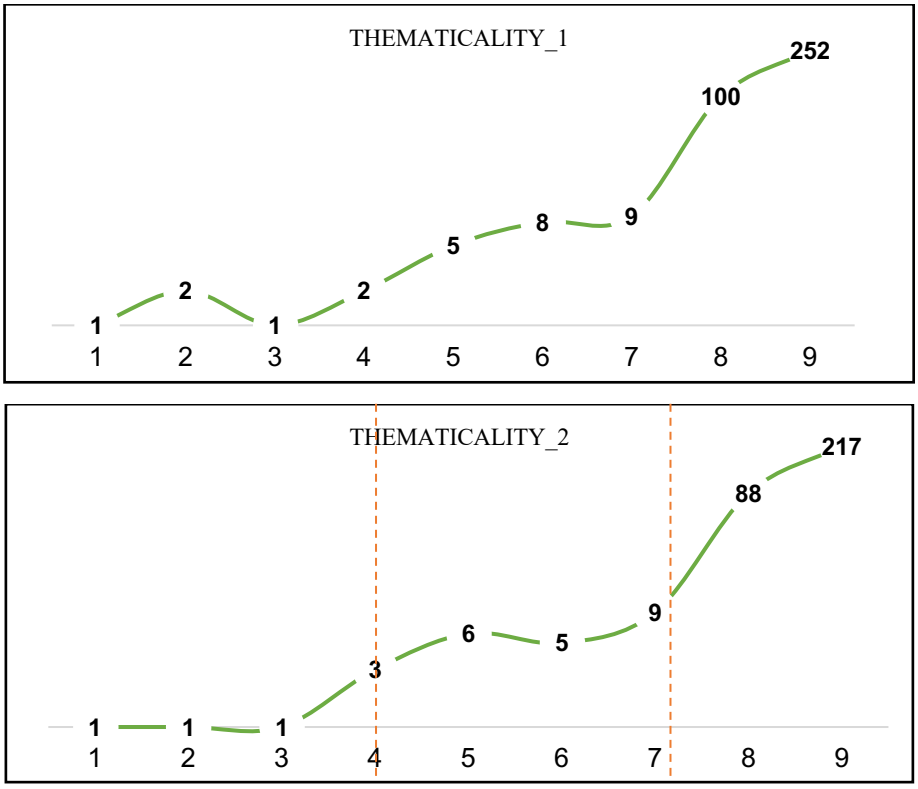


Рис. 3 *- Интервальные графики для текста 4*

На рисунке 3 два графика отображают неоднородное распределение темы по тексту. На обоих графиках видно наслоение мод друг на друга, тема развивается динамично. На графике Thematicality_1 присутствует ярко выраженная мода, но после морфемных преобразований она нивелируется, то есть происходит смещение высокотематичных слов в сторону менее показательной среднечастотной группы.

Итак, для данного текста также можно выделить три частотные группы, в самую ядерную, низкочастотную попадают слова из первого, второго, третьего и четвертого интервалов, а именно: *звукоизоляция, гаджет, плацебо, ультразвук, прогревание, стетоскоп*. Среднечастотная группа представлена словами из интервалов пять, шесть и девять: *фокусировать, аудиозапись,*

какофония, контент, звукозапись, отдалить, расслабление, регенерация, гипертония, знать, загрязнять, усугублять, эколог, вибрация, навредить, раковй, коммуникативный, целебный..

Текст 5

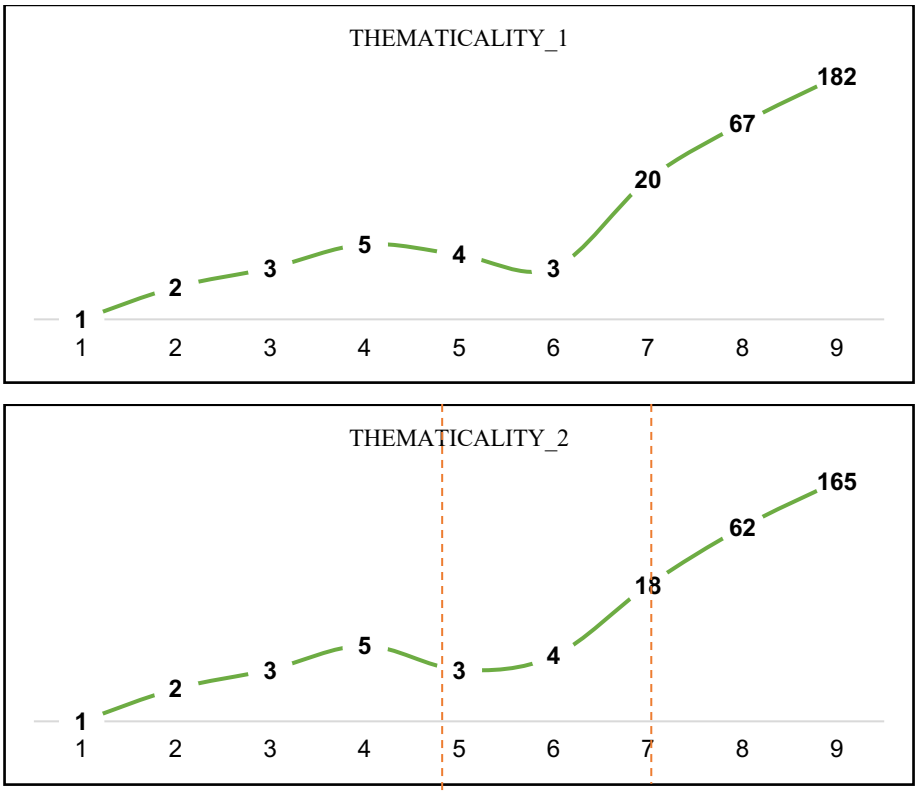


***Рис. 4** - Интервальные графики для текста 5*

На рисунке 5 наблюдается наложение распределений друг на друга, что говорит о неоднородности текста. Однако, такое наложение происходит не в начале кривой, как в предыдущих политематичных текстах, а в центре. На наш взгляд, это связано с тем, что первая часть текста 5 представлена одной темой, которая сменяется другой только ближе к концу, поэтому наиболее тематичные слова отражают тематику, которая пронизывает большую часть повествования.

Характерной особенностью данного текста является трудность разделения значений графика на три частотных группы, так как они накладываются друг на друга. Граница была проведена по точкам пересечения мод, то есть в интервалах пять и семь. Таким образом, низкочастотную часть, в тематически маркированную лексику попали слова из интервалов один, два,

три. Сюда вошли слова такие как *живой, триггер, электрокардиограмма, испытуемый, некомфортно, проигрываться, саундтрек, аудиосистема, вживую, клик, учащаться*. «Среднетематичную» группу составили слова, попавшие в интервалы пять, шесть и семь, среди которых: *синтезатор, прослушать, студийный, струнный, наушник, вдумчивый, частичка, микроскопический, слушатель, пульс, проделывать, акустический, продюсер, импровизация, файл, музыкант, композиция, комфортный, композитор*. Соответственно, общая лексика представлена в интервалах восемь и девять, и является самой частотной группой.

Текст 9

Текст 9 вне всяких сомнений является политематичным, и тому подтверждением служит распределение тематически маркированных слов на рисунке 5. В первую очередь, оба графика имеют вид неоднородного распределения с накладывающимися друг на друга модами.

Что касается морфемных преобразований, то значительных изменений в нижнем графике не наблюдается, за исключением видоизменения кривой в начале, которое произошло в результате перераспределения тематичных слов.

График Thematicality_2 с легкостью можно разделить на три группы. Низкочастотная группа включает в себя слова, находящиеся в интервалах один, два и три, в которые входят слова реципиент, веселие, теннисист, синхронизировать. Среднечастотная группа представлена интервалами четыре, пять, шесть и семь и словами *фагот, глухой, медитативный, озвучание, виниловый, ракетка, раздражитель, знать, юниор, кларнет, меломан, живой, струнный, таинственность, формат, барабанить, распознавать, транслировать, звучание, саксофон, теннис, феномен, трек*. Интервалы восемь и девять содержат наиболее частотную лексику, не обладающей предметной спецификой.

