Исследование ассоциативных связей слов в корпусе социальных сетей с помощью дистрибутивно-семантических моделей

А.А. Антипенко[1, 2], О.А. Митрофанова[1]

[1] Санкт-Петербургский государственный университет, [2] Just AI

anne.morke@gmail.com, o.mitrofanova@spbu.ru

Аннотация

Данная статья отражает результаты эксперимента по автоматическому извлечению ассоциативных связей из корпуса русскоязычных текстов социальной сети Facebook с помощью алгоритмов и инструментов дистрибутивной семантики. Для лексем, выражающих ключевые понятия русскоязычной картины мира, автоматически извлечены ассоциаты из корпуса Facebook с использованием нейросетевых архитектур Word2Vec (CBOW и Skip-gram). Был проведен сопоставительный анализ полученных данных и данных Русского ассоциативного словаря и Русского дистрибутивного тезауруса. Полученные результаты позволяют провести лингвистический анализ языкового сознания современных пользователей социальных сетей.

Ключевые слова: дистрибутивная семантика, корпусная лингвистика, социальные сети, ассоциативный эксперимент, языковое сознание, русский язык

Библиографическая ссылка: Антипенко А.А., Митрофанова О.А. Исследование ассоциативных связей слов в корпусе социальных сетей с помощью дистрибутивно-семантических моделей // Компьютерная лингвистика и вычислительные онтологии. Выпуск 3 (Труды XXII Международной объединенной научной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19 – 22 июня 2019 г. Сборник научных трудов). — СПб: Университет ИТМО, 2019. С. 77 –91. DOI: 10.17586/2541-9781-2019-3-77-91

1. Введение

Традиционные исследования языкового сознания в психолингвистике основываются на свободном ассоциативном эксперименте, при котором группе испытуемых предъявляются слова-стимулы и фиксируются реакции. Полученные данные используются для создания ассоциативных словарей, которые представляют собой отражение языковой картины мира «стандартного» носителя языка.

Как пишет Н.В. Уфимцева, по ассоциативным реакциям «можно судить о речевой синтагматике» [1, с. 229]: велика вероятность совместной встречаемости ассоциатов в потоке речи. Можно предположить, что слова схожей семантики, встречающиеся в одних и тех же контекстах, связаны в языковом сознании человека, как и ассоциаты, а связи между ними сопоставимы со связями между стимулом и реакцией при ассоциативном эксперименте. Существующие инструменты анализа естественного языка позволяют выявлять семантически близкие слова, основываясь на векторных представлениях слов в дистрибутивно-семантических моделях. Целью нашего исследования стало автоматическое выделение ассоциативных связей ключевых концептов русскоязычной картины мира из корпуса постов социальной сети Facebook. Мы предлагаем решение,

основанное на методах корпусной лингвистики, психолингвистики и дистрибутивной семантики. Новизна нашего исследования заключается в адаптации методологии очной работы с испытуемыми к условиям исследования текстового общения носителей русского языка в социальных сетях, в получении новых данных о динамике языкового сознания носителей русского языка, основанных на корпусе постов Facebook, в верификации этих данных по данным лексикографических источников.

Материалом данного исследования являются тексты корпуса русскоязычного сегмента социальной сети Facebook, собранного в рамках проекта «Стресс, здоровье и психологическое благополучие в социальных сетях: кросс-культурное исследование». Авторы статьи выражают благодарность П.В. Паничевой за предоставление данных корпуса, консультации и помощь в работе.

2. Тексты Интернета и социальных сетей

В настоящее время исследователи имеют возможность работать с многомиллионными корпусами, в том числе и с массивами текстов различных типов, хранящихся в Интернете. Например, новостные корпусы могут использоваться для прогнозирования курсов валют, результатов выборов и других политических событий, отзывы о товарах и услугах – для извлечения тональных слов и конструкций, маркирующих положительную, отрицательную или нейтральную оценку, параллельные корпусы могут использоваться при создании систем машинного перевода, а также представлять интерес для исследователей-лингвистов, переводчиков, преподавателей и учащихся [2]. Тексты социальных сетей востребованы как источник данных не только для лингвистов, но и для психологов, социологов, специалистов по рекламе.

В рамках данного исследования наибольший интерес представляют социальные сети, поскольку сообщения в них могут организовывать диалогический (или полилогический) дискурс, что максимально приближает такие тексты к спонтанной речи и повышает их ценность для целей нашего исследования [3].

С развитием Интернета сформировались новые формы коммуникации, для которых характерно смешение черт устной и письменной речи. Такую промежуточную форму речи называют интернет-дискурсом [4]. Несмотря на существенное преобладание письменных текстов в Интернете, при анализе интернет-дискурса вопрос о соотношении устного и письменного дискурса является одним из ключевых. В то время как многие исследователи считают, что данные типы дискурса представляют собой континуум, в середине которого находится интернет-дискурс, выдвигаются предположения о том, что дискурс Интернета представляет собой отдельную форму, обладающую собственными характеристиками [5].

Таким образом, тексты социальных сетей представляют собой особый материал для исследования [6, 7], который позволяет выявить языковые признаки, о которых нельзя судить по другим письменным жанрам, при этом примеры коммуникации в сети доступны широкому кругу пользователей и исследователей, в отличие от устных текстов, фиксация и анализ которых сопряжены с некоторыми трудностями.

3. Ассоциативный эксперимент и ассоциативные словари русского языка

3.1. Ассоциативный эксперимент

Наше исследование предполагает автоматизированную процедуру, по существу воспроизводящую схему ассоциативного эксперимента, где в качестве стимулов рассматриваются слова, реализующие основные концепты русскоязычной картины мира, а реакции автоматически извлекаются из корпуса текстов социальных сетей, отражающих особенности полилогической интернет-коммуникации [8]. В данном случае

ассоциативные словари могут рассматриваться как источники эталонных данных об ассоциативных реакциях носителей русского языка. Ниже мы приводим краткое описание ассоциативного эксперимента и основные характеристики использованных ассоциативных словарей.

