

Кочетова Л. А. Исследование аксиологических доминант в жанре пресс-релиза на основе методов автоматического извлечения ключевых слов корпуса текстов / Л. А. Кочетова, В. В. Попов // Научный диалог. — 2019. — № 6. — С. 32—49. — DOI: 10.24224/2227-1295-2019-6-32-49.

Kochetova, L. A., Popov, V. V. (2019). Research of Axiological Dominants in Press Release Genre based on Automatic Extraction of Key Words from Corpus. *Nauchnyi dialog*, 6: 32-49. DOI: 10.24224/2227-1295-2019-6-32-49. (In Russ.).



УДК 811.161.1'42:316.752+004.4'412:659.4

DOI: 10.24224/2227-1295-2019-6-32-49

Исследование аксиологических доминант в жанре пресс-релиза на основе методов автоматического извлечения ключевых слов корпуса текстов¹

© **Кочетова Лариса Анатольевна (2019)**, orcid.org/0000-0002-5278-7373, SPIN-code 5806-7213, Researcher ID H-8598-2015, Scopus Author ID 56712699400, доктор филологических наук, профессор, кафедра английской филологии, ФГАОУ ВО «Волгоградский государственный университет» (Волгоград, Россия), kochetova@volsu.ru.

© **Попов Владимир Валентинович (2019)**, orcid.org/0000-0003-0419-2874, SPIN-code 4310-7945, Researcher ID B-4758-2018, кандидат физико-математических наук, доцент, кафедра компьютерных наук и экспериментальной математики, ФГАОУ ВО «Волгоградский государственный университет» (Волгоград, Россия) popov.vlaval@volsu.ru.

Актуальность работы обусловлена интересом к изучению дискурсивных практик методами корпусной лингвистики. Предметом анализа выступают аксиологические доминанты жанра, объективированные ключевыми словами, которые извлекаются из корпуса текстов статистическими методами. Авторы ставят целью показать специфику ценностных доминант русскоязычного жанра пресс-релиза на основе использования метода ключевых слов корпусной методологии и алгоритма, разработанного с применением языка программирования Python, основанного на мере tf-idf, которая позволяет объективно выявить лексико-семантические доминанты, объективирующие ценностные смыслы. Исследование проводилось на материале репрезентативного специализированного корпуса, собранного из текстов пресс-релизов, размещенных на официальных сайтах ведущих российских компаний в период с 2017 по 2019 годы. Установлено, что тексты пресс-релизов компаний адресованы профессиональным инвесторам, а также широкому дискурсивному сообществу. По-

1 Исследование выполнено при поддержке Российского фонда фундаментальных исследований и администрации Волгоградской области (грант РФФИ № 15-04-00134).

казано, что дискурсивные практики компаний энергетического сектора акцентируют перспективы развития, заботу об окружающей среде, значимость своей деятельности для социума в целом. Компании производственного сектора акцентируют достижения, качество и отечественное происхождение продукции. Сетевые компании ритейлера ориентированы на утилитарные ценности выбора, удобства и экономии.

Ключевые слова: корпусная лингвистика; жанр; жанр «пресс-релиз»; аксиологическая доминанта; ключевые слова.

1. Введение

Автоматическая обработка текстов естественного языка, в задачи которой входит разработка и апробация алгоритмов в целях извлечения структурированной информации из больших массивов текстовых данных, является основным направлением развития компьютерной лингвистики. Обращение к методам, основанным на количественных данных, становится в последнее время отличительной характеристикой дискурсивного анализа. Основной методологией теории дискурса выступает корпусная лингвистика — область лингвистических исследований, использующая компьютерные технологии в целях статистической обработки текстов. Синергизм двух направлений, теории дискурса и корпусной лингвистики, отражает внепарадигмальный подход к методологии, в котором метод является не отражением теории, а инструментом анализа, способом достижения цели исследования и средством верификации научных гипотез.

Сопоставляя корпусно-ориентированный и дискурсивный анализ, В. Е. Чернявская указывает, что корпусные методы связываются с достижением доказательности в интерпретативных концепциях, преодолением тенденциозности и избирательности подходов [Чернявская, 2018]. На основе корпусных данных выявляются общие закономерности тематической организации текстов, строятся когнитивно-дискурсивные модели, характеризующие смысловую организацию текстов единой жанровой принадлежности [Ильинова и др., 2017; Baker, 2006; Partington, 2010], определяются дискурсивные способы конструирования социальной реальности. Активное использование корпусных технологий в анализе дискурсивных практик, осуществляющихся в текстах, позволяет актуализировать и репрезентанты аксиологических смыслов, объективно выявить социально значимую текстовую информацию, основываясь на применении количественных данных [Кочетова и др., 2019].

