С. О. Савчук
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
savsvetlana@mail.ru

КОРПУС РЕГИОНАЛЬНЫХ ГАЗЕТ РОССИИ И ЗАРУБЕЖЬЯ[1]

В статье описывается опыт создания нового модуля в составе НКРЯ — корпуса региональной и зарубежной прессы. Главную задачу этого корпуса мы видим в том, чтобы предоставить исследователям инструмент для изучения региональной вариативности русского языка. В состав корпуса в настоящий момент входят: 1) корпус газет гродненского региона (7 изданий, 1,9 млн словоупотреблений), подготовленный совместно со специалистами Гродненского университета; 2) подкорпус российских региональных газет (13 изданий, около 2 млн словоупотреблений); 3) корпус региональных российских газет 1990–2000-х гг. (40 газет, 2,6 млн словоупотреблений); 4) региональные выпуски газеты «Комсомольская правда» (6,5 млн словоупотреблений). Тексты корпуса снабжены морфологической, семантической аннотацией и подробной метаразметкой. В статье описываются принципы отбора текстов, стандарт и программные средства обработки текстов, организация поиска по корпусу, перспективы развития состава и корпусного инструментария.

Ключевые слова: Национальный корпус русского языка, газетный корпус, корпус региональной прессы, состав и структура, аннотация текстов.

[1] Работа выполнена при поддержке РГНФ (гранты № 13-24-01004 и № 15-04-12018) и РФФИ (грант № 15-06-04334).

I. Введение

Корпус региональной и зарубежной прессы представляет собой самостоятельный модуль в составе Национального корпуса русского языка. Он открыт для свободного доступа с октября 2014 года по адресу http://www.ruscorpora.ru/search-regional.html. Найти его в составе НКРЯ несложно: он объединен с большим корпусом СМИ[1] в один газетный блок, вход осуществляется с боковой панели главной страницы Корпуса.

Главную задачу этого корпуса мы видим в том, чтобы предоставить исследователям инструмент для изучения региональной вариативности русского языка. Вероятность возникновения вариантов узуса повышается в том случае, если язык функционирует в иноязычном окружении, в условиях языковых контактов, выступает как язык межнационального или международного общения. Для таких языков, как английский, французский, немецкий, установлен состав национальных языковых вариантов (в частности британский, американский, канадский, австралийский, новозеландский варианты английского языка; варианты французского языка Франции, Бельгии, Швейцарии и Канады; немецкий, австрийский, швейцарский варианты немецкого языка), описаны особенности их словаря и грамматики. Материал о национальных вариантах включается в состав учебников и пособий для изучающих эти языки, создаются корпуса вариантов того или иного языка, которые используются не только в качестве исследовательских ресурсов, но имеют большую популярность в педагогической практике.

Что касается русского языка, то изучение его как средства межнационального общения на большой территории, населенной разными народами, было начато еще в СССР. Значительный материал о взаимодействии русского и родного языка был накоплен специалистами в области преподавания русского языка как иностранного. В последние десятилетия в изменившихся геополитических условиях для России стала актуальной тема вариативности русского языка в ближнем и дальнем зарубежье (в Белоруссии, Казахстане, Киргизии, Молдавии, Узбекистане, Америке, Финляндии, Австралии, Израиле) и в национальных республиках Российской

[1] Принципы формирования, состав и структура и первый опыт использования Корпуса СМИ 2000-х годов (газетного корпуса) см. в [Савчук 2011].

Федерации. Приводит ли новая языковая ситуация в государствах постсоветского пространства к усилению процессов дивергенции региональных вариантов и кодифицированного языка метрополии, к формированию национальных вариантов русского языка за пределами России? Для того чтобы ответить на эти вопросы, нужна репрезентативная экспериментальная база и продуманная методика анализа. Учитывая то обстоятельство, что изменения в русском языковом узусе того или иного региона могут накапливаться постепенно и быть незаметными невооруженным глазом, эффективным в данном случае может оказаться использование лингвостатистических методов, применяемых на большом массиве текстов. Все это делает создание такого массива настоятельно необходимым и актуальным.

Работы по формированию корпуса региональных СМИ начались в рамках международного проекта, выполняемого коллективами исследователей ИРЯ им. В. В. Виноградова РАН и Гродненского государственного университета им. Я. Купалы[1]. Целью проекта было изучение лексико-семантических и культурно-специфических особенностей русской речи на территории Гродненской области Республики Беларусь[2] на материале областных и районных газет.

Были согласованы общие принципы и методика организации корпусов, программное обеспечение и технология подготовки и описания текстов, используемые в НКРЯ, были модернизированы и настроены под новые задачи. В течение 2013 года белорусской стороной был подготовлен корпус русскоязычных газет Гродненского региона, а российской стороной — корпус российских региональных СМИ 2010-х годов. Пилотные версии этих корпусов были размещены на сайте ruscorpora.ru и стали ядром создаваемого корпуса региональных СМИ. Рассмотрим его состав более подробно.

---

[1] При поддержке РГНФ (грант № 13-24-01004, руководитель А.Я. Шайкевич) и БРФФИ (договор № Г13Р-050, руководитель Л.В. Рычкова).

[2] Ситуация в Гродненском регионе представляет особый интерес в том плане, что, во-первых, русский язык функционирует в нем, как и на всей территории Белоруссии, в качестве государственного, то есть обслуживает все сферы речевой коммуникации, а во-вторых, существует в контакте не только с белорусским, но и с польским и литовским языками.

Состав корпуса

Русскоязычные газеты Гродненского региона

• «Вечерний Гродно» (http://www.vgr.by), городская вечерняя газета г. Гродно, издается с 2000 г., выходит 1 раз в неделю, тираж около 23000–25000 экз., в корпус вошло 696 статей за 2012 г.

• «Перспектива» (http://www.perspektiva-info.by), региональная газета, г. Гродно и Гродненский район, издается с 1939 г. (первоначально — «Свободная Беларусь», затем — «За свободную Беларусь», «Сельская новь» и «Сельская навіна»), выходит 2 раза в неделю, тираж около 11000 экз., в корпусе 750 статей за 2012 г.

• «Островецкая правда»/«Астравецкая праўда» (http://www.ostrovets.by) районная газета г. п. Островец, издается с 1941 г. (первоначально — «Бальшавіцкі арганізатар»), выходит 2 раза в неделю, в корпусе 915 статей за 2012 г.

• «Берестовицкая газета»/«Бераставіцкая газета» (http://www.beresta.by), районная газета г. п. Б. Берестовица, издается с 1944 г. («Знамя Советов», с 1966 г. «За камунізм», с 1991 г. «Бераставіцкая газета»), выходит 2 раза в неделю, тираж около 3000 экземпляров, в корпусе 1235 статей за 2012 г.

• «Ивьевский край»/«Іўеўскі край» (http://ivyenews.by), общественно-политическая районная газета Ивьевского района, г. Ивье, выходит 2 раза в неделю, тираж более 5600 экз., в корпусе 198 статей за 2012 г.

• «Праца» (http://www.praca.zelva.by), районная газета г. п. Зельва, под таким названием издается с 1967 года, выходит 2 раза в неделю, тираж более 4000 экз., в корпусе 926 статей за 2012 г.

