

Т. И. Резникова, М. В. Кополев

ЛИНГВИСТИЧЕСКИ АННОТИРОВАННЫЕ КОРПУСА РУССКОГО ЯЗЫКА (ОБЗОР ОБЩЕДОСТУПНЫХ РЕСУРСОВ)

На протяжении многих лет русский язык оставался одним из немногих крупнейших языков, не имеющих собственного корпуса текстов, доступного для исследователей всего мира. Между тем в последнее время, когда появление корпусов привело к существенным изменениям исследовательских практик, если не к серьезным сдвигам в осознании феномена языка [Fillmore 1992], необходимость создания корпуса русских текстов была ясно осознана как российскими, так и зарубежными лингвистами. Неудивительно, что почти одновременно в разных странах возникли научные проекты по разработке корпусов русского языка для размещения в Интернете.

В 1999 г. в университете г. Тюбинген (Германия) началась работа над корпусом, в основу которого лег Уппсальский корпус русских текстов. В Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ с 2000 г. создается Корпус текстов русских газет конца XX века. Через год на Отделении славянских и балтийских языков и литератур Хельсинкского университета началась разработка Хельсинкского аннотированного корпуса ХАНКО. Почти одновременно в рамках программы РАН «Филология и информатика» стартовал проект по созданию Национального корпуса русского языка.

Конечно, существующие русские корпуса не исчерпываются вышеперечисленными ресурсами. Самым известным корпусом русского языка долгие годы оставался уже упомянутый Уппсальский корпус, созданный шведскими русистами. В первоначальном виде его объем составлял 1 млн. словоупотреблений, в нем отсутствовала лемматизация и морфологическая разметка. Тем самым ни по размеру, ни по составу информации, которой снабжены тексты, Уппсальский корпус, к сожалению, не отвечает современным стандартам составления корпусов. Во многом именно осознание его недостаточности для адекватного представления языка, ограниченной применимости для лингвистических исследований, а

также устарелости материалов привело к активизации работ по созданию альтернативных корпусов. Из других проектов, предшествовавших эре интернет-корпусов, следует упомянуть Машинный фонд русского языка, создававшийся с конца 1980-х гг. под руководством В. М. Андрющенко и А. П. Ершова [Андрющенко 1989]. Хотя проект не был завершен и не привел к разработке представительного корпуса, в его рамках были собраны коллекции текстов разного типа. К проектам недавнего времени относятся корпус текстов по современной публицистике (1990-е гг.), создававшийся в Отделе экспериментальной лексикографии Института русского языка А. Н. Барановым, М. Н. Михайловым и Г. О. Сидоровым [1998], параллельный русско-финский корпус в университете г. Тампере в Финляндии, разрабатываемый М. Н. Михайловым и Х. Томмолой [Михайлов 2003], а также корпус СПбГУ, создаваемый под руководством В. Б. Касевича [Венцов, Касевич 1998]¹. Указанные корпуса не являются доступными для широкого круга пользователей, информацию об их исследовательском потенциале, их характеристики можно извлечь пока только из созданных разработчиками описаний.

В настоящем обзоре будут рассматриваться существующие общедоступные корпуса современного русского языка, сведения о которых, как нам кажется, представляют наибольший практический интерес для исследователей языка и всех интересующихся теми или иными аспектами его функционирования². Кроме НКРЯ, это следующие корпуса:

Хельсинкский аннотированный корпус (ХАНКО)	(http://www.ling.helsinki.fi/projects/hanco/);
Тюбингенский корпус русского языка (ТК)	(http://www.sfb441.uni-tuebingen.de/b1/rus/korpora.html);
Корпус газетных текстов русского языка конца XX века (КГТ)	(http://www.philol.msu.ru/~lex/corpus/)

Поскольку Национальному корпусу русского языка посвящены почти все статьи настоящего сборника, в нашем обзоре сведения о нем будут привлекаться лишь в сопоставительном аспекте.

¹ Об истории корпусных разработок в России см. также [Шаров 2003].

² В настоящем обзоре не учитываются два *диахронических* корпуса, которые, в силу особенностей электронных публикаций древних текстов, обладают целым рядом существенных отличий. Укажем их адреса: Регенсбургский диахронический корпус русского языка (http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/Corpus/kiss/index-ru.htm); российский проект «Манускрипт» (<http://mns.udsu.ru:1300/>).

1. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОЕКТОВ

1.1. Тюбингенский корпус (ТК). Тюбингенский корпус русского языка создавался в рамках междисциплинарной исследовательской программы, посвященной теоретическим и эмпирическим основам грамматического исследования («Linguistische Datenstrukturen: Theoretische und empirische Grundlagen der Grammatikforschung»). В проекте участвуют различные исследовательские группы, каждая из которых занимается решением конкретных научных задач, связанных с общей тематикой программы. Одна из таких групп с 1999 по 2004 г. работала над описанием форм обращения и вежливости в славянских языках, прежде всего в русском и чешском, на основе корпусов этих языков. К началу проекта корпусной анализ языковых явлений чешского языка представлялся вполне выполнимой задачей благодаря существованию Чешского национального корпуса — одного из первых по времени создания представительных корпусов славянских языков. Для русского языка аналогичная задача оказалась более трудоемкой: из-за отсутствия корпуса, который бы отражал современное языковое употребление, изучение того или иного явления требовало предварительного создания базы для исследования. Тем самым построение по возможности большого и представительного корпуса русского языка стало одной из целей проекта.

Общее руководство проектом осуществлял профессор Отделения славистики университета г. Тюбинген Т. Бергер, в проекте приняли участие преподаватели и аспиранты отделения. ТК стал первым корпусом русского языка, появившимся в открытом доступе в сети Интернет.

В настоящее время проект по исследованию форм обращения и вежливости и, соответственно, работа над корпусом завершены, поэтому основные параметры ТК, которые будут описываться в настоящей статье, в будущем, по всей вероятности, не претерпят существенных изменений. Заметим, что остальные три корпуса продолжают активно развиваться.

1.2. Корпус газетных текстов (КГТ). «Компьютерный корпус газетных текстов русского языка конца XX века» был подготовлен в течение 2000–2002 гг. в Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ. Создание корпусов и других коллекций текстов является одним из

основных направлений деятельности лаборатории. Здесь в разное время разрабатывались, например, корпус русских литературно-художественных текстов XIX-XX вв., корпуса русских рассказов, психологических самоописаний, сочинений иностранцев на русском языке, параллельных текстов на русском и английском языках. В отличие от газетного корпуса, все прочие текстовые коллекции не являются общедоступными.

Проект по созданию корпуса газетных текстов осуществляется под руководством заведующего Лабораторией общей и компьютерной лексикологии и лексикографии А. А. Поликарпова, в разработке корпуса принимают участие сотрудники Лаборатории [Виноградова и др. 2001].

В настоящее время в Интернете доступен небольшой тестовый фрагмент корпуса, более полная версия готовится к представлению.

1.3. Хельсинкский аннотированный корпус (ХАНКО). Корпус ХАНКО задуман как составная часть проекта «Функциональный синтаксис русского языка», работа над которым ведется на Отделении славянских и балтийских языков и литератур Хельсинкского университета (см. [Мустайоки 2005]). Постановка задачи предполагает необходимость внесения в корпус подробных лингвистических данных о входящих в его состав единицах и, соответственно, в значительной степени требует ручной обработки текстов. Это определило один из основных принципов построения корпуса — его направленность на максимальный охват грамматической информации, а не на объем материала.

Проектом по созданию корпуса ХАНКО руководит профессор Отделения славистики и балтистики А. Мустайоки, главным разработчиком корпуса является М. Копотев. В работе принимают участие сотрудники Хельсинкского и Тартуского университетов, а также русскоязычные студенты [Мустайоки, Копотев 2003].

В настоящее время осуществление проекта продолжается. В Интернете доступны результаты первого этапа работы — морфологически аннотированный корпус. В ближайшее время появится и синтаксическая разметка.