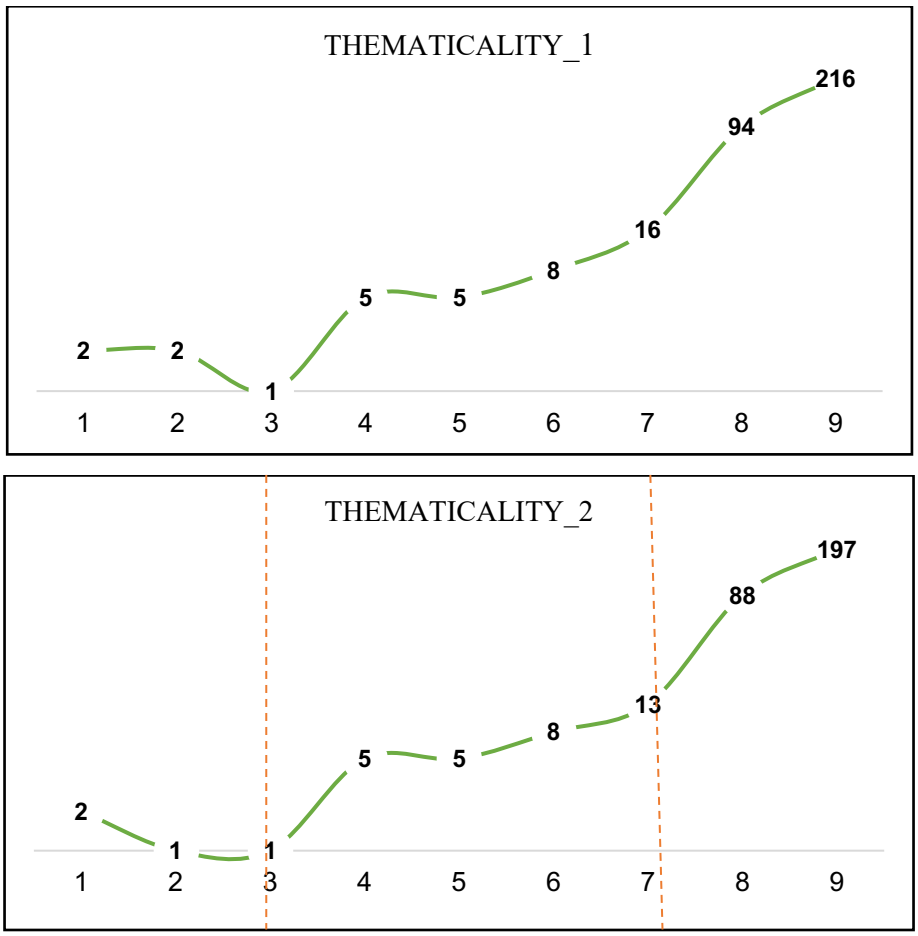


Рис. 5 *- Интервальные графики для текста 9*

Текст 10

Текст 10 представлен бимодальным распределением (рисунок 6) с ярко выраженными модами с вершинами в интервалах два и четыре. В целом, для данного текста кривая после проведенного эксперимента кривая практически не изменилась за исключением появления в нижнем графике моды, граничащей с горизонтальной осью. Текст 10 можно охарактеризовать как политематичный в силу его большого объема, так как основная тема плавно сменяется другой, пусть и хвост распределения выглядит достаточно однородным. Но колебание тематичных слов в начале говорит о наличии тематических маркеров одновременно для нескольких тем, что и вызывает наложение мод.

На графике Thematicality_2 наиболее тематичная группа и низкочастотная группа представлена словами из интервалов с первого по

пятый, причем границу нам позволяет провести место наложения высокочастотного хвоста на низкочастотную группу. Итак, в наиболее специфичные слова попадают верхушки мод – слова из интервалов под номерами два и четыре: сигарный, чаровать, фрустрация и одержимость, импровизатор, репродуктивный, а также слова на пересечении данных интервалов – *нейронаука, ноктюрн, мистический, музицирование*. В среднечастотную группу попали слова из интервалов шесть, семь, восемь и девять. К ним относятся следующие тематически маркированные слова: *льстивый, помешательство, менструальный, информативный, сыгранный, усыплять, рокер, фуга, фурор, минор, приравниваться, маркер, фортепианный, неосознанно, обуздать, неистовый, прослушивание, коррелировать, экстаз, пианист и другие.*

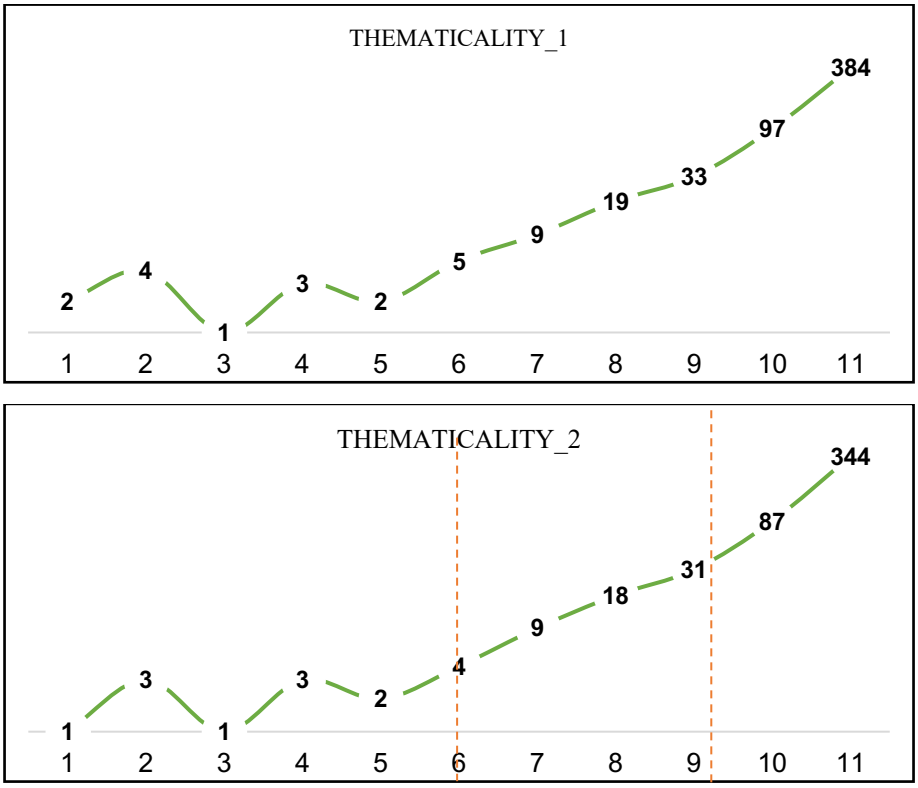


***Рис. 6** - Интервальные графики для текста 10*

Монотематичные тексты

Текст 3

На рис. 2 изображены графики распределений тематичных слов по тексту 3. Первичный график Thematicality_1 представляет собой однородное распределение слов без ярко выраженных мод, что подтверждает монотематичность данного текста.

Нижний график Thematicality_2 демонстрирует более смещенные данные, слегка меняя верхний рисунок. При этом после описанного выше алгоритма отсечения аффиксов произошло перераспределение слов по интервалам, образовав подобие мод.

Если рассмотреть график с точки зрения деления на частотные и низкочастотные группы, то в низкочастотную, но самую тематически маркированную лексику входят слова из первого, второго, третьего, четвертого и пятого интервалов: *урчание, сенсор, осязательный, расслаблять, пощекотать*.

В среднечастотную группу вошли слова из групп шесть и семь: *кузнечик, питомец, виолончелист, сдвоить, плавательный, обтекаемый, ушной, мяукать, сердцебиение, вибрация, сверчок, перепонка, фоновый, рецептор, обрабатываться, преобразоваться, мурлыкать, муравей, мембрана, усилитель, виолончель, крокодил, ветеринар, травмировать, ускорять, амплитуда*.

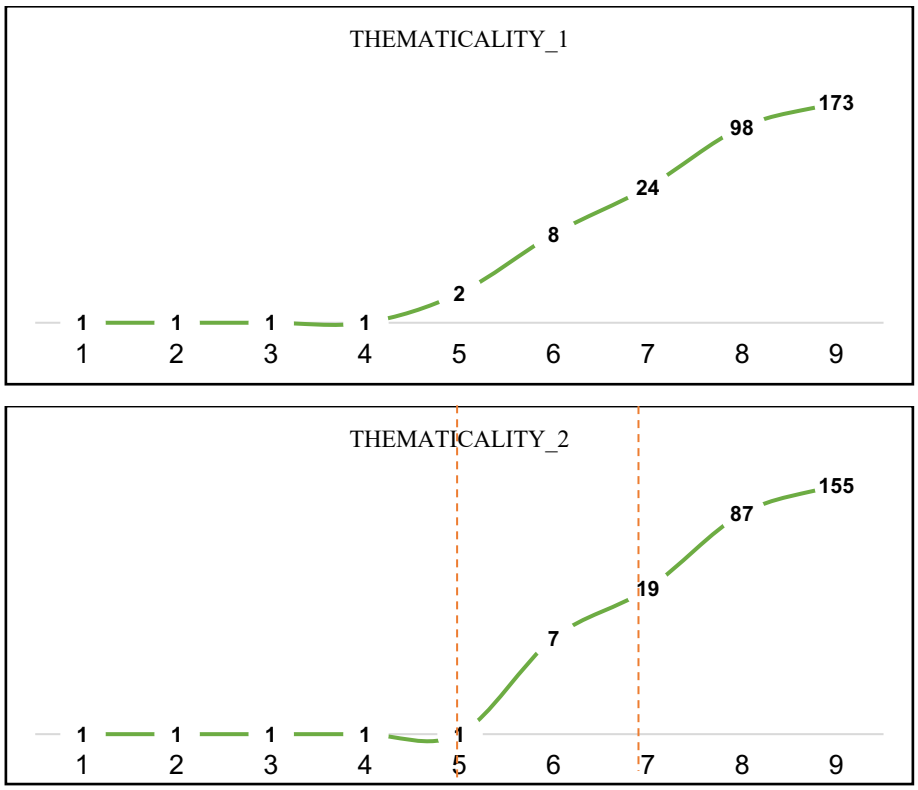


***Рис. 7** - Интервальные графики для текста 3*

Текст 6

Распределение слов по тексту 6 подтверждает отнесение его к монотематичным. Оба графика отражают статичное распределение темы по тексту, при этом после работы программы значения интервалов изменились лишь слегка изменились.

Разделив кривую на несколько отдельных участков, мы получили три частотных группы, в первую из которых вошли слова, попавшие в первые три интервала, во вторую – слова из интервалов 4-6, и в третью – слова из интервалов семь и восемь.

В низкочастотную группу вошли слова: *щебетание, саундтрек, ритмичный*.

В среднечастотную группу попали слова: *храп, знать, раздражительность, меломан, раздражающий, продуктивность, колыбельный, наушник, расслаблять, бессонница, негативно, ограждать, комфортный, приложение, динамик, прослушивание, датчик, звук, отслеживать, насладиться*.