• «Свислочская газета»/«Свіслацкая газета» (http://www.svisloch.info), районная газета г. Свислочь, издается с 1943 г., выходит 2 раза в неделю, тираж более 4000 экз., в корпусе 903 статьи за 2012 г.

В таблице 1 дается характеристика газет по типу аудитории и приводится объем текстов каждого издания, выраженный в количестве словоупотреблений.

Таблица 1

Название	**Тип газеты**	**Год издания**	**Количество с/у**
«Вечерний Гродно»	региональная вечерняя	2012	182 504
«Перспектива»	региональная	2012	184 970
«Бераставіцкая газета»	районная	2012	414 545
«Іўеўскі край»	районная	2012	133 580
«Астравецкая праўда»	районная	2012	467 517
«Праца»	районная	2012	352 597
«Свіслацкая газета»	районная	2012	330 007
ИТОГО			2 074 720

Отбор источников производился на основе анализа СМИ Гродненского региона, который был проведен белорусскими коллегами [см. Рычкова и др. 2012]. С учетом опыта формирования газетных корпусов в составе НКРЯ были разработаны критерии отбора. Их можно разделить на технические и содержательные. К техническим критериям относятся наличие сайта газеты и качество его технической поддержки. Техническая поддержка обеспечивает существование архива газетных номеров, его регулярные пополнения, логическую организацию сайта, аннотацию текстов, формат их представления. От этих показателей зависит скорость и качество обработки текстов соответствующих газет для размещения в корпусе. Поскольку время, отпущенное на формирование корпуса, ограничено рамками договора по проекту, скорость и качество подготовки корпуса как экспериментальной базы исследования приобретает решающее значение.

К содержательным критериям отбора относится разнообразие источников, обеспечивающее представительность корпуса. Источники должны варьироваться по географии, по типу газет, типу аудитории, политической принадлежности и др. В рамках конкретного проекта важно найти баланс между максимальными содержательными запросами и ограниченными техническими возможностями. Как представляется, сформированный пилотный вариант газетного корпуса СМИ Гродненщины вполне отвечает требованию представительности.

Рис. 1.

Во-первых, обеспечено географическое разнообразие: в корпус включены газеты шести районов Гродненской области, расположенных на севере (Островецкий район), западе (Гродненский, Берестовицкий районы), юго-западе (Свислочский, Зельвенский районы) и на востоке области (Ивьевский район).

Во-вторых, представлены газеты разных типов: две региональные («Перспектива», «Вечерний Гродно») и пять районных газет («Берестовицкая газета», «Островецкая правда», «Ивьевский край», «Свислочская газета», «Праца» (г. п. Зельва).

В-третьих, газеты различаются по типу аудитории: районные ориентированы на сельских жителей, они освещают в основном местные новости и проблемы локального уровня, в то время как региональные газеты ориентируются на жителей города Гродно и области, наряду с местными новостями уделяют внимание проблемам регионального, государственного и международного уровня. Кроме того, две гродненские газеты представляют другое типологическое различие: *«Перспектива»* относится к типу общественно-политических газет, с большой долей деловой и социально значимой информации, а *«Вечерний Гродно»* — к типу вечерних городских газет, предназначенных для досуга и семейного чтения.

В процессе подготовки текстов обнаружилось, что часть текстов районных газет, озаглавленных по-русски, оказалась на белорусском языке[1]. Это не только комментарии читателей, но и корреспондентские материалы. В отдельных номерах доля белорусскоязычных текстов достигала 40%. Было принято решение включить эти тексты в состав корпуса, введя еще один признак метаописания — «язык текста» — и снабдив их соответствующей пометой. Статьи на белорусском языке (общим объемом около 500000 словоупотреблений) доступны для точного поиска в составе общего корпуса, но их мож-

[1] Использование русского языка во всех заголовках электронной версии газеты, в том числе и в статьях на белорусском языке, обусловлено требованиями поисковой системы.

но выделить в отдельный подкорпус, выбрав значение опции «язык текста» — белорусский. Подкорпус газетных текстов на белорусском языке в русскоязычных газетах пяти районов, представленных в корпусе, свидетельствует, с одной стороны, о реальном двуязычии жителей этих районов, а с другой стороны, о стремлении газет быть ближе к читателям, ориентироваться на разные типы аудитории, в том числе и в языковом плане.

Ориентации газет на читателя, развитию их интерактивности способствует электронная форма изданий: во многих газетах бывшие отделы писем преобразованы в отделы обратной связи в виде комментариев, форумов и под. Тексты, помещенные в этих разделах, принадлежат непрофессиональным авторам и стилистически отличаются от основных публикаций: в отличие от них, комментарии читателей ближе к разговорной речи, чем к публицистике. Они представляют не меньший интерес для исследования региональных особенностей русского языка, чем основные газетные жанры, так как могут обнаружить еще один аспект регионального варьирования: если основной корпус газетных текстов мы используем для изучения вариантов нормы, то в комментариях мы будем иметь дело с вариантами разговорного узуса. Комментарии читателей к публикациям помещены в корпус, оформлены как самостоятельные тексты, каждый из которых включает все отклики на статью (их может быть от одного до нескольких десятков в зависимости от актуальности темы). Они имеют в составе заголовка название основной публикации (например, *У озера Свитязь появится хозяин / Комментарии, Лидские автобусы изменят дизайн / Комментарии*) и аннотированы как «электронная коммуникация».

Корпус-эталон российских газет

Согласно основной гипотезе проекта, региональные особенности в текстах гродненских газет могут быть выявлены при сравнении с корпусом-эталоном российских СМИ. Приступая к его формированию, мы учитывали, что этот корпус должен быть сопоставимым по объему и по временному периоду с гродненским корпусом. Что касается конкретного состава газет, здесь открывался веер возможностей. Можно было бы выбрать какую-нибудь область, например, Новгородскую или Смоленскую, и собрать местные газеты разных

типов; можно было действовать так же на уровне какого-либо административного образования, например, собрать газеты субъектов РФ, входящих в Северо-Западный административный округ (Ненецкий АО, Республика Коми, Республика Карелия, Архангельская, Мурманская, Вологодская, Ленинградская, Новгородская, Псковская, Калининградская области). В этом случае, вероятно, следовало бы ожидать значительного присутствия в корпусе-эталоне российских СМИ текстов с местными региональными особенностями. Мы решили нивелировать зависимость от регионального фактора. Сделать это можно было бы либо путем случайной выборки газет, либо путем отбора в корпус достаточно большого количества изданий из разных регионов (с учетом технических критериев, о которых говорилось выше). Для пилотной версии корпуса был выбран второй способ.

В настоящее время в корпус-эталон российских СМИ включены 13 газет, представляющих семь регионов России: Центральный, Северо-Западный, Южный, Приволжский, Уральский, Сибирский, Дальневосточный. Время выхода газет — в основном 2012–2013 гг. Общий объем корпуса — около 2 млн словоупотреблений.

• «Амурская правда» (http://www.ampravda.ru) — ежедневная региональная общественно-политическая газета Амурской области, г. Благовещенск. Издается с 24 февраля 1918 года. В советскую эпоху — печатный орган Амурского областного комитета КПСС и областного Совета народных депутатов. Тираж — 8000 экз., в корпус включено 174 текста за 2012–2013 г.