В поле зрения исследователей, использующих корпусные методы анализа текстового материала, как правило, попадают тексты публицистического, в том числе новостного, и художественного дискурсов. Вместе с тем интерес для лингвистики представляет также изучение жанров професси-

ональной коммуникации, порождаемых в определенных институциональных рамках и функционирующих в конкретных социальных контекстах, к числу которых относятся пресс-релизы. Количественные данные, автоматически полученные из корпусов текстов естественного языка, позволяют более полно осмыслить роль языковых структур в процессе конструирования социального пространства, раскрывая аксиологическое содержание дискурсивных практик и социальные аспекты дискурсивной репрезентации деятельности экономических субъектов, и могут быть использованы региональными институтами и средствами массовой информации в целях оптимизации коммуникативно-текстового пространства.

Объектом нашего исследования являются тексты жанра пресс-релиза, опубликованные на официальных сайтах российских компаний. Пресс-релизы, которые в типологическом аспекте относят к PR-текстам (о классификации PR-текстов см. [Кривоносов, 2002]), жанрам профессиональной коммуникации [Bhatia, 1993, 2004], используются определенным дискурсивным сообществом в конкретных социальных контекстах и направлены на достижение коммуникативных целей [Swales, 1990]. Жанру пресс-релиза свойственен ряд конвенций, которые характеризуют профессионально-ориентированные практики данного дискурсивного сообщества, направленные на достижение определенных коммуникативных целей. По выражению А. Д. Кривоносова, тексты пресс-релизов содержат «оптимизированную информацию» [Кривоносов, 2002], релевантную социальным запросам дискурсивного сообщества, являющуюся способом дискурсивного позиционирования компании в коммуникативно-текстовом пространстве и повышения ее узнаваемости и аттрактивности, что в конечном итоге способствует достижению ее главной цели — социальному одобрению своей деятельности. В содержательном аспекте жанр пресс-релиза включает «тип социальной информации, которая производится в процессе деятельности социального субъекта (фирмы, организации, персоны), распространяется по его же инициативе, отражает в оптимизированном виде значимые факты деятельности данного субъекта, предназначена для определенного сегмента общественности и служит целям формирования эффективной коммуникационной среды данного социального субъекта» [Там же, с. 14]. Как отмечает М. В. Бусыгина, «тематический центр пресс-релиза сосредоточен на компании, поэтому текст отражает ее корпоративные ценности, приверженность определенным идеалам: защита окружающей среды, социальная ответственность и пр.» [Бусыгина, 2010, с. 12].

Природа жанра определяет специфику его адресатной специализации, под которой понимается дифференцированное воздействие субъ-

екта речи на разные типы адресатов, при этом такое воздействие часто осуществляется синкретично в одном и том же высказывании [Карасик, 2018, с. 32]. Тексты пресс-релиза предназначены для представителей средств массовой информации, поскольку известно, что эта жанровая форма текстовой коммуникации составляет информационную основу для создания публичных медийных текстов различных жанров и стилей [Володина, 2007; Кривоносов, 2012]; профессионального сообщества, инвесторов, которые ориентируются на содержательно-фактуальную информацию, включающуюся в пресс-релизы в целях принятия инвестиционных решений; на массового адресата, потребителей и представителей широкой общественности, которые узнают о деятельности компании через средства массовой информации.

Цель данной статьи состоит в выявлении аксиологических доминант в корпусе текстов жанра пресс-релиза на основе анализа ключевых слов, объективированных *n*-граммами, которые извлекаются из корпуса при помощи компьютерного инструментария, основанного на использовании определенных алгоритмов.

Прокомментируем понимание термина *ключевые слова*, первоначальная методика идентификации которых в отечественной лингвистике была разработана Л. В. Мурзиным и А. С. Штерн. Критерием выделения ключевых слов текста являлась частотность, которая применялась в целях определения наиболее значимых с точки зрения содержания раздела / главы / целого текста слов и / или словосочетаний [Мурзин и др., 1991, с. 75]. В русле данной теории ключевым считается слово или словосочетание, которое несет существенную смысловую нагрузку, служит средством содержательной организации текста и частотность его употребления в тексте превышает частотность использования по данным частотного словаря [Там же, с. 74], [Воронина и др., 2009, с. 69]. В корпусной лингвистике трактовка термина *ключевые слова* принципиально отличается от изложенной выше, поскольку для их выделения необходимо использование референциального корпуса, позволяющего провести сопоставительный анализ частотностей использования слов в обоих корпусах. Ключевыми считаются слова, частотность употребления которых в некотором корпусе является необычно высокой по сравнению с частотностью в референциальном корпусе текстов [Scott, 1997, с. 236]. Отметим, что наиболее близким к корпусной методологии является понимание ключевых слов, предложенное Л. В. Сахарным, который считал, что употребление ключевых слов в конкретном тексте обычно выше их лингвистической вероятности [Сахарный и др., 1984, с. 82].

В корпусной лингвистике для ключевых слов рассчитывается специальный коэффициент (Keuness), который является индикатором значимости ключевого слова как дескриптора содержания текста и рассчитывается по формуле, учитывающей частотность использования и ранг слова в сравниваемых корпусах. Значимость (*p* value) обозначает степень вероятности того, что данная величина является случайной [Biber et al., 2007, с. 138]. Пороговой величиной статистической значимости для ключевых слов принято считать $p \leq 0.01$, $LL \geq 6.63$. Она позволяет отсекаать лексемы, частотность которых в сравниваемых корпусах носит случайный характер, свидетельствуя о проявлении языковой вариативности [Scot, 2011].

