ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

Использование лингвостатистического инструментария в выявлении специфики концептов СМИ Гродненщины

Мохань Елена Николаевна, старший преподаватель
Гродненский государственный университет имени Янки Купалы (г. Гродно, Беларусь)

В статье с использованием лингвостатистического инструментария выявлены значимые отличия в рекуррентности концептов в русскоязычных текстах СМИ Гродненщины по отношению к текстам корпуса-эталона.

Ключевые слова: *концепт, лингвостатистика, рекуррентность, корпус.*

Известно, что функционирование языка в иноязычном окружении имеет свои особенности, которые при определенных условиях могут закрепляться в узуальной норме, приспосабливаясь к функционированию в разных условиях, например, в речи билингвов. Соответственно, актуализируется проблема кодификации норм при описании функционирования языков за пределами метрополий, в частности, русского языка на территории Республики Беларусь.

В решении вышеназванных проблем особую роль могут и должны сыграть «региональные» корпусы русского языка. Такой корпус был создан в рамках проекта «Лексико-семантическая и социокультурная специфика русской речи Гродненщины (на материале текстов СМИ)» (договор с БРФФИ № Г13Р-050 от 16 апреля 2013 г.) как оригинальный языковой ресурс на аутентичном языковом материале, отражающем узус русского и белорусского языков, сосуществующих в условиях белорусско-русской лингвокультурной общности. Иллюстративный лингвистический корпус СМИ Гродненщины – РМ(Г) положил начало формированию «в составе НКРЯ нового модуля, представляющего региональные варианты литературного стандарта» [7, с. 352].

Дискурс русскоязычных масс-медиа Беларуси представляет собой уникальный языковой материал не только для изучения ценностей белорусской культуры, но и для изучения вербализации ценностей русской культуры: «Признаки культуры «Своего» особенно наглядно выступают в контрасте с культурой «Другого», служащей фоном для национально-культурного и социально-культурного дистанцирования» [5, с. 3].

Корпус СМИ Гродненщины, имеющий объем более 2,5 млн. словоупотреблений, составлен на основе региональных СМИ («Берестовицкая газета», «Ивьевский край», «Островецкая правда», «Праца», «Свислочская газета», «Вечерний Гродно», «Перспектива») и соответствует требованию репрезентативности, что позволило использовать его для решения задачи выявления не только семантических отличий и регионально специфичной лексики, отражающей, в частности, страноведческие реалии, но и для выявления социокультурной специфики русскоязычного дискурса СМИ Гродненщины, для чего была использована методика корпуса-эталона. Для этого на материале российских СМИ, отобранных в соответствии с характеристиками, аналогичными для изданий СМИ Гродненщины, тексты которых вошли в РМ(Г), был сформирован сопоставимый подкорпус, сыгравший роль корпуса-эталона.

Лингвостатистические исследования имеют продолжительные традиции, тем не менее, они приобретают особую значимость именно в последнее время в связи с доступностью для обработки больших массивов языковых данных в виде лингвистических корпусов и специализированных баз данных.

Так, в докладах, представленных в разные годы на Международной конференции «Диалог», которая является ведущей российской конференцией по компьютерной лингвистике и единственным в мире форумом, посвященным, прежде всего, проблемам компьютерного анализа русского языка (были проанализированы труды данной конференции за 2000–2014 гг. [10]), представлены следующие направления лингвостатистических исследований: проведение первичной обработки лингвистических данных (статистическое оценивание, проверка статистических гипотез, построение теоретических моделей [2; 4]); решение систематико-таксономических задач (обработка данных с помощью кластерного, многомерного анализа, дистрибутивно-статистического метода, методов атрибуции, типологии текста [14; 11; 12]), а также информационно-поисковых задач (автоматический поиск текстов, текстовых единиц, обладающих определенным набором качественных и количественных характеристик для решения стилистических, грамматических, фонетических проблем [1; 8; 9; 3]).