Основоположники психолингвистики исходили из предположения о том, что психику человека можно изучать через связь стимула и реакции. Так, исследуя ассоциативные связи, Ч. Осгуд выявил уровни порождения и восприятия речи, речь в его представлении также является реакцией на различные стимулы. Через ассоциативные связи Ч. Осгуд делал заключения о семантике слов, их коннотациях и месте в семантическом поле, об организации высказываний [9].

Ассоциативный эксперимент является основным методом исследования языкового сознания в психолингвистике [10]. При проведении эксперимента группе испытуемых предъявляются слова-стимулы и фиксируются слова-реакции. Полученные материалы позволяют делать выводы о картине мира носителей языка, о «семантических законах языка, о соотношении семантики и синтаксиса в речи и языке» [1, с. 228], что находит отражение в ассоциативных словарях. По данным ассоциативных словарей «можно судить об особенностях функционирования языкового сознания человека и способах построения речевого высказывания, обычно не осознаваемых носителями языка и не выявляемых другими способами исследования» [1, с. 229].

А.Р. Лурия выделяет две большие группы реакций: внешние (по смежности) и внутренние (по сходству/контрасту) ассоциативные связи, что примерно соответствует синтагматическим и парадигматическим отношениям между понятиями [11].

Н.В. Уфимцева предлагает выделять следующие типы реакций, чтобы проследить динамику развития языкового сознания в зависимости от возраста испытуемых, а также проанализировать реакции, которые возникают на стимулы, относящиеся к различным частям речи [1]: парадигматические; синтагматические; номинатор-операторные; ономатопеитические. К парадигматическим реакциям Н.В. Уфимцева относит слова из одного семантического поля, слова-синонимы и слова, описывающие какой-либо признак стимула; к синтагматическим – реакции, содержащие оценку, а также слова, отражающие синтаксическую сочетаемость стимула; к номинатор-операторным – дериваты; к ономатопеитическим – слова, характеризующиеся фонетическим подобием, реакции-рифмы.

Дж. Диз ввел понятие ассоциативного значения слова [12], которое выводится из реакций на данное слово, выступающее в качестве стимула. Ассоциативное значение слова отражает его семантику, особенности употребления, а также указывает на его место в ассоциативном поле. В ядре ассоциативного поля содержатся наиболее частотные, наиболее повторяемые реакции на слово-стимул.

Материалы ассоциативных словарей основываются на результатах массовых ассоциативных экспериментов, они представляют собой сеть понятий, связанных в сознании среднего носителя языка, и отражают образ мира и культуру определенной группы носителей. Особенностью ассоциативных словарей является то, что они не носят нормативный характер и не устанавливают правила, а отражают функционирование языка и описывают ту языковую картину, которая сложилась в действительности.

3.2. Словарь ассоциативных норм русского языка (САНРЯ)

Первым ассоциативным словарем русского языка является «Словарь ассоциативных норм русского языка» [13], который составлялся и обрабатывался в течение 1969−1972 гг. Словарь вышел в 1977 г., изначально в него планировалось включить 500 словарных статей, однако был издан только первый выпуск, в который вошло 196 слов-стимулов. При ассоциативном эксперименте испытуемыми являлись носители русского языка с высшим или неоконченным высшим образованием в возрасте от 16 до 50 лет.

3.3. Русский ассоциативный словарь (РАС)

Русский ассоциативный словарь (РАС) [14] представляет собой ассоциативно-вербальную сеть. Словарь состоит из двух частей: прямой словарь, где заглавными словами являются стимулы и представлены реакции на них, упорядоченные по частоте, и обратный словарь, где в качестве заглавных слов выступают реакции и приводятся стимулы, которыми они были вызваны. При создании РАС ассоциативный эксперимент проводился среди студентов-носителей русского языка в возрасте от 17 до 25 лет. В исходный набор стимулов вошли 1277 единиц, взятых из следующих источников: «Словарь ассоциативных норм русского языка» [13], «Частотный словарь русского языка» [15], «Русский семантический словарь» [16], для некоторых слов из списка были подобраны синонимы и антонимы, образованы дериваты, а также в список были включены некоторые предлоги, частицы, междометия, числительные, союзы, местоимения для наиболее системного отражения русской лексики.

3.4. Русский дистрибутивный тезаурус (РДТ)

Русский дистрибутивный тезаурус (РДТ) представляет собой первый свободно доступный дистрибутивный тезаурус русского языка [https://nlpub.ru/Russian_Distributional_Thesaurus]. Данный лингвистический ресурс покрывает около миллиона наиболее частотных слов русского языка и представляет значительный интерес для задач автоматической обработки текстов. РДТ основан на модели Skip-gram (Word2Vec), обученной на корпусе книг, собранных в электронной библиотеке lib.rus.ec, объемом 12.9 млрд с/у. При создании тезауруса над корпусом не производилось какой-либо предварительной обработки, за исключением токенизации, отсутствуют частеречная разметка, лемматизация или стемминг. Были созданы модели с различными параметрами, их точность оценивалась вручную на основе краудсорсинга. В соответствии с ручной оценкой, точность лучшей модели составляет 97.1% для первых пяти соседей и 91.2% для первых двадцати соседей. Для данной модели были заданы следующие параметры: контекстное окно размера 10, векторное пространство размерности 500, три итерации, минимальная частота слова в корпусе – 5. РДТ представляет собой граф подобия слов, который содержит почти 932 тыс. входов, 4,5 млн выходов и 194 млн семантических отношений. Полученные данные находятся в свободном доступе и могут быть использованы для различных лингвистических исследований [17].

4. Дистрибутивная семантика и дистрибутивные методы в лингвистике

Дистрибутивная семантика – отрасль лингвистики, которая позволяет исследовать значение слов и определять степень семантической близости слов на основании их распределения в корпусах текстов. Согласно дистрибутивной гипотезе, сформулированной З. Харрисом, слова, встречающиеся в одних и тех же контекстах, обладают схожим значением [18]. Существует сильная и слабая версия дистрибутивной гипотезы, которые рассматривают зависимость между распределением слова и его семантикой либо как корреляцию, либо как причинно-следственные отношения соответственно.