Как показывает анализ литературы, ключевые слова, извлеченные автоматическими методами, позволяют определить тематику текстов, выявить содержащиеся в них концепты [Кононова, 2015; Ульянова и др., 2017], охарактеризовать жанрово-стилистические особенности текстов исследуемого корпуса, а также проанализировать конструируемые в дискурсе социально и культурно обусловленные ценностные смыслы.

В нашей работе ключевые слова исследуемого корпуса извлекаются с помощью алгоритма, созданного в рамках реализуемого проекта, посвященного автоматической обработке русскоязычных текстов (об алгоритмах выделения ключевых слов см: [Григорьева и др., 2017]). Е. Г. Григорьева и ее коллеги представляют графовую модель текста, позволяющую вычислять частотные характеристики слов с учетом расположения пар слов в какой-либо общей части текста, например, в одном предложении. На основе данной модели разработан алгоритм определения ключевых слов текста, основанный на мере *tf-idf*, характеризующей частоту термов в документе и уникальность термов для данного документа, реализация которого выполнена на языке программирования Python. В результате вычислений создается матрица свойств размерности $n * m$, где n — количество терминов, а m — размер корпуса, то есть число документов в этом корпусе [Там же, с. 59].

Ключевые слова, установленные при помощи предлагаемого алгоритма, позволяют выявить наиболее частотные n -граммы, отражающие способы дискурсивного конструирования деятельности институционального субъекта в коммуникативно-текстовом пространстве и определяющие специфику профессиональных дискурсивных практик, формирующих аттрактивность компании для инвесторов и конструирующих ее социальную значимость в сознании представителей дискурсивного сообщества.

2. Характеристика корпуса текстов жанра «пресс-релиз»

Исследование проводилось на основе специализированного корпуса текстов пресс-релизов¹, являющегося эмпирической базой исследования. Он состоит из трех параллельных корпусов, включающих тексты жанра, опубликованные в период с 2017 по 2018 годы и размещенные на официальных сайтах известных российских компаний. Созданный корпус содержит отобранные методом сплошной выборки рекламные тексты, размещенные на официальных сайтах дилеров в период с 2017 по 2019 годы, включает 197 497 слов и содержит 733 текста, средний размер которых варьируется в пределах 19,69 до 414,89 слов. Структура корпуса образована тремя подкорпусами текстов: подкорпус текстов, размещенных ведущими компаниями энергетического сектора (далее — ЭК), включает 347 текстов общим объемом 92 397 слов; подкорпус текстов, размещенных ведущими производственными компаниями, содержит 106 текстов (83 632 слова) (далее — ПК); подкорпус текстов, размещенных крупнейшими компаниями розничной торговли (далее — РК) — 27 текстов из 21 468 слов (табл. 1).

Таблица 1

Структура корпуса текстов жанра «пресс-релиз»

Корпус	Кол-во слов в корпусе	Кол-во текстов	Среднее кол-во слов в тексте
Компании энергетического сектора	92,397	347	266,27
«Газпромнефть»	1,9471	58	335,7
«Лукойл»	34,756	197	176,42
«Роснефть»	38,170	92	414,89
Компании производственного сектора	83,632	280	298,68
«Черкизово»	33,681	100	367,57
«Мираторг»	41,964	160	280,66
«Русagro»	1,752	20	19,69
Компании розничной торговли	21,468	106	202,52
«Магнит»	16,395	85	192,88
X5 Retail Group	2,009	8	251,12
«Лента»	3064	13	235,69
Всего	197,497	733	269,43

1 Используемый в исследовании корпус текстов пресс-релизов отечественных компаний создавался в 2017—2018 годах студентами 3-го курса направления подготовки «Фундаментальная и прикладная лингвистика» Волгоградского государственного университета в рамках производственной практики.

Как видно из таблицы 1, репрезентация корпоративного сектора в коммуникативно-дискурсивном пространстве представлена неравномерно, так как количество текстов пресс-релизов, размещенных компаниями за один и тот же промежуток времени, значительно варьируется. Некоторые компании публикуют пространные пресс-релизы, но делают это достаточно редко. Другие, напротив, размещают значительное количество коротких по объему текстов. Среднее количество слов на один текст в подкорпусе ЭК составляет 266,27, в то время как в подкорпусе ПК это значение — 298,68 слов, а в подкорпусе РК — 202,52 слова.

3. Методика выявления ключевых слов с использованием программных средств

В ходе исследования с помощью компьютерной программы были идентифицированы ключевые слова для каждого из трех анализируемых подкорпусов, то есть были определены лексические единицы, отличающиеся необычно высокой частотностью (уникальностью). Выявление ключевых слов корпусов позволяет определить рекуррентные для данного типа дискурсивной практики смысловые модели, которые невозможно определить традиционными методами [Scot, 1997]. Анализ ключевых слов позволяет выявить лексемы, отражающие ценностные доминанты, провести их семантическую категоризацию и на основе сравнительного анализа определить ценностную специфику текстов, отражающую стратегии позиционирования компании на рынке и в социуме.