• «Байкальские Вести» (http://www.baikvesti.ru) — общественно-политическая еженедельная газета Иркутской области, г. Иркутск, издается с 2001 г. Первоначально газета называлась «Восточно-Сибирские вести», а в декабре 2003 года переименована в «Байкальские вести», причем коллектив остался прежним. С 1 января 2010 года газета выходила дважды в неделю, с июля 2012 года выходит 1 раз в неделю по понедельникам, тираж 10000 экз., в корпусе 490 текстов в основном за 2012 г.

• «Богатей» (http://bogatej.ru) — еженедельная независимая деловая газета, г. Саратов, издается с 1996 г., тираж 4500 экз. Читателями газеты, как сказано на сайте газеты, "являются бизнесмены, чиновники, политики, люди интеллектуальных профессий, которые уча-

ствуют в управлении хозяйством и обществом". В корпусе 270 текстов в основном за 2013 г.

• «Вечерний Орел» (http://vechorel.ru) — первая орловская электронная газета, общественно-политическое издание, организовано в 2010 г. как городская еженедельная вечерняя газета, тираж составлял 8000 экз. В корпус включено 138 текстов за 2012–2013 г.

• «Вятский край» (http://vk.ru) — ежедневная общественно-политическая газета, одна из четырёх областных государственных газет Кировской области. Газета с таким названием выходила еще в дореволюционной Вятке, но сегодняшний «Вятский край» свой отсчет ведет с 5 октября 1990 г., тираж составляет 60000 еженедельно, в корпусе 272 текста за 2012–2013 г.

• «Ижевская газета» (http://www.izhtime.ru/izhgazeta) — городская газета, г. Ижевск, интернет-версия издавалась с 2006 по 2010 г. В корпусе 192 текста за 2010 г.

• «Красное знамя» (http://komikz.ru/) — старейшая газета Республики Коми. Ведет своё начало от газеты «Зырянская жизнь» (вышла 10 июня 1918 года в г. Усть-Сысольске Вологодской губернии). До 1955 года выпускалась также под названиями «В Зырянском краю», «Удж» (труд), «Югыд туй» (светлый путь), «За новый Север». До 1991 года являлась официальным органом Коми обкома КПСС, Верховного Совета и Совета Министров Коми АССР. В настоящее время — независимая частная газета, выходит 1 раз в неделю, тираж около 10000 экз., регион распространения — Республика Коми, г. Сыктывкар; в корпусе 248 текстов за 2012–2013 г.

• «Новая газета Кубани» (http://ngkub.ru/) является региональным партнером федерального печатного общественно-политического издания «Новая газета». Первый номер «Новой газеты Кубани» вышел в свет 20 мая 2004 г. Целевая аудитория газеты — политики, предприниматели, бизнесмены, работники культуры и искусства. Издатель — ООО «Центр политического мониторинга». Газета выходит 2 раза в неделю. Тираж издания — 15000 экз. Регион распространения — Южный федеральный округ. Соотношение региональных и федеральных материалов составляет 40% : 60%. В корпусе 389 текстов за 2009–2013 г.

• «Новости Саратовской губернии» (http://sarnovosti.ru/) — региональное сетевое общественно-политическое издание, выходит с 2000 г., г. Саратов. Учредитель — министерство информации и пе-

чати Саратовской области. Тематика издания — освещение происходящих в области событий, деятельности органов власти всех уровней, общественных организаций. 8000–10000 посещений в сутки.

• «Рабочий путь» (http://www.rabochy-put.ru) — еженедельная общественно-политическая региональная газета, издается с марта 1917 года. Как сказано на сайте газеты, «Это не только летописец, но и участник славной истории Смоленщины и России. Газета не прекращала издаваться и во время оккупации территории области фашистскими войсками — «РП» печатался в Москве и распространялся с воздуха. «Рабочий путь» — политический тяжеловес. Он был и остается любимой газетой четырех поколений смолян». Тираж газеты 25000 экз. В корпусе 290 статей за 2008–2013 г.

• «Уральский рабочий» (http://газета-уральский-рабочий.рф) — ежедневная общественно-политическая газета, г. Екатеринбург. Главная и старейшая газета Среднего Урала издается с 14 февраля 1907 года. В советскую эпоху — орган Свердловского обкома КПСС и областного Совета депутатов трудящихся. Тираж около 20000 экземпляров, выходит пять раз в неделю. В корпус включено 112 текстов в основном за 2012–2013 г.

• VL.ru (Владивосток) — информационно-справочный портал города Владивостока, существует с 1997 г. Имеет разделы: Новости Владивостока (http://www.newsvl.ru) и справочные разделы (Афиша, Недвижимость, Транспорт, Базы отдыха, Работа, Форумы и пр.). В корпусе 229 текстов за 2011–2013 г.

• «Новгородские ведомости» (http://novved.ru) — единственная областная газета Новгородской области, издается с 29 декабря 1990 г. Распространяется во всех районах области, в Великом Новгороде и Северо-западном регионе. Выходит 3 раза в неделю. Номер в среду посвящён публикации официальных документов (областные законы, постановления и распоряжения администрации Новгородской области), в номерах во вторник и субботу публикуются материалы о политике, общественной жизни, культуре, спорте и др. Тираж газеты более 21000 экз. В корпус включено 849 текстов за 2012–2013 г.

Количественный вклад каждого издания в корпус-эталон российских СМИ 2010-х годов отражен в таблице 2.

Таблица 2

Название	**Регион**	**Тип газеты**	**Год издания**	**Кол- во с/у**
«Амурская правда» (Благовещенск)	Дальневосточный	региональная	2012-2013	114460
VL.ru (Владивосток)	Дальневосточный	городское электронное издание	2011-2013	44238
«Байкальские Вести» (Иркутск)	Сибирский	региональная	2010-2013	541580
«Уральский рабочий» (Екатеринбург)	Уральский	региональная	2011-2013	66215
«Богатей» (Саратов)	Приволжский	региональная	2013	136438
«Новости Саратовской губернии»	Приволжский	региональное электронное издание	2013	13912
«Ижевская газета» (Ижевск)	Приволжский	городская	2010	90516
«Вятский край» (Киров)	Приволжский	региональная	2012-2013	111826
«Вечерний Орел» (Орел)	Центральный	городское электронное издание	2012-2013	33587
«Рабочий путь» (Смоленск)	Центральный	региональная	2008-2013	112885
«Красное знамя» (Республика Коми)	Северо-Западный	региональная	2012-2013	158071
«Новгородские ведомости» (В. Новгород)	Северо-Западный	региональная	2012-2013	351679
«Новая газета Кубани» (Краснодар)	Южный	региональная/ федеральная	2009-2013	251323
ИТОГО				2026730

Рис. 2.