Один из наиболее распространенных подходов к моделированию лексического значения в дистрибутивно-семантических моделях – это построение многомерных векторов, отражающих значения слов, на основании совместной встречаемости в больших корпусах текстов [19]. Векторные представления слов представляют интерес с точки зрения когнитивного моделирования, поскольку они обладают сходством с наборами нейронов [20]. С помощью векторных моделей решаются такие задачи, как извлечение отношений и фактов, идентификация конструкций, назначение семантических ролей,

более того, это наиболее распространенный метод вычисления семантической близости слов [21].

В дистрибутивной семантике разрабатываются как синтагматические, так и парадигматические модели [22]. Одним из важных параметров, который необходимо учитывать в таких моделях, является размер контекстного окна. В качестве контекста можно рассматривать словосочетание, предложение или документ, при этом чем меньше размер контекстного окна, тем сильнее проблема разреженности данных влияет на результат. Строится матрица совместной встречаемости, в которой ячейки отражают частоту встречаемости слова в контексте. Имея такое представление данных, можно сравнить контекстные векторы слов таким образом, что слова, встретившиеся в одних и тех же контекстах, получат высокий коэффициент сходства.

На сегодняшний день наиболее распространенная дистрибутивно-семантическая модель, позволяющая определять синтагматические и парадигматические отношения в корпусе текстов, это Word2Vec [https://code.google.com/archive/p/Word2Vec/] – нейросетевая архитектура, позволяющая получать векторные представления слов, образующие многомерное векторное пространство. Для получения векторов используется машинное обучение. Векторы слов расположены в этом пространстве таким образом, что слова, встречающиеся в одних и тех же контекстах, находятся на близком расстоянии.

Word2Vec использует два алгоритма обучения: Continious Bag of Words (CBOW), предсказывающий слово по его окружению, и Skip-gram, предугадывающий по слову его окружение. Архитектура CBOW работает быстрее, однако модели, основанные на архитектуре Skip-gram, позволяют получить более точный результат, особенно для редких слов. Алгоритмы CBOW и Skip-gram были предложены Т. Миколовым и коллегами [23, 24]. Данные алгоритмы нацелены на минимизирование вычислительной сложности.

Существуют модели типа Word2Vec для русского языка, созданные в рамках проекта RusVectōrēs [https://rusvectores.org/ru/] на основе разнообразных корпусных источников (НКРЯ, Википедия, новостной корпус и т.д.) [25, 26]. Наряду с этим, возможно построение аналогичных моделей на специализированных корпусах, семантические отношения в которых являются предметом экспериментального исследования, что и было осуществлено в нашей работе.

5. Анализ экспериментальных данных об ассоциатах в корпусе социальных сетей

5.1. Корпусные данные и отбор лексем для эксперимента

Материалом данного исследования является корпус текстов русскоязычного сегмента социальной сети Facebook без метаданных о пользователях [27]. В исследовании приняли участие 8367 русскоговорящих пользователей Facebook. Данные были собраны в октябре 2016 года. Участники исследования предоставили согласие на обработку постов, находящихся в открытом доступе. 3973 пользователя (47%) написали более десяти постов, эти данные послужили материалом исследования, то есть в корпус вошли посты тех авторов, участие которых в интернет-дискурсе представлено не менее, чем десятью записями. Объем необработанных данных составил 28953525 с/у.

Была произведена предварительная обработка корпуса. Тексты всех постов были объединены в один файл. Лемматизация корпуса производилась программой MyStem [https://tech.yandex.ru/mystem/], разработанной компанией «Яндекс». Программа MyStem является морфологическим анализатором для русского языка, работающим на основе словаря, которая позволяет разрешать морфологическую неоднозначность. Также в программе MyStem есть возможность предсказания морфологической разметки и исходной формы слов, отсутствующих в словаре. Было принято решение исключить из корпуса 651015 лексем (2,2%), которые не встретились в словаре анализатора и о которых

были сделаны предположения при лемматизации, поскольку их число достаточно небольшое и не сказалось существенно на объеме корпуса.

В предобработке был использован словарь стоп-слов объемом 1104 слова, составленный на основе словарей служебной лексики и оборотов НКРЯ [28], в который вошли прежде всего междометия, предлоги, местоимения, частицы и союзы. Данным словам присуще в первую очередь грамматическое, а не лексическое значение, поэтому их можно не учитывать при анализе семантической близости слов. После удаления стоп-слов объем корпуса сократился на 46% и составил 15552981 с/у.

Сначала было принято решение отбирать слова на основании ключевых концептов, составляющих ядро языкового сознания носителей русского языка, выделенных Н.В. Уфимцевой [29]. Однако наиболее важным критерием отбора было наличие отобранных лексем во всех рассматриваемых источниках: РАС, РДТ, корпус текстов социальной сети Facebook; кроме того, некоторые из ключевых концептов попали в список стоп-слов, поэтому слова для исследования отбирались следующим образом.

Был составлен частотный список слов в корпусе текстов социальной сети Facebook после предварительной обработки и отобрана первая тысяча лексем. Для слов из начала списка проверялась частота по «Частотному словарю современного русского языка» [30]. Частота в данном словаре указана в ipm (instances per million – количество употреблений на миллион слов корпуса). Для двух отобранных слов – имен собственных Россия и Москва – данные о частоте в частотном словаре отсутствуют, но их можно восстановить по Национальному корпусу русского языка (НКРЯ). В НКРЯ есть данные об абсолютных частотах, чтобы получить частоту в ipm, нужно разделить абсолютные частоты на 92 – сумма орфографических слов корпуса, принятая за единицу вычисления ipm в словаре О.Н. Ляшевской. Таким образом, получается, что все отобранные лексемы являются достаточно частотными, диапазон частот в ipm – от 89,5 (находить) до 3727,5 (год), средняя частота составляет 618. Также проверялось наличие всех лексем из списка в РАС и РДТ, были оставлены те лексемы, которые есть во всех источниках.