Алгоритм определения ключевых слов, используемый в данной работе, основан на мере $tf-idf$ и позволяет выделить приблизительно 40 лексических единиц с наибольшей мерой $tf-idf$, а из них — 20 слов, которые чаще других употребляются в данном тексте. Каждое слово анализируемого документа приводится к лемме, и из этих лемм формируется словарь DICT. Далее определяется $tf-idf$ — мера каждого слова из словаря DICT — и формируется список List1, состоящий не менее чем из 80 слов с наибольшей мерой $tf-idf$. С этой целью находится такое наибольшее число F_r , что не менее 80 слов из словаря DICT имеют меру $tf-idf$, большую или равную F_r . Все эти слова включаются в список List1. Из списка List1 формируется список List2, содержащий не менее 40 слов, которые чаще всего встречаются в данном документе. Для каждого слова w из списка List2 находится величина $deg(w)$, которая равна числу предложений в анализируемом тексте, содержащим слово w и какое-либо другое слово w' из списка List2. Эти слова и принимаются за ключевые слова рассматриваемого документа.

3. Пресс-релизы компаний разных видов: различия в ключевых сочетаниях

Полученные с применением вышеописанного алгоритма данные для каждого из исследуемых подкорпусов сведены в таблицы, представленные ниже. Выбор лексических единиц в текстах жанра отражает дискурсивные практики, свойственные конкретным дискурсивным сообществам. Анализ списка ключевых слов показывает, что тексты пресс-релизов компаний включают общеупотребительные лексемы, бизнес-термины и технические термины. Вместе с тем в процентном отношении данные типы лексических единиц распределяются в трех подкорпусах неравномерно, что свидетельствует о различиях в дискурсивных практиках компаний. Так, в подкорпусе текстов ЭК лексические единицы распределяются следующим образом: большую часть ключевых слов составляет бизнес терминология 42,5 %, технические термины насчитывают 22,5 %, общеупотребительная лексика — 17,5 %. В подкорпусе текстов ПК: общеупотребительная лексика и бизнес-термины распределяются приблизительно одинаково — 40 % и 32,5 % соответственно, технические термины составляют 22 %. В подкорпусе текстов РК доминирует общеупотребительная лексика — 55 %, бизнес-термины составляют 35 %, технические термины — всего 7,5 %. Характер лексических единиц и их количественные подсчеты свидетельствуют об ориентации дискурсивных практик институциональных субъектов на различные группы адресатов, что находит отражение в системе аксиологических характеристик исследуемых подкорпусов.

Рассмотрим ключевые слова и связанные с ними ценностные признаки, репрезентированные в текстах жанра каждого из подкорпусов. Тексты пресс-релизов энергетических компаний (табл. 2) часто используют автореференцию, репрезентируемую ключевыми словами, которые являются именами собственными — «Роснефть», «Лукойл», «Газпром», — что отражает их ориентацию на представителей средств массовой информации, использующих материал пресс-релизов для создания собственных медийных сообщений. Дискурсивная практика жанра предполагает использование лексических единиц, предназначенных для профессионального сообщества инвесторов. Анализ списка ключевых слов показывает, что тексты пресс-релизов энергетических компаний включают в себя лексические единицы *добыча, год, объем, результат, рост, реализация, миллиард*, акцентирующие положительные результаты деятельности компаний; в список ключевых слов также попадают лексемы *программа, проект, разработка, строительство, новый, развитие*, семантика которых ориентирует

Таблица 2

Список ключевых сочетаний, извлеченных из корпуса текстов пресс-релизов энергетических компаний, с указанием меры tf-idf и абсолютной частотности употребления (АЧ)

Компании сырьевого сектора			
Ранг	Ключевое словосочетание	Tf-idf	АЧ
1	месторождение	0.02180	498
2	добыча	0.02176	497
3	нефть	0.01900	833
4	компания	0.01683	266
5	год	0.01450	994
6	роснефть	0.01379	630
7	лукойл	0.01252	858
8	проект	0.00424	653
9	скважина	0.00276	197
10	углеводород	0.00372	183
11	нефтепродукт	0.00354	181
12	также	0.0024	493
13	газпром	0.0037	475
14	развитие	0.00751	413
15	бурение	0.00218	122
16	реализация	0.01592	361
17	один	0.00336	349
18	который	0.00407	344
19	работа	0.00211	342
20	соглашение	0.00461	316
21	квартал	0.00451	206
22	сотрудничество	0.00445	305
23	российский	0.00443	304
24	миллиард	0.00432	296
25	рамка	0.00414	284
26	газовый	0.00407	93
27	директор	0.00402	276
28	новый	0.00395	249
29	область	0.00380	256
30	составить	0.00363	249
31	рост	0.00353	242
32	Россия	0.00350	240