Особое направление в лингвостатистических исследованиях связано с составлением и использованием частотных словарей, эффективность применения которых для решения различных прикладных и исследовательских задач сегодня не вызывает сомнения. Так, например, стратификация лексики по признаку частотности имеет большое значение для выявления количественных характеристик слов, коррелирующих с их качественными характеристиками. Особую роль могут и должны играть частотные словари в исследовании рекуррентности концептов, с которой связана степень актуализации тех или иных социокультурных факторов в рамках определенной лингвокультуры [6, с. 149]. Рекуррентность концепта (частотность его языковых репрезентаций в речи) является важным показателем актуальности концепта в когнитивном сознании народа и отражает не только языковую, но и когнитивную, лингвосоциальную актуальность концепта [13].

Анализируя наиболее частотные единицы языка в определенный период его развития, можно выявить концепты, обладающие наибольшей актуаль-

ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

ностью для данного периода развития национального сознания.

Такое исследование было проведено относительно СМИ Гродненщины путем сопоставительного анализа частотных списков словоформ РМ(Г) и корпуса-эталона российских СМИ. С этой целью были выявлены первая тысяча наиболее и последняя тысяча наименее частотных элементов в частотном полном списке словоформ РМ(Г) и в аналогичном списке корпуса-эталона. Служебные и уникальные для белорусского корпуса слова не подвергались анализу.

Доля покрытия высокочастотными единицами в РМ(Г) (14,2 %) и в корпусе-эталоне (13,8 %) приблизительно одинакова.

Путем наложения частотных списков первой тысячи словоформ РМ(Г) (772) и корпуса-эталона (773) были выявлены 478 единиц, встречающихся в текстах как белорусских, так и российских СМИ, и 294 единицы, не имеющие соответствий в первой тысяче корпуса-эталона. Из них 17 являются уникальными экспликаторами социокультурных особенностей речи Гродненщины.

Анализ состава лексических единиц СМИ Гродненщины, превышающих частоту в сопоставлении с данными корпуса-эталона, – свидетельствует, что в РМ(Г) более рекуррентными являются такие концепты, как:

Беларусь (*Беларусь, белорусский, белорусских, белорусского, белорусской, Гродно, национального*); Власть (*председателя, председатель, облисполкома, райисполкома, отдел, отдела, начальник, начальника, комитета, заместитель, представителей, вопросы, вопросам, меры, учреждения, учреждениях, учреждений, организации, организаций, ЖКХ, обратиться, обращение, обращения, заявления, исполнительного, осуществляется, обязанности, ответ, планируется, идеологической, объединения, отделение, отделения, общественного, общественных*); Административное деление (*деревне, деревни, районного, районной, районном, районный, района, району, районе, сельсовета, республиканского, республики, республике, житель, жители, жителей, жительства, населения*); Социальная политика (*помощи, помощь, внимание, внимания, социальной, защиты, защите, условий, базовых, благополучия, занятости, инвалидов, пенсии, пособие, размер, размере, помочь, обслуживания, акция, акции*); Здоровье и медицина (*здоровье, здоровья, здравоохранения, заболевания, питания*); Сельское хозяйство (*сельского, сельских, хозяйство, хозяйства, хозяйстве, земле, зерна, молока, площадь, совхоз, поле, СПК, га, животных*); Семья, школа, дети (*брака, семья, семьи, семье, семей, родителей, родители, отец, мама, матери, взрослых; школа, школы, школе, школу, школ, школьников, образования, гимназии, классов, книги, обучения, СШ, учащихся, учитель; детей, дети, детьми, детям, ребенок, ребенка, ребят, ребята, возраста, возрасте, молодежи, молодой, молодые, молодых*); Охрана порядка (*безопасности, нарушение, нарушения, милиции, преступлений, профилактики, РОВД, МЧС, несовершеннолетних, внутренних, правил, правила, кодекса, пожара, пожарной, имущества*); Труд (*работников, работники, труд, труду, труда, трудовой, работа, работе, работы, работу, ответственность, вместе, коллектив*); Патриотизм (*Великой, Отечественной, войны, память, памяти, ветеранов*); Праздник (*праздник, праздника, праздником*); Культура (*культуры, творчества*); Выборы (*депутатов, участие*); Постройки (*доме, дома, ремонт, жилых, магазин*); Речь (*рассказал*); Оценка (*лучшие, лучших, любви, любовь, радость, спасибо, счастья, уважаемые*); Индивиды (*женщин, женщина, женщины, граждане, граждан, гражданам, людей, люди, людьми, людям*); Время (*дни, день, дня, дней, ежегодно, будущем*); Спорт (*команда, соревнования, спорта*); Информация (*информацию, комментарии*); Досуг (*мероприятий, мероприятия, отдыха, песни, туризма, встречи, гостей, гости, традиции*); Энергетика (*АЭС*); Контроль (*проверки, регистрации, инспекции*); Транспорт (*водитель, ГАИ, дороги, движения*); Право (*договор*).