Сравнение гродненского корпуса и корпуса-эталона российских СМИ показывает, что они отличаются по составу газет: гродненский корпус на 80% состоит из текстов районных газет (причем районов с сельскохозяйственной специализацией), в российском корпусе собраны 8 региональных (областных) газет, 3 городские и 1 федеральная газета представлена своим региональным партнером. В гродненский корпус вошли электронные версии газет, которые выпускаются также и в печатной форме. В российском корпусе таких газет 10, а 3 издания представляют собой сетевые (электронные) СМИ. Издания гродненского корпуса локализованы в одном регионе Беларуси, география российского корпуса — максимально широка. Для контрастивных исследований такое отличие корпусов — скорее плюс, но для проведения сопоставительных исследований необходимо развитие экспериментальной базы в нескольких направлениях: в первую очередь включение районных газет в российский корпус, увеличение доли текстов, представляющих собой комментарии читателей. В долгосрочной перспективе предстоит пополнение обоих корпусов новыми газетами — в гродненский корпус войдут газеты других районов Гродненщины, студенческие газеты и малотиражные издания, в российский корпус будут включены газеты новых регионов, а также другие типы изданий, в том числе электронные.

Корпус региональных российских газет 1990–2000-х гг.

Гродненский корпус и корпус российских газет 2010-х годов формируют ядро регионального корпуса, но не единственную его составляющую. Концепция корпуса, который планируется создавать в течение нескольких лет, а затем постоянно пополнять, состоит в следующем. Корпус планируется сделать принципиально открытым, в том числе для сотрудничества с разными коллективами исследователей и заинтересованными организациями. Его можно будет расширять, присоединяя к нему новые модули, например, русскоязычные газеты Казахстана или Прибалтики, региональные газеты советского периода или начала XX в. и под. В настоящее время разработан общий дизайн корпуса, параметры описания текстов и стандарты их обработки. Внутри каждого модуля может быть своя специфика, которая определяется особенностями материала, но единственным общим требованием является соблюдение стандартов аннотации и обработки текстов, которое обеспечит возможность их интеграции в составе целого.

Одним из таких модулей, включенных в состав корпуса, является корпус региональных СМИ 1990–2000-х годов. В нем представлены тексты 40 изданий (в том числе районных и городских) из 7 федеральных округов России, и по объему он сопоставим с корпусом российских газет 2010-х годов. Состав этого подкорпуса приведен в таблице 3.

Таблица 3

Название	**Регион**	**Тип газеты**	**Год издания**	**Кол- во с/у**
«Амурский Меридиан» (Хабаровск)	Дальневосточный	региональная	2004	21937
«Биробиджанер Штерн» (Биробиджан)	Дальневосточный	региональная	2004	15194
«Владивосток» (Владивосток)	Дальневосточный	региональная	2003	32906
«Ежедневные новости» (Владивосток)	Дальневосточный	региональная	2003	13080

Название	Регион	Тип газеты	Год издания	Кол- во с/у
«Республика Саха» (Якутск)	Дальневосточный	региональная	1996	8484
«Рыбак Приморья» (Владивосток)	Дальневосточный	деловая региональная	2003	60103
«Биржа плюс свой дом» (Н. Новгород)	Приволжский	региональная	2002	152914
«Богатей» (Саратов)	Приволжский	региональная	2003	268266
«Вечерняя Казань» (Казань)	Приволжский	региональная	2003	52109
«Дело» (Самара)	Приволжский	деловой региональный журн.	2001-2002	421307
«Марийская правда» (Йошкар-Ола)	Приволжский	региональная	2003	46071
«Московский комсомолец» в Нижнем Новгороде	Приволжский	федеральная региональный выпуск	2004	2387
«Московский комсомолец» в Саранске	Приволжский	федеральная региональный выпуск	2004	2659
«Нефтяник» (Пермь)	Приволжский	региональная	2003	59741
«Нижегородские губернские ведомости» (Н. Новгород)	Приволжский	региональная	1998, 2003	23795
«Оренбуржье» (Оренбург)	Приволжский	региональная	1997	26884
«Пермский строитель» (Пермь)	Приволжский	региональная	2003-2004	196941
«Петербургский Час пик» (С.-Петербург)	Северо-Западный	региональная	2003	130765

Название	**Регион**	**Тип газеты**	**Год издания**	**Кол- во с/у**
«Санкт-Петербургские ведомости» (С.-Петербург)	Северо-Западный	региональная	2003	63790
«Смена» (С.-Петербург)	Северо-Западный	региональная	2003	1173
«Калининградская правда» (Калининград)	Северо-Западный	региональная	2003	11813
«Калининградские Новые колеса» (Калининград)	Северо-Западный	региональная	2004	37695
«Московский комсомолец» в Сыктывкаре	Северо-Западный	федеральная региональный выпуск	2003	1954
«Восточно-Сибирская правда» (Иркутск)	Сибирский	региональная	2003	194482
«Континент Сибирь» (Новосибирск)	Сибирский	региональная	2004	50873
«Красноярский рабочий» (Красноярск)	Сибирский	региональная	2003	77306
«Свободный курс» (Барнаул)	Сибирский	региональная	1997	25837
«Деловой квартал» (Екатеринбург)	Уральский	городской еженедельный журнал	2003	12671
«Вечерний Екатеринбург»	Уральский	вечерняя	2004	12554
«Сургутская трибуна» (Сургут)	Уральский	региональная	2001	9431
«Уральский автомобиль» (Миасс)	Уральский	районная	2004	62006
«Весть» (Калуга)	Центральный	региональная	2002	108438

Название	Регион	Тип газеты	Год издания	Кол- во с/у
«Воронежские вести» (Воронеж)	Центральный	региональная	2003	21903
«Встреча» (Дубна)	Центральный	районная	2003	301243
«Наша жизнь» (с. Перемышль, Калужская обл.)	Центральный	районная	2003	6674
«Северный край» (Ярославль)	Центральный	региональная	1997	5605
«Дагестанская правда» (Махачкала)	Северо-Кавказ-ский	региональная	2003-2005	49126
«Краевые новости» (Краснодар)	Южный	региональная	2003	6383
«Новороссийский рабочий» (Новороссийск)	Южный	городская	2003	9654
«Приазовский край» (Ростов-на-Дону)	Южный	региональная	2004	17769
«Сочи» (Сочи)	Южный	городская	2002	2502
ИТОГО				2626425

Таким образом, два корпуса газет — 1990–2000-х и 2010-х годов — представляют российскую региональную прессу в двух временных срезах и дают возможность проводить сравнительные микродиахронические исследования, отслеживать изменения в тематическом наполнении, в жанровом составе и в самих жанровых формах, стилистике взаимодействия с читателем и пр.

Региональные выпуски газеты «Комсомольская правда»

Другой модуль, включенный в состав текущей версии регионального корпуса,— это региональные выпуски КП. Вкладки с местными новостями издаются более чем в 40 регионах России (Абакан, Барнаул, Белгород, Благовещенск, Брянск и т. д.) и в зарубежных странах (Беларусь, Кыргызстан, Молдова, Северная Европа, Бал-

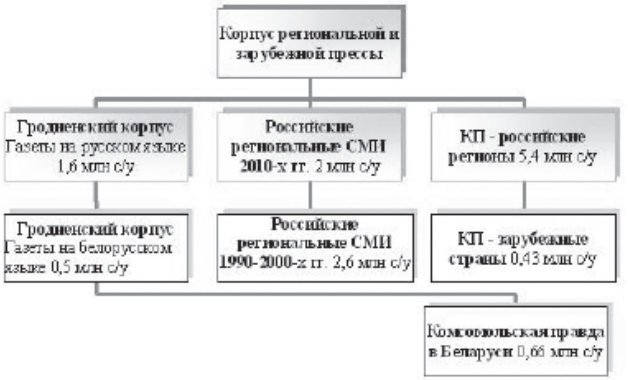
каны, Кипр). Материалы пишутся сотрудниками на местах, потому эти выпуски КП можно рассматривать как региональные издания. Однако будучи самостоятельными, редакции и издатели этих выпусков не могут не придерживаться общей политики центрального издания, принципов освещения и приемов подачи материала и др. Поэтому имеет смысл рассматривать местные выпуски КП и других федеральных газет как самостоятельный тип изданий. Состав подкорпуса приведен в таблице 4.