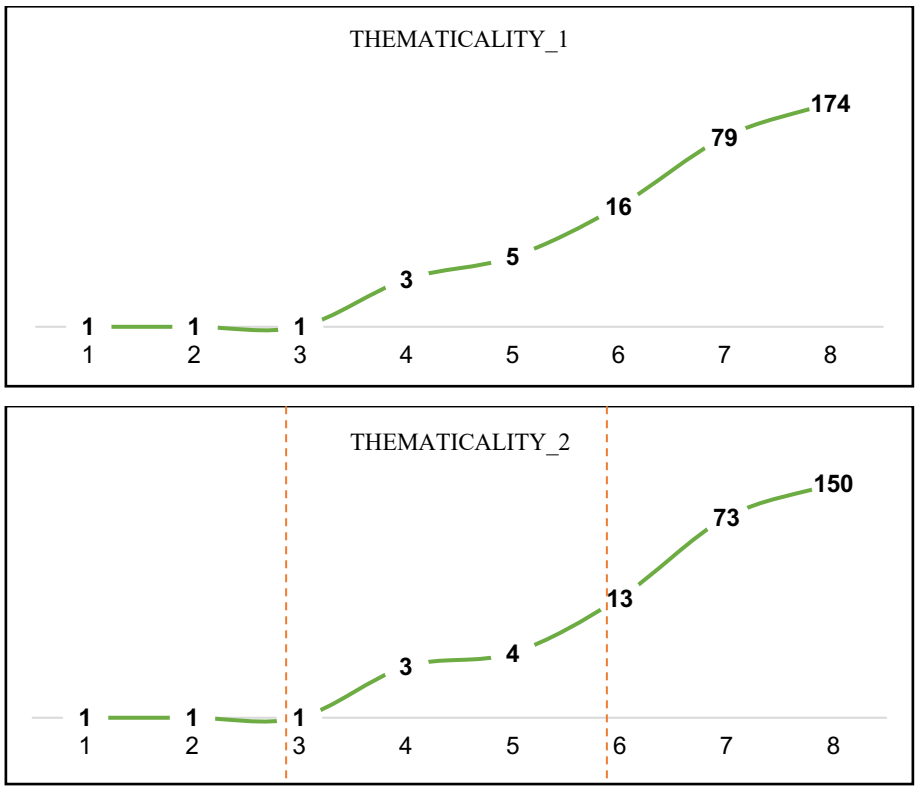


Рис. 8 *- Интервальные графики для текста 6*

Текст 7

Текст 7 также содержит одну основную тему, что можно увидеть на рисунке 5. График представлен одномодальным распределением с ярко выраженной модой. Обратим внимание на различное количество интервалов на обоих графиках. Дело в том, что объем подкорпуса для текста 7 до и после деривационных преобразований отличался, отсюда и различия в количестве интервалов и общем виде распределения.

Стоит уточнить, что объем текста 7 довольно велик, что не может не оказывать влияние на общий вид кривой, поскольку даже в монотематичном тексте имеют место некоторые подтемы, которые объединены общей, преобладающей темой, поэтому наличие моды нас нисколько не удивляет.

Можно увидеть, что площадь участка под графиком Thematicality_2 стала значительно меньше, чем у графика Thematicality_1, что говорит о перераспределении слов между интервальными группами. Тем не менее, основная интересующая нас группа (вершина мода) представлена равным

количеством слов до и после отсечения суффиксов. В такую ядерную группу попали четыре слова, среди которых *ужастик, наиграть, озвучивание, фавн*.

Итак, в низкочастотную и наиболее специфичную лексическую группу слов попали элементы первого, второго, третьего и четвертого интервалов: *испытуемый, ужастик, наиграть, озвучивание, фавн, понервничать, видеоряд, дерби, тиканье.* Среднечастотную группу составляют слова из интервалов пять и шесть*: озвучивать, сердцебиение, саванна, техасский, утрировать, животрепещущий, пробирать, гармоника, плеер, кинолента, подкидыш, спецэффект, оттачивать, эльф, головоломка, негодность, нагнетать, резня, ножевой, гуру, контроллер, аккомпанировать*. И высокочастотную группу составили слова из интервалов семь и восемь, таких слов большинство.

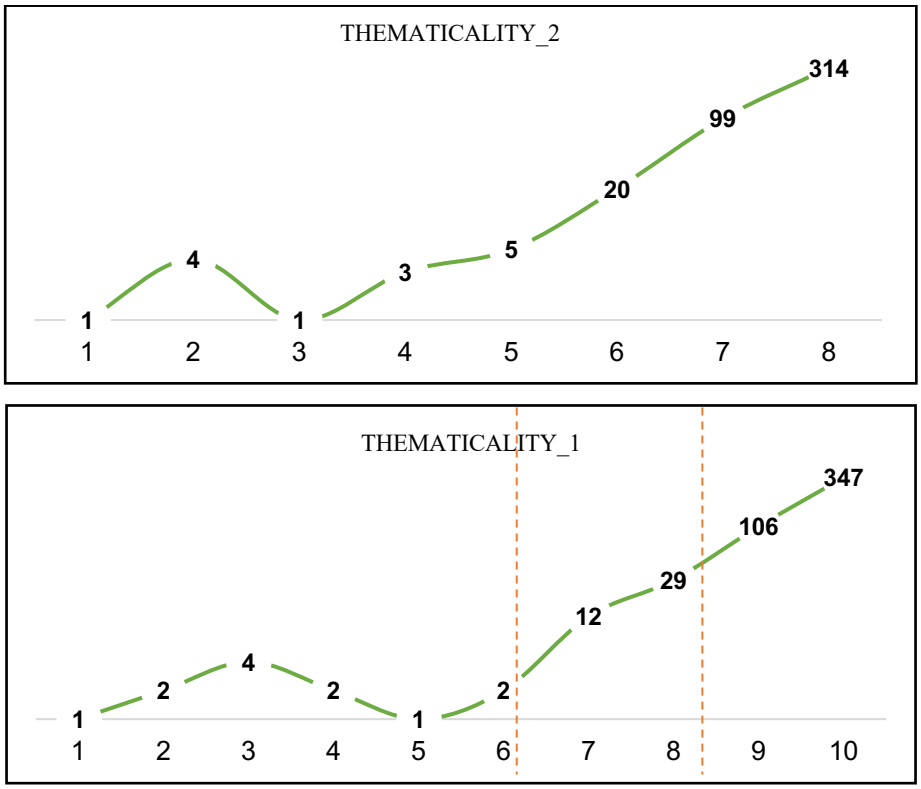


Рис. 9 *- Интервальные графики для текста 7*

Текст 8

На рисунке 9 изображены графики для текста 8. В целом, графики имеют вид однородного распределения без выраженных мод, что является характерным для монотематичного текста.

Процедура отсечения аффиксов не дала значительных изменений кривой, кроме некоторого изменения количества слов в интервалах, что связано с сокращением числа слов в итоговом подкорпусе. Сложным оказался и вопрос разделения графика Thematicality_2 на три частотных группы, поэтому для данного текста было решено оставить лишь две: низкочастотная группа – слова из интервалов с первого по пятый, и высокочастотная группа – интервалы шесть и семь. Итак, в первую группу вошли слова *испытуемый, томография, гиппокамп, серотонин, сканирование, миндалевидный, стэнфордский, деньга, модальный, веселие, восприимчивый, подкорка, прошедшее, разочаровываться*. Во вторую группу попали слова из интервалов шесть и семь.

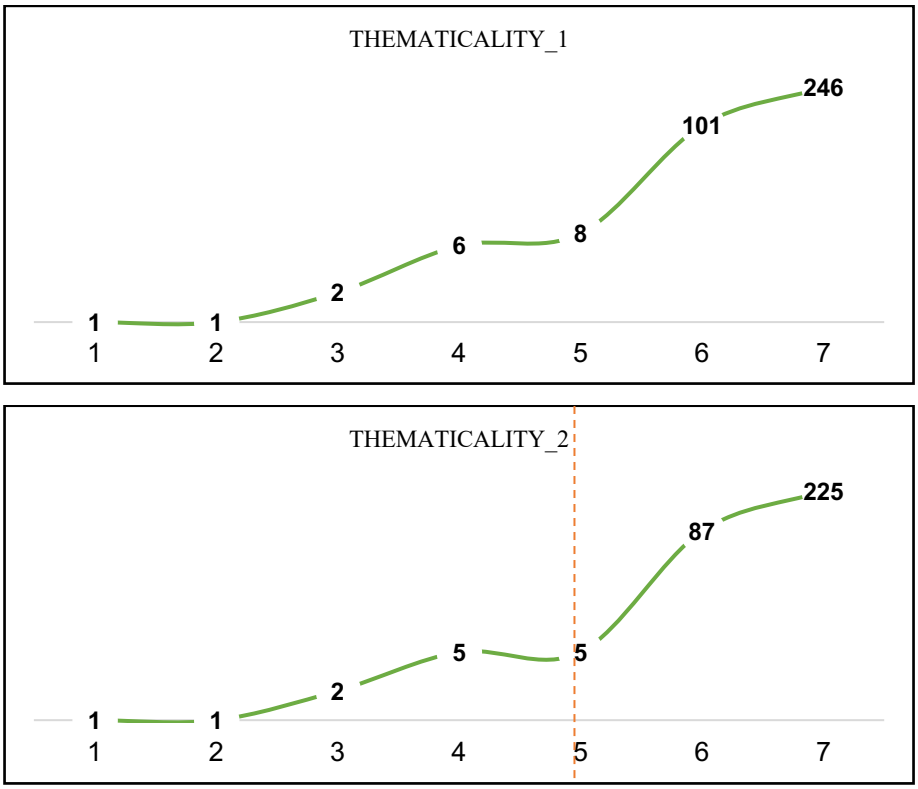


***Рис. 10** - Интервальные графики для текста 8*

Количественная оценка полученных результатов

Сразу оговоримся, что количественно оценить работу данного подхода трудно в силу, во-первых, частых расхождений в оценках экспертов, во-вторых, в силу субъективности этих оценок. Полученные истинные списки ключевых слов, выделенных экспертами, с одной стороны, позволяют утверждать о наличии некоторого «золотого стандарта», но, с другой стороны, при наличии большего количества экспертов были бы изменены и, возможно, получили бы больший процент согласия.