В корпусе-эталоне наиболее частотны по отношению к РМ(Г) 119 единиц, которые являются вербализаторами следующих концептов:

Россия (*Владимир, Путин, России*); Административное деление (*стране, областной, городской, области*); Время (*сначала, потом, прошлого, сегодня, давно, сейчас*); Финансы (*средства, миллионов, млрд., деньги, денег, долларов, бюджета, цены, фонда*); Власть (*управление, президента, директора, директор, власти, решение, главного, обеспечения, государственной, проект*); Экономика (*роста, уровень, уровня, ОАО, производства, продукции, оборудования, предприятия, предприятий, строительства, специалистов*); Суд (*суд, дела, дело*); Оценка (*проблема, проблемы, проблем, активно, выше, опыт, правда, нового, новых, новой, большинство, просто, меньше*); Речь (*говорят*).

Количественные характеристики наиболее рекуррентных концептов в текстах гродненских и российских СМИ во многом коррелируют с их качественными свойствами: тематичность, информативность, нейтральность, а также указывают на то, что является актуальным, социально и культурно значимым в данный период развития национального сознания белорусов и россиян.

Если относительная частота употребления словоформы в РМ(Г) меньше или равна частоте употребления данной словоформы в корпусе-эталоне, то такую разницу частот будем считать статистически не значимой.

В зоне элементов со статистически не значимой разницей частот каждого корпуса можно выделить две подзоны. В первую подзону входят элементы (121), которые не являются высокочастотными ни в корпусе-эталоне, ни в РМ(Г).

Вторую подзону образуют единицы, частоты которых в каждом из корпусов приближаются к вербализаторам высоко рекуррентных концептов. Так, в РМ(Г) 56 элементов по частоте приближаются к вербализаторам высоко рекуррентных концептов СМИ Гродненщины. Это элементы, вербализирующие следующие концепты:

Время (*года, месяц*); Индивиды (*участников*); Оценка (*интересно, важно, немного*); Охрана порядка (*порядок*).

В корпусе-эталоне 59 элементов являются номи-

нантами следующих высоко рекуррентных концептов российских СМИ:

Оценка (*лучше, многие, большой);* Время (*раньше, сразу);* Власть (*органов, комиссии, управления, деятельность);* Административное деление (*страны)* и другие.

В частотном списке РМ(Г) представляют интерес и низкочастотные единицы, то есть те единицы, которые либо отсутствуют в частотном списке корпуса-эталона, либо относительная частота употребления которых меньше, чем в корпусе-эталоне.

Наложив частотные списки последней тысячи единиц в частотном полном списке РМ(Г) и в аналогичном списке корпуса-эталона друг на друга, мы не выявили ни одного пересечения элементов, т.е. «хвосты» частотных списков не совпадают.