Таблица 4

Название	**Регион**	**Кол- во с/у**
«Комсомольская правда» (российские регионы)	40 субъектов РФ	5 422 196
«Комсомольская правда» в Беларуси	Беларусь	662 451
«Комсомольская правда» – Северная Европа	Эстония, Латвия, Литва	65 535
«Комсомольская правда» в Кыргызстане	Кыргызстан	67 193
«Комсомольская правда» в Молдове	Молдова	300 119
ИТОГО		6 517 494

Таким образом, в текущей версии регионального газетного корпуса выделяются четыре относительно самостоятельные коллекции (см. Рис. 3): 1) тексты русскоязычных газет Гродненщины на русском и белорусском языке (Гродненский корпус), две коллекции региональных СМИ России с дистанцией в 10 лет — 2) издания 1990–2000-х годов и 3) СМИ 2010-х годов и 4) коллекция региональных выпусков «Комсомольской правды». С ними можно работать как с единым массивом, так и с каждой коллекцией в отдельности, а также использовать коллекции в различных комбинациях. В частности, можно объединить в один подкорпус газеты Гродненщины и белорусские выпуски «Комсомольской правды». Эти и многие другие возможности обеспечиваются поиском по корпусу.

Рис. 3.



II. Подготовка корпуса. Стандарт обработки текстов

Превращение текста со страницы интернет-издания в элемент корпуса — долгий процесс преобразований, на отдельных этапах которого текст должен быть освобожден от одной аннотации, встраивающей его в электронный газетный гипертекст, и снабжен новой метатекстовой и лингвистической аннотацией, обеспечивающей поиск в корпусе. При подготовке текстов мы опирались на опыт разработки большого газетного корпуса, которая осуществлялась с использованием программных средств. Полностью отказаться от ручной обработки текстов не удалось (о причинах будет сказано ниже), но как представляется, был достигнут разумный компромисс в сочетании автоматических средств и ручного труда.

Первый и самый важный этап переформатирования текстов выполнялся с помощью специальной программы пакетной обработки файлов (автор Л. Д. Алексеевский). В программу встроен базовый шаблон преобразования текста из исходного формата в выходной формат, что позволяет обрабатывать файлы, имеющие одинаковый формат (например, тексты одного издания), в сколь угодно большом количестве. Однако поскольку в разных интернет-изданиях используются разная программная поддержка, для каждого издания приходится модифицировать шаблон замены с помощью ручной настройки. Изменения в шаблоне касаются в основном извлечения

метаинформации, поэтому справиться с задачей ручной настройки можно даже не владея специальными навыками программирования. После отладки нового модифицированного шаблона на нескольких текстах запускается автоматический режим работы. Шаблоны рекомендуется сохранять, поскольку они могут пригодиться для обработки новых текстов того же издания, а также подойти для текстов какого-то другого издания (в нашей практике такое случилось дважды).

Программная обработка файлов осуществляется в два приема: первый запуск программы обеспечивает извлечение нужной текстовой и метатекстовой информации (автор, название текста, дата, рубрика, если есть, адрес страницы, если есть) и очистку от лишней html-разметки. В результате второго запуска программы текст приводится к правильному xml-формату. В данном случае используется единый шаблон, поэтому производить эту операцию можно не по отдельным изданиям, а со всей обрабатываемой коллекцией. В полученном файле имеются две зоны — текст с размеченным заголовком и абзацами и зона метаданных, оформленная специальными тегами.

```
<?xml version="1.0" encoding="UTF-8"?><html><head>
<title>До первой звезды</title>
<meta content="До первой звезды" name="title">
</meta>
<meta content="15.03.2013" name="date"></meta>
<meta content="Старая Русса" name="topic"></meta>
<meta content="Новгородские ведомости"
name="source"></meta>
<meta content="http://novved.ru/staraya-russa/1419-
do-pervoj-zvezdy.html" name="url"></meta>
</head>
<body>

<p class="h1">До первой звезды</p>
<p>Наши ветераны — сильнее всех </p>
<p>В начале марта в Валдае прошел традиционный
турнир по мини-футболу, посвященный памяти одного
из основателей команды «Юпитер» Александра
Белоусова. В играх участвовала и команда
```

```
старорусских ветеранов, возраст которых перешагнул
50-летнюю отметку. </p>
<p>Как сообщил вратарь старорусских спортсменов, а
по совместительству и директор МАУ «Спорткомплекс»
Борис Васильев, наши на турнире были лучшими. </p>
<p>В финале мы встречались с командой хозяев и
выиграли со счетом 5:4. В баталиях участвовали
команды из Валдая, Старой Руссы, Парфина и
Крестец.</p>
</body>
</html>
```

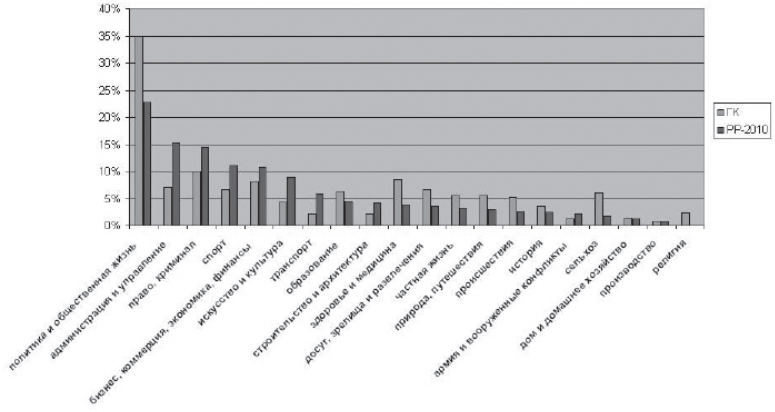
Следующий этап обработки — метатекстовая аннотация — выполняется программными средствами с последующим редактированием. Для этого используется программа QTMetabaseEditor (автор Л. Д. Алексеевский), которая считывает метаданные из файлов в соответствующие поля базы данных. Дальнейшее заполнение полей (прежде всего определение текстовых характеристик) производится вручную, причем удобнее это делать с помощью той же программы, которая по каждому полю предлагает выбор из списка значений, что, несомненно, сокращает количество формальных ошибок при заполнении. Экспортируется содержимое базы в формат таблицы. csv, которая сопровождает каталоги с текстовыми файлами и хранит всю метаинформацию о текстах.