В то же время коэффициент тематичности позволяет оценивать тематичные слова с определенной спецификой. Такие слова хотя и отражают тему текста, но являются низкочастотными, поэтому в большинстве случаев пропускаются экспертами.

Тем не менее, на основании полученных данных были посчитаны две базовых для оценки алгоритма метрики – точность (sion) и полнота (Recall), которые считаются по следующим формулам:

$$\text{Precision} = \frac{\text{Число совпадений}}{\text{Общее число выделенных слов}}$$

$$\text{Recall} = \frac{\text{Число совпадений}}{\text{Общее количество правильно выделенных слов}}$$

Оценка проводилась по тем же десяти текстам. Слова с высоким коэффициентом тематичности, занимающие первые тридцать позиций в таблице, после работы программы по отсечению аффиксов сравнивались со словами, выделенными экспертами в качестве ключевых. Заметим, что в оценку включались только существительные и прилагательные, так как эксперты выделяли только эти части речи. В результате подсчета метрик получились следующие значения:

***Таблица 29** - количественная оценка работы алгоритма*

	Количество слов, выделенных экспертами	Топ 30 слов по коэффициенту тематичности	Количество совпадений	Точность	Полнота
Текст 1	12	30	6	20 %	50 %
Текст 2	12	30	3	10 %	25 %
Текст 3	8	30	2	7 %	25 %
Текст 4	8	30	6	20 %	75 %
Текст 5	7	30	3	10 %	43 %
Текст 6	11	30	5	16 %	45 %
Текст 7	15	30	4	13 %	27 %
Текст 8	15	30	3	10 %	20 %
Текст 9	12	30	6	20 %	50 %
Текст 10	12	30	3	10 %	25 %
Среднее значение:				13,6 %	38,5 %

Средние показатели точности и полноты очень малы, однако, как мы уже объяснили, такие маленькие значения связаны с высокой степенью субъективности выделения ключевых слов, а также со специфичностью самого коэффициента тематичности, который, в первую очередь, позволяет выявить именно низкочастотную лексику.

Результаты количественной оценки алгоритма значительно улучшились, когда экспертам были предложены уже готовые списки ключевых слов, составленные на основе программной выдачи тематичных слов, то есть списки, составленные из слов с максимальными значениями коэффициента тематичности. Экспертам было предложено оценить, какие из предложенных тематичных слов могут быть выделены в качестве ключевых. Однокоренные слова считались за одно слово.

Приведем пример такой оценки для текста 4.

Таблица 30 *- повторная экспертная оценка для текста 4*

Топ 30 тематичных слов в выдаче	Эксперт 1	Эксперт 2	Эксперт 3	Эксперт 4	Общая оценка (> 70% согласия) КС да (+), нет(-)
звукоизоляция	+	+	+	+	+
гаджет	+	-	-	-	-
плацебо	+	-	+	+	+
ультразвук	+	+	+	+	+
стетоскоп	-	-	-	-	-
фокусировать	-	-	-	-	-
аудиозапись	+	+	+	-	+
какофония	+	-	+	+	+
контент	-	-	-	-	-
звукозапись	+	+	+	+	+
отдалить	-	-	-	-	-
расслабление	+	+	+	+	+
регенерация	+	-	+	+	+
гипертония	-	-	-	-	-
знать	-	-	-	-	-
загрязнять	-	-	-	-	-
усугублять	-	+	+	+	+
эколог	+	+	-	+	+
вибрация	+	-	+	+	+
навредить	+	+	+	-	+
раковый	-	-	+	-	-
коммуникативный	-	-	-	-	-
целебный	+	-	+	+	+
мегаполис	-	-	-	-	-

терапия	+	-	+	+	+
звук	+	+	+	+	+
наушник	+	+	+	+	+
экосистема	-	+	+	-	-
хирургия	-	-	-	-	-

Получив подобным образом оценки выданных алгоритмом тематичных слов, для каждого текста аналогично была посчитана точность выдачи. Таким образом, получились следующие значения по данной метрике:

Таблица 31 *- количественная оценка экспертных оценок по программной выдаче*

	Топ 30 слов по коэффициенту тематичности	**Количество совпадений с экспертными оценками**	**Точность**
Текст 1	30	20	66 %
Текст 2	30	18	60 %
Текст 3	30	22	73 %
Текст 4	30	17	56 %
Текст 5	30	24	80 %
Текст 6	30	21	70 %
Текст 7	30	17	56 %
Текст 8	30	21	70 %
Текст 9	30	19	63 %
Текст 10	30	16	53 %
Усредненная точность:			**65 %**

Итак, нам удалось поднять значение метрики до 65%. Тем не менее, этого недостаточно для утверждения о высокой эффективности алгоритма. Количественные методы оценивания естественного языка не позволяют в

полной мере охватить ситуативный контекст исследуемого явления. К тому же, количественные методы требуют строгости инструкций, которые сложно применить к ситуации с индивидуальным выбором ключевых слов. Данный метод оценки позволяет объективно оценить лишь общие, «нормальные» явления, но не справляются с оценкой отклонений от нормы, что подтверждается сильными различиями в оценках экспертов.

Качественная оценка полученных результатов

Качественная оценка также базировалась на оценках экспертов, но на этот раз экспертам требовалось оценить принадлежность выделенных алгоритмом ключевых слов к теме текста по шкале их тематичности и терминологичности. Под терминологичностью мы понимаем специфичное значение слова, то есть употребление данного слова в более узком значении, подчеркивающим значимость данного слова в конкретном тексте.

Таким образом, эксперт выставлял оценку 0, если конкретное слово не имело никакого отношения к теме, представленной в данном тексте; оценка 1 выставлялась, если слово отражало тему текста, но не отличалось специфичностью значения (треминологичностью); наконец, оценка 2 выставлялась, если слово имело более узкое, специфичное значение и было значимо для отражения темы текста.

Получив для каждого слова в тексте среднее значение «тематичности» по оценкам четырех экспертов, мы посчитали среднее арифметическое значение тематичности для всего текста. Усреднив полученные оценки, мы получили результат в 1,495, который составляет 75 % от максимального значения. Таким образом, результативность данного алгоритма по качественной методике оценивания составила 75%.

Чтобы получить объективную оценку работы описанного в данной работе алгоритма, мы взяли среднее значение между полученными оценками по статистическим и качественным показателям и получили значение в 70 %.

Нельзя утверждать, что этого значения достаточно для применения данного алгоритма к большему объему текстов, однако, при доработке программы в будущем увеличить эффективность метода представляется возможным.

Выводы

На основе полученных результатов мы можем сделать несколько выводов. Во-первых, процедура тематической атрибуции позволяет выделить три группы тематичных слов: ядерную и низкочастотную группу, в которую попадают самые специфичные в конкретном тексте слова, среднечастотную группу, куда также попадают высокотематичные слова, но не имеющие никакой специфики, то есть общие слова, характеризующие данную тему, и, наконец, высокочастотная группа, в которую попадают слова с минимальными значениями коэффициента тематичности.

Во-вторых, графики распределений слов по текстам в зависимости от их монотематичности или политематичности разнятся. Монотематичный текст представлен на графике однородным распределением монотонно возрастающей кривой в большинстве случаев с только одной четко выраженной модой или ее отсутствием. Для политематичного текста, наоборот, характерна неоднородность распределения, наличие нескольких ярко выраженных мод или их частое наложение друг на друга.

В-третьих, слова, отмеченные экспертами как ключевые или тематичные, в большинстве случаев попадают в низкочастотную и среднечастотную группы, что означает, что эксперты склонны выделять в качестве ключевых и тематически маркированных слов не столько специфичную лексику (термины) с низкой частотой встречаемости, сколько тематичные слова, которые не имеют какого-либо «специального» значения и довольно часто встречаются в тексте.

Коэффициент тематичности позволяет выделить наиболее специфичные, терминологичные, редкие слова, но частотные для данного

корпуса слова выносит на более низкие позиции в таблице, следовательно, необходимо доработать формулу подсчета данного коэффициента.

В слова, отмеченные экспертами как терминологичные, попали слова, действительно представляющие собой термины, характерные для определенных предметных областей: *частота, вибрация, ритм, темп, пульс, проводимость, амплитуда, диапазон, сенсор, ультразвук, стресс, гармоника, раздражитель*. Слова, такие как *музыка, звук, композиция, трек, мелодия* отмечались как тематичные слова без специфики значения.

Итак, разработанный нами алгоритм пока не может заменить методы выделения ключевых тематичных слов с помощью экспертных оценок, поэтому говорить об улучшенных результатах работы программы не представляется возможным. Однако, наш подход позволяет выделять тематичные слова, наделенные спецификой значения и обладающие терминологичностью. Таким образом, тема может быть представима не только при помощи общеупотребительной лексики с широкой семантикой, но и с помощью редкой терминологичной лексики, отражающей более узкое значение, а значит и позволяющей находить подтемы для общей темы.

Заключение

Данная работа посвящена актуальной проблеме поиска оптимального метода определения темы текста. Среди множества подходов нелегко найти универсальный алгоритм, применимый к любому типу текста.

В данном исследовании была предпринята попытка решения данной проблемы с помощью деривационного анализа. Результаты работы алгоритма показали необходимость доработки. Во-первых, перерассмотрения требует коэффициент тематичности, для рассчета которого необходимо будет учесть регулярность появления конкретного слова, чтобы высокое значение получали не только низкочастотные слова. Во-вторых, данная работа предполагает нахождение полной основы слова, следовательно, в перспективе дальнейшего исследования и попыток улучшения эффективности метода автор данной работы намерен учитывать не только суффиксы и флексии, но и префиксы. Эта задача осложнена малой дифференцируемостью префиксов для отдельных частей речи. К тому же, данная методика потребует составления списка слов исключений из-за наличия в языке слов, у которых префикс неотделим от корня.