2. СОСТАВ И ОБЪЕМ КОРПУСОВ

2.1. Тюбингенский корпус. Из трех описанных выше корпусов самую большую и разнородную коллекцию текстов представляет

собой Тюбингенский корпус. Он включает в себя ряд подкорпусов, к которым применялись различные процедуры лингвистического анализа.

В основу ТК лег Уппсальский корпус, который благодаря тюбингенскому проекту стал доступен онлайн. Его объем составляет 1 млн. словоупотреблений, он состоит из 600 текстовых фрагментов, примерно в равной пропорции распределенных между художественной прозой и публицистикой. Художественные тексты включают отрывки из произведений 40 авторов, созданных в период 1960–88 гг. Публицистика охватывает период с 1985 по 1988 гг. и отражает различные тематические сферы. Второй подкорпус в составе ТК образуют тексты интервью. Построение этого подкорпуса очевидным образом было продиктовано тематикой исследования, приведшего к созданию всего корпуса — изучение форм обращения и вежливости. В данный подкорпус вошли тексты интервью из газет и журналов, свободно доступных в Интернете («Аргументы и факты», «Аргументы и факты (Владивосток)», «Арт-Петербург», «Биржа труда», «Ваша газета», «Ведомости», «Вестник», «Иностранец», «Интербизнес», «Киевские новости», «Литературная газета», «Мир денег», «Музыкальная газета», «На дне», «Натали», «Новая газета», «Новости Петербурга», «Огонек», «Отдыхай», «Психологическая газета», «Пять углов», «Пчела», «Сегодня», «Странник»), а также записи интервью на радиостанции «Эхо Москвы». Тексты относятся к периоду с 1996 г. Тематика интервью охватывает следующие сферы: политика и общественная жизнь, экономика, музыка, литература, жизнь молодежи, спорт. Объем этого подкорпуса — около 290 тыс. словоупотреблений.

Следующий подкорпус включает в себя полные тексты всех доступных в Интернете статей из журнала «Огонек» с 20 номера за 1996 год по 17-ый номер за 2002 г. Объем подкорпуса — около 9,19 млн. словоупотреблений.

Остальные подкорпуса образуют тексты художественной литературы. В отличие от Уппсальской части Тюбингенского корпуса, которая состоит из текстовых фрагментов, в эти подкорпуса были помещены целые тексты. Это, во-первых, литература XX в., куда относятся подкорпус детективных романов (тексты Ч. Абдуллаева, А. Адамова, Б. Акунина, С. Алексеева, М. Бакониной, В. Богомолова, Г. Брянцева, братьев Вайнеров, Д. Вересова, Г. Горпожакса, Ю. Дольд-Михайлика, В. Доценко, В. Клюевой, Л. Кожевникова,

Д. Корецкого, А. Левина, Р. Мир-Хайдарова, Л. Овалова, А. Ольбика, Т. Поляковой, Л. Пучкова, А. Рассказова, А. Рыбина, Е. Сартинова, Ю. Семенова, В. Сеницыной, А. Таманцева, В. Тевекеляна, В. Угрюмова, А. Щеголева, А. Щелокова, М. Юденич (общий объем — свыше 8 млн. словоупотреблений); подкорпуса произведений А. Марининой (свыше 200 тыс. словоупотреблений); А. Н. Рыбакова (свыше 170 тыс. словоупотреблений); А. и Б. Стругацких (свыше 550 тыс. словоупотреблений); И. Ильфа и Е. Петрова (около 200 тыс. словоупотреблений), М. А. Булгакова (свыше 840 тыс. словоупотреблений). Кроме того, поскольку в теоретические задачи тюбингенского проекта входит не только описание функционирования форм обращения и вежливости в современном языке, но и их исследование в диахроническом аспекте, в корпус были включены некоторые произведения художественной литературы 19 в. Это подкорпуса произведений Л. Н. Толстого (свыше 810 тыс. словоупотреблений), Ф. М. Достоевского (около 1,77 млн. словоупотреблений), И. С. Тургенева (около 450 тыс. словоупотреблений) и Н. С. Лескова (около 850 тыс. словоупотреблений). Общий объем подкорпусов художественной литературы (не считая uppsalsких художественных текстов) составляет свыше 14 млн. словоупотреблений.

2.2. Корпус газетных текстов. В КГТ вошли полные тексты выбранных номеров ряда российских газет на русском языке, опубликованных в 1994-1997 гг. При отборе материала авторы ориентировались на то, чтобы в корпусе были представлены издания различного типа: ежедневные и неежедневные, «левые» и «правые», центральные и местные, общие и профессионально ориентированные; иными словами, ставилась задача создания репрезентативной газетной выборки. В результате было отобрано 13 газет: «Завтра», «Известия», «Литературная газета», «Московский комсомолец», «Московские новости», «Независимая газета», «Новая газета (Понедельник)», «Новгородские ведомости», «Новгород», «Правда», «Правда-5», «Свободный Сахалин», «Томская неделя». Принципы отбора материала позволяют надеяться, что корпус сможет давать объективную картину соотношения в газетном материале текстов различных типов и жанров. Общий объем корпуса — свыше 11 млн. словоупотреблений.

Версия, доступная в настоящий момент в Интернете, в значительной степени отличается от целого корпуса. В нее вошли 446

текстов из 8 различных газет за 1997 г. общим объемом свыше 200 тыс. словоупотреблений. Предполагается, что интернет-версия корпуса будет расширена до размера в 1 млн. словоупотреблений.

2.3. Корпус ХАНКО. Самым небольшим объемом из всех рассматриваемых текстовых коллекций характеризуется корпус ХАНКО. В качестве источника текстового материала для корпуса был выбран один журнал, достаточно полно представляющий современную публицистику. Критериями при отборе послужили отражение в журнале широкого спектра публицистических жанров, тематическое разнообразие статей, высокий уровень владения стилистическими ресурсами русского языка их авторов. В результате был выбран журнал «Итоги»: в корпус вошли все крупные статьи из нескольких номеров за январь 2001 г. Общий объем корпуса составляет 100 тыс. словоупотреблений.

2.4. Национальный корпус русского языка. По замыслу создателей, основная часть Национального корпуса русского языка будет состоять из двух подкорпусов: корпус ранних текстов (начало XIX — середина XX века) и корпус современных текстов (середина XX — начало XXI века). Последний задуман как представительный корпус современного языка, включающий тексты самых различных жанров и типов, объемом 100 млн. словоупотреблений¹.

По состоянию на начало октября 2005 г. в Интернете была доступна часть корпуса современных текстов объемом свыше 60 млн. словоупотреблений, а также фрагмент корпуса ранних текстов объемом свыше 5 млн.

3. РАЗМЕТКА КОРПУСОВ

Один из постулатов аннотирования, сформулированных в 1993 году одним из создателей корпуса Lancaster-Oslo/Bergen (LOB) и Британского национального корпуса (BNC) Джеффри Личем, предполагает существование эксплицированных и доступных описаний лингвистической разметки: «*The scheme of analysis presupposed by the annotations — the annotation scheme — should be based on principles or guidelines accessible to the end-user*» [Leech 1993: 275].

Этот постулат Дж. Лича кажется очевидным, однако он реализуется не во всех анализируемых корпусах. Конечно, все четыре корпуса снабжены более или менее подробной справкой, но полу-

¹ Подробно о составе НКРЯ см. статью С. О. Савчук в настоящем сборнике.

чение точных сведений о том, что означает каждый параметр описания, какие теоретические или чисто практические мотивировки стоят за выделением той или иной единицы — все это часто остается неэксплицированным или комментируется в узкоспециализированных сборниках статей. В силу этого приведенное ниже сравнение принципов аннотирования корпусов базируется в ряде случаев не на описаниях разработчиков, а на собственных разысканиях авторов настоящего обзора.