При разработке формата таблицы метаданных, в отличие от большого газетного корпуса, мы решили не отказываться от использования ручного редактирования при заполнении полей по нескольким причинам. Во-первых, ядро корпуса — это белорусско-русский проект, который предполагал подробную аннотацию текстов с целью изучения лексико-семантических и культурно-специфических особенностей газетных текстов, в том числе и в жанровом отношении. Во-вторых, корпус сравнительно невелик по объему, и такой объем работы оказался по силам даже небольшому коллективу. Кроме того, в большинстве изданий удалось извлечь информацию о рубрике, что облегчило определение типа текста и тематики. Всего лишь облегчило, хотя первоначальный расчет был на то, что наличие рубрики позволит автоматически заполнить поле тематики. Однако на практике оказалось, что организация текстов в рубрики не кор-

релирует с тематическим членением текстов. Действительно, чаще всего рубрика в самом деле соответствует тематическому содержанию публикации (см., например, такие рубрики, как «Политика», «Общество», «Бизнес», «Происшествия», «Здоровье», «Спорт»). Реже она дает информацию о жанре и помогает определить тип текста (например, в рубриках «Новости», «Комментарии» или «Обзор прессы» размещаются соответственно информационные сообщения, комментарии и обзоры; в рубрике «Персона» — в основном интервью с известными людьми, «Наш репортаж» — репортажи). Иногда название рубрики отражает логику внутренней организации сайта, но не несет нужной нам информации о жанре и тематике текста (например, такие рубрики, как «Старая Русса» в «Новгородских ведомостях» или «Киров и область» в «Вятском крае», или «Без рубрики» в «Вечернем Гродно»). Кроме того, в исходных файлах некоторых изданий информация о рубрике отсутствовала. Поэтому полностью отказаться от ручного редактирования метатекстовой аннотации оказалось невозможным. Наконец, в-третьих, качественная ручная разметка пилотного корпуса позволит в дальнейшем использовать его в качестве эталона для обучения программ автоматического определения жанров и тематики текстов.

Тематическое распределение текстов в Гродненском и эталонном корпусах представлено на Рис. 4.

Рис. 4.



Учитывая, что в последнее время исследования по автоматической атрибуции жанровой и тематической принадлежности текстов на больших массивах данных ведутся с большой интенсивностью, в будущем мы надеемся автоматически разметить тексты «Комсомольской правды» и освоить автоматический способ разметки тематики и типов текста в пополнениях корпуса.

III. Поиск по корпусу

При разработке стандартов аннотации и организации поиска по корпусу был учтен опыт создания основного и газетного корпусов НКРЯ. Метатекстовая аннотация состоит в приписывании каждому тексту атрибутов, по которым в дальнейшем будет осуществляться поиск в корпусе. В метатекстовой разметке регионального корпуса есть значительные отличия как от основного, так и от газетного корпуса. Набор метатекстовых признаков и основных значений приведен ниже.

0. Название файла и путь к нему

I. Информация об авторе

1. **Имя автора**
2. **Пол автора**
3. **Возраст автора**

II. Информация о тексте

4. **Название текста**
5. **Дата создания текста**
6. **Сфера функционирования текста:** публицистика, к которой относятся основная масса газетных текстов, реклама, деловая, художественная.
7. **Тема текста**, или предметная область*
8. **Хронотоп**, или место и время описываемых событий: только для художественных текстов
9. **Тип текста***
10. **Жанр художественной литературы**
11. Стиль текста

III. Информация об аудитории

12. Возраст аудитории: н-возраст, если это незначимый фактор, детский, подростковый, взрослый
13. Уровень образования аудитории: для газет как средства массовой информации, в основном н-уровень

14. **Размер аудитории**: национальная (для федеральных газет), региональная (для областных газет), районная (для районных газет, а также городских, аудитория которых ограничена конкретным языковым поселением).

IV. Библиографическое описание текста

15. Источник текста — адрес страницы в интернете

16. **Название издания** — название газеты или электронного издания

17. Тип источника — газета, журнал, электронный текст

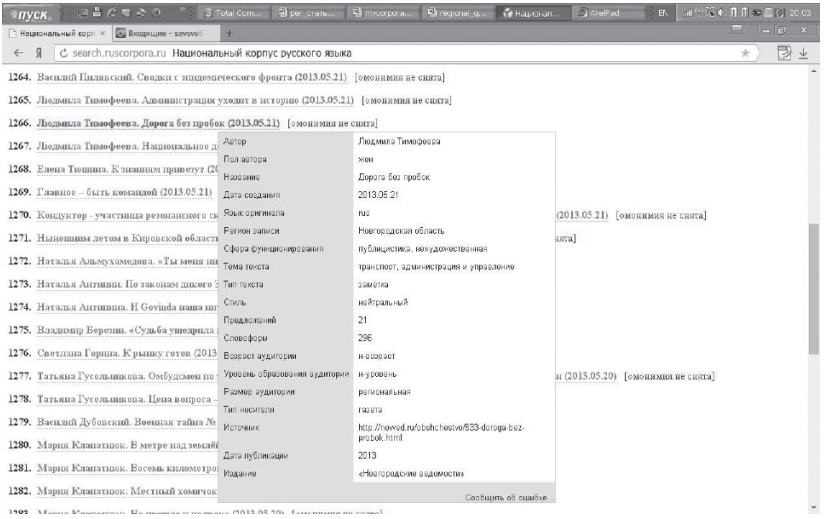
18. **Страна**

19. **Регион**

20. **Язык**

Тексты регионального корпуса, за исключением региональных выпусков КП, имеют полный набор метапризнаков, который приводится в «паспорте» текста, доступном при нажатии на активную строку с названием текста на странице выдачи подкорпуса или выдачи контекстов (см. Рис. 5). В текстах КП пока нет разметки по тематике и типу текста (эти признаки в приведенном выше списке отмечены звездочкой).

Рис. 5.

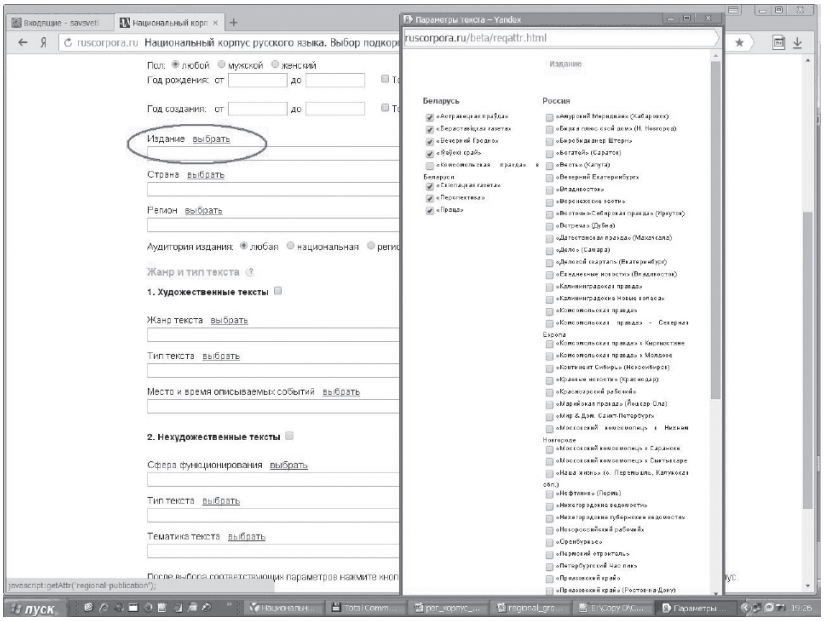


Почти все метаатрибуты (они выделены в списке жирным шрифтом) представлены на странице http://ruscorpora.ru/mycorpora-regional.html, и по ним организован выбор подкорпуса. При выборе рабочего подкорпуса мы имеем возможность задавать комбинации признаков и конструировать корпус в соответствии с поставленными задачами. Спектр возможностей широк — можно работать как со всем массивом текстов, так и с текстами отдельной газеты. Удобный и понятный интерфейс позволяет делать это с минимальными усилиями — установка большинства опций предполагает выбор из списков.