Продолжение табл. 2			
Компании сырьевого сектора			
Ранг	Ключевое словосочетание	Tf-idf	АЧ
33	результат	0.00322	221
34	объем	0.00316	217
35	строительство	0.00303	208
36	программа	0.00275	189
37	международный	0.00267	183
38	предприятие	0.00266	221
39	совместный	0.00259	178
40	экологический	0.00218	194

вана на будущее и акцентирует перспективы развития компаний, оцениваемые в положительном ключе. В целом ЭК конструируются в дискурсе как крупные экономические субъекты, ориентированные на международный рынок, и как активные участники общественно значимых мероприятий на региональном уровне (*международный, соглашение, совместный, сотрудничество, подписать*), что позволяет конструировать дискурсивный образ социально значимой компании, которая вносит вклад в социально-экономическое развитие региона. Использование данных лексем свидетельствует о том, что адресатом текстов пресс-релизов выступает широкое сообщество. Присутствие в списке ключевых слов лексемы *экологический* отражает тот факт, что в дискурсивных практиках компании апеллируют к общезначимым человеческим ценностям, акцентируя заботу об окружающей среде. Это свидетельствует о стремлении институциональных корпоративных субъектов получить социальное одобрение своей деятельности со стороны общества.

Как показывают ключевые слова подкорпуса текстов производственных компаний (табл. 3), в дискурсивных практиках самопрезентации данных корпоративных субъектов автореферентные стратегии используются реже, исключение составляет только компания «Мираторг». Ключевые слова являются наименованиями производимой продукции (*свинина, индейка, говядина*), они в большей степени акцентируют достигнутые результаты, поскольку список включает количественные лексемы *миллиард, тысяча, рубль, тонна, год*. В дискурсивной практике конструируется значимость компаний для национальной экономики, акцентируется ориентация на отечественного потребителя, высокое качество производимой продукции.

Таблица 3

Список ключевых сочетаний, извлеченных из корпуса текстов пресс-релизов производственных компаний, с указанием меры tf-idf и абсолютной частотности употребления (АЧ)

Компании пищевой промышленности			
Ранг	Ключевое словосочетание	Tf-idf	АЧ
1	мираторг	0.02557	755
2	год	0.02056	1238
3	компания	0.01375	828
4	говядина	0.01166	234
5	производство	0.01093	658
6	ферма	0.01016	204
7	мясной	0.00927	372
8	группа	0.00927	558
9	продукция	0.00893	538
10	область	0.00888	535
11	рубль	0.00817	492
11	свинина	0.00767	154
12	мясо	0.00765	307
13	Россия	0.00752	453
14	тонна	0.00689	415
15	миллиард	0.00639	385
16	тысяча	0.00636	383
17	крупный	0.00623	375
18	проект	0.00564	340
19	индейка	0.00558	112
20	квартал	0.00535	215
21	более	0.00501	302
22	новый	0.00475	286
23	мясоперерабатывающий	0.00468	94
24	мощность	0.00466	187
25	один	0.00455	274
26	предприятие	0.00423	255
27	производитель	0.00417	251
28	качество	0.00355	214
29	рынок	0.00348	210
30	полуфабрикат	0.00314	126
31	рост	0.00310	187

Продолжение табл. 3			
Компании пищевой промышленности			
Ранг	Ключевое словосочетание	Tf-idf	АЧ
32	высококачественный	0.00294	118
33	свиноводство	0.00289	58
34	растениеводство	0.00274	55
35	увеличить	0.00270	163
36	продукт	0.00265	160
37	показатель	0.00264	159
38	агрохолдинг	0.00264	146
39	развитие	0.00257	155
40	комплекс	0.00257	155

Ключевые слова подкорпуса текстов пресс-релизов компаний розничной торговли (табл. 4) показывают, что специфической характеристикой дискурсивных практик в этом сегменте является информация, ориентированная на потребителя. В аксиологическом плане пресс-релизы сетевых компаний ритейлеров отражают утилитарные ценности, актуализирующие выбор, качество, удобство и экономию для покупателей. Ключевые слова передают фактуальную информацию об адресах вновь открытых магазинов, проведении акций и предоставлении скидок.