3.1. Морфологическая разметка. Ко всем рассматриваемым корпусам (на всем их объеме или в некоторой их части) применялась морфологическая разметка. В Тюбингском корпусе морфологической разметкой снабжены следующие подкорпуса: Уппсальский корпус (тем самым именно в рамках тюбингского проекта изначально неаннотированный Уппсальский корпус получил морфологическую разметку), тексты М. А. Булгакова и И. С. Тургенева. Общий объем морфологически аннотированных текстов — 2,3 млн. словоупотреблений. Разметка осуществлялась при помощи статистического морфологического анализатора TnT, разработанного Т. Брантсом (университет г. Саарбрюкен) [Brants]. Программа не предназначена для анализа текстов на каком-либо конкретном языке, она может применяться к различным языкам и настраиваться с различными наборами параметров. Обучение программы для ее применения к русскому корпусу производилось на материале списков текстоформ и тэгов, а также с помощью размеченного вручную текста объемом 165 тыс. словоупотреблений. По оценке разработчиков корпуса, эффективность анализатора составляет 93-94% [Betsch, Meyer 2003].

В КГТ морфологическая разметка осуществлялась автоматически и полуавтоматически при помощи системы Диктум-1, разработанной в Лаборатории общей и компьютерной лексикологии и лексикографии под руководством А. А. Поликарпова [Kukushkina, Polikarpov 1996]. Процесс приписывания текстоформам грамматических показателей соответствует в этой системе, как правило, их разбиению на непересекающиеся классы. Так, признак *см* приписывается существительным мужского, *сж* — женского и *сс* — среднего рода. При этом признак *с* получают не все существительные, а только существительные с неустановленным родовым оформлением. Аналогичным образом трактуются и омонимичные формы. Им приписываются особые кластерные признаки. Так, например, де-

скриптор *e-ив* получают имена, у которых совпадают формы именительного и винительного падежа единственного числа. При этом дескриптор *e-и* присваивается только тем именам, у которых форма именительного падежа единственного числа не омонимична какой-либо другой или же была однозначно распознана как таковая. Сходным образом приписывается признак *e-в*. В результате применения такой системы разметки поиск, например, по лексеме *сердце* и признаку *e-н* (предложный падеж единственного числа) дает пустой результат, поскольку всем употреблениям лексики *сердце* в форме предложного падежа единственного числа, равно как и в формах именительного и винительного падежа присвоен дескриптор *e-ивн*. Таким образом, разработчики используют оригинальную систему аннотирования, последовательно выдерживая ее на материале всего корпуса. Вместе с тем, использование корпуса практически невозможно без предварительного знакомства с этой системой, однако сайт КГТ не содержит ее детального описания, что существенно затрудняет работу с корпусом для неподготовленного пользователя.

В ХАНКО процедура автоматической морфологической разметки осуществлялась при помощи системы автоматического распознавания русских слов RUSTWOL [Vilki 1997]. Затем средствами отдельного модуля проводилось автоматическое снятие омонимии, при этом использовались лишь те правила, которые обеспечивают стопроцентную точность при устранении некоторого варианта морфологического разбора (например, случаи управления предлога определенным падежом). На следующем этапе осуществлялась ручная проверка результатов анализа.

Разные стратегии, выбранные разработчиками трех корпусов, порождают принципиальные различия в соотношении запрашиваемой пользователем и выдаваемой системой информации. В ХАНКО пользователь получает все те примеры, которые точно соответствуют его запросу, но поиск при этом ведется на сравнительно небольшом объеме текстов (100 тыс. словоупотреблений). В доступной на сегодняшний день версии КГТ примеры ищутся по текстам объемом 200 тыс. словоупотреблений, но при этом пользователь зачастую должен формулировать запрос по кластерному признаку, следовательно, кроме нужных контекстов система выдает и те, в которых встречаются формы, омонимичные искомым. В ТК система может не только выдать «лишние» контексты (случаи не-

правильного разбора), но и пропустить те формы, которые подходят под запрос пользователя, в том случае, если они были неправильно разобраны анализатором. Однако благодаря значительному объему корпуса (2,3 млн. словоупотреблений) даже при наличии «потерянных» контекстов количество примеров, которые пользователь получает из ТК после отсеивания лишних, значительно выше, чем в случае поиска по корпусам ХАНКО и КГТ. Например, по запросу на формы страдательных причастий настоящего времени в именительном падеже множественного числа ХАНКО выдает 8 соответствующих примеров, КГТ — 22 контекста с формами именительного или винительного падежа, Тюбингенский корпус — 165 контекстов, из них 105 отвечающих запросу и 60 неправильно разобранных, в том числе 33 с омонимичными искомым формами винительного падежа. Что касается остальных 27 случаев неправильного разбора, то они в основном связаны с несклоняемыми именами и именами собственными (как страдательные причастия настоящего времени в именительном падеже множественного числа разбираются, например, следующие текстоформы: *Париже, Херсонесе, Десне, Ставрополье, Иоффе, Круазе, эссе*)².

На основном массиве текстов НКРЯ морфологическая разметка осуществляется автоматически, однако в части корпуса (на настоящий момент объемом свыше 4 млн. словоупотреблений) произведено ручное снятие омонимии и дополнительная коррекция результатов работы программы автоматического морфологического анализа³. Более подробно инвентари грамматических параметров, используемых при морфологической разметке, будут сопоставляться в разделе 4.3.

² Эти два пересекающиеся класса имен — собственные и несклоняемые — представляют, по всей вероятности, один из основных источников ошибок разметки в Тюбингенском корпусе. Так, фамилии на *-дзе* встречаются в аннотированном корпусе 37 раз. В зависимости от контекста они получают следующую разметку: отадъективное наречие, сравнительная степень наречия или прилагательного, существительное женского рода в форме единственного числа предложного или дательного падежа или среднего рода в форме единственного числа именительного падежа, прилагательное или причастие настоящего времени в форме именительного или винительного падежа единственного числа среднего рода или множественного числа, императив 1-го или 2-го лица множественного числа и форма 2-го лица множественного числа настоящего времени.

³ См. статью О. Н. Ляшевской и др. в настоящем сборнике.

3.2. Синтаксическая разметка. Как известно, процесс внесения синтаксической информации в корпус в меньшей степени поддается автоматизации и требует больше ручной работы, чем морфологическая разметка. Поэтому на текстовых коллекциях большого объема задача детального синтаксического анализа обычно не ставится.

В ТК к элементам синтаксической разметки можно отнести аннотирование устойчивых синтаксических оборотов, выступающих в качестве эквивалента слова (например, *потому что, все равно, как раз, прежде всего*). В КГТ размечены предложные группы, т. е. можно осуществлять поиск по сочетанию предлога с именной группой в заданном падеже, определяя при этом входящее в ее состав существительное по признаку одушевленность/неодушевленность (запросы по «именным синтаксемам» формулируются при помощи предлога и соответствующего вопросительного местоимения: *в ком, в чем, в кого, во что* и т. п.; всего в корпусе 109 типов таких групп).

Закономерно, что разработчики хельсинкского проекта, ограничившись небольшим объемом текстов с целью максимального охвата грамматической информации, предполагают включить в корпус детальную синтаксическую разметку. На настоящий момент результаты этой работы еще не доступны в интернет-версии ХАНКО, однако их размещение планируется осуществить в ближайшее время. По замыслу разработчиков, синтаксическая разметка будет нетривиальным образом совмещать (точнее, предлагать параллельно) две схемы аннотирования: в терминах членов предложения (что позволит описать синтаксические узлы на доступном широкому кругу пользователей метаязыке) и в терминах деревьев зависимостей (что в свою очередь позволит аккуратно представить прежде всего типы и иерархию синтаксических связей). Более подробную информацию о будущей синтаксической разметке в ХАНКО можно найти в [Мустайоки и др. 2005]. Кроме того, в доступной версии ресурса, как и в ТК, размечаются неоднословные устойчивые обороты: на основе списков из [Ефремова 2004, Рогожникова 2003] выделяется приблизительно 2000 таких единиц.

