С.О. Савчук, ИРЯ им. В.В. Виноградова РАН

КОРПУС СОВРЕМЕННОЙ РУССКОЙ ПРЕССЫ: ИЗ ОПЫТА СОЗДАНИЯ И ИСПОЛЬЗОВАНИЯ [1]

1. Газетный текст как объект корпусной лингвистики

Газетный материал обладает целым рядом особенностей, по достоинству оцененных создателями текстовых корпусов. Во-первых, это *стилистическая неоднородность:* наряду с публицистическими текстами газета включает тексты официально-деловые, научно-популярные, художественные. Во-вторых, *жанровая неоднородность*, обеспеченная многообразием публицистических жанров, – от строгих, максимально стандартизованных (информационное сообщение, хроника) до свободных жанров (репортаж, очерк, эссе). В-третьих, *тематическое разнообразие* текстов. В-четвертых, ориентированность газет на разные целевые аудитории отражается на тематическом, жанровом, языковом составе как разных изданий в целом, так и разных рубрик внутри одного издания. *Язык газеты* отражает текущее языковое употребление и оказывает обратное влияние на языковую практику. Как уже многократно отмечалось, в настоящее время языковой вкус эпохи формируется не художественной литературой, а газетным текстом. Наконец немаловажным положительным фактором является доступность газетного материала. Все эти особенности объясняют привлекательность газетных текстов для создателей

[1] Работа выполнена при поддержке: Программы ОИФН РАН «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации»; Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика»; РФФИ (грант 10-06-00151-а).

корпусов, включающих газетный материал в состав всех больших корпусов.

Газетные тексты на русском языке представлены в составе нескольких электронных корпусов. Самая большая коллекция входит в состав полнотекстовой **базы «Интегрум»**. Содержит тексты российских газет за период с середины 1990-х гг. до настоящего времени, список которых постоянно пополняется. Поисковые средства, обеспечивающие выполнение лексических запросов, а также дополнительные сервисы делают базы «Интегрума» привлекательным исследовательским ресурсом [1]. Однако доступ к текстам возможен только по платной подписке.

Компьютерный корпус текстов русских газет конца XX-го века (Лаборатории общей и компьютерной лексикологии и лексикографии МГУ им. М.В. Ломоносова)[2] содержит тексты 13 газет за 1994-1997 год. В корпусе представлены газеты разных типов – ежедневные и еженедельные, центральные и местные, общие и профессионально ориентированные, различной политической принадлежности. В корпус включены полные номера газет, что создает относительно объективную картину реального соотношения текстов различных типов и жанров в газетном материале. Общий объем корпуса превышает 11 млн словоупотреблений. Тексты и словоформы снабжены метатекстовой, морфологической, лексико-семантической, морфемной аннотацией. Глубоко аннотированная ядерная часть составляет более 1 млн словоупотреблений, фрагмент объемом 200 тыс. словоупотреблений находятся в открытом доступе.

Корпус российских газет 90-х годов XX в. (отдел Машинного фонда ИРЯ им. В.В. Виноградова РАН)[3]. Содержит тексты 9 центральных газет, небольшие тексты представлены

---

[1] Никипорец-Такигава Г. Integrum: точные методы и гуманитарные науки. М.: Летний Сад, 2006 // РЯЗР, 2007, №3

[2] **http://www.philol.msu.ru/~lex/corpus/corp_descr.html**

[3] **http://cfrl.ru/newspap.shtm**

целиком, большие по объему – во фрагментах. Всего в корпусе содержится более 29 тыс. фрагментов, общим объемом около 7,5 млн словоупотреблений. Тексты индексированы по метатекстовым признакам, включая жанровые и тематические характеристики. Обеспечен лексический поиск и поиск по морфемам и словообразовательной модели. Корпус размещен в открытом доступе.

Подкорпус газетных текстов в составе основного корпуса письменных текстов **НКРЯ** первоначально был сформирован как представительный корпус современной прессы и включал тексты более 100 газет, относящихся к периоду 2000-2004 годов и сбалансированных по таким параметрам, как место выхода издания, периодичность выхода, тематическая и политическая ориентированность, возрастные, количественные и профессиональные признаки аудитории. В соответствии с принципами формирования НКРЯ подкорпус содержит только целые тексты, как правило, газеты включаются целиком. Все тексты имеют метатекстовую, морфологическую и семантическую разметку (часть текстов вошла в корпус со снятой омонимией).