В состав последней тысячи наименее частотных словоформ РМ(Г) входит 582 уникальные словоформы, эксплицирующие социокультурные особенности СМИ Гродненщины, и 418 словоформ, присутствующих в корпусе-эталоне, но не входящих в последнюю тысячу частотного списка.

Сопоставив относительные частоты этих 418 низкочастотных единиц последней тысячи частотного списка РМ(Г) с их аналогами в корпусе-эталоне, находящимися за пределами последней тысячи низкочастотных единиц, мы пришли к выводу, что наименее частотными (118) в РМ(Г) по сравнению с корпусом-эталоном, являются вербализаторы следующих концептов:

Политика (*геополитические, вожди);* Экономика (*внешнеторговых, газовики, гастарбайтеров, водоотведения, водозабор, водке*, *генеральную, возводимое);* Россия (*Волге, волжский, волжские);* История (*генерал-губернатора);* Искусство (*галерею);* Компьютерные технологии (*гарнитура);* Армия (*гвардейского);* География (*географической, возвышенность)* и другие.

В корпусе-эталоне к низкочастотным (161) относятся вербализаторы следующих концептов:

Выборы (*выборщиков, выбирай);* Еда (*выпекает, галушки);* Семья (*внучата);* Культура (*вокалистка);* Транспорт (*водилы, вокзальной, въездной, возу);* Здоровье и медицина (*внутриутробном);* Мифология (*Гера)* и другие.

К низкочастотным элементам (81) в обоих корпусах относятся представители следующих концептов:

Цветы (*гвоздика);* Религия (*Воздвижение);* Зарубежье (*Гвинет, Гамбурга);* Досуг (*гармошке);* Криминал (*гангстерами)* и другие.

Анализ рекуррентности лексических единиц позволил выявить коммуникативно релевантные и нерелевантные номинанты концептов в концептосфере белорусских и российских СМИ.

Наиболее рекуррентными в текстах как белорусских, так и российских СМИ являются следующие концепты: «Власть», «Административное деление», «Речь», «Оценка», «Время».

К наиболее рекуррентным концептам РМ(Г) относятся: «Беларусь», «Социальная политика», «Здоровье и медицина», «Сельское хозяйство», «Семья, школа, дети», «Охрана порядка», «Труд», «Патриотизм», «Праздник», «Культура», «Выборы», «Постройки», «Индивиды», «Время», «Спорт», «Информация», «Досуг», «Энергетика», «Контроль», «Транспорт».

В корпусе-эталоне наиболее рекуррентными концептами являются «Россия», «Финансы», «Экономика», «Суд».

В номинативное поле концепта входят как релевантные единицы, образующие ядро концепта, так и среднечастотные единицы, входящие в околоядерную зону, которые по частоте приближаются к вербализаторам рекуррентных концептов как в белорусских, так и в российских СМИ.

К менее рекуррентным концептам в текстах обоих корпусов относятся: «Еда», «Цветы», «Религия», «Зарубежье», «Досуг», «Криминал».

Менее рекуррентными концептами в РМ(Г) являются: «Политика», «Экономика», «Россия», «История», «Искусство», «Компьютерные технологии», «Армия», «География».

В корпусе-эталоне к нерекуррентным концептам относятся: «Выборы», «Еда», «Семья», «Культура», «Транспорт», «Здоровье и медицина», «Мифология».

Рекуррентность языковых средств объективации концепта свидетельствует о процессах актуализации и деактуализации концептов в общественном сознании и отражает различия социальных аксиологий, поскольку свидетельствует о степени актуальности и важности для общества тех или иных проблем, т. е. отражает социокультурную специфику национальных одноязычных дискурсов.

Литература:

1. Антонов, Е. С., Курзинер Е. С. Автоматическое определение тематики большого необработанного текстового массива / Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара «Диалог 2002» по лингвистике и ее приложениям. URL: http://www.dialog-21.ru/digest/archive/2002/?year=2002&vol=22725&id=7516 (дата обращения: 18.06.2014).