Таблица 4

Список ключевых сочетаний, извлеченных из корпуса текстов пресс-релизов компаний розничной торговли, с указанием меры tf-idf и абсолютной частотности употребления (АЧ)

Компании розничной торговли			
Ранг	Ключевое словосочетание	Tf-idf	АЧ
1	магнит	0.07266	1267
2	магазин	0.01730	973
3	компания	0.01398	587
4	касса	0.01254	243
5	сеть	0.01239	743
6	торговый	0.01203	689
7	товар	0.01002	624
8	продукция	0.00965	470
9	розничный	0.00951	598

Продолжение табл. 4			
Компании розничной торговли			
Ранг	Ключевое словосочетание	Tf-idf	АЧ
10	гипермаркет	0.00865	353
11	насчитывать	0.00713	215
12	покупатель	0.00699	434
13	один	0.00699	391
14	быть	0.00612	286
15	новый	0.00576	318
16	посетитель	0.00569	329
17	более	0.00562	347
18	год	0.00533	221
19	парковка	0.00497	86
20	ассортимент	0.00425	336
21	акция	0.00425	211
22	формат	0.00418	269
23	Россия	0.00403	180
24	семейный	0.00396	284
25	скидка	0.00389	50
26	крупный	0.00360	176
27	производитель	0.00353	202
28	продукт	0.00331	279
29	вместительный	0.00324	60
30	работа	0.00324	118
31	покупка	0.00302	208
32	открыть	0.00288	217
33	самообслуживание	0.00281	72
34	большой	0.00273	139
35	позиция	0.00259	249
36	адрес	0.00266	37
37	необходимый	0.00245	118
38	собственный	0.00245	78
39	открытие	0.00223	40
40	повседневный	0.00216	75

4. Выводы

Таким образом, при помощи автоматической обработки текстов получены необходимые структурированные данные из больших объемов тек-

стовой информации для дальнейшего использования в производственных целях. Как показывает анализ, предложенный алгоритм извлечения ключевых слов позволяет надежно объективировать аксиологическую специфику дискурсивных практик корпоративной коммуникации методами компьютерной лингвистики. Сопоставительный анализ ключевых слов, полученных на основе разработанного алгоритма, обнаруживает уникальные для дискурсивной репрезентации различных компаний лексемы, отражающие различные коммуникативные способы осуществления дискурсивных практик, и позволяет сделать выводы о позиционировании корпоративных субъектов в коммуникативно-текстовом, экономическом и социальном пространствах, определить специфику адресатной специализации текстов.

Литература

1. Бусыгина М. В. Жанровые и функционально-семантические характеристики пресс-релиза в современном медиадискурсе : автореферат диссертации ... кандидата филологических наук : 10.02.19 / М. В. Бусыгина. — Волгоград, 2010. — 19 с.

2. Володина С. В. PR-текст в системе дискурсивных взаимодействий / С. В. Володина // Вестник Московского университета. Серия 10 : Журналистика. — 2007. — № 5. — С. 28—33.

3. Воронина И. Е. Алгоритмы определения семантической близости ключевых слов по их окружению в тексте / И. Е. Воронина, А. А. Кретов, И. В. Попова // Вестник Воронежского государственного университета. Серия «Системный анализ и информационные технологии». — 2010. — № 1. — С. 148—153.

4. Григорьева Е. Г. Алгоритм выделения ключевых слов на основе графовой модели лингвистического корпуса / Е. Г. Григорьева, В. А. Клячин Ю. В. Помельников, В. В. Попов // Вестник Волгоградского государственного университета. Серия 2 : Языкознание. — 2017. — Т. 16. — № 2. — С. 58—67.

5. Ильинова Е. Ю. Динамика репрезентации процессуальной составляющей спортивного события в диахроническом корпусе / Е. Ю. Ильинова, Л. А. Кочетова // Вестник Волгоградского государственного университета. Серия 2 : Языкознание. — 2017. — Т. 16. — № 2. — С. 47—57.

6. Карасик В. И. Адресатная специализация в публичном политическом дискурсе // В. И. Карасик // Вестник РУДН. Серия : теория языка, семиотика, семантика. — 2018. — Т. 9. — № 1. — С. 32—49.

7. Кононова И. В. Корпусные данные в описании структуры лингвокультурных концептов / И. В. Кононова // Когнитивные исследования языка. Выпуск XXII : Язык и сознание в междисциплинарной парадигме исследований : материалы Международного конгресса по когнитивной лингвистике. 30 сентября — 2 октября 2015 г. / ответственный редактор выпуска Т. А. Клепикова. — Москва : Ин-т языкознания ; Тамбов : Тамбовский гос. ун-т им. Г. Р. Державина, 2015. — С. 547—549.

8. *Кочетова Л. А.* Когнитивно-корпусный подход к анализу конструирования ценностных смыслов в рекламном дискурсе / Л. А. Кочетова, И. В. Кононова // Вопросы когнитивной лингвистики. — 2019. — № 2. — С. 65—74.

9. *Кривонос А. Д.* PR-текст как инструмент публичных коммуникаций : автореферат диссертации ... доктора филологических наук : 10.01.10 / А. Д. Кривонос. — Санкт-Петербург, 2002. — 42 с.

10. *Мурзин Л. Н.* Текст и его восприятие / Л. Н. Мурзин, А. С. Штерн. — Свердловск : Издательство Уральского университета, 1991. — 169 с.

11. *Сахарный Л. В.* Набор ключевых слов как текст / Л. В. Сахарный, С. А. Сиротко-Сибирский, А. С. Штерн // Психолого-педагогические и лингвистические проблемы исследования текста. — Пермь : [б. и.], 1984. — С. 81—83.