В целом, примененные в данной работе методы могут послужить основой для разработки полноценного алгоритма тематической атрибуции.

Список литературы:

1. Гринев-Гриневич С.В. Терминоведение. – М., 2008. – С. 304
2. Бодуэн де Куртенэ И.А. «Об отношении русского письма к русскому языку» / И.А.Бодуэн де Куртенэ // Избранные труды по общему языкознанию. – Т. 2. – М.: Изд-во АН СССР, 1963
3. Головин Б.Н. Лингвистические основы учения о терминах. – М., 1987. – С. 103
4. Крушевский Н.В. Избранные работы по языкознанию. - М., 1998.
5. Адамец, П. Порядок слов в современном русском языке / П. Адамец. – Прага: Academia, 1966 – С. 20-39.
6. Адамец, П. Порядок слов в современном русском языке / П. Адамец. – Прага: Academia, 1966 – С. 20-39.
7. Виноградов В.В. Избранные труды. Исследования по русской грамматике. — М.: Наука, 1975.
8. Винокур Г. О. О некоторых явлениях словообразования в русской технической терминологии // Труды Московского института истории, философии и литературы. М.: ЛИТЕРА, 1939. Т. 5. Сборник статей по языковедению. С. 3-54.
9. Винокур Г.О. Заметки по русскому словообразованию, 1959
10. Воронцов К. В. Вероятностное тематическое моделирование. 2013. [Электронный ресурс URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf]
11. Земская Е. А., Кубрякова Е. С., Проблемы словообразования на современном этапе, «Вопросы языкознания», 1978, No 6;
12. И.В.Евсеева, Т.А.Лузгина, И.А.Славкина, Ф.В.Степанова. Современный русский язык: Курс лекций / И.В.Евсеева, Т.А.Лузгина, И.А.Славкина, Ф.В.Степанова; Под ред. И.А.Славкиной; Сибирский федеральный ун-т. - Красноярск,2007. – С. 642

13. Ковтунова, И.И. Современный русский язык. Порядок слов и актуальное членение предложения / И.И.Ковтунова – М.: Просвещение, 1976 – С. 239
14. Кольцов С.Н., Кольцова О.Ю., Митрофанова О.А., Шиморина А.С. Интерпретация семантических связей в текстах русскоязычного сегмента Живого Журнала на основе тематической модели LDA // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS–2014, Санкт-Петербург, 19 – 20 ноября 2014 г. СПб., 2014. С. 135–142
15. Крушельницкая, К.Г. К вопросу о смысловом членении предложения / К.Г.Крушельницкая // Вопросы языкознания. – 1956 – №5. – С. 55-67.
16. Кубрякова Е.С. Основы морфологического анализа (на материале германских языков). Файлы. Академическая и специальная литература. ... Издательство "Наука", Москва, 1974 – С. 324
17. Кубрякова Е.С. Основы морфологического анализа: На матер. германск.
18. Кузьмина И.А. Принципы морфемного членения слов в отечественной лингвистике, 2010
19. Курилович Е., Деривация лексическая и деривация синтаксическая, в его кн.: Очерки по лингвистике, М., 1962
20. Лопатин В.В. Русское словообразование и морфемика. Проблемы и принципы описания. - М.: Наука, 1977. - С. 41-62, 287-310, 109, 106-107.
21. Маслов Ю. С., О некоторых расхождениях в понимании термина «морфема», «Учёные записки ЛГУ», 1961, № 301, сер. филол. наук, в. 60, С. 140—52;
22. Михайлов М.А. Вопросы фонологического анализа: Выделимость основ и формантов. – М.: 1974. – С.8
23. Николаев Г.А. Проблемы теории словообразования в трудах Н.В. Крушевского // Николай Крушевский: научное наследие и современность. - Казань, 2001.

24. Николина Н.А. Вопросы словообразования в трудах И.А. Бодуэна де Куртенэ и Н.В. Крушевского.
25. Панфилов В.З. Грамматика и логика: Грамматическое и логико-грамматическое членение простого предложения. М. ; Л., 1963
26. Плунгян 2003 — В. А. Плунгян. Общая морфология: введение в пробле-

матику. М.: УРСС, 2000 (2 изд. 2003, 3 изд., испр. и доп. 2009).

Прикладной анализ текстовых данных на Python. Бенгфорт, Билбро, Охеда – С. 146-147

27. Распопов, И.П. Актуальное членение предложения: простого повествования преимущественно в монологической речи / И. П. Распопов.— Уфа : Изд-во Башк. ун-та, 1961 – С. 27-80
28. Савенкова, Е.Д. Этрусская морфемика: Опыт формал. моделирования / Е. Д. Савенкова; С.-Петерб. гос. ун-т. - СПб.
29. Соколова Г.Г. Транспозиция прилагательных и существительных. – М.: Высшая школа, 1973. – С. 175
30. Суперанская А. В. , Н. В. Подольская,Н.В. Васильева Общая терминология. Вопросы теории. 2012
31. Трубачев О. Н. Приемы семантической реконструкции / О. Н Трубачев // Сравнительно-историческое изучение языков разных семей. Теория лингвистической реконструкции. — М., 1988. — С. 197–222.
32. Фирбас, Я. Функция вопроса в процессе коммуникации / Я.Фирбас //Вопросы языкознания. – 1972. – №2.
33. Хасанов Э. Р. О специфике семантической и лексической деривации в современном русском языке Вестник Челябинского государственного университета, 2018. No 6 (416). Филологические науки. Вып. 113. С. 198—203. 4) Баранов А.Н. «Введение в прикладную лингвистику». – М., 2001. – С. 360

34. Хохлова М.В. Сопоставительный анализ статистических мер на примере частеречных преференций сочетаемости существительных//Компьютерная лингвистика и вычислительные онтологии. 2017
35. Шанский Н.М. Очерки по русскому словообразованию. - М., 1968. яз. – М.: 2008. – С. 41 – 43
36. Boost, K. Neue Untersuchingen zum Wesen and zur Struktur des deutschen Satzes [Text] / K. Boost. – Berlin: Akademie-Verlag, 1955. – C. 88
37. Dokulil, Miloš. Tvoření slov v češtině. 1, Teorie odvozování slov. Vyd. 1. Praha: Nakladatelství Československé akademie věd, 1962.
38. Daneˇs F. Functional sentence perspective and the organization of the text // Papers on functional sentence perspective. – Prague, 1974. – p. 106–128.
39. Danes F., A Three-Level Approach to Syntax. Travaux linguistiques de Prague (TLP), Academia, Prague, 1966, 1; 225-241
40. Firbas I. On some basic issiues of the theory of functional sentence perspective: comments on Alexander Szwedek`s critique //Brno studies in English. 1983. Vol.15. P. 9–36.
41. Indexing by latent semantic analysis / S. Deerwester, T. D. Susan, G. W. Furnas et al. // Journal of the American Society for Information Science. — 1990 — Vol. 41

Приложение 1

Топ-30 первых слов по коэффициенту тематичности после работы алгоритма

Текст 1			
слово	**частота_ЛШ**	**частота_корпус**	**thematicality**
трек	4,2	0,035830619	85,31099736
испытуемый	0,5	0,003257329	65,1465798
засыпание	0,6	0,003257329	54,2888165
метроном	0,6	0,003257329	54,2888165
расслаблять	2,5	0,013029316	52,11726384
симпатический	0,9	0,003257329	36,19254434
артериальный	2,8	0,009771987	34,89995347
усыплять	1,2	0,003257329	27,14440825
головоломка	1,6	0,003257329	20,35830619
минимум	1,7	0,003257329	19,16075877
замедляться	1,9	0,003257329	17,14383679
композиция	26,8	0,045602606	17,01589771
бессонница	7,9	0,013029316	16,49280501
прослушивание	4,3	0,006514658	15,1503674
подстраиваться	2,2	0,003257329	14,80604086
участница	4,6	0,006514658	14,16229996
сосудистый	2,5	0,003257329	13,02931596
вегетативный	2,7	0,003257329	12,06418145
расслабление	2,8	0,003257329	11,63331782
терапия	9,4	0,009771987	10,39573082
доработать	3,3	0,003257329	9,87069391
замедление	3,5	0,003257329	9,306654258
мелодия	25	0,022801303	9,120521173
сердечный	20,3	0,016286645	8,022977808
частота	22,3	0,016286645	7,303428229
настрой	6,4	0,003257329	5,089576547
пробуждение	6,6	0,003257329	4,935346955
ритм	30,6	0,013029316	4,257946392
снижать	17,9	0,006514658	3,639473732
Текст 2			
тактильный	0,6	0,008888889	148,1481481
гаджет	0,4	0,004444444	111,1111111
испытуемый	0,5	0,004444444	88,88888889
карканье	0,5	0,004444444	88,88888889
проводимость	2,6	0,017777778	68,37606838
живой	1,8	0,008888889	49,38271605
сенсор	0,9	0,004444444	49,38271605
заставка	1,3	0,004444444	34,18803419
фортепьяно	1,8	0,004444444	24,69135802