2. Беликов, В. И., Ахметова М. В. Статистическая оценка функциональных свойств лексики по материалам Интернета / Компьютерная лингвистика и интеллектуальные технологии: Материалы Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). URL: http://www.dialog-21.ru/digests/dialog2009/materials/html/05.htm (дата обращения: 17. 06.2014).

3. Браславский, П. И., Соколов Е. А. Автоматическое извлечение двухсловных терминов с использованием поисковых машин Интернета // Компьютерная лингвистика и интеллектуальные технологии: Материалы Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2007 г.). М.: Изд-во РГГУ, 2007. – С. 89–94.

4. Захаров, В. П., Хохлова М. В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке / Компьютерная лингвистика и интеллектуальные технологии: Материалы Международной конференции «Диалог 2010» (Бекасово, 26–30 мая 2010 г.). URL: http://www.dialog-

21.ru/digests/dialog2010/materials/html/22.htm (дата обращения: 15.06.2014).

5. Клинкова, А. А. Вербализация американских ценностей в дискурсе масс-медиа о Японии: автореф. дис. ... канд. филолог. наук. – Иркутск: ГОУ ВПО «Иркутский государственный лингвистический университет», 2009. – 18 с.

6. Когнитивная лингвистика / З. Д. Попова, И. А. Стернин. М: АСТ: Восток–Запад, 2007. – 314 с.

7. Кустова, Г. И., Савчук С. О. Изучение лексико-семантической и социокультурной специфики русской речи на территории Республики Беларусь (на материале текстов СМИ) // Труды международной конференции «Корпусная лингвистика – 2013». – СПб.: С.-Петербургский гос. университет, Филологический факультет, 2013. – С. 344–352.

8. Лукашевич, Н. В., Добров Б. В. Исследование тематической структуры текста на основе большого лингвистического ресурса / Труды Международного семинара «Диалог 2000» по лингвистике и ее приложениям. URL: http://www.dialog-21.ru/digest/archive/2000/?year=2000&vol=22725&id=6521 (дата обращения: 15.06.2014).

9. Лукашевич, Н. В., Добров Б. В. Автоматическое аннотирование новостных кластеров на основе тематического представления / Компьютерная лингвистика и интеллектуальные технологии: Материалы Международной конференции «Диалог» (Бекасово, 27–31 мая 2009 г.). URL: http://www.dialog-21.ru/digests/dialog2009/materials/html/46.htm (дата обращения: 16.06.2014).

10. Международная конференция по компьютерной лингвистике «Диалог» URL: http://www.dialog-21.ru/conference-last/ (дата обращения: 16.06.2014).

11. Михеев, М. Ю. Компиляция или ... клише? Сравнивая характерные для авторского стиля наборы словосочетаний / Компьютерная лингвистика и интеллектуальные технологии: Материалы Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). URL: http://www.dialog-21.ru/digests/dialog2010/materials/html/52.htm (дата обращения: 14.06.2014).

12. Романов, А. С., Мещеряков Р. В. Определение пола автора короткого электронного сообщения / Компьютерная лингвистика и интеллектуальные технологии: Материалы Международной конференции «Диалог» (Бекасово, 27–31 мая 2009 г.). URL: http://www.dialog-1.ru/digests/dialog2009/materials/html/67.htm (дата обращения: 16.06.2014).

13. Титова, О. И. Перспективы лингвистического исследования рекуррентных единиц лексикона // Филологические науки. – 2003. – № 2. – С. 79–86.

14. Шайкевич, А. Я., Савчук С. О. Анализ лексико-семантических особенностей региональной прессы (на примере газет Гродненского региона Беларуси / Компьютерная лингвистика и интеллектуальные технологии: Материалы Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). URL: http://www.dialog-21.ru/digest/2014/pdf/ (дата обращения: 14.06.2014).