12. *Светлов А. В.* Автоматизация процесса получения лингвистической информации : современные возможности / А. В. Светлов, А. С. Комендантов // Вестник Волгоградского государственного университета. Серия 2 : Языкознание. — 2017. — Т. 16. — № 2. — С. 39—46.

13. *Ульянова У. А.* Ключевые слова в тексте «Missing Manual» : проблемы выявления / У. А. Ульянова, Л. А. Петроченко // Вестник Волгоградского государственного университета. Серия 2 : Языкознание. — 2017. — Т. 16. — № 2. — С. 68—81.

14. *Чернявская В. Е.* Дискурсивный анализ и корпусные методы : необходимое доказательное звено? Объяснительные возможности качественных и количественных подходов / В. Е. Чернявская // Вопросы когнитивной лингвистики. — 2018. — № 2. — С. 31—37.

15. *Baker P.* Using Corpora in Discourse Analysis / P. Baker. — London & New York : Continuum, 2006. — 198 p.

16. *Biber D.* Corpus Linguistics : Investigating Language structure and Use / D. Biber, S. Conrad, R. Reppen. — Cambridge University Press, 1998. — 311 p.

17. *Bhatia V.* Analysing Genre : Language Use in Professional Settings / V. Bhatia. — Harlow : Longman, 1993. — 453 p.

18. *Bhatia V.* Worlds of Written Discourse / V. Bhatia. — London : Continuum, 2004. — 248 p.

19. *Partington A.* Modern diachronic corpus-assisted discourse studies on UK newspapers : an overview of the project / A. Partington // Corpora. — 2010. — Vol. 5. Iss. 2. — Pp. 83—108.

20. *Scott M.* PC analysis of key words — and key key words / M. Scott // System. — 1997. — Vol. 25. Iss. 2. — Pp. 233—245.

21. *Scott M.* WordSmith Tools Manual [Electronic resource] / M. Scott. — Liverpool : Lexical Analysis Software Ltd. 2015. — Version 6. — Access mode : <http://lexically.net/downloads/version6/wordsmith6.pdf>.

22. *Stubbs M.* Three concepts of keywords / M. Stubbs // Keyness in Texts / M. Bondi, M. Scott (eds.). — Amsterdam : John Benjamins, 2010. — Pp. 21—42.

23. *Swales J. M.* Genre Analysis. English in Academic and Research Settings / J. M. Swales. — Cambridge : Cambridge University Press, 1990. — 274 p.

Research of Axiological Dominants in Press Release Genre based on Automatic Extraction of Key Words from Corpus¹

© **Larisa A. Kochetova (2019)**, orcid.org/0000-0002-5278-7373, SPIN-code 5806-7213, Researcher ID H-8598-2015, Scopus Author ID 56712699400, Doctor of Philology, Department of English Philology, Volgograd State University (Volgograd, Russia), kochetova@volsu.ru.

© **Vladimir V. Popov (2019)**, orcid.org/0000-0003-0419-2874, SPIN-code 4310-7945, Researcher ID B-4758-2018, PhD in physical and mathematical sciences, associate professor, Department of Computer Science and Experimental Mathematics, Volgograd State University (Volgograd, Russia), popov.vlaval@volsu.ru.

The relevance of the work is determined by the interest in the study of discursive practices by corpus linguistics methods. The subject of the analysis is the axiological dominants of a genre that are objectified by key words, extracted from the corpus by statistical methods. The authors aim to show the specificity of the value dominants of the Russian-language genre of press release based on the use of the keyword method of the corpus methodology and an algorithm developed using the programming language Python, based on the measure tf-idf, which allows to objectively identify lexical and semantic dominants that represent value meanings. The study was conducted on the material of a representative specialized corpus, that combine texts of press releases posted on the official websites of major Russian companies between 2017 and 2019. Based on the analysis of lexical choices, it is established that the texts of press releases of the companies are addressed to professional investors, as well as to a wide discursive community. It is shown that the discursive practices of energy sector companies emphasize the prospects of development, care for the environment, the importance of their activities for society in general. Companies in the manufacturing sector emphasize the achievements and highlight quality and domestic origin of products. Retail network companies are focused on utilitarian values of choice, convenience and economy.

Key words: corpus linguistics; genre; genre of "press release"; axiological dominant; key words.

References

- Baker, P. (2006). *Using Corpora in Discourse Analysis*. London & New York: Continuum.
- Biber, D., Conrad, S., Reppen, R. (1998). *Corpus Linguistics: Investigating Language structure and Use*. Cambridge University Press.
- Bhatia, V. (1993). *Analysing Genre: Language Use in Professional Settings*. Harlow: Longman.
- Bhatia, V. (2004). *Worlds of Written Discourse*. London: Continuum.
- Busygina, M. V. (2010). *Zhanrovyye i funktsionalno-semanticheskiye kharakteristiki press-reliza v sovremennom mediadiskurse: avtoreferat dissertatsii... kandi-data filologicheskikh nauk*. Volgograd. (In Russ.).
- Chernyavskaya, V. E. (2018). Diskursivnyy analiz i korpusnyye metody: neobkhodimoye dokazatelnoye zveno? Obyasnitelnyye vozmozhnosti kachestvennykh

1 The study is supported by the Russian Foundation for Basic Research and Volgograd region administration (grant RFBR No. 15-04-00134).