В итоге для ассоциативного эксперимента был составлен список из 96 слов:

— 61 существительное (*человек, год, день, жизнь, друг, время, Россия, мир, ребенок, слово, дело, работа, страна, место, дом, город, любовь, рука, вопрос, женщина, душа, бог, утро, глаз, сила, народ, Москва, история, деньги, земля, война, ночь, голова, сердце, власть, лицо, конец, сторона, час, свет, мама, праздник, случай, вода, мужчина, отношение, часть, фильм, книга, семья, проблема, школа, вечер, путь, имя, право, результат, мысль, дорога, сын, язык*),
— 29 глаголов (*знать, говорить, хотеть, давать, понимать, любить, жить, думать, видеть, начинать, оставаться, приходить, иметь, смотреть, получать, находить, проходить, принимать, написать, ждать, пойти, стоять, помогать, решать, читать, отвечать, помнить, сидеть, называть*),
— 5 прилагательных (*новый, хороший, большой, добрый, нужный*),
— 1 наречие (*сегодня*).

5.1. Извлечение ассоциатов методами дистрибутивной семантики

В данной работе использовалась реализация Word2Vec в библиотеке gensim [https://radimrehurek.com/gensim/] через дистрибутив Anaconda [https://anaconda.org], включающий в себя ряд библиотек для анализа данных и машинного обучения. Программа была реализована на языке программирования Python (ver. 3.6) в среде разработки Spyder. Для извлечения ассоциатов было создано две модели при помощи инструмента Word2Vec: модель, основанная на алгоритме обучения CBOW (Continious Bag of Words) и модель, основанная на алгоритме обучения Skip-gram. При создании обеих моделей игнорировались слова, которые встретились в корпусе менее пяти раз. Размерность векторов составила 100, поскольку несмотря на то, что большие значения могут давать более точный результат, для них требуется очень большой объем исходных данных.

Использовалось контекстное окно размера 5, что обусловлено объемом корпуса, а также характером текстов – для текстов социальных сетей характерно изложение мыслей в краткой, сжатой форме.

Word2Vec позволяет сохранять созданные модели и использовать их для таких задач, как оценка сходства между словами в запросе, исключение лишнего слова из запроса, поиск наиболее близких по значению слов. В качестве запроса использовались слова, отобранные ранее, извлекались первые 20 близких слов. Для примера рассмотрим лексему «время», ср. таблицу 1.

Таблица 1. Ассоциаты для слова время (первые 20 близких слов)

Близкие слова в модели CBOW	Близкие слова в модели Skip-gram
десятилетие, 0.515	впустую', 0.705
момент', 0.511	документированный', 0.680
впустую', 0.503	фрагментарно', 0.675
период', 0.470	менеджерский', 0.662
прошлое', 0.445	отлагательство', 0.661
сутки', 0.443	наслаиваться', 0.658
промежуток', 0.427	скоротечный', 0.658
тратить', 0.413	лекционный', 0.658
черед', 0.413	мэнди', 0.657
мгновение', 0.407	непрерывно', 0.654
будущее', 0.400	текучка', 0.653
циферблат', 0.397	отсеваться', 0.651
постоянно', 0.397	отрезвление', 0.651
минута', 0.393	гринвич', 0.649
полгода', 0.383	загранпоездка', 0.648
плакальщица', 0.388	интервал', 0.647
эпоха', 0.382	прокручиваться', 0.647
час', 0.378	циклично', 0.646
быстро', 0.377	окукливаться', 0.646
недолго', 0.371	остепеняться', 0.646

5.2. Создание сравнительной таблицы ассоциаций

Для анализа результатов все полученные данные (близкие слова из моделей CBOW и Skip-gram, а также первые 20 слов-ассоциатов из РАС и РДТ) были собраны в сравнительную таблицу, в которой производился подсчет совпадений между ассоциатами из различных источников. Фрагмент таблицы представлен в таблице 2. Наборы слов сравнивались попарно: модель CBOW и РАС, модель Skip-gram и РАС, модель CBOW и РДТ, модель Skip-gram и РДТ. Были подсчитаны точные совпадения для всех лексем, а также коэффициенты, отражающие степень сходства между двумя наборами слов, для первых пятидесяти лексем.

Для подсчета коэффициентов ряды слов сравнивались попарно и приписывались следующие веса:

— точные совпадения – 5;
— дериваты (не являющиеся антонимами) – 4;
— синонимы – 3;
— гипонимы, гиперонимы, согипонимы, холонимы, меронимы (слова, связанные отношениями род/вид либо часть/целое) – 2;
— антонимы – 1.

5.3. Статистический и лингвистический анализ экспериментальных данных

Было подсчитано среднее число точных совпадений (ТС1), средний процент точных совпадений (ТС2) – их абсолютное и относительное значение, а также минимальное и

максимальное количество совпадений для каждой пары источников (разброс – Р). Данные о точных совпадениях представлены в Таблице 3.

Таблица 2. Фрагмент сравнительной таблицы для слова год

РАС	РДТ	CBOW	SG
високосный	месяц	полгода	полвека
рождение	полгода	месяц	полгода
Змея	день	полвека	помета
месяц	неделя	неделя	предвоенный
новый	период	десятилетие	годок
день	январь	столетие	ориентировочно
время	сентябрь	тысячелетие	месяц
век	осень	годик	виолка
Дракон	полвека	ежегодно	пятидесятилетие
Лошадь	апрель	го	хиджра
прошедший	декабрь	григорианский	зеленогорск
удачный	ежегодно	летний	семидесятилетие
назад	невисокосный	период	бобби
спустя	август	рождаться	григорианский
счастливый	июнь	родиться	проц
учебный	февраль	январь	трудодень
1988	полмесяца	календарный	дарина
долго	ноябрь	семестр	огайо
1989	июль	доживать	студенчество
длинный	май	каникулы	биофак

Таблица 3. Данные о точных совпадениях

	Среднее число точных совпадений (ТС1)	Средний процент точных совпадений (ТС2)	Разброс (Р)
CBOW&РАС	1,66	8,28	от 0 до 6
Skip-gram&РАС	0,98	4,9	от 0 до 3
CBOW&РДТ	3,81	19,06	от 0 до 10
Skip-gram&РДТ	2,92	14,58	от 0 до 10

Причины небольшого количества совпадений могут быть следующими:

— данные получены из разноплановых источников: данные РАС получены методом ассоциативного эксперимента при очной работе с испытуемыми, данные РДТ извлечены автоматически из корпуса книг, данные из корпуса постов Facebook также извлекались автоматически;

— использованные методы ориентированы на разные типы связей: наиболее разнообразные слова-реакции наблюдаются в РАС (синтагматические, парадигматические, фонетические, идиоматические, коннотативные), при извлечении методом CBOW выявляется больше синтагматических реакций, а при использовании метода Skip-gram – парадигматических, однако оба метода позволяют получить разные типы реакций в зависимости от части речи и лексического значения слов;

— ассоциации специфичны для хронологического среза и круга носителей языка: в РДТ представлены близкие слова из текстов авторов русской литературы, в РАС – ассоциации русских студентов в возрасте 17-25 лет 90-х годов, в корпусе Facebook – ассоциации современных носителей русского языка – пользователей социальных сетей.