Выбирая значение «Язык» = «белорусский», мы задаем небольшой подкорпус текстов на белорусском языке в гродненских газетах. Выбор значения «Язык» = «русский» позволяет нам работать со всем 13-миллионным корпусом текстов на русском языке.

Для выбора Гродненского корпуса текстов на русском языке можно использовать следующий набор атрибутов: «Язык» = «русский», «Страна» = «Беларусь», «Регион» = «Гродненская область». Другой путь к тому же корпусу — через атрибут «Издание»: в открывшемся

Рис. 6.

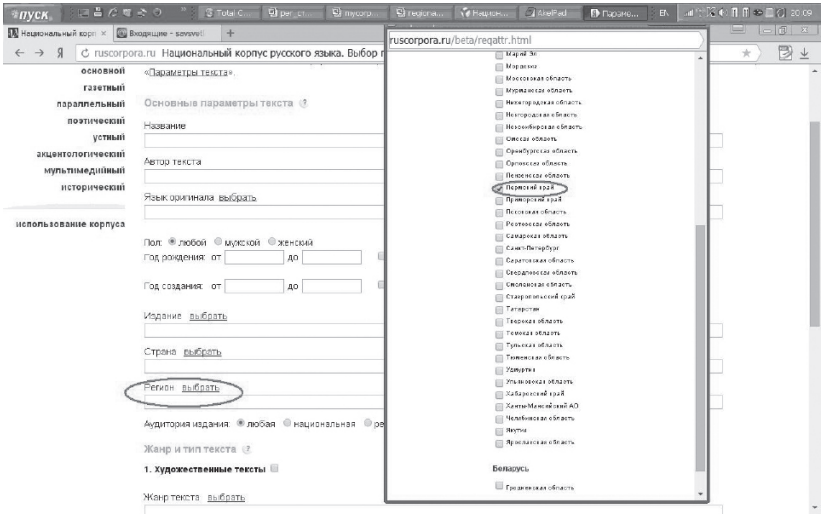


списке нужно отметить галочками все белорусские газеты (кроме «Комсомольской правды» в Беларуси) или выбрать нужные.

Коллекция российских СМИ отбирается аналогично: «Страна» = «Россия» (язык «русский» можно не устанавливать), газеты выбираются из списка «Издание». Если нужен весь эталонный корпус российских СМИ, используем следующие параметры: «Страна» = «Россия», «Год создания» от 2009 до 2013, из списка газет исключаем «Комсомольскую правду». Если нужен ранний подкорпус российских газет, устанавливаем дату 1996–2004 (выпуски «Комсомольской правды» в него не попадают, поскольку они датированы 2011–2012 годами). Если предстоит работать только с текстами «Комсомольской правды», подкорпус формируется из списка «Издание».

Корпус предоставляет возможность работы с газетами какого-либо региона (для Российской федерации можно выбрать уровень отдельного субъекта федерации — области, края, республики, или административного округа, например Сибирского, Уральского, Южного). Для этого нужно использовать параметр «Регион», выбрав нужное значение или несколько значений из списка.

Рис. 7.



Например, Пермский край представлен четырьмя газетами — «Пермский строитель», «Нефтяник», «Уральский автомобиль»

и местные выпуски «Комсомольской правды», всего в подкорпусе 925 текстов, почти 495 тыс. словоупотреблений. Доля газет Иркутской области («Байкальские Вести», «Восточно-Сибирская правда» и региональные выпуски КП) составляет 1102 текста, более 845 тыс. словоупотреблений, Краснодарского края — 713 текстов (около 456 тыс. словоупотреблений) из пяти газет («Краевые новости», «Новая газета Кубани», «Новороссийский рабочий», «Сочи», региональные выпуски КП). По мере пополнения корпуса новыми газетами доля газет отдельного региона будет также расти, и это даст возможность исследовать с помощью корпуса особенности функционирования русского языка в отдельных российских регионах и странах ближнего и дальнего зарубежья.

IV. Перспективы развития корпуса

В настоящее время по адресу http://ruscorpora.ru/search-regional.html доступна рабочая версия регионального корпуса. В нем представлены газеты нескольких уровней — региональные выпуски центральных газет, газеты регионального уровня и местные издания — районные и городские. Временные рамки текстов 1996–2013 годы. География печатных изданий широка и охватывает все федеральные округа России, а также страны ближнего зарубежья (Беларусь, Молдова, Кыргызстан), страны Балтии. В корпус включены издания, разнообразные во многих отношениях: общественно-политические и отраслевые, деловые и вечерние, предназначенные для досуга, ежедневные и еженедельные, печатные издания и электронные СМИ. Помимо корреспондентских материалов в корпусе присутствуют отклики и комментарии непрофессиональных авторов, в которых реализуется обратная связь газеты с читателем. Корпус используется в качестве экспериментальной базы исследования региональных особенностей русской речи, что отражено в публикациях [Кустова, Савчук 2013; Рычкова 2014; Рычкова, Станкевич 2014; Шайкевич, Савчук 2014 и др.].

Корпус можно рассматривать как пилотный проект с заданными параметрами развития будущего большого корпуса региональных СМИ. Основное направление в развитии корпуса — формирование полноценных газетных подкорпусов для разных регионов, так чтобы все регионы были в нем равномерно представлены. Ориентиром

могут служить параметры Гродненского корпуса, то есть объем для каждого региона должен быть не менее 1,5–2 млн словоупотреблений, и в нем должны присутствовать газеты разных уровней. Исходя из этого задачей ближайшего будущего является увеличение доли текстов районных газет и читательских комментариев в корпусе российских СМИ. Другой первоочередной задачей является пополнение корпуса за счет изданий тех регионов, которые пока слабо в нем представлены. Русскоязычные газеты ближнего и дальнего зарубежья пополнят блок зарубежных СМИ. Гродненский корпус, в случае продолжения совместного проекта, будет пополняться газетами других районов области.

Другим вектором развития корпуса, наряду с территориальным, является временной. В настоящее время корпус содержит тексты только современных газет, но в него можно включать коллекции региональных газет 1960–1980-х годов или более ранних периодов. Однако старые газеты менее доступны и подготовка их потребует больших усилий, чем работа с современными электронными изданиями.

Требование разнообразия обусловливает поиск и включение в корпус газет с разными целевыми аудиториями, которые различаются по величине, возрасту, профессиональной и политической ориентации, — разумеется, с учетом реального положения дел в сфере региональных СМИ. Таким образом, целью дальнейшего развития регионального корпуса является создание сбалансированного ресурса, в котором разные регионы были бы представлены с достаточной полнотой, необходимой для проведения лексико-семантических и грамматических исследований.