В дальнейшем, по мере развития диахронической составляющей НКРЯ, газетный подкорпус пополнялся (и продолжает пополняться) материалами газет, относящихся к разным периодам истории. Наиболее значительные коллекции относятся к периодам 1990-х годов, 1960-1980-х годов, 1900-1910 гг. Газетные тексты XX в. попадают в корпус либо в оцифрованном виде (в основном в формате djvu), либо проходят полный цикл обработки от сканирования бумажных оригиналов до перевода файлов в корпусной формат xml. Трудоемкостью и дороговизной процессов обработки «исторического» газетного материала объясняется то обстоятельство, что мы еще далеки от идеально сбалансированного корпуса газет XX века. Отчасти отсутствие газетного материала за какие-то периоды

компенсируется текстами журнальной публицистики, но в планы развития НКРЯ постоянным пунктом включено и заполнение лакун в газетном подкорпусе.

2. Опыт создания Корпуса современной русской прессы

Корпус СМИ 2000-х годов, в отличие от газетного подкорпуса НКРЯ, был создан очень быстро, в течение 2009 г.[1]. Первоначальная цель создания нового газетного корпуса была чисто утилитарной – увеличить в НКРЯ долю текстов новейшего периода (после 2005 г.). Разработка корпуса была во многом экспериментальной: был составлен предварительный дизайн, задающий параметры текстов, широко использовались технологии автоматической обработки текстов, сводящие к минимуму ручной труд, применялась упрощенная метатекстовая аннотация. Следует признать, что в целом эксперимент оказался удачным: по уровню посещаемости Корпус СМИ 2000-х годов занимает 2-е место после основного корпуса письменных текстов, каких-то серьезных ошибок в представлении материала отмечено не было, по количеству ошибок, о которых сообщают пользователи, газетный корпус не отличается от других корпусов.

2.1. Принципы отбора текстов, состав и структура корпуса

Тексты для корпуса были предоставлены компанией Corpus Technologies. Было отобрано 7 изданий: центральные газеты общего содержания «Известия», «Труд-7», «Комсомольская правда»; деловая газета «РБК Daily», спортивная газета «Советский спорт», а также материалы информационных агентств – РИА «Новости» и «Новый регион». Включение материалов информационных агентств отражает реальное

[1] Официальное сообщение об открытии доступа к корпусу было сделано в феврале 2010 года.

положение дел в СМИ, при котором бумажная пресса все больше уступает свою аудиторию интернет-изданиям (например, сайт газеты РБК Daily ежедневно посещают 80-100 тыс. пользователей, что сопоставимо с тиражом бумажной версии, составляющим 80 тыс. экземпляров).

Согласно разработанному проекту корпуса все издания представлены приблизительно в равных пропорциях за каждый период[1]. В отличие от основного корпуса, в который включались газеты и журналы целиком, для корпуса СМИ отбирались тексты объемом не менее 200 словоупотреблений, что должно было обеспечить жанровое и языковое разнообразие текстов. При установленном ограничении на объем в корпус попадает больше текстов, относящихся к свободным жанрам (авторская заметка, корреспонденция, интервью, статья, очерк и пр.), и меньше более стандартизованных в языковом отношении мелких текстов (новостных сообщений, объявлений и т.д.). Распределение текстов корпуса по источнику и дате представлено в табл. 1.

В связи с необходимостью создания большого корпуса в короткий срок было решено отказаться от ручной метаразметки газетных текстов и использовать только ту метаинформацию, которая содержалась в файлах. Упрощенная метатекстовая аннотация газетного корпуса обеспечивает отбор подкорпуса по следующим параметрам: имя автора, название текста, название издания, год издания.

Таблица 1. Количественное распределение текстов в корпусе современной прессы

Название издания	**2000-2003**	**2004**	**2005**	**2006**	**2007**	**2008**
Известия		1 530 975	993 069	2 702 044	2 396 735	-

[1] Исключение было сделано для газет «Комсомольская правда» и «Труд-7»: поскольку они слабо представлены в основном корпусе, в газетный корпус были включены тексты этих газет, начиная с 2000 г.

КП	11009071	3 688 145	4 531 914	5 003 650	5 822 360	15 627
Труд-7	18626621	3 965 547	4 269 787	3 964 161	4 905 784	23 692
РБК Daily		2 924 667	2 837 808	4 264 547	6 325 340	136 186
Советский спорт		1 871 121	1 631 639	1 589 181	1 712 326	-
РИА Новости		896 616	1 658 919	1 848 540	2 229 665	666 689
Новый регион		2 674 301	3 185 938	1 626 594	1 518 399	244 345
Итого	**29635692**	**17551372**	**19109074**	**20998717**	**24910609**	**1086539**

2.2. Технология обработки текстов

Как показал опыт подготовки газетных текстов для НКРЯ, применение стандартной технологии потребовало бы большой затраты времени и ресурсов. Поэтому единственно возможным решением было сведение к минимуму ручного труда при конвертации файлов в формат корпуса, метатекстовой аннотации, морфологической и семантической разметки. Корректура текстов не производилась, так что газетный корпус в полной мере отражает практику печати.