отрывистый	2,1	0,004444444	21,16402116
обрабатываться	2,4	0,004444444	18,51851852
гарнитур	2,7	0,004444444	16,46090535
расшифровываться	5,8	0,008888889	15,3256705
наименее	4,3	0,004444444	10,33591731
прослушивание	4,3	0,004444444	10,33591731
наушник	4,9	0,004444444	9,070294785
восстанавливаться	5,8	0,004444444	7,662835249
сигнализация	5,8	0,004444444	7,662835249
отклик	11,8	0,008888889	7,532956685
вибрация	6,1	0,004444444	7,285974499
крепиться	6,5	0,004444444	6,837606838
восприятие	39,6	0,026666667	6,734006734
перекрывать	6,6	0,004444444	6,734006734
визуальный	8,4	0,004444444	5,291005291
отворачиваться	8,4	0,004444444	5,291005291
кора	18,8	0,008888889	4,728132388
симфония	10,3	0,004444444	4,314994606
феномен	20,8	0,008888889	4,273504274
композитор	35,8	0,013333333	3,724394786
Текст 3			
урчание	1,1	0,009836066	89,41877794
сенсор	0,9	0,006557377	72,85974499
осязательный	0,6	0,003278689	54,64480874
расслаблять	2,5	0,009836066	39,3442623
пощекотать	1,1	0,003278689	29,80625931
кузнечик	4,8	0,013114754	27,32240437
питомец	3,9	0,009836066	25,22068096
виолончелист	1,3	0,003278689	25,22068096
ушной	1,7	0,003278689	19,28640309
мяукать	1,8	0,003278689	18,21493625
сердцебиение	1,8	0,003278689	18,21493625
вибрация	6,1	0,009836066	16,12469766
сверчок	2,1	0,003278689	15,6128025
фоновый	2,3	0,003278689	14,2551675
рецептор	4,8	0,006557377	13,66120219
обрабатываться	2,4	0,003278689	13,66120219
преобразоваться	2,5	0,003278689	13,1147541
мурлыкать	2,9	0,003278689	11,3058225
муравей	12,1	0,013114754	10,83863975
мембрана	6,6	0,006557377	9,935419771
усилитель	3,4	0,003278689	9,643201543
виолончель	3,5	0,003278689	9,367681499
крокодил	10,7	0,009836066	9,192584648
ветеринар	4	0,003278689	8,196721311
травмировать	4,4	0,003278689	7,451564829

ускорять	4,5	0,003278689	7,285974499
амплитуда	4,6	0,003278689	7,127583749
кошка	57,4	0,039344262	6,854401097
диапазон	19,7	0,013114754	6,657235583
Текст 4			
звукоизоляция	0,8	0,007894737	98,68421053
гаджет	0,4	0,002631579	65,78947368
плацебо	1,1	0,005263158	47,84688995
ультразвук	4,1	0,018421053	44,92939666
стетоскоп	0,6	0,002631579	43,85964912
фокусировать	0,7	0,002631579	37,59398496
аудиозапись	0,8	0,002631579	32,89473684
какофония	0,8	0,002631579	32,89473684
контент	0,8	0,002631579	32,89473684
звукозапись	1,4	0,002631579	18,79699248
отдалить	2,8	0,005263158	18,79699248
расслабление	2,8	0,005263158	18,79699248
регенерация	1,6	0,002631579	16,44736842
гипертония	2,2	0,002631579	11,96172249
знать	2,5	0,002631579	10,52631579
загрязнять	2,7	0,002631579	9,746588694
усугублять	2,8	0,002631579	9,398496241
эколог	2,9	0,002631579	9,074410163
вибрация	6,1	0,005263158	8,628127696
навредить	3,4	0,002631579	7,73993808
раковый	3,4	0,002631579	7,73993808
коммуникативный	3,6	0,002631579	7,30994152
целебный	4,6	0,002631579	5,720823799
мегаполис	4,7	0,002631579	5,599104143
терапия	9,4	0,005263158	5,599104143
звук	123,7	0,068421053	5,531208782
наушник	4,9	0,002631579	5,37056928
экосистема	5,5	0,002631579	4,784688995
хирургия	6,3	0,002631579	4,17710944
Текст 5			
живой	1,8	0,027874564	154,8586914
триггер	0,4	0,003484321	87,10801394
электрокардиограмма	0,4	0,003484321	87,10801394
испытуемый	0,5	0,003484321	69,68641115
проигрываться	0,5	0,003484321	69,68641115
саундтрек	0,7	0,003484321	49,77600796
аудиосистема	0,8	0,003484321	43,55400697
вживую	0,8	0,003484321	43,55400697
учащаться	0,9	0,003484321	38,71467286
прослушать	9,8	0,017421603	17,7771457
синтезатор	1,3	0,003484321	26,80246583

студийный	2,1	0,003484321	16,59200265
струнный	2,2	0,003484321	15,83782072
наушник	4,9	0,006968641	14,22171656
вдумчивый	3	0,003484321	11,61440186
микроскопический	3,7	0,003484321	9,417082588
слушатель	24,4	0,020905923	8,568001371
пульс	9,4	0,006968641	7,413447995
проделывать	5,4	0,003484321	6,452445477
акустический	5,6	0,003484321	6,222000996
продюсер	12	0,006968641	5,807200929
импровизация	6,5	0,003484321	5,360493165
музыкант	40,1	0,020905923	5,213447218
композиция	26,8	0,013937282	5,200478444
композитор	35,8	0,017421603	4,866369494
ассоциироваться	7,2	0,003484321	4,839334108
настройка	7,5	0,003484321	4,645760743
погружение	7,5	0,003484321	4,645760743
музыка	241,6	0,083623693	3,461245587
Текст 6			
щебетание	0,5	0,003571429	71,42857143
саундтрек	0,7	0,003571429	51,02040816
ритмичный	2,3	0,007142857	31,05590062
храп	3,7	0,010714286	28,95752896
знать	1,3	0,003571429	27,47252747
раздражительность	1,5	0,003571429	23,80952381
меломан	1,6	0,003571429	22,32142857
продуктивность	6	0,010714286	17,85714286
колыбельный	2,1	0,003571429	17,00680272
наушник	4,9	0,007142857	14,57725948
расслаблять	2,5	0,003571429	14,28571429
бессонница	7,9	0,010714286	13,56238698
негативно	5,4	0,007142857	13,22751323
ограждать	2,8	0,003571429	12,75510204
комфортный	6,8	0,007142857	10,50420168
приложение	28,2	0,025	8,865248227
динамик	4,3	0,003571429	8,305647841
прослушивание	4,3	0,003571429	8,305647841
датчик	10	0,007142857	7,142857143
звук	112,4	0,075	6,672597865
насладиться	5,4	0,003571429	6,613756614
отслеживать	5,5	0,003571429	6,493506494
сигнализация	5,8	0,003571429	6,157635468
планировка	6	0,003571429	5,952380952
вибрация	6,1	0,003571429	5,854800937
успокаивать	18,4	0,010714286	5,822981366
стресс	12,4	0,007142857	5,760368664

пробуждение	6,6	0,003571429	5,411255411
производительность	20,1	0,010714286	5,330490405
Текст 7			
испытуемый	0,5	0,003952569	79,0513834
ужастик	1,1	0,005928854	53,8986705
наиграть	0,4	0,001976285	49,40711462
озвучивание	0,4	0,001976285	49,40711462
фавн	0,4	0,001976285	49,40711462
понервничать	0,5	0,001976285	39,5256917
видеоряд	0,6	0,001976285	32,93807642
тиканье	0,7	0,001976285	28,23263693
озвучивать	3,4	0,007905138	23,25040688
сердцебиение	1,8	0,003952569	21,95871761
нелинейный	4,8	0,009467556	20,58629776
утрировать	1,1	0,001976285	17,9662235
животрепещущий	1,2	0,001976285	16,46903821
пробирать	1,2	0,001976285	16,46903821
гармоника	1,3	0,001976285	15,20218912
плеер	1,3	0,001976285	15,20218912
кинолента	1,4	0,001976285	14,11631846
подкидыш	1,4	0,001976285	14,11631846
спецэффект	1,4	0,001976285	14,11631846
оттачивать	1,5	0,001976285	13,17523057
головоломка	1,6	0,001976285	12,35177866
негодность	1,7	0,001976285	11,62520344
нагнетать	1,8	0,001976285	10,97935881
резня	1,8	0,001976285	10,97935881
ножевой	1,9	0,001976285	10,40149782
гуру	2	0,001976285	9,881422925
контроллер	2,3	0,001976285	8,592541674
додумать	2,4	0,001976285	8,234519104
аккомпанировать	2,5	0,001976285	7,90513834
раздражитель	2,6	0,001976285	7,601094558
Текст 8			
испытуемый	0,5	0,008219178	164,3835616
томография	0,4	0,005479452	136,9863014
гиппокамп	0,4	0,002739726	68,49315068
серотонин	0,4	0,002739726	68,49315068
сканирование	1,5	0,008219178	54,79452055
модальный	0,7	0,002739726	39,13894325
веселие	0,9	0,002739726	30,4414003
восприимчивый	2	0,005479452	27,39726027
подкорка	1	0,002739726	27,39726027
разочаровываться	1	0,002739726	27,39726027
мурашки	3,2	0,005479452	17,12328767
прослушать	9,8	0,016438356	16,77383282

резонансный	1,8	0,002739726	15,22070015
социализация	2	0,002739726	13,69863014
сенсорный	2,1	0,002739726	13,04631442
созвучие	2,1	0,002739726	13,04631442
кросс	2,2	0,002739726	12,45330012
релаксация	2,3	0,002739726	11,91185229
меланхолия	2,6	0,002739726	10,5374078
отслеживать	5,5	0,005479452	9,9626401
ассоциировать	9	0,008219178	9,132420091
осчастливить	3,1	0,002739726	8,837825895
расслабить	3,7	0,002739726	7,404664939
эмоция	32	0,021917808	6,849315068
мимика	4	0,002739726	6,849315068
мелодия	25	0,016438356	6,575342466
трек	4,2	0,002739726	6,523157208
гормон	9,7	0,005479452	5,648919644
музыка	241,6	0,123287671	5,102966525
Текст 9			
реципиент	0,8	0,005730659	71,63323782
веселие	0,9	0,005730659	63,67398918
теннисист	3,4	0,017191977	50,56463846
синхронизировать	0,7	0,00286533	40,93327876
фагот	0,8	0,00286533	35,81661891
глухой	2,7	0,008595989	31,83699459
медитативный	0,9	0,00286533	31,83699459
озвучание	0,9	0,00286533	31,83699459
виниловый	1	0,00286533	28,65329513
абстрагироваться	1	0,00286533	28,65329513
ракетка	3,3	0,008595989	26,04845012
раздражитель	2,6	0,005730659	22,04099625
знать	1,3	0,00286533	22,04099625
кларнет	1,5	0,00286533	19,10219675
меломан	1,6	0,00286533	17,90830946
живой	1,8	0,00286533	15,91849729
струнный	2,2	0,00286533	13,02422506
таинственность	2,4	0,00286533	11,93887297
формат	15,5	0,017191977	11,09159811
барабанить	2,7	0,00286533	10,61233153
распознавать	2,8	0,00286533	10,23331969
саксофон	3,6	0,00286533	7,959248647
теннис	12,4	0,008595989	6,932248822
феномен	20,8	0,014326648	6,887811329
визуальный	8,4	0,005730659	6,822213126
информатика	4,2	0,00286533	6,822213126
трек	4,2	0,00286533	6,822213126
динамик	4,3	0,00286533	6,663557007