Предполагается также дальнейшее развитие корпусного инструментария: организация поиска по n-граммам и графическое представление статистической информации — распределения частот употребления словоформ по разным осям — по годам или по регионам. В целях совершенствования интерфейса планируется разработать справочные страницы с информацией о регионах, включая карты, с кратким описанием периодических изданий — все это упростит и облегчит выбор подкорпуса и сделает работу с ним более занимательной. Справочный раздел корпуса можно будет расширять и путем размещения на сайте работ с результатами исследований, проведенных с использованием корпуса.

Мы надеемся на сотрудничество с коллективами региональных исследовательских центров — объединение усилий поможет осуществить проект.

Литература

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». Санкт-Петербург, 2013. С. 344–352.

Рычкова, Л. В., Станкевич, А. Ю., Бодак Ю. А. Принципы создания газетного корпуса СМИ Гродненщины // Компьютерная лингвистика: научное направление и учебная дисциплина [Текст] сборник научных статей. Вып. 2. / Отв. ред. В. И. Коваль [и др.] М-во образования РБ, ГГУ им. Ф. Скорины; Научно-методический центр русистики. Гомель: ГГУ им. Ф. Скорины, 2012. С. 88–92.

Рычкова Л. В. Специальная лексика в языке региональных СМИ // Терминология и знание. Вып. IV: материалы IV Междунар. симпозиума, Москва, 6–8 июня 2014 г. / Под ред. С. Д. Шелова. М., 2014. С. 157–172.

Рычкова Л. В., Станкевич А. Ю. «Диалог» языков на страницах СМИ Гродненщины // Актуальные проблемы теории дискурса: материалы Междунар. электрон. конф., Актюбинск, 30 мая 2014 г. / «Actual problems of the theory of discourse» proceedings of the international scientific e-conference, May 30, 2014. Актобе: Актюбинский регион. гос. ун-т им. К. Журбанова. С. 51–61.

Рычкова Л. В. Изучение социально-коммуникативных особенностей русского языка на основе использования лингвистических корпусов региональных средств массовой информации (на примере корпуса Гродненщины). // Universitas Catholica Rosenbergensis Studia Russico-Slovaca. Verbum. Ruzomberok, 2014. С. 48–58.

Савчук С. О. Корпус современной русской прессы: из опыта создания и использования // Труды Международной конференции «Корпусная лингвистика — 2011». СПб: Санкт-Петербургский государственный университет, 2011. С. 149–154

Савчук С. О. Корпусное исследование вариантов родовой при-

надлежности имен существительных в русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». М.: РГГУ, 2011. С. 562–580

Станкевич А. Ю. Технология сбора и систематизации электронного контента для корпуса русскоязычных СМИ Гродненщины // Карповские научные чтения: сб. науч. ст. Вып. 8: в 2 ч. / редкол.: А. И. Головня (отв. ред.) [и др.]. Минск: Белорус. Дом печати, 2014. Ч. 2. С. 6–10.

Шайкевич А. Я., Савчук С. О. Анализ лексико-семантических особенностей региональной прессы (на примере газет Гродненского региона Беларуси) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). М.: Изд-во РГГУ, 2014. С. 585–598.

S. O. Savchuk
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
savsvetlana@mail.ru

THE CORPUS OF REGIONAL RUSSIAN-LANGUAGE NEWSPAPERS IN RUSSIA AND ABROAD

The article describes the corpus of regional and foreign press as a new module within the RNC. Unlike newspaper corpus including only materials of the national press, the new corpus focuses on texts from regional publications. The corpus presently consists of the following subcorpora: 1) newspapers of the Grodno region (7 editions, 1.9 million tokens in total), prepared jointly with the specialists of the Grodno University, 2) Russian regional press of 2010s (13 newspapers, about 2 million tokens), 3) Russian regional press of 1990-2000 (40 newspapers, 2.6 million tokens), 4) regional releases of the national newspaper "Komsomolskaya Pravda" (6.5 million tokens). Our approach to selecting texts, the standards and software for word processing, organization of the corpus inter-

face are described in the article as well as the prospects of the development of corpus composition and searching instruments.

The texts of the regional corpus are provided with morphological, semantic and detailed metatextual annotation. The corpus is freely available at the address http://www.ruscorpora.ru/search-regional.html. The main purpose of this corpus as we see it is to provide researchers with a tool of studying the regional variation of the Russian language.

Key words: newspaper corpus, regional press, composition and structure of corpus, text annotation

References

Kustova G. I., Savchuk S. O. [The study of lexical-semantic and socio-cultural specifics of the Russian language on the territory of Belarus (on the material of mass media texts)]. *Trudy Mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2013"* [Proceedings of International Conference "Corpus Linguistics – 2013"]. St. Petersburg, 2013. pp. 344–352. (In Russ.)

Rychkova, L. V., Stankevich, A. Yu., Bodak Yu. A. [The principles of creating newspaper corpus of the Grodno region]. *Komp'yuternaya lingvistika: nauchnoe napravlenie i uchebnaya distsiplina.* [Computational linguistics: research area and academic discipline]. Vol. 2. V. I. Koval' et al. (eds.). Gomel, GSU them. F. Skorina, 2012, pp. 88–92. (In Russ.)

Rychkova, L. V. [Special vocabulary in regional media]. *Terminologiya i znanie. Vyp. IV: materialy IV Mezhdunar. simpoziuma* [Terminology and knowledge. Vol. IV: proceedings of the IV Intern. Symposium]. S. D. Shelov (ed.). Moscow, 2014, pp. 157–172. (In Russ.)

Rychkova L. V. [Studying the socio-communicative features of the Russian language at the base of regional mass-media text corpora: The case of the Corpus of Grodno region media]. *Universitas Catholica Rosenbergensis Studia Russico-Slovaca. Verbum.* Ruzomberok, 2014, pp. 48–58. (In Russ.)

Rychkova L. V., Stankevich A. Yu. ["Dialogue" of languages in the media of Grodno region]. *Aktual'nye problemy teorii diskursa: materialy Mezhdunar. elektron. konf.* [Actual problems of the theory of discourse: proceedings of the international scientific e-conference, May 30, 2014]. Aktobe, Aktobe region. state Univ. K. Gurbanov, pp. 51–61. (In Russ.)

Savchuk S. O. [The Corpus of Modern Russian Press: Compilation and Use]. *Trudy Mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2011"* [Proceedings of International Conference "Corpus Linguistics – 2011"]. St. Petersburg, St. Petersburg State Univ. Publ., 2011, pp. 149–154. (In Russ.)

Savchuk S. O. [Corpus-based study of morphological variability: Variation in gender forms of Russian nouns]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog – 2011"]. Moscow, RSUH Publ., 2011, pp. 562–580. (In Russ.)

Stankevich A. Yu. [Technology of collecting and organizing electronic content for Russian-language mass-media of Grodno region]. *Karpovskie nauchnye chteniya: sb. nauch. st.* [Karpovsky readings: collection of scientific articles]. Vol. 8: in 2 parts. A. I. Golovnya (ed.). Minsk, Belarusian printing house, 2014, p. 2, pp. 6–10. (In Russ.)

Shajkevich A. Ya., Savchuk S. O. [Distributional-statistical analysis of regional press (newspapers of Grodno region of Belarus)]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog – 2014"]. Moscow, RSUH Publ., 2014, pp. 585–598. (In Russ.)