Тестирование бета-версии корпуса сразу же обнаружило огромное количество повторяющихся контекстов при ответе на многие запросы. Анализ повторов выявил основные причины появления дублетных текстов в корпусе. Все они так или иначе связаны с периодичностью СМИ, но часть повторов объясняется особенностями организации текстов в интернет-изданиях. Перечислим основные типы выявленных повторов.

1. Один и тот же текст повторяется в разных разделах сайта. Это может быть связано либо с политематичностью текста, следствием чего является его размещение в нескольких тематических рубриках, либо может быть обусловлено структурой сайта. В частности, сайт «Новый регион» организован таким образом, что региональные разделы наряду с текстами, посвященными местным новостям, содержат некоторую общую

часть, повторяющуюся во всех или в значительной части разделов. Таким образом, при загрузке текстов с сайта один и тот же текст может попасть в корпус из разных региональных разделов. Кроме того, дублеты могут создаваться при одновременном сохранении двух версий одного и того же текста – веб-страницы и версии для печати.

2. Другая группа повторов также связана с особенностями организации интернет-изданий. Если газета выходит 1 раз в сутки, то интернет издание существует в режиме реального времени, так что один и тот же новостной текст может несколько раз повторяться в течение одного дня: как отдельная новость, как новость в составе обзора, как часть итогового новостного выпуска в конце дня. Словесный состав текстов при этом практически неизменен.

3. Третий источник дублетных контекстов имеет отношение к особенностям организации текстов в газете: одно и то же событие может стать темой разножанровых текстов – анонса, заметки, корреспонденции, комментария, статьи и др. Если все эти тексты написаны одним автором, то повторы обширных фрагментов неизбежны. Кроме того, довольно часты случаи, когда в издании на протяжении 2-3 лет дословно повторяются тексты, приуроченные к какой-нибудь памятной дате. Наконец, мы столкнулись со случаями заимствования текстов из одного издания в другое или публикации одинаковых текстов одного автора в разных изданиях. Все эти подводные камни следует учитывать при составлении корпусов на основе интернет-источников.

Дублеты 1-го и 2-го типов были устранены программными средствами (автор программ Л. Алексеевский). Некоторое количество повторов 3-го типа сохранилось в корпусе: поскольку они определяются спецификой газетных текстов, их устранение означало бы вмешательство в текст, что не входит в задачу составителей корпусов.

2.3. Перспективы развития корпуса современной русской прессы

Корпус СМИ проектировался как актуальный, пополняемый, поэтому в процессе его создания разрабатывались и принципы его информационной поддержки и развития. Оптимальным было признано следующее решение: сохраняя существующие пропорции текстов различных изданий, обеспечить ежегодное пополнение корпуса в объеме 20 млн словоупотреблений. Таким образом, к 2013 г. планируется удвоить объем корпуса и довести его до 200 млн словоупотреблений. В настоящее время идет подготовка второй очереди корпуса, к концу 2011 г. планируется пополнить его текстами за 2008-2010 гг. в объеме 50 млн с/у.

3. Опыт использования Корпуса современной русской прессы в лингвистических исследованиях

Отдельной самостоятельной задачей является использование газетного корпуса в различных лингвистических исследованиях с целью анализа данных, полученных на сопоставимых по объему корпусах различных типов – сравнительно однородном газетном корпусе и сбалансированном корпусе современных текстов в составе НКРЯ[1]. Кратко остановимся на некоторых результатах.

---

[1] См., например: Савчук С.О. Корпусное исследование вариантов родовой принадлежности имен существительных в русском языке // Компьютерная лингвистика и Интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (2011) (в печати); Савчук С.О. Опыт корпусного исследования морфологической вариативности: варианты родительного падежа множественного числа существительных мужского рода // Компьютерная лингвистика и Интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (2010). Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 419-429

Газетный корпус привлекался к исследованию соотношения морфологических вариантов. Основные наблюдения следующие.

1. Газетный корпус содержит однородные тексты главным образом современной тематики, потому частотность слов пассивного словаря в них заведомо ниже, так что снижается вероятность обнаружения конкретного падежного варианта.

2. Распределение частотных слов в целом аналогично существующему в корпусе 2-ой половины XX в.