В НКРЯ задача сплошной синтаксической разметки в объеме всего корпуса не ставится. К элементам синтаксического анализа, который предполагается реализовать на материале всего корпуса,

относится аннотирование устойчивых оборотов (ср. ТК и ХАНКО), установление списка которых производится с учетом частотности в корпусе. Кроме того, на сайте Национального корпуса планируется поместить фрагмент синтаксически размеченного корпуса (общим объемом свыше 30 тыс. предложений) с использованием аппарата грамматики зависимостей⁴.

3.3. Семантическая разметка. ТК не содержит дополнительной семантической информации.

В КГТ включены элементы семантической разметки. Во-первых, некоторым словам приписаны семантические признаки. Таким образом размечены прежде всего имена, обозначающие лиц и животных, которые разбиваются на 60 классов, организованных по семантическим и словообразовательным принципам. В остальные 10 классов попадают имена со значением действия (единый класс), имена речи, глаголы речи, прилагательные цвета и несколько других словообразовательно-лексических классов. Во-вторых, в корпусе размечены синонимические отношения между отдельными лексическими единицами. Тем самым пользователь может по одному синониму получить контексты употребления для всего синонимического ряда.

В ХАНКО семантическая разметка находится в стадии первичной разработки и будет осуществлена после завершения синтаксической части. Планируется, что корпус будет содержать информацию о семантических категориях, список которых разрабатывается научным коллективом под руководством А. Мустайоки [Мустайоки 2005]. Как предполагают авторы, на базе морфологической и синтаксической разметки создание функциональной части корпуса можно будет частично автоматизировать, однако планируется осуществлять и ручную работу.

В НКРЯ семантическая разметка осуществляется автоматически: большинству лемм в тексте приписывается один или несколько семантических и словообразовательных признаков. При этом подробная классификация охватывает не только предметные имена, но и не предметную лексику, прилагательные, глаголы и наречия. Важно отметить, что одна лемма может попадать одновременно в разные классы. Вместе с тем, поскольку все семантические

⁴ Подробнее см. статью Ю. Д. Апресяна и др. в настоящем сборнике.

пометы, присвоенные вокабуле в словаре, автоматически переносятся на любое ее вхождение в корпус, лексические омонимы не разводятся, совмещаясь в одной лемме. В силу этого поиск только одного члена полной омонимичной пары невозможен. Так, например, запрос «ЛУК: существительное: 'оружие'» выдает и контексты такого рода: «*Золотистые связки лука над крыльцом*». [Сергей Довлатов. *Заповедник (1983)*]. Однако в настоящее время ведется работа по созданию и внедрению в корпус семантических фильтров, которые позволят по заданному контексту или конструкции автоматически снимать лексическую многозначность⁵.

3.4. Метаразметка. Рассматриваемые корпуса, значительно различаясь по содержанию, в разной степени нуждаются в метаописании входящих в их состав текстов. Корпус ХАНКО, будучи довольно однородным по внешним текстовым параметрам — напомним, в него вошли статьи из журнала «Итоги» за один месяц — содержит минимальную метаинформацию: номер журнала и тип текста (статья, рецензия, интервью).

ТК содержит разнородные текстовые коллекции. Система поиска допускает запросы не только по всему корпусу, но и по его части, однако формирование пользовательского подкорпуса возможно только путем простого выбора одной или нескольких из выделяемых в составе корпуса коллекций. Эти коллекции организованы по разным параметрам: это может быть автор текста (таковы подкорпуса текстов А. Марининой, М. А. Булгакова, А. Н. Рыбакова, бр. Стругацких, Л. Н. Толстого, Ф. М. Достоевского, И. С. Тургенева, Н. С. Лескова), жанр текста (детективы), тип текста (интервью), источник текста (журнал «Огонек» за определенные годы). Кроме того, особую коллекцию образует Уппсальский корпус, поиск в котором можно вести отдельно по художественным и публицистическим текстам.

Метаописания газетных текстов, образующих КГТ, включают в себя детальную жанровую классификацию статей. На основе анализа материала был выявлен круг основных жанрообразующих факторов, характеризующих предмет сообщения, его коммуникативную цель и композиционно-стилевую форму. По этим параметрам было выделено 9 жанровых типов (собственно информа-

⁵ Подробнее о семантической разметке в НКРЯ см. статью Г. И. Кустовой и др. в настоящем сборнике.

ционные, информационно-публицистические, собственно публицистические, художественно-публицистические, рекламные жанры и др.), которые распределяются между 96 конкретными жанрами. Таким образом, пользователь может для своих исследовательских целей формировать подкорпус на основе жанровых признаков текстов (задавая жанр и/или жанровый тип текстов). Кроме того, параметром при определении собственного подмассива текстов в КГТ может служить наименование издания в сочетании с датой его выпуска (например, поиск будет вестись только по газете «Новгородские ведомости» за 28.10.1997). Использование такой подробной жанровой классификации представляется небесспорным. Во-первых, для исследования лингвистических особенностей того или иного жанра необходимо, чтобы каждому из них соответствовало значительное количество статей в корпусе. Очевидно, что при нынешнем числе статей (446) разбиение на 96 жанров не имеет практического смысла для пользователей. Во-вторых, в этом случае, как кажется, трудно избежать произвольных решений при отнесении той или иной статьи к конкретному жанру. Так, например, не вполне понятно, можно ли провести четкую границу между жанрами «Очерк проблемный + Репортаж» и «Репортаж + Очерк проблемный» или «Статья аналитическая» и «Статья аналитическая + Статья проблемная». Кроме того, метаразметка КГТ не учитывает ряд параметров, традиционно используемых для классификации текстов; и если, например, характеристика по полу и возрасту автора действительно не столь существенна в применении к газетным текстам, то тематика статьи (политика, спорт и т. п.) в некоторой степени определяет ее лингвистические особенности.

Значительный объем и разнородность текстов в НКРЯ потребовали разработки подробной системы метаразметки. За ее основу была взята классификация, предложенная в рекомендациях EAGLES, которая затем была адаптирована к русскому материалу с учетом отечественных традиций в области стилистики и типологии текстов. В итоге все тексты, входящие в НКРЯ, характеризуются с точки зрения множества различных параметров, в том числе по полу и возрасту автора, году создания текста, его объему, сфере его функционирования, его жанру, типу, тематике и др.⁶

⁶ См. статью С. О. Савчук в настоящем сборнике.

3.5. Другие типы разметки. Наличием некоторых других типов лингвистической и экстралингвистической разметки характеризуется КГТ. Большинству лемм в корпусе приписана их морфемная модель, т. е. схема с заполненными аффиксальными позициями и переменной для корня (например, *про-Р-и-ть-ся* или *Р-о-Р-ств-о*). Таких схем на материале интернет-версии корпуса объемом свыше 200 тыс. словоупотреблений обнаруживается более 5,5 тыс. Словообразовательная разметка позволяет осуществлять поиск слов (или включающих их контекстов), которые отвечают заданной морфемной модели или содержат заданные аффиксы.

Кроме того, большинству слов приписаны ранги в соответствии с их частотностью в корпусе. По уровню частотности все встречающиеся в корпусе леммы разбиты на 20 групп: первую образует слово с частотным рангом 1, последнюю — с частотными рангами 32769–65536. Тем самым одним из параметров при поиске могут служить частотно-ранговые характеристики лемм: можно определять уровень частотности той или иной единицы в корпусе (по принадлежности к одной из 20 групп), частично просматривать списки лемм, относящихся к определенной группе, а также ограничивать поиск какого-либо лингвистического явления словами определенной частотности.

Завершая обзор типов аннотирования, хотелось бы обратить внимание на справедливое замечание Дж. Лич о том, что разметка корпуса должна быть по возможности теоретически нейтральной: «*To avoid misapplication, annotation schemes should preferably be based as far as possible on 'consensual', theory-neutral analyses of the data*» [Leech 1993: 275].