Те совпадения, которые были зарегистрированы, отражают наиболее устойчивые синтагматические связи, как правило, отражающие лексическую сочетаемость. Учитывая условия, описанные выше, а также наличие довольно сильных связей при учете

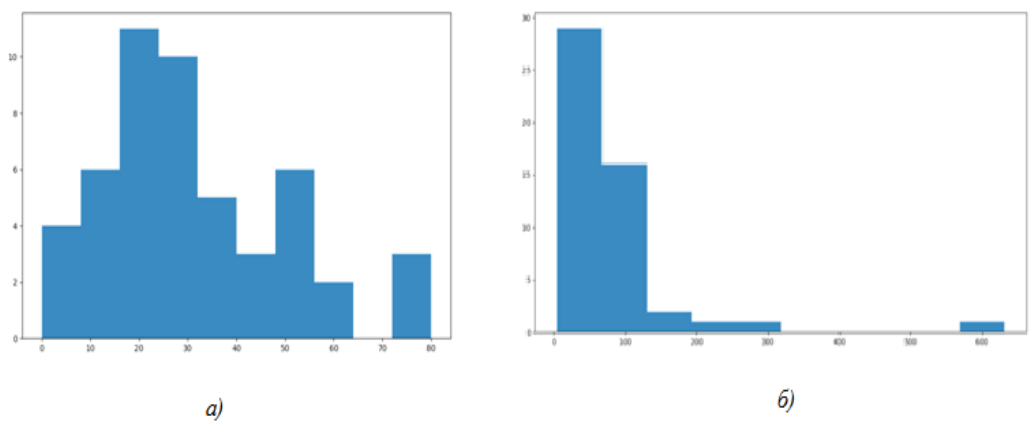
парадигматических отношений, можно сделать вывод, что совпадения не случайны – они информативны и находятся в ядре языкового сознания. Например, во всех источниках есть такие связи, как хороший – отличный, хороший – плохой, любить – ненавидеть, ребенок – взрослый, друг – закадычный, Россия – страна, видеть – слышать, женщина – девушка.

Для коэффициентов, отражающих силу связи (КС), были посчитаны медиана – М1 –и мода – М2, а также построены гистограммы распределений значений по частоте, где на оси x представлены значения коэффициентов, а на оси y – их абсолютные частоты. Гистограммы позволяют увидеть силу связи между ассоциатами из различных источников, таким образом, подтверждается гипотеза о том, что данные сопоставимы, то есть между ними существует связь. Наиболее частотные значения коэффициентов, отраженные на графиках, соответствуют полученным значениям М2, что позволяет судить о приблизительной средней силе связи. Также на гистограммах наблюдаются выбросы – максимальные значения коэффициентов, встречающиеся с минимальной частотой. Данные значения обусловлены специфичностью связей отдельных лексем, связанной с их лексическим значением: наиболее высокие коэффициенты характерны для имени-собственного Москва, с которым связаны названия других русских городов – Санкт-Петербург, Иркутск, Воронеж, Омск, Челябинск, Екатеринбург, Волгоград, Ярославль, Красноярск и др. (см. Таблица 4, Рисунки 1, 2).

В результате статистического анализа совпадений между рядами ассоциатов, полученных из разных источников, выяснилось, что процент расхождений достаточно велик. Легко заметить, что, как при учете точных совпадений, так и при анализе коэффициентов, отражающих силу связи, сходство и данных РАС, и данных РДТ с моделью, основанной на алгоритме CBOW, примерно вдвое выше, чем с моделью, основанной на алгоритме Skip-gram.

Таблица 4. Данные о силе связи

	Медиана (М1)	**Мода (М2)**
CBOW&РАС	27,5	21,45
Skip-gram&РАС	15,5	11,32
CBOW&РДТ	63	39,93
Skip-gram&РДТ	41,5	23,22



а) *б)*

Рис. 1. Распределение коэффициентов для модели CBOW и *а)* РАС, *б)* РДТ

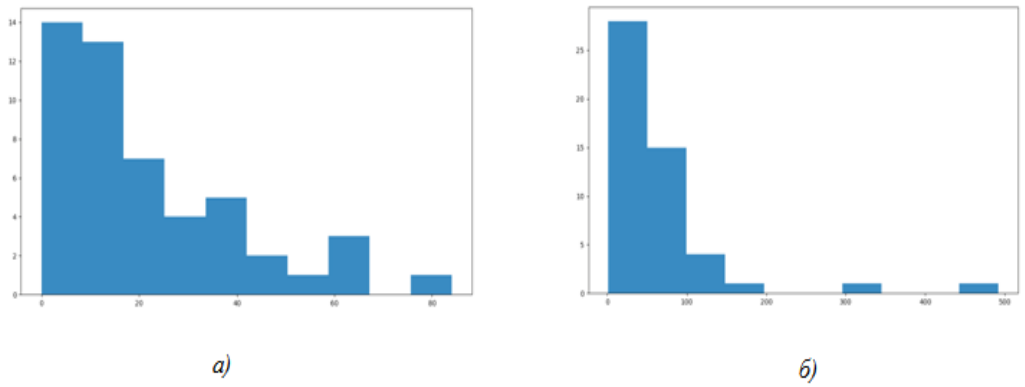


Рис. 2. Распределение коэффициентов для Skip-gram и *а)* РАС, *б)* РДТ

Поскольку размер корпуса относительно небольшой, при использовании архитектуры CBOW, когда рассматриваются многие слова из контекста для предсказания одного слова, происходит сглаживание распределения, то есть данный метод работает как регуляризация. При этом сходство обеих моделей с РДТ существенно выше, чем с РАС. Это обусловлено тем, что для создания РДТ использовался инструмент Word2Vec, как и для моделей корпуса Facebook, тогда как при создании РАС была задействована совершенно иная методика – ассоциативный эксперимент.