прослушать	9,8	0,005730659	5,847611251
наушник	4,9	0,00286533	5,847611251
мелодия	25	0,011461318	4,584527221
Текст 10			
нейронаука	0,4	0,001788909	44,7227191
сигарный	0,5	0,001788909	35,7781753
чаровать	0,5	0,001788909	35,7781753
фрустрация	0,5	0,001788909	35,7781753
ноктюрн	0,6	0,001788909	29,8151461
одержимость	2	0,005366726	26,8336315
импровизатор	0,7	0,001788909	25,5558395
репродуктивный	2,2	0,005366726	24,3942104
музицирование	0,8	0,001788909	22,3613596
льстивый	0,9	0,001788909	19,8767641
помешательство	1,9	0,003577818	18,8306186
менструальный	1,1	0,001788909	16,262807
информативный	2,3	0,003577818	15,5557284
сыгранный	1,2	0,001788909	14,907573
усыплять	1,2	0,001788909	14,907573
рокер	1,2	0,001788909	14,907573
фуга	1,3	0,001788909	13,7608367
фурор	1,3	0,001788909	13,7608367
минор	1,4	0,001788909	12,7779198
инкрустировать	1,5	0,001788909	11,9260584
харизма	1,6	0,001788909	11,1806798
приравниваться	1,7	0,001788909	10,5229927
маркер	3,6	0,003577818	9,93838203
фортепианный	1,8	0,001788909	9,93838203
неосознанно	1,8	0,001788909	9,93838203
обуздать	1,9	0,001788909	9,41530929
прослушивание	4,3	0,003577818	8,32050589
экстаз	4,9	0,003577818	7,30166843
пианист	10,1	0,007155635	7,08478719
демонстративный	2,6	0,001788909	6,88041833

Приложение 2

Экспертные оценки ключевых слов

Текст 1

	музыка	**исследование**	**ученый**	**бессонница**	**расслабление**	**снижение**	**влияние**	**стресс**
эксперт 1	1	1	1	1	1	1	1	1
эксперт 2	1	1	0	1	1	1	0	1
эксперт 3	1	1	1	1	0	1	1	1

эксперт 4	0	1	0	0	1	0	1	1
Sketch Engine	1	0	0	1	1	1	0	1
Программа TF_IDF	0	0	1	1	0	1	0	1
Rutermextract	1	1	1	1	0	0	0	1
Процент совпадений	**71,4285714**	**71,4285714**	**57,142857**	**85,7142857**	**57,1428571**	**71,4285714**	**42,8571429**	**100**

	пробуждение	**эффект**	**ритм**	**давление**	**частота**	**тревога**	**комфорт**	**темп**
эксперт 1	1	0	0	0	0	0	0	0
эксперт 2	0	1	1	1	1	1	1	1
эксперт 3	1	0	1	0	0	0	0	1
эксперт 4	1	1	1	1	1	1	1	1
Sketch Engine	0	1	1	1	1	0	0	0
Программа TF_IDF	0	0	1	1	1	0	0	0
Rutermextract	0	0	1	1	1	0	0	0
Процент совпадений	**42,8**	**57,1**	**85,7**	**71,4**	**71,4**	**28,5**	**28,5**	**42,8**

	пульс	**мелодия**	**трек**	**композиция**	**нервная система**	**эксперимент**	**сердце**	**артериальный**
эксперт 1	0	0	0	0	0	0	0	0
эксперт 2	1	0	0	0	0	0	1	1
эксперт 3	0	1	1	0	1	0	1	0
эксперт 4	0	1	1	1	0	1	1	1
Sketch Engine	0	0	0	0	0	1	1	1
Программа TF_IDF	0	1	1	1	0	0	1	0
Rutermextract	0	0	0	1	0	0	1	1
Процент совпадений	**14,2**	**42,8**	**42,8**	**42,8**	**14,2**	**28,5**	**85,7**	**57,1**

Текст 2

	орган	**звук**	**синестезия**	**цвет**	**мозг**	**кость**	**проводимость**	**сигнал**
эксперт 1	1	1	1	1	1	1	1	1
эксперт 2	1	0	1	1	1	1	1	1
эксперт 3	1	0	1	0	0	1	1	0
эксперт 4	1	0	1	1	1	1	1	1

Sketch Engine	1	0	1	0	1	1	1	0
Программа TF-IDF	1	1	1	0	1	1	1	0
Rutermextract	1	1	1	0	1	1	1	1
Процент совпадений	**100**	**42,8**	**100**	**42,8**	**85,7**	**100**	**100**	**57,14**

	музыка	**хромостезия**	**ассоциация**	**информация**	**способность**	**синсет**	**восприятие**	**слух**
эксперт 1	0	0	0	0	0	0	0	0
эксперт 2	1	1	1	1	1	1	0	0
эксперт 3	0	1	0	0	0	0	0	1
эксперт 4	1	1	0	1	1	0	1	0
Sketch Engine	0	1	0	1	1	0	1	1
Программа	0	0	0	1	1	0	1	1
Rutermextract	0	0	0	1	1	0	1	1
Процент совпадений	28,5	57,14	16,6	71,4	71,4	14,2	57,14	57,1

	ухо	**зрение**	**сигнал**	**чувство**	**ощущение**	**слышать**	**участок**	**кора**
эксперт 1	0	0	0	1	1	1	0	0
эксперт 2	0	0	1	1	0	0	0	0
эксперт 3	1	1	0	1	0	0	0	0
эксперт 4	0	0	1	1	1	0	0	1
Sketch Engine	1	0	0	1	0	0	1	1
Программа	1	0	0	0	0	0	1	0
Rutermextract	1		0	1	0	1	1	1
Процент совпадений	**57,14**	**0**	**33,3**	**85,7**	**28,5**	**28,6**	**42,8**	**42,8**

Текст 3

	Звук	**ухо**	**диапазон**	**колебание**	**слух**	**канал**	**восприятие**	**музыка**
эксперт 1	1	1	1	1	1	1	1	1
эксперт 2	1	1	1	1	1	0	1	0
эксперт 3	1	1	1	0	1	1	1	1
эксперт 4	1	1	0	1	1	1	1	1
Sketch Engine	1	0	1	1	1	1	1	0
Программа	1	1	1	0	1	0	0	1
Rutermextract	1	1	0	0	1	0	1	1
Корреляция по совпадениям	100	85,71	71,4	57,1	100	57,1	85,7	71,4

	эмоции	мозг	импульс	вибрации	животные	темп	кузнечик	рыбы
эксперт 1	1	0	0	0	0	0	0	0
эксперт 2	1	1	1	1	1	1	0	0
эксперт 3	1	1	0	0	1	0	1	1
эксперт 4	1	0	0	1	1	1	1	0
Sketch Engine	0	0	0	0	1	0	1	1
Программа TF-IDF	0	0	0	0	1	0	1	0
Rutermextract	0	1	0	1	1	0	1	0
Корреляция по совпадениям	**57,1**	**42,85**	**14,2**	**42,8**	**85,7**	**28,5**	**71,4**	**28,5**

кошка	эксперимент	мурлыканье	человек	сенсор	отверстие	аудиоинформация	питомец	мурчание
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	1	1	0	0	1	0	0	1
1	1	1	1	1	1	1	0	0
1	0	0	1	1	1	1	0	1
1	0	0	0	0	1	0	1	1
1	1	1	1	0	0	0	1	1
71,42	**42,85**	**42,85**	**42,8**	**28,5**	**57,14**	**28,5**	**28,57**	**57,1**

Текст 4

	воздействие	шум	тишина	звук	звукозапись	здоровье	вред	расслабление	эффект
эксперт 1	1	1	1	1	1	1	1	0	0
эксперт 2	1	1	1	1	1	0	0	0	1
эксперт 3	1	1	1	1	1	1	1	1	1
эксперт 4	1	1	1	1	0	1	1	1	1
Sketch Engine	1	1	0	1	0	0	0	1	1
Программа TF-IDF	1	1	1	1	0	1	0	0	0
Rutermextract	0	1	0	1	0	1	0	0	0
Корреляция по совпадениям	**85,7**	**100,0**	**71,4**	**100,0**	**42,9**	**71,4**	**42,9**	**42,9**	**57,1**

	плацебо	восстановление	звукоизоляция	пациент	система	метод	настроение	негативный	ультразвук
эксперт 1	0	0	0	0	0	0	0	0	0
эксперт 2	0	0	1	0	0	0	1	1	1
эксперт 3	1	0	1	0	1	1	1	0	1
эксперт 4	1	0	1	0	1	1	0	1	0
Sketch Engine	1	1	1	1	1	0	1	0	1
Программа	0	1	1	1	0	0	0	0	1
Rutermextract	0	0	1	1	0	1	1	0	1
Корреляция по совпадениям	**42,9**	**28,6**	**85,7**	**42,9**	**42,9**	**42,9**	**57,1**	**28,6**	**71,4**