Однако на практике выполнение этого постулата сталкивается с серьезными трудностями. Дело в том, что степень полноты и общепризнанность классификаций языковых уровней существенно различается. Например, в научной литературе по морфологии могут дискутироваться вопросы о количестве русских падежей, но не вызывает сомнения сам факт существования категории падежа. В области синтаксиса, как известно, такого единства нет. Широко распространенная в практике преподавания классификация, опирающаяся на представление о главных и второстепенных членах предложения, не может считаться общепризнанной; современные синтаксические теории, описывающие синтаксические отношения в виде структуры составляющих, не имеют столь же широкого рас-

пространения, особенно в учебной практике; подходы функционального синтаксиса плохо согласуются с положениями «Русской грамматики» 1980-го года, и т. д.

Еще больше проблем связано с семантической разметкой корпуса, поскольку такого рода аннотирование связано не только с выбором определенного формализма, но и с решением вопроса, информация какого типа должна маркироваться семантическими тегами. В мировой практике эксперименты по реализации семантической разметки можно разделить на три класса. Под семантической разметкой может пониматься, во-первых, маркирование частных значений многозначных лексем (в этом случае разработчикам необходимо выбрать словарь, на основе которого будет производиться дифференциация значений), во-вторых — приписывание лексемам обобщенных семантических признаков на основе некоторой классификации лексики (в этом случае возникает проблема выбора стандартной, общепринятой классификации) и в-третьих — отражение семантических отношений между словами в тексте (при этом нужно определить набор семантических ролей для каждой лексической единицы).

Собственно, неравномерная представленность разных языковых уровней в разметке анализируемых корпусов объясняется не только тем, что синтаксическая и семантическая разметка обычно подразумевают трудоемкую ручную работу, но и отсутствием стандартных классификационных схем для представления синтаксических и семантических данных в корпусе. К этому добавляется и техническая проблема: если большинство программ автоматического морфологического аннотирования русского языка базируются на общепринятом стандарте — «Грамматическом словаре русского языка» А. А. Зализняка, то в основе алгоритмов синтаксических парсеров часто лежат совершенно разные синтаксические теории. Не меньше сложностей технического порядка возникает при описании семантического и словообразовательного компонента языковых единиц. Можно не сомневаться, что эта работа связана со множеством сложных вопросов, на многие из которых современная лингвистика еще не нашла ответа.

Таким образом, неравномерность аннотирования разных языковых уровней выявляет, среди прочего, две существенные проблемы современной русистики: отсутствие *полных* теоретически обоснованных и общепринятых классификаций, с одной стороны,

и сложность (граничащая с невозможностью) автоматического аннотирования на основе этих классификаций, — с другой. В этом смысле всякий языковой корпус в силу необходимости тотального описания материала кристаллизует проблемные области в описании того или иного языка. Он оказывается не только инструментом для быстрого поиска примеров, но и источником совершенствования теоретических и чисто дескриптивных подходов к языку.

4. ПОИСК В КОРПУСЕ

4.1. Поисковый интерфейс и выдача контекстов. Пользователю, обращающемуся к ТК, предлагается выбрать один из трех типов поиска: простой, комплексный и поиск по морфологически аннотированному корпусу. При простом поиске доступны два подкорпуса: Уппсальский корпус и тексты интервью. ТК в полном объеме (см. 2.1) открыт для комплексного поиска. Третий тип поиска естественным образом доступен на материале только тех подкорпусов, которые содержат морфологическую разметку (см. 3.1). Во всех типах поиска все параметры запроса (языковые выражения, грамматические признаки) задаются в одном поисковом окне, тем самым для работы с корпусом пользователю необходимо детально изучить синтаксические правила построения запросов. Кроме того, если поиск осуществляется по грамматическим признакам, пользователь не может выбрать их из заданного списка, а должен сам внести их в поисковое окно. Однако в открытом доступе списка этих обозначений нет. Строго говоря, у пользователя существует только одна возможность построить запрос по грамматическим признакам: просмотреть некоторое количество размеченных текстов, чтобы понять, какие обозначения соответствуют нужным грамматическим показателям.

Параметры выдачи различаются для разных типов поиска. При простом поиске максимальный контекст составляет по одному предложению слева и справа от того, в котором содержится найденное выражение. При других типах поиска пользователь может сам определить объем выдаваемого контекста (в знаках, текстформах или предложениях). Максимальный контекст ограничен 120 текстформами слева и справа от искомого выражения (если размер контекста задается в текстформах) или 6 предложениями (если его размер задается в предложениях). При поиске по морфологически аннотированному корпусу существует возможность

отображения при каждой текстоформе в выдаваемом контексте ее грамматических характеристик.

В КГТ для формирования запроса пользователю предлагается четыре ряда полей. Каждый ряд соответствует одному типу запрашиваемой информации и включает три поля. В первом поле из заданного перечня выбирается тип информации, по которой будет вестись поиск (словоформа, исходная форма слова, знак препинания справа, лексико-грамматический разряд, постоянные признаки, переменные признаки, корень, морфемная модель, жанровый тип, жанр, именные синтаксемы, семантический класс, доминанты членов синонимических групп и др.). Во втором поле задается значение признака, выбранного в первом поле, т. е. конкретная текстоформа (например, *коллегу*), лемма (например, *коллега*) или морфологическая, синтаксическая, семантическая или метатекстовая характеристика (например, *неодуш* для неодушевленных существительных, *2ie* для глаголов в форме повелительного наклонения 2-го лица единственного числа, *на ком* для именных групп с предлогом *на* и одушевленным существительным в предложном падеже, «животн» для существительных, обозначающих животных, *Обз+Рец#* для статей в жанре «Обзор+Рецензия»). В третьем поле определяется метод поиска: «Подстрока», если заданная последовательность символов должна входить в слово или его характеристику или «Буквальное совпадение», если она в точности задает искомое слово/характеристику. Четыре ряда полей позволяют таким образом определить до четырех признаков поискового выражения, связав их между собой логическими операторами (*и/или/и нет/или нет*). Определенное неудобство системы поиска в КГТ заключается в том, что для формирования запроса по какой-либо характеристике текстоформы, леммы или текста необходимо знать, какие значения могут принимать те параметры, которые задаются в первом поле (лексико-грамматический разряд, постоянные признаки и т. д.), а также какие сокращения используются для их обозначения. В отличие от ТК, списки этих сокращений можно найти в открытом доступе, однако для этого нужно хорошо знать систему поиска в корпусе.

В КГТ предлагается два типа выдачи найденной единицы. При традиционной, «контекстуальной», выдаче пользователь получает все контексты, которые содержат отвечающие запросу текстоформы (объемом не более 30 текстоформ слева и справа от искомого

выражения). При «словарном» типе выдачи результатом поиска является список текстовых единиц или признаков определенной категории, которые могут сочетаться с заданным словом или признаком. Например, задав лексему *вопрос* и в разделе «Выдача» выбрав параметр «Группировка: Именные синтаксемы», пользователь получит список предложных групп, в составе которых в корпусе встречается слово *вопрос* с указанием количества примеров для каждой. Задав переменный признак *дн* (действительное причастие настоящего времени) и выбрав параметр «Группировка: Жанровый тип», пользователь получит статистику распределения заданной формы по жанровым типам, и т. п. Существенный недостаток системы поиска в КГТ для пользователя заключаются в ограничении на число выдаваемых примеров: система выводит не более 30 контекстов и не более 200 элементов списка (если в разделе «Выдача» выбрана одна из группировок), хотя пользователю сообщается и общее число найденных в корпусе примеров. Грамматическая информация в результатах поиска не выводится.

Как кажется, в ХАНКО и НКРЯ реализована более удобная система поиска, чем в рассмотренных выше корпусах. В ХАНКО каждой текстовой единице соответствует две строки запроса. Первая строка, состоящая из трех полей, предназначена для поиска по заданному сочетанию букв. В первом поле пользователь выбирает, ведется ли поиск по текстоформе или по лемме («нач. форме»); второе поле определяет отношение задаваемой буквенной последовательности и искомой единицы («равна», «содержит», «начинается», «заканчивается», «не равна», «не содержит», «не начинается», «не заканчивается»); наконец, третье поле предназначено для ввода искомой последовательности. Вторая строка запроса предназначена для поиска по грамматическим признакам: пользователю предоставляется список используемых в системе грамматических параметров и их конкретных значений, выбранные признаки автоматически переносятся в поисковое окно.