С лингвистической точки зрения интерес представляют типы связей, как между словом-стимулом и полученными для него лексемами-ассоциатами, так и между наборами ассоциатов из различных источников. В данных из всех источников наблюдаются как синтагматические, так и парадигматические связи. Тот факт, что данные РАС сильно разнятся с данными других источников, обусловлен спецификой реакций, получаемых при ассоциативном эксперименте.

Например, в РАС встречаются устойчивые фразеологические реакции (принимать – близко к сердцу, жизнь – прожить не поле перейти, мир – труд май, любовь – с первого взгляда, бог – весть что, сила – есть ума не надо, ждать – у моря погоды), а также реакции-словосочетания (время – друг без друга, мир – дому твоему, сегодня – в клубе танцы, понимать – друг друга, слово – о полку Игореве, место – встречи изменить нельзя, вопрос – загнавший в угол, женщина – которая поет, оставаться – самим собой), которые в рамках нашего эксперимента невозможно сопоставить с данными, полученными автоматически.

Сходство лексем, извлеченных из корпуса Facebook моделями CBOW и Skip-gram, с данными РАС для существительных выше, чем для глаголов и прилагательных, поскольку при анализе существительных наблюдается наибольший разброс различных типов синтагматических и парадигматических реакций и в РАС, и в полученных моделях, тогда как для глаголов и прилагательных характерно преобладание синтагматических реакций в РАС и парадигматических реакций в моделях.

В РДТ соотношение синтагматических и парадигматических соседей сходно с соотношением в моделях, поэтому сила связи соседей в моделях для прилагательных и глаголов с соседями в РДТ всегда больше, чем с ассоциатами в РАС, однако при сравнении существительных обнаруживается большой разброс результатов, и связь ближайших слов в моделях может быть сильнее как с реакциями в РАС, так и с соседями в РДТ. Фрагмент данных приведен в таблице 5.

Таблица 5. Сравнение типов реакций для существительного *жизнь*, глагола *говорить* и прилагательного *большой* по данным разных источников

	РАС и CBOW		**РАС и Skip-gram**		**РДТ и CBOW**		**РДТ и Skip-gram**	
	ТС1	**КС**	**ТС1**	**КС**	**ТС1**	**КС**	**ТС1**	**КС**
жизнь	0	7	0	6	1	28	0	13
говорить	2	21	0	3	8	66	4	30
большой	3	72	3	84	10	289	10	340

5.4. Анализ данных социальной сети Facebook

В условиях применения различных методов получения ассоциатов, использования данных из различных источников, наличия информативных совпадений, описанных в предыдущем разделе, можно предположить, что полученные различия обусловлены не только различиями методов и лингвистических источников, но и динамикой языкового сознания, то есть по близким словам, извлеченным из корпуса текстов социальной сети Facebook, можно судить о языковом сознании современных пользователей социальных сетей. Поскольку объем данных был относительно небольшой, модель, основанная на архитектуре CBOW, сработала лучше и дала большее число совпадений с РАС. Рассмотрим совпадения и расхождения данных РАС и результатов работы этой модели.

Наибольшее число совпадений – 20−30% – наблюдается для следующих лексем: сын (дочь, отец, мать, старший, блудный, брат), результат (итог, оценка, анализ, положительный), сидеть (стоять, лежать, диван, кресло), нужный (необходимый, важный, ненужный, полезный). Это указывает на существование устойчивых синтагматических (в меньшей степени) и парадигматических (в большей степени) связей, присутствующих в различных источниках. Больше всего совпадений зарегистрировано для слова сын, что свидетельствует об устойчивости, стабильности и широком употреблении лексико-семантической группы родственных отношений, а также наблюдается идиоматизированная связь сын – блудный.

Для значительной части лексем обнаруживается совпадение на 15% (человек, друг, хороший, сегодня, ребенок, слово, большой, любовь, женщина, бог, утро, сила, история и др.) и на 10% (время, знать, говорить, работа, жить, страна, город, оставаться, душа, война, сердце, лицо, стоять и др.). Наиболее устойчивы такие парадигматические связи слова хороший, как плохой, отличный, приятный, при этом в РАС наблюдается преобладание синтагматических реакций (человек, друг, муж, товарищ, ученик, фильм, поступок и др.), а в данных социальной сети Facebook преобладают синонимы, среди которых встречаются слова разговорного стиля, характерные для общения в социальных сетях (замечательный, классный, прекрасный, позитивный, чудесный, идеальный, великолепный, круто и др.). При сравнении данных для слова страна выявляются следующие пересечения: государство, Россия. В РАС присутствует лексика, связанная с советской эпохой (советов, советский, СССР), тогда как в данных корпуса Facebook наблюдаются связи, которые могут свидетельствовать о мировой глобализации и о взаимодействии государств (Америка, Евросоюз, ЕС, США, Европа, Латвия, цивилизованный). Интерес представляют связи слова женщина, общими являются ассоциаты мужчина, девушка, жена. В РАС присутствует большое количество реакций с положительной коннотацией (красивая, молодая, милая, умная, добрая, любимая, привлекательная, интересная, мудрая), тогда как в данных корпуса Facebook большинство связанных лексем нейтральны, присутствуют слова, которые стали более широко употребляться в последние годы (замужний, женатый, феминистка, блондинка, беременный, сексуально, сорокалетний).

О динамике языкового сознания также свидетельствует наличие в данных корпуса Facebook новой лексики, обусловленное изменениями, происходящими в языке, и возникновением новых реалий и понятий: искать – авито, получать – шенген, деньги –

ипотека, друг – роднуля, работа – офис, клиент, дом – съемный, город – мегаполис, вопрос – респондент, получать – перечислять, принимать – пролонгация, написать – эссе, власть – оппозиция, олигарх, коррупционер, фильм – трейлер, язык – субтитр.