Текст 5

	музыка	живой	эмоции	слышать/слушать	эксперимент	влияние	положительный	композитор	восприятие
эксперт 1	1	1	1	1	0	1	1	0	1
эксперт 2	1	1	1	0	1	1	1	0	1
эксперт 3	1	1	0	1	1	0	0	1	0
эксперт 4	1	1	1	1	1	1	1	0	1
Sketch Engine	1	1	0	1	1	0	0	1	1
Программа TF-IDF	1	1	0	1	1	1	0	1	0
Rutermextract	1	1	1	1	0	0	0	1	1
Корреляция по совпадениям	**100**	**100**	**57,1**	**85,7**	**71,4**	**57,1**	**42,9**	**57,1**	**71,4**

	исполнитель	триггер	память	стресс
эксперт 1	1	0	0	0
эксперт 2	1	0	0	0
эксперт 3	1	0	0	0
эксперт 4	1	1	1	1
Sketch Engine	1	0	0	0
Программа	1	1	1	0
Rutermextract	1	0	0	0
Корреляция по совпадениям	**100,0**	**28,6**	**28,6**	**14,3**

Текст 6

	бессонница	**слушать**	**музыка**	**стресс**	**приложение**	**сон**	**храп**	**сигнал**	**наушник**
эксперт 1	0	1	1	0	1	1	0	1	1
эксперт 2	1	0	1	1	0	1	0	0	0
эксперт 3	1	0	1	1	1	1	1	0	1
эксперт 4	1	1	1	1	1	1	0	1	1
Sketch Engine	1	0	1	0	0	0	0	0	0
Программа TF-IDF	0	1	1	1	1	1	1	0	0
Rutermextract	1	0	1	0	1	1	0	0	0
Корреляция по совпадениям	**71,4**	**42,9**	**100,0**	**57,1**	**71,4**	**85,7**	**28,6**	**28,6**	**42,9**

	звук	**работать**	**шум**	**продуктивность**	**производительность**	**здоровье**	**положительный**	**влияние**	**повышать**
эксперт 1	1	0	1	1	0	1	1	1	1
эксперт 2	1	1	1	1	1	0	1	1	1
эксперт 3	1	1	1	0	0	1	0	0	0
эксперт 4	1	1	1	1	1	1	0	1	0
Sketch Engine	1	0	1	1	1	0	0	0	0
Программа	1	1	1	1	1	1	0	0	0
Rutermextract	1	1	1	1	0	0	0	0	0
Корреляция по совпадениям	**100,0**	**71,4**	**100,0**	**85,7**	**57,1**	**57,1**	**28,6**	**42,9**	**28,6**

Текст 7

	ужас	**атмосфера**	**звук**	**музыка**	**композитор**	**слух**	**хоррор**	**исследование**	**напряжение**
эксперт 1	1	1	1	1	1	1	0	1	1
эксперт 2	1	1	1	1	0	1	1	0	1
эксперт 3	1	1	1	1	0	1	1	0	0
эксперт 4	1	1	1	1	0	1	1	0	1
Sketch Engine	1	1	1	1	1	0	1	0	1
Программа TF-IDF	1	1	1	1	0	0	1	0	0
Rutermextract	1	1	1	1	0	1	0	1	0
Корреляция по совпадениям	**100**	**100**	**100**	**100**	**28,6**	**71,4**	**71,4**	**28,6**	**57,1**

	кино	фильм	игра	страх	нелинейный	частота	сердцебиение	страшный	пугать
эксперт 1	0	0	1	0	1	1	0	1	1
эксперт 2	1	0	1	1	1	0	1	1	1
эксперт 3	0	1	0	1	1	1	0	1	0
эксперт 4	1	1	1	1	1	1	1	1	1
Sketch Engine	0	1	0	1	1	0	1	1	1
Программа TF-IDF	0	1	1	1	1	0	0	1	1
Rutermextract	0	1	1	0	0	0	0	0	0
Корреляция по совпадениям	**28,6**	**71,4**	**71,4**	**71,4**	**85,7**	**42,9**	**42,9**	**85,7**	**71,4**

тишина	слышать	актер	озвучание	зловещий	нервный	окончания	эмоции	положительный	сказаться
0	0	1	1	1	1	1	0	1	1
1	0	0	1	0	0	0	1	0	0
1	1	1	1	0	1	1	1	1	1
0	1	0	1	0	1	1	1	1	1
0	0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
28,6	**42,9**	**28,6**	**71,4**	**28,6**	**42,9**	**42,9**	**42,9**	**42,9**	**42,9**

Текст 8

	музыка	область	мозг	влиять	социальный	связь	удовольствие	ассоциация	память
эксперт 1	1	1	1	1	1	1	1	1	1
эксперт 2	1	0	1	1	1	1	1	1	1
эксперт 3	0	1	1	1	1	1	1	0	1
эксперт 4	1	1	1	1	1	1	1	0	1
Sketch Engine	1	1	1	0	0	0	1	0	0
Программа	0	0	1	0	0	0	1	0	1
Rutermextract	1	1	1	0	0	0	0	0	1
Корреляция по совпадениям	**71,4**	**71,4**	**100,0**	**57,1**	**57,1**	**57,1**	**85,7**	**28,6**	**85,7**

	слушать	эксперимент	эмоции	Настроение	чувство	физический
эксперт 1	1	0	1	1	0	1
эксперт 2	0	0	1	1	1	0
эксперт 3	1	0	1	1	1	1
эксперт 4	1	1	1	1	0	1
Sketch Engine	1	1	0	0	0	0
Программа TF-IDF	0	0	1	0	0	0
Rutermextract	1	1	1	0	0	0
Корреляция по совпадениям	**71,4**	**42,9**	**85,7**	**57,1**	**28,6**	**42,9**

	форма	улучшать	испытуемый	опыт	сканирование	центр	песня
эксперт 1	1	1	0	0	0	1	0
эксперт 2	0	0	0	1	0	1	1
эксперт 3	1	1	0	0	0	1	1
эксперт 4	1	1	1	0	1	1	0
Sketch Engine	0	0	1	0	1	1	1
Программа TF-IDF	0	0	0	0	0	1	1
Rutermextract	0	0	0	0	1	0	0
Корреляция по совпадениям	**42,9**	**42,9**	**28,6**	**14,3**	**42,9**	**85,7**	**57,1**

Текст 9

	звук	база	данные	раздражитель	технис	эмоциональный (эмоция)	мелодия	феномен
эксперт 1	1	1	1	1	1	1	1	1
эксперт 2	1	1	1	1	1	1	1	1
эксперт 3	1	1	1	0	1	0	1	1
эксперт 4	1	1	1	1	1	1	1	1
Sketch Engine	1	0	0	1	1	0	1	1
Программа TF-IDF	1	1	1	1	1	1	1	1
Rutermextract	1	0	0	0	1	0	1	1
Корреляция по совпадениям	**100**	**71,4**	**71,42**	**71,42**	**100**	**57,1**	**100**	**100**

	музыка	**инструмент**	**слушать**	**звучать**	**восприятие**	**Иоганн Себастьян Бах**	**фильм**	**медитативный**	**ракетка**
эксперт 1	1	0	0	0	0	1	0	1	0
эксперт 2	1	1	1	1	1	0	0	1	0
эксперт 3	1	1	1	1	0	1	1	0	0
эксперт 4	1	0	1	1	1	1	1	0	0
Sketch Engine	1	1	0	1	0	0	0	0	1
Программа	1	0	0	1	0	1	0	1	0
Rutermextract	1	0	0	1	0	0	1	0	0
Корреляция по совпадениям	**100**	**42,8**	**42,8**	**85,7**	**28,5**	**57,14**	**42,85**	**42,8**	**14,3**

Текст 10

	Франц Лист	**листомания**	**феномен**	**чарующий**	**эффект**	**музыкант**	**нейронака**	**влиять**	**предпочтение**
эксперт 1	1	1	1	0	0	1	0	1	0
эксперт 2	1	1	1	0	0	1	1	1	1
эксперт 3	1	1	1	1	1	1	0	1	1
эксперт 4	1	1	1	0	1	1	1	1	1
Sketch Engine	1	1	0	0	1	1	0	0	0
Программа	1	1	0	0	0	1	0	0	1
Rutermextract	1	0	0	0	0	1	0	0	0
Корреляция по совпадениям	**100,0**	**85,7**	**57,1**	**14,3**	**42,9**	**100,0**	**28,6**	**57,1**	**57,1**

	женщина	**мужчина**	**пианист**	**композитор**	**ген**	**способность**	**партнер**	**черта**	**характер**
эксперт 1	0	0	1	0	0	0	1	0	0
эксперт 2	1	1	1	1	1	1	0	0	0
эксперт 3	1	0	1	0	1	1	1	1	1
эксперт 4	1	1	1	1	0	1	0	1	1
Sketch Engine	0	0	1	0	0	0	0	0	0
Программа	1	0	1	0	0	1	1	0	0
Rutermextract	1	0	0	0	0	1	1	0	0
Корреляция по	**71,4**	**28,6**	**85,7**	**28,6**	**28,6**	**71,4**	**57,1**	**28,6**	**28,6**

совпадения м									

	исследование	**экстаз**	**талант**	**привлекательность**	**качество**	**величайший**	**вкус**
эксперт 1	0	0	0	1	0	0	0
эксперт 2	0	1	1	1	0	1	1
эксперт 3	1	1	0	0	1	1	1
эксперт 4	1	1	1	1	1	0	1
Sketch Engine	0	0	1	0	0	0	1
Программа TF-IDF	0	0	0	1	0	0	0
Rutermextract	1	0	1	0	0	0	1
Корреляция по совпадениям	**42,9**	**42,9**	**57,1**	**57,1**	**28,6**	**28,6**	**71,4**