Контекст выдачи равен одному предложению, однако по запросу пользователя он может быть расширен до 11 (по 5 предложений слева и справа от того, в котором встретилось искомое выражение). Существует возможность получить информацию о грамматических признаках каждой из текстоформ в выдаваемых контекстах.

В НКРЯ для поиска по текстоформе и по лемме используются разные поисковые окна. Запрос по текстоформе (или их последо-

вательности) задается в окне «Поиск точных форм», остальные типы запросов — в разделе «Лексико-грамматический поиск»⁷. В последнем для ввода поисковых параметров, относящихся к одной текстовой единице, используется три поля: «слово» (в это поле вводится искомая лемма), «грамматические признаки» и «семантические признаки». При двух последних полях имеются ссылки «выбрать», нажав на которые пользователь получает перечень морфологических или, соответственно, семантических и словообразовательных характеристик. Выбранные признаки автоматически переносятся в поисковое окно.

Контекст выдачи равен одному предложению, однако по запросу пользователя он может быть расширен до 7 (по 3 предложения слева и справа от того, в котором встретилось искомое выражение). Информация о грамматических признаках каждой текстоформы в выдаваемых контекстах появляется во всплывающей подсказке.

В ТК и в ХАНКО реализована возможность ввода запроса в латинской транслитерации, которая представляется несомненным достоинством для лингвистов, работающих за рубежом. Планируется, что в ближайшее время эта опция будет добавлена и в НКРЯ.

Итак, предлагая пользователю интерфейсы разной степени удобства, все корпуса позволяют осуществлять поиск по заданному языковому выражению (сочетанию букв) и по грамматическим признакам. Реализация этих типов поиска будет сопоставляться в двух последующих разделах. Кроме того, в НКРЯ на базе всей лексической системы и в КГТ на базе ее части (см. 3.3) поиск может вестись по семантическим характеристикам леммы. Наконец, в КГТ возможны некоторые другие типы поиска — по корню слова или его части, по морфемной модели слова, по синонимической группе (т. е., например, задав в поисковом окне лемму *бездельник*, пользователь получит контексты употребления слов *бездельник* и *лентяй*), по именованным группам вида «предлог+существительное», по частотно-ранговым характеристикам лемм.

4.2. Поиск по языковому выражению. Можно выделить три типа поисковых запросов с заданным сочетанием букв: поиск по текстоформе, по лемме и по последовательности текстоформ/лемм.

⁷ В разделе «Лексико-грамматический поиск» может задаваться и поиск по текстоформе: искомое языковое выражение в этом случае помещается в кавычки.

Самый простой случай — поиск по текстоформе или ее части — реализуется во всех рассматриваемых корпусах.

Поиск по лемме или ее части всех ее текстоформ из-за отсутствия лемматизации невозможен в ТК. Конечно, при формировании запроса можно использовать символы «.*» (точка+звездочка), замещающие любую буквенную последовательность в текстоформе, однако очевидно, что такой поиск может порождать значительное количество «шума», кроме того, его применимость ограничена в случае чередований и супплетивных форм.

Запрос на последовательность текстоформ/лемм невозможен в КГТ. Все характеристики, задаваемые пользователем в четырех строках поисковой формы и соединяемые логическим И, могут относиться только к одной текстоформе. Так, можно искать лемму *экономический* ИЛИ лемму *политика*: система выдаст контексты, в которых встречаются текстоформы первой или второй леммы, но запрос по тем же леммам, соединенным логическим И, даст пустой результат, поскольку нет текстоформ, одновременно принадлежащих лемме *экономический* и лемме *политика*. Исключение в этом отношении составляют запросы по «именным синтаксемам», в которых результатом поиска являются словосочетания, однако и в этом случае остальные параметры можно формулировать только по одной лемме — входящему в состав предложной группы существительному.

Отсутствие лемматизации в ТК определяет недоступность поиска по заданной последовательности лемм, однако запрос может формулироваться по последовательности текстоформ. Широкие возможности в этом отношении открывают комплексный поиск и поиск по морфологически аннотированному корпусу (см. 4.1), при которых используется поисковая программа CQP, разработанная в Институте машинной обработки речи в Штутгарте и применяемая в различных корпусах на материале разных языков. Кроме тех типов запросов, которые доступны при простом поиске, она позволяет, например, искать текстоформы, находящиеся на заданном расстоянии друг от друга или же в пределах одного предложения. Однако, как уже указывалось выше, неудобство системы поиска состоит в необходимости предварительного знакомства с синтаксисом запросов. Так, поиск сочетаний текстоформ глагола *играть* с предлогом *на*, расположенных на расстоянии от 0 до 3 слов друг

от друга, задается при помощи следующей последовательности символов:

«игра.*» [0,3] «на»

По этому запросу, очевидно, найдутся не только формы глагола *играть* с предлогом *на*, но и сочетания, включающие словоформу *игра* соответствующего существительного. Если же в морфологически аннотированном корпусе мы попытаемся ограничить выдачу, определив первое слово как глагол, запрос, который должен собственноручно ввести пользователь, будет выглядеть так:

[word=«игра.*» & tag=«verb.*»] [0,3] «на»

В ХАНКО и НКРЯ поиск может вестись по последовательности как текстоформ, так и лемм. В ХАНКО по умолчанию выдается поисковая форма, позволяющая задать характеристики одной текстоформы/леммы, в НКРЯ — двух, однако в обоих корпусах форма при необходимости легко может быть расширена до произвольного числа характеризующихся текстоформ. В обоих корпусах предусмотрено специальное окно для определения расстояния между искомыми текстоформами.

Подчеркнем, что все сказанное выше о поиске по словосочетанию релевантно не только в том случае, если элементы словосочетания задаются в виде буквенных последовательностей, но и в тех случаях, когда все его компоненты или некоторые из них определяются по грамматическим характеристикам или же по другим параметрам, предлагаемым поисковой системой данного корпуса.

Следует отдельно отметить, что в ТК, ХАНКО и КГТ существует возможность поиска контекстов, содержащих определенные знаки пунктуации. В НКРЯ такой поиск пока не поддерживается.

4.3. Поиск по грамматическим признакам. Как обсуждалось выше, тот факт, что морфологическая разметка представляет собой более простую задачу, чем синтаксическая или семантическая, объясняется среди прочего наличием в русистике традиционной, общепризнанной классификации грамматических категорий и их значений. Соответственно, большинство словоформ анализируются в разных корпусах сходным образом. В то же время некоторые частные аспекты грамматической системы по-разному отражаются в системе разметки анализируемых ресурсов, что приводит к различиям в наборе грамматических признаков, по которым может вестись поиск в корпусе. Эти различия объясняются не только су-

существованием в русской грамматике ряда спорных вопросов, но и некоторыми техническими особенностями представления грамматической информации. Известно, что серьезная проблема, встающая перед создателями аннотированного корпуса, связана с дилеммой объем материала vs точность его обработки. Создание анализатора, безошибочно производящего анализ русского текста, по-видимому, невозможно. На сегодняшний день качественное аннотирование русского текста всегда связано с ручной постобработкой. В этом смысле при относительной ограниченности организационных возможностей перед создателями любого корпуса всегда стоит выбор: сравнительно небольшой, но выверенный корпус или объемный, но аннотированный автоматически. Представляется, что оба принципа имеют право на существование.