3. Тенденция к распределению окончаний *–ов* и ∅, характерная для современных текстов, выражена в корпусе СМИ более ярко, чем в современном сбалансированном корпусе, что может объясняться большей однородностью корпуса прессы. Слова актуальные подчиняются общим правилам, закрепляют за собой стандартное окончание *–ов*, утрачивая нулевой вариант. Слова пассивного запаса вариативность сохраняют или употребляются с окончанием ∅. Так, например, варианты род. п. мн. ч. представлены в корпусе СМИ в полном составе с такими показателями: *кадетов*[1] – *кадет*[1] 66/1 (ср. в сбалансированном корпусе 25/10), то есть слово даже увеличило частотность, вероятно, отчасти в связи с выходом на телеэкраны сериала «Кадеты»; *кадетов*[2] – *кадет*[2] 16/1 (в сбалансированном корпусе 48/3); в значимых количествах представлена форма род. мн. лексемы *кадет*[3] ‘спортсмен-юниор до 16 лет’: *кадетов*[3] – *кадет*[3] 16/0. Отсутствие нулевых форм весьма красноречиво. Соотношение *гренадеров* – *гренадер* 25/0 (против 13/9 в сбалансированном корпусе), при этом слово употребляется в переносном значении ‘человек высокого роста’, главным образом по отношению к спорсменам.

Таким образом, корпус современной русской прессы представляет собой новый ресурс для исследования процессов в современном русском языке; он может использоваться как самостоятельно, так и в сопоставлении с данными корпуса современных текстов в составе НКРЯ.

Дополнительным подтверждением высказанного предположения послужил анализ употребления слов рассматриваемой группы в новом корпусе – СМИ 2000-х годов. Он содержит однородные тексты главным образом современной тематики, потому частотность слов пассивного словаря в них заведомо ниже, так что снижается вероятность обнаружения конкретного падежного варианта. Соотношение вариантов род. мн. существительных разных подгрупп выглядит следующим образом.

У слов подгруппы 1 (*солдат* и *партизан*) соотношение вариантов не отличается от данных, полученных по корпусу 2-ой пол. XX в. (далее – сбалансированный корпус).

Слова 2-й подгруппы представлены в корпусе СМИ в полном составе с такими показателями: *кадетов*[1] – *кадет*[1] 66/1 (ср. в сбалансированном корпусе 25/10), то есть слово даже увеличило частотность, вероятно, отчасти в связи с выходом на телеэкраны сериала «Кадеты»; *кадетов*[2] – *кадет*[2] 16/1 (в сбалансированном корпусе 48/3); в значимых количествах представлена форма род. мн. лексемы *кадет*[3] ‘спортсмен-юниор до 16 лет’: *кадетов*[3] – *кадет*[3] 16/0. Отсутствие нулевых форм весьма красноречиво. Соотношение *гренадеров* – *гренадер* 25/0 (против 13/9 в сбалансированном корпусе), при этом слово употребляется в переносном значении ‘человек высокого роста’, главным образом по отношению к спорсменам. Остальные формы представлены в следующих пропорциях: *гардемаринов* – *гардемарин* 21/0 (против 12/2 в сбалансированном корпусе); *карабинеров* – *карабинер* 29/0 (против 11/0), во всех контекстах речь идет об итальянских карабинерах; *рекрутов* – *рекрут* 30/0 (против 26/1), слово употребляется в расширенном значении не только по отношению к солдатам-новобранцам, но и по отношению к работникам, спортсменам, бандитам, и даже государствам («прием новеньких государств-рекрутов» в ЕС).

Формы слова *гусар* представлены в пропорции *гусаров – гусар* 5/25 (против 10/23 в сбалансированном корпусе), из них 13 случаев употребления в контексте «Эскадрон гусар летучих».

Из 3-й подгруппы 2 слова (*рейтар* и *улан*) вообще не представлены в корпусе СМИ в форме род. мн. У других слов группы интересующие нас формы соотносятся следующим образом: *драгун – драгунов* 6/0 (против 9/0), *кирасир – кирасиров* 1/1 (против 9/7), *янычар – янычаров* 7/2 (против 14/0).

Как видим, тенденция к распределению окончаний *–ов* и ∅, характерная для современных текстов, выражена в корпусе СМИ более ярко, чем в современном сбалансированном корпусе, что может объясняться большей однородностью корпуса прессы. Слова актуальные подчиняются общим правилам, закрепляют за собой стандартное окончание *–ов*, утрачивая нулевой вариант. Слова пассивного запаса вариативность сохраняют или употребляются с окончанием ∅. Не соответствуют общей тенденции слово *гусар*, у которого ∅ вариант поддерживается устойчивым контекстом, и слово *янычар*, у которого несколько увеличилось количество вариантов *–ов*.