6. Заключение

В данной работе было проведено исследование по автоматическому извлечению ассоциативных связей для слов, выражающих ключевые концепты русскоязычной картины мира, из корпуса текстов социальной сети Facebook. Для извлечения близких по значению слов использовался инструмент для автоматического анализа естественного языка Word2Vec, основанный на нейронных сетях. Было реализовано два варианта алгоритма: с использованием архитектуры Continious Bag of Words и архитектуры Skip-gram. Полученные данные сравнивались с данными РАС и РДТ.

Было выявлено существенное расхождение данных, полученных из различных источников с использованием различных методов, однако были обнаружены совпадения, являющиеся информативными. В условиях эксперимента можно считать, что расхождения обусловлены не только методологическими различиями, но и динамикой языкового сознания носителей русского языка.

В ходе исследования были получены и обработаны эмпирические данные, свидетельствующие о динамике языкового сознания носителей русского языка и о ядре языкового сознания современных пользователей социальных сетей.

Литература

[1] Уфимцева Н.В. Языковое сознание: динамика и вариативность. М.: Институт языкознания РАН, 2011.

[2] Анализ социальных сетей: методы и приложения / Коршунов А.В. [и др.] // Труды Института системного программирования РАН. 2014. Т. 26. Вып. 1. С. 439−456.

[3] Матусевич А.А. Общение в социальных сетях: прагматический, коммуникативный, лингвостилистический аспекты характеристики: Дисс. ... канд. филол. наук. Киров, 2016.

[4] Ушаков А.А. Интернет-дискурс как особый тип речи // Вестник Адыгейского государственного университета. Серия 2: Филология и искусствоведение. 2010. № 4. С.170-174.

[5] Crystal D. Language and the Internet. Cambridge, UK: Cambridge University Press, 2006.

[6] Hogan B. Analyzing Social Networks via the Internet // The SAGE Handbook of Online Research Methods: SAGE Publications. 2008. P. 141−160.

[7] Pérez-Sabater C. The Linguistics of Social Networking: A Study of Writing Conventions on Facebook // Linguistik Online. 2012. Вып. 56 (6). P. 81−93. URL: http://linguistik-online.net/56_12/perez-sabater.html.

[8] 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.

[9] Psycholinguistics: A Survey of Theory and Research Problems / Osgood Ch. [et al.]. Waverly Press, 1954.

[10] Ягунова Е.В. Эксперимент в психолингвистике: Конспекты лекций и методические рекомендации. Учебное пособие для вузов. СПб: Издательство «Остров», 2005.

[11] Лурия А.Р. Язык и сознание. / Под ред. Е.Д. Хомской. М.: Изд-во Московского университета, 1979.

[12] Deese J. The Structure of Associations in Language and Thought. Baltimore: The Johns Hopkins Press, 1965.
[13] Словарь ассоциативных норм русского языка / А.А. Леонтьев [и др.]. М.: Издательство Московского университета, 1977.
[14] Русский ассоциативный словарь: в 4 т. / Ю.Н. Караулов [и др.]. М., 1994–1996. Т. 1.
[15] Частотный словарь русского языка / под ред. Л.Н. Засориной. М.: Русский язык, 1977.
[16] Русский семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову / Ю.Н. Караулов [и др.]. М.: Наука, 1983.
[17] Human and Machine Judgements for Russian Semantic Relatedness / Panchenko A. et al. // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts: AIST–2016. Communications in Computer and Information Science. Springer, Cham, 2017. Vol. 661.
[18] Harris Z. Distributional Structure // Word. 1954. № 10(23). P. 146−162.
[19] Baroni M., Lenci A. Distributional Memory: A General Framework for Corpus-Based Semantics // Computational Linguistics. 2010. Vol. 36(4). P. 673–721.
[20] Rohde D., Gonnerman L., Plaut D. An Improved Model of Semantic Similarity Based on Lexical Co-occurrence // Communications of the ACM. 2006. № 8. P. 627–633.
[21] Jurafsky D., Martin H. Speech and Language Processing (Third Edition Draft). 2017. URL: https://web.stanford.edu/~jurafsky/slp3/.
[22] Sahlgren M. The Distributional Hypothesis. From Context to Leaning // Distributional Models of the Lexicon in Linguistics and Cognitive Science (Special Issue of the Italian Journal of Linguistics). Rivista di Linguistica. 2008. Vol. 20(1). P. 33−53.
[23] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. 2013. URL: http://arxiv.org/abs /1301.3781.
[24] Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT. 2013. P. 746-751.
[25]Kutuzov A., Kuzmenko EComparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: The Case for Russian // A. Gelbukh (ed.) CICLing 2015. Part I. Springer LNCS 9041. 2015. P. 47–58.
[26]Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Springer, Cham. 2017. Vol. 661. P. 155-161. DOI: 10.1007/978-3-319-52920-2_15.
[27] Panicheva P., Erofeeva A., Ledovaya Ja. Semantic Feature Aggregation for Gender Identification in Russian Facebook // Artificial Intelligence and Natural Language 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers. Communications in Computer and Information Science. Springer, 2017. Vol. 789. P. 3–15.
[28] Митрофанова О.А. Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика−2015». СПб.: Издательство Санкт-Петербургского университета. 2015. С. 332–343.
[29] Уфимцева Н.В. Образ мира русских: системность и содержание // Язык и культура. М., 2009. С. 98–111.
[30] Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

The Study of Word Associations in the Social Networks Corpus by means of Distributional Semantics Models

A.A. Antipenko[1, 2], O.A. Mitrofanova[1]

[1] St. Petersburg State University, [2] Just AI

The paper discusses results of the experiment on automatic extraction of associative relations from the corpus of Russian texts from Facebook social network. Experiments were carried out with the help of algorithms and tools of Distributional Semantics. We extracted associations for lexemes expressing key concepts of Russian-specific world view. The procedure was performed by means of Word2Vec (CBOW and Skip-gram) neural network architectures. We carried out linguistic analysis of the output data and compared it with the associations described in the Russian Associative Dictionary and the Russian Distributional Thesaurus. Results achieved in course of experiments allow to make conclusions on the dynamic of Russian-specific language consciousness of contemporary social network users.