Ниже мы обсудим основные расхождения корпусов по морфологическим поисковым параметрам. Однако сначала кратко остановимся на принципах их классификации, реализованных в каждом из корпусов. Наиболее традиционную и тем самым удобную для пользователя систему представляют ХАНКО и НКРЯ: общий перечень морфологических характеристик разбит по грамматическим категориям — часть речи, число, падеж, время и т. д., в каждой из которых пользователь может выбрать нужные значения. В КГТ грамматические параметры распределены между тремя типами информации, каждый из которых задается в отдельной строке запроса: лексико-грамматический разряд, постоянные признаки, переменные признаки. Возможные значения этих признаков выдаются сплошным списком, без дальнейшей классификации по грамматическим категориям. При этом перечни включают значительное число наименований, поскольку одной строке соответствует не конкретное значение одной категории (например, *женский род* или *прошедшее время*), а полная характеристика текстоформы — в случае переменных признаков (например, *будущее время, множ. число, 2-ое лицо*) или леммы — в случае постоянных признаков (например, существительное женского рода) или лексико-грамматического разряда (например, *несов. вид + переходность*). В ТК классификация грамматических характеристик, по которым может вестись поиск, — в том смысле, в котором это обсуждалось выше в связи с остальными корпусами, — отсутствует, поскольку в открытом доступе нет и самих списков используемых морфологических параметров (см. 4.1).

Перейдем собственно к обсуждению различий в системе грамматических признаков. Набор частеречных характеристик в четырех корпусах в значительной степени совпадает. Расхождения касаются, во-первых, подробности в членении местоимений: в НКРЯ выделяются местоимения-существительные, местоимения-прилагательные, местоимения-предикативы и местоимения-наречия; в КГТ не учитываются местоименные предикативы и наречия, но вводятся местоимения-числительные; в ТК категориальное противопоставление существительное *vs* прилагательное затрагивает только некоторые разряды местоимений (указательные, вопросительные, относительные и др.); в ХАНКО местоимения рассматриваются как единый частеречный класс, однако ряд местоименных наречий (например, *где-то, когда-то, никогда*) в большинстве употреблений получают два варианта грамматического разбора (местоимение и наречие). Во-вторых, в НКРЯ и КГТ выделяются вводные слова (в ХАНКО и ТК они размечаются в зависимости от типа как наречия, частицы или глаголы).

Различия между корпусами затрагивают и падежную систему. Кроме шести традиционных падежей, учитываемых во всех четырех схемах аннотирования, в ХАНКО и НКРЯ выделяются второй родительный и второй предложный. Кроме того, в НКРЯ, в корпусе со снятой омонимией, вводятся звательный, второй винительный (ср. *идти в солдаты*) и счетная форма (ср. *два часа*).

В ТК и ХАНКО глагольный вид трактуется как категория с тремя значениями: совершенный, несовершенный и двувидовой глагол. Соответственно, для глаголов, причисляемых к двувидовым, не приводится указание на вид по контексту. В КГТ и НКРЯ характеристика «двувидовой глагол» не выделяется в качестве отдельного поискового параметра, однако по сути принимается сходное решение: текстоформы глаголов такого типа трактуются как омонимичные между лексемой совершенного и несовершенного вида. Исключение в этом отношении представляет часть НКРЯ со снятой омонимией: здесь в тех случаях, когда контекст однозначно определяет видовую характеристику текстоформы, сохраняется только один вариант разбора, ср.:

На уроке учитель постоянно **использует** [использовать =V,ipf,tran =sg,act,praes,3p,indic,act] преимущества химического эксперимента <...>. [Характеристика учителя химии (2000)].

В сфере залоговых противопоставлений корпуса реализуют следующие решения. В КГТ залоговыми значениями (действительное/страдательное) характеризуются только причастия. В ТК для причастий предусмотрено три параметра: действительное, действительное+возвратное (для причастий на *-ся*) и страдательное. Остальным глагольным формам присваивается признак возвратности/невозвратности. Трактовка залога в НКРЯ идейно близка тюбингенской. Список поисковых параметров включает три залоговых значения — «действительный», «страдательный» и «медиальный» (для глаголов на *-ся*). При этом речь идет фактически о двух независимых противопоставлениях: форм действительного и страдательного залога у причастий и действительных и медиальных глагольных лексем. Таким образом, как в ТК, так и в НКРЯ разметка основана на формальном принципе. Другой подход представлен в ХАНКО. Рассматриваемое грамматико-семантическое поле покрывается здесь двумя категориями: залог (со значениями «действительный» и «страдательный») и возвратность (значения — «невозвратный» и «возвратный»). При этом как формы страдательного залога могут характеризоваться не только причастия, но и личные формы глагола, ср.:

<...> лес **вырубался** [вырубать, глагол, личн. ф., несов. вид, прош. вр., невозвр., изъявит., **страдат.**, мужск. р., ед. ч.] русской армией, чтобы лишить горцев возможности прятаться и вести партизанскую войну <...>

Тем самым текстоформы на *-ся* в ХАНКО трактуются либо как формы страдательного залога (для них исходной формой является инфинитив без *-ся*), либо как возвратные (они возводятся к инфинитиву с *-ся*). Очевидно, что различить эти случаи можно только при ручной постобработке морфологической разметки. В то же время отказ от их противопоставления в НКРЯ был вызван не только техническими сложностями и невозможностью автоматизировать этот процесс, но и наличием значительного числа спорных случаев. Выбрать ту или иную трактовку означало бы навязать пользователю исследовательскую позицию разработчиков корпуса. В ХАНКО эта проблема частично преодолевается допустимостью множественных разборов. Так, как страдательная форма невозвратного глагола *закрывать* или действительная форма возвратного глагола *закрывается* трактуется текстоформа *закрывается* в следующем контексте:

Кремль все больше **закрывается** от журналистов <...>.

Однако, как кажется, спорных случаев больше, чем это отмечается в ХАНКО, ср. различную разметку сходных употреблений:

В России до сих пор **продаются** [продавать, глагол, личн. ф., несов.вид, наст.вр., **невозвр.**, изъявит., **страдат.**, Зл., мн. ч.] преимущественно турецкие дублинки <...>

В целом за последние год два качество дубленок, **продающихся** [продаваться, глагол, причастие, несов.вид, наст.вр., **возвр.**, **действит.**, мн. ч., 1род. п., полная ф.] в России, заметно улучшилось, а ассортимент стал гораздо более разнообразным.

Тем не менее следует отметить, что в целом ручная постобработка ХАНКО обеспечивает более подробную и аккуратную по сравнению с остальными корпусами систему морфологической аннотации. В первую очередь это касается трактовки аналитических форм. В ХАНКО размечаются и, соответственно, являются доступными для поиска формы сослагательного наклонения (*сходил бы*), сложного будущего времени (*буду ходить*), аналитические формы сравнительной и превосходной степеней сравнения прилагательных и наречий (*более быстрый/быстро, самый быстрый, быстрее всего*). Формы сослагательного наклонения аннотируются и в КГТ, однако другие аналитические формы, не учитываемые в разметке, недоступны для поиска, поскольку запрос в корпусе может формулироваться только по характеристикам *одной* текстоформы (см. 4.2). В ТК и НКРЯ на данном этапе разработки аналитические формы не размечаются, но поиск может вестись по словосочетанию с заданными морфологическими характеристиками входящих в него элементов. Так, в НКРЯ запрос на аналитическую форму будущего времени можно построить следующим образом:

Слово 1: *быть* в форме будущего времени

Слово 2: глагол в форме инфинитива несовершенного вида.

Недостаток данного способа поиска заключается в выдаче некоторого количества «шума»: ведь компоненты аналитической формы (например, формы будущего времени) могут находиться на значительном расстоянии друг от друга, соответственно, для учета всех случаев необходимо задавать большое расстояние между искомыми текстоформами, а это неизбежно порождает лишние контексты.

К другим проявлениям характерной для ХАНКО разметки «второго уровня» — уровня неоднословных грамматических и лексических единиц — относится аннотирование составных и дробных числительных. Список поисковых параметров включает две группы признаков, определяющих числительное: тип числительного по

структуре (со значениями «простое» и «составное») и разряд числительного («порядковое», «количественное», «собирательное», «дробное»). В остальных корпусах различие проводится только между порядковыми и количественными числительными. Кроме того, в ХАНКО учитываются разрывные формы местоимений (*ни от кого*). Соответственно, поиск по лемме *никто* выдаст и контексты, включающие формы с предлогами (*ни с кем, ни у кого* и др.).