- i kolichestvennykh podkhodov. *Voprosy kognitivnoy lingvistiki*, 2: 31—37. (In Russ.).
- Grigoryeva, E. G., Klyachin, V. A., Pomelnikov, Yu. V., Popov, V. V. (2017). Algoritm vydeleniya klyuchevykh slov na osnove grafovoy modeli lingvisticheskogo korpusa. *Vestnik Volgogradskogo gosudarstvennogo universiteta*, 2: *Yazykoznanie*, 16 (2): 58—67. (In Russ.).
- Ilinova, E. Yu., Kochetova, L. A. (2017). Dinamika reprezentatsii protsessualnoy sostavlyayushchey sportivnogo sobyitiya v diakhronicheskom korpuse. *Vestnik Volgogradskogo gosudarstvennogo universiteta*, 2: *Yazykoznanie*, 16 (2): 47—57. (In Russ.).
- Karasik, V. I. (2018). Adresatnaya spetsializatsiya v publichnom politicheskom diskurse. *Vestnik RUDN. Seriya: teoriya yazyka, semiotika, semantika*, 9 (1): 32—49. (In Russ.).
- Kochetova, L. A., Kononova, I. V. (2019). Kognitivno-korpusnyy podkhod k analizu konstruirovaniya tsennostnykh smyslov v reklamnom diskurse. *Voprosy kognitivnoy lingvistiki*, 2: 65—74. (In Russ.).
- Kononova, I. V. (2015). Korpusnyye dannyye v opisani structure lingvokulturnykh kontseptov. In: Klepikova, T. A. (ed.). *Kognitivnyye issledovaniya yazyka, XXII: Yazyk i soznaniye v mezhdistsiplinarnoy paradigme issledovaniy: materialy Mezhdunarodnogo kongressa po kognitivnoy lingvistike*. 30 sentyabrya — 2 oktyabrya 2015 g. / otvetstvennyy redaktor vypusk. Moskva: In-t yazykoznaniya; Tambov: Tambovskiy gos. un-t im. G. R. Derzhavina. 547—549. (In Russ.).
- Krivososov, A. D. (2002). *PR-tekst kak instrument publichnykh kommunikatsiy: avtoreferat dissertatsii ... doktora filologicheskikh nauk*. Sankt-Peterburg. (In Russ.).
- Murzin, L. N., Shtern, A. S. (1991). *Tekst i yego vospriyatiye*. Sverdlovsk: Izdatelstvo Uralskogo universiteta. (In Russ.).
- Partington, A. (2010). Modern diachronic corpus-assisted discourse studies on UK newspapers: an overview of the project. *Corpora*, 5 (2): 83—108.
- Sakharnyy, L. V., Sirotko-Sibirskiy, S. A., Shtern, A. S. (1984). Nabor klyuchevykh slov kak tekst. In: *Psikhologo-pedagogicheskiye i lingvisticheskiye problemy issledovaniya teksta*. Perm: [b. i.]. 81—83. (In Russ.).
- Scott, M. (1997). PC analysis of key words — and key key words. *System*, 25 (2): 233—245.
- Scott, M. (2015). *WordSmith Tools Manual*, 6. Liverpool: Lexical Analysis Software Ltd. Available at: <http://lexically.net/downloads/version6/wordsmith6.pdf>.
- Stubbs, M. (2010). Three concepts of keywords. In: Bondi, M., Scott, M. (eds.). *Keyness in Texts*. Amsterdam: John Benjamins. 21—42.
- Svetlov, A. V., Komendantov, A. S. (2017). Avtomatizatsiya protsessa polucheniya lingvisticheskoy informatsii: sovremennyye vozmozhnosti. *Vestnik Volgogradskogo gosudarstvennogo universiteta*, 2: *Yazykoznanie*, 16 (2): 39—46. (In Russ.).

- Swales, J. M. (1990). *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.
- Ulyanova, U. A., Petrochenko, L. A. (2017). Klyuchevyye slova v tekste «Missing Manual»: problemy vyyavleniya. *Vestnik Volgogradskogo gosudarstvennogo universiteta, 2: Yazykoznaniye, 16 (2)*: 68—81. (In Russ.).
- Volodina, S. V. (2007). PR-tekst v sisteme diskursivnykh vzaimodeystviy. *Vestnik Moskovskogo universiteta, 10: Zhurnalistika, 5*: 28—33. (In Russ.).
- Voronina, I. E., Kretov, A. A., Popova, I. V. (2010). Algoritmy opredeleniya semanticheskoy blizosti klyuchevykh slov po ikh okruzheniyu v tekste. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya «Sistemnyy analiz i informatsionnyye tekhnologii», 1*: 148—153. (In Russ.).