Keywords: Distributional Semantics, Corpus Linguistics, social networks, associative experiment, language consciousness, Russian

Reference for citation: Antipenko A.A., Mitrofanova O.A. The Study of Word Associations in the Social Networks Corpus by means of Distributional Semantics Models // Computer Linguistics and Computing Ontologies. Vol. 3 (Proceedings of the XXII International Joint Scientific Conference «Internet and Modern Society», IMS-2019, St. Petersburg, June 19-22, 2019). - St. Petersburg: ITMO University, 2019. P. 77 – 91. DOI: 10.17586/2541-9781-2019-3-77-91

References

[1] Ufimtseva N.V. Jazykovoje soznanije: dinamika i variativnost. M.: Institut jazykoznanija RAN, 2011.
[2] Analiz sotsialnyh setej: metody i prilozhenija / Korshunov A.V. [i dr.] // Trudy Instituta sistemnogo programmirovanija RAN. T. 26. Vyp. 1. 2014. P. 439−456.
[3] Matusevich A.A. Obschenije v sotsialnyh setyah: pragmaticheskij, kommunikativnyj, lingvostilisticheskij aspekty harakteristiki: Diss. … kand. filol. nauk. Kirov, 2016.
[4] Ushakov A.A. Internet-diskurs kak osobyj tip rechi // Vestnik Adygejskogo gosudarstvennogo universiteta. Serija 2: Filologija i iskusstvovedenije. № 4. Adygejsk, 2010.
[5] Crystal D. Language and the Internet. Cambridge, UK: Cambridge University Press, 2006.
[6] Hogan B. Analyzing Social Networks via the Internet // The SAGE Handbook of Online Research Methods: SAGE Publications, 2008. P. 141−160.
[7] Pérez-Sabater C. The Linguistics of Social Networking: A Study of Writing Conventions on Facebook // Linguistik Online. Issue 56 (6), 2012. P. 81−93. URL: http://linguistik-online.net/56_12/perez-sabater.html.
[8] Koltsov S.N., Koltsova O.Ju., Mitrofanova O.A., Shimorina A.S. Interpretatsija semanticheskih svyazej v tekstah russkojazychnogo segmenta Jzivogo Zhurnala na osnove tematicheskoj modeli LDA // Tehnologii informatsionnogo obschestva v nauke, obrazovanii i culture: sbornik nauchnyh statej. Materialy XVII Vserossijskoj objedinennoj konfrencii «Internet i sovremennoje obschestvo» IMS–2014, Sankt-Peterburg, 19–20 nojabrya 2014 r. SPb, 2014. P. 135–142.
[9] Psycholinguistics: A Survey of Theory and Research Problems / Osgood Ch. [et al.]. Waverly Press, 1954.

[10] Iagounova E.V. Eksperiment v psiholingvistike: Konspekty lektsij i metodicheskije rekomendacii. Uchebnoje posobije dlya vyzov. SPb.: Izdatelstvo «Ostrov», 2005.
[11] Lurija A.R. Jazyk i soznanije. / Pod red. E.D. Homskoj. M.: Izdatelstvo Moskovskogo universiteta, 1979.
[12] Deese J. The Structure of Associations in Language and Thought. Baltimore: The Johns Hopkins Press, 1965.
[13] Slovar assotsiativnyh norm russkogo jazyka / A.A. Leontjev [i dr.]. M.: Izdatelstvo Moskovskogo universiteta, 1977.
[14] Russkij assotsiativnyj slovar: v 4 t. / Ju.N. Karaulov [i dr.]. M., 1994–1996. T. 1.
[15] Chastotnyj slovar ruskogo jazyka / pod red. L.N. Zasorinoj. M.: Russkij jazyk, 1977.
[16] Ruskij semanticheskij slovar: Opyt avtomaticheskogo postrojenija tezaurusa: ot ponyatija k slovu / Ju.N. Karaulov [i dr.]. M., Nauka, 1983.
[17] Human and Machine Judgements for Russian Semantic Relatedness / Panchenko A. et al. // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts: AIST–2016. Communications in Computer and Information Science. Vol. 661. Springer, Cham, 2017.
[18] Harris Z. Distributional Structure // Word. № 10(23), 1954. P. 146−162.
[19] Baroni M., Lenci A. Distributional Memory: A General Framework for Corpus-Based Semantics // Computational Linguistics. Vol. 36(4). 2010. P. 673–721.
[20] Rohde D., Gonnerman L., Plaut D. An Improved Model of Semantic Similarity Based on Lexical Co-occurrence // Communications of the ACM. № 8. 2006. P. 627–633.
[21] Jurafsky D., Martin H. Speech and Language Processing (Third Edition Draft). 2017. URL: https://web.stanford.edu/~jurafsky/slp3/.
[22] Sahlgren M. The Distributional Hypothesis. From Context to Leaning // Distributional Models of the Lexicon in Linguistics and Cognitive Science (Special Issue of the Italian Journal of Linguistics). Rivista di Linguistica. 2008. Vol. 20(1). P. 33−53.
[23] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. 2013a.
[24] Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT. 2013b.
[25] Kutuzov A., Kuzmenko E. (2015) Comparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: The Case for Russian // A. Gelbukh (ed.) CICLing 2015. Part I. Springer LNCS 9041. P. 47–58.
[26] Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Springer, Cham. Vol. 661. P. 155-161. DOI: 10.1007/978-3-319-52920-2_15.
[27] Panicheva P., Erofeeva A., Ledovaya Ja. Semantic Feature Aggregation for Gender Identification in Russian Facebook // Artificial Intelligence and Natural Language 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers. Communications in Computer and Information Science. Vol. 789. Springer, 2017. P. 3–15.
[28] Mitrofanova O.A. Verojatnostnoje modelirovanije tematiki russkojazychnyh korpusov tekstov s ispolzovanijem kompjuternogo instrumenta GenSim // Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika−2015». SPb.: Izdatelstvo Sankt-Peterburgskogo universiteta, 2015. P. 332–343.
[29] Ufimtseva N.V. Obraz mira russkih: sistemnost i soderzhanije // Jazyk i kultura. M., 2009. P. 98–111.
[30] Lyashevskaja O.N., Sharov S.A. Chastotnyj slovar sovremennogo russkogo jazyka (na materialah Natsionalnogo korpusa russkogo jazyka). M.: Azbukovnik, 2009.