Планируется, что в НКРЯ все рассмотренные выше типы неоднословных единиц (а также некоторые формы, не учитываемые в морфологической части ХАНКО, например, аналитическая форма прошедшего времени страдательного залога глаголов совершенного вида, ср. *был сделан*) будут размечаться как особый тип устойчивых оборотов (ср. 3.2), а пользователю будет предоставляться возможность поиска текстоформ как в составе оборотов, так и вне их. Тем самым, например, задав поиск форм инфинитива вне оборотов, пользователь не будет получать на выдаче аналитические формы будущего времени.

Ручная постобработка обуславливает и ряд других особенностей морфологической разметки ХАНКО, связанных с формами, трудными для автоматического анализа. Так, проводится различие между сравнительной степенью прилагательных и наречий для текстоформ типа *красивее* (*КРАСИВО, КРАСИВЫЙ*). В ТК эти формы тоже получают разную разметку, однако в силу того, что аннотация производится автоматически, порождается значительное число неправильных разборов. Кроме того, в ХАНКО возможен поиск существительных *pluralia tantum*; предусмотрена эта опция и в КГТ. В НКРЯ морфологический разбор существительных данного типа отличается от разметки других имен (исходной считается форма множественного числа, число помещается к словоклассифицирующим, а не к словоизменительным пометам), однако поиск всех имен *pluralia tantum* не поддерживается.

Среди особенностей в наборе грамматических поисковых параметров НКРЯ следует отметить вторую форму повелительного наклонения — инклюзивную форму с суффиксом *-те* (*пойдемте*).

В целом, как можно понять из приведенных характеристик, построение системы грамматического аннотирования корпуса представляет собой серьезную исследовательскую задачу. Ее решение

Сводная таблица характеристик аннотированных лингвистических корпусов русского языка в Интернете

	ТК	КГТ (интернет-версия)	ХАНКО	НКРЯ
Состав корпуса	Упсальский корпус; публицистика (1996-2002гг.); художественная литература XIX-XX вв.	газетные тексты (1997 г.)	журнал «Итоги» (январь 2001 г.)	1) сбалансированный корпус текстов 2-ой половины XX в. 2) тексты XIX в.
Объем корпуса (в словоупотреблениях)	ок. 25 млн.	>200 тыс.; планируется 1 млн.	100 тыс.	1) более 60 млн., планируется 100 млн. 2) > 5 млн., планируется 30 млн.
Типы разметки				
морфологич.	автоматич. в части корпуса (2,3 млн.), нет лемматизации	автоматич.	автоматич. +ручная постобработка	автоматич. (весь корпус) + ручная (>4 млн.)
словообразов.	—	+	—	элементы
синтаксич.	элементы	элементы	элементы, планируется подробная	элементы, планируется подробная в части корпуса
семантич.	—	элементы	планируется	+
метаразметка	—	+	элементы	+
Поисковые возможности — поиск по:				
текстоформе	+	+	+	+
лемме	—	+	+	+
последов-ти текстоформ	+	—	+	+
грамм. признакам	+	+	+	+
сем. признакам	—	+	планируется	+
знакам пунктуации	+	+	+	планируется
Выдача контекстов				
ограничения на количество контекстов	нет	не более 30	нет	нет
максимальный объем контекста	13 предложений/241 слово	61 слово	11 предложений	7 предложений

требует как определения теоретических принципов разметки, так и их соизмерения с объемом и спецификой анализируемого материала. Очевидно, скажем, что степень детальности аннотирования, характерная для ХАНКО, не представляется возможной для превышающего его в тысячу раз по объему НКРЯ, и, напротив, морфологические признаки, существенные для разнородной текстовой коллекции большого объема (например, звательный падеж, 2-я форма повелительного наклонения), нерелевантны для небольшого собрания публицистических текстов, где подобные формы практически отсутствуют.

* * *

Итак, в настоящем обзоре обсуждались основные характеристики и поисковые возможности общедоступных корпусов русского языка, прежде всего, ТК, корпуса газетных текстов Лаборатории общей и компьютерной лексикографии и лексикологии МГУ и Хельсинкского аннотированного корпуса ХАНКО. Данные Национального корпуса русского языка приводились для возможности сравнения его основных параметров с остальными рассмотренными корпусами. Характеристики представленных корпусов можно обобщить в виде таблицы на с. 58.

Очевидно, что эффективность и удобство того или иного корпуса для пользователя определяется типом исследования, которое проводится на его основе, а также характером и частотностью изучаемого на его материале явления. Надеемся, что настоящий обзор дал представление о возможностях, преимуществах и потенциале развития каждого из корпусов.

Литература

- Андрюшенко В. М. Концепция и архитектура машинного фонда русского языка. — М.: Наука, 1989.
- Баранов А. Н., Михайлов М. Н., Сидоров Г. О. «Динамический корпус текстов» как новая технология прикладной лингвистики // Труды международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. — Казань, 1998. — Т. 2.

Т. И. Резникова, М. В. Копотев

- Венцов А. В., Касевич В. Б. Словарь для модели восприятия речи // Вестн. С.-Петербург. ун-та. 1998. — Сер. 2, вып. 3. — С. 32-39.
- Виноградова В. Б., Кукушкина О. В., Поликарпов А. А., Савчук С. О. Компьютерный корпус текстов русских газет конца XX века: создание, категоризация, автоматизированный анализ языковых особенностей // «Русский язык: исторические судьбы и современность». Международный конгресс русистов-исследователей. Москва, филологический ф-т МГУ им. М. В. Ломоносова 13-16 марта 2001 г. Труды и материалы. — М.: Изд-во Моск. ун-та, 2001. — С. 398.
- Ефремова Т. Ф. Толковый словарь служебных частей речи русского языка. — М.: Астрель-АСТ, 2004.
- Михайлов М. Н. Параллельные корпуса художественных текстов. — Тампере, 2003.
- Мустайоки А. Теория функционального синтаксиса. — М., 2005.
- Мустайоки, А. Копотев, М. В. Принципы создания Хельсинкского аннотированного корпуса русских текстов ХАНКО в сети Интернет // Научно-техническая информация. — 2003. — Сер. 2, № 6. — С. 33-37.
- Мустайоки А., Копотев М. В., Гурин Г. Б., Саломатина М. С. Принципы синтаксической разметки Хельсинкского аннотированного корпуса русских текстов ХАНКО // Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей». — СПб., 2005. — С. 90-95.
- Рогожникова Р. П. Толковый словарь сочетаний, эквивалентных слову. — М.: Астрель-АСТ. 2003.
- Шаров С. А. Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. — 2003. — Сер. 2, № 6. — С. 9-18.
- Betsch, M., Meyer, R. Automatic Annotation of Russian texts: Evaluation of Different Tagging Methods // P. Kosta, J. Błaszczak, J. Frasek, L. Geist, M. Żygis (eds.). Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages (FDSL IV) held at Potsdam University, November 28-30, 2001. — Frankfurt/Main, 2003.
- Brants, T. TnT — Statistical Part-of-Speech Tagging. www.coli.uni-sb.de/~thorsten/publications/Brants-TR-TnT.pdf
- Fillmore Ch. 'Corpus linguistics' vs. 'Computer-aided armchair linguistics' // Directions in Corpus Linguistics. Proceedings from a 1992 Nobel Symposium on Corpus Linguistics. — Berlin, New York, 1992. — P. 35-60.
- Kukushkina O., Polikarpov A. DicTUM-1 — A System for Dictionary-Text Universal Manipulations and Analysis // L. Borodkin (ed.) Data Modelling,

Корпуса русского языка в Интернете

Modelling History. XI International Conference of the Association for History and Computing. Moscow State University, August 20-24 1996. Abstracts. — Moscow, 1996. — P. 50-52.

Leech G. Corpus annotation schemes // Literary and Linguistic Computing, 1993. — 8/4. — P. 275-281.

Vilki L. RUSTWOL: A System for Automatic Recognition of Russian Words. — 1997. <http://www.lingsoft.fi/doc/rustwol/rustwol.txt>