С. О. Савчук

МЕТАТЕКСТОВАЯ РАЗМЕТКА В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: БАЗОВЫЕ ПРИНЦИПЫ И ОСНОВНЫЕ ФУНКЦИИ

Создание размеченного корпуса предполагает снабжение целых текстов и отдельных словоформ дополнительной информацией — лингвистической (морфологической, синтаксической, семантической, стилистической), социологической, библиографической и др. Описание целых текстов по определенным параметрам называется в Национального корпуса русского языка (НКРЯ) метаразметкой.

Метаразметка выполняет в корпусе несколько функций:

- служит для формирования архитектуры корпуса;
- позволяет контролировать процесс информационного наполнения корпуса, оценивать его представительность и сбалансированность;
- обеспечивает возможность поиска и отбора текстов пользователем для составления подкорпусов с заданными свойствами.

Отсюда следует, что чем больше набор параметров, по которым характеризуется каждый текст, тем шире возможности поиска текстов для решения различных лингвистических задач[1].

Первоначально в основу описания текстов корпуса была положена классификация, предложенная в рекомендациях EAGLES [Sinclair, 1996], как наиболее приближенная к решению практических задач и использованная при разметке ряда корпусов[2]; она была адаптирована С. А. Шаровым к описанию материала русского языка.

---

[1] Существует мнение, что тщательно продуманная метатекстовая разметка способна снять проблему балансировки текстов при составлении репрезентативного корпуса: разработчикам достаточно составить корпус максимально большого объема, снабдив тексты метаописанием, и предоставить пользователю возможность самому отбирать необходимые подкорпуса по предложенным метапризнакам (см. Шимкова 2005, с. 55].

[2] Британского национального корпуса (см.: BNC: The BNC Users Reference Guide, 2000. http://www.natcorp.ox.ac.uk/World/HTML/); Чешского национального корпуса (см.: Kopřivová M. Český národní korpus na přelomu tisíciletí. // Český národní korpus/ Praha 2000 <http://ucnk.ff.cuni.cz>); Американского национального корпуса (см.: Randi Reppen and Nancy Ide The American National Corpus: Overall Goals and the First

Принципы этой классификации, в том числе и в применении к русскому языку, подробно рассматривались в работах [Шаров; 2003; Sharoff; 2004; Шаров, Савчук, 2004]; в работе [Sharoff, 2005] обсуждается вариант метаразметки текстов, основанный на упрощенной классификации, оптимальной при составления интернет-корпусов.

Общая схема классификации отражает структуру акта коммуникации; признаки, по которым описывается текст, группируются вокруг основных компонентов речевой ситуации:

- **Автор** (тип автора, пол, возраст);
- **Адресат**, или аудитория (размер, пол, возраст, уровень образования, подготовленность, степень знакомства автора с аудиторией);
- **Цель** коммуникации (информирование, рекомендация, дискуссия, инструктирование, развлечение) и набор речевых жанров, обусловленный каждой из целей;
- **Предмет** коммуникации (предметные или тематические области);
- **Обстановка** коммуникации (степень официальности и характер контактности общающихся);
- **Канал** коммуникации (формы речи — устная, письменная, электронная, написанная для устного произнесения; виды устной и письменной речи с подразделением на типы, стили речи);

Подобный ситуативный подход к описанию текстов (с большей или меньшей детализацией в характеристике отдельных компонентов) представлен и в отечественных работах по социолингвистике [Швейцер, Никольский, 1987; Современный русский язык, 2003], типологии текста [Чебанов, Мартыненко, 1999], стилистике [Васильева, 1976; Кожина, 1993][3].

Release //Journal of English Linguistics 2004 32: 105-113); Словацкого национального корпуса (см.: Гарабик Р. Словацкий национальный корпус// Труды международной конференции «Корпрусная лингвистика-2004». — СПб., 2004. С. 99-121) и др.

[3] Так, А. Н. Васильева среди экстралингвистических стилеобразующих факторов, оказывающих влияние на стилистическое оформление речи, называет следующие: 1) степень официальности/ неофициальности обстановки общения; 2) характер контактности общающихся; 3) форма существования речи; 3) характер субъекта речи (индивидуальный, собирательный, абстрагированный); 5) характер адресата речи; 6) жанр; 7) степень предварительной подготовки речи.

Очевидно, близость классификаций, независимо разрабатываемых в зарубежной и отечественной типологии речи, свидетельствует об универсальности принципов, на которых они строятся, что позволяет использовать эти принципы для описания текстов на любом языке, но с учетом специфики языка и традиций его изучения.

Для отечественной лингвистики речи такой традицией является использование категорий «сфера функционирования текста» и «речевой жанр» как тип текста. При этом речевые жанры не принято «привязывать» напрямую к коммуникативным целям (как это делается в схеме Синклера), они определяются комплексом признаков, хотя цель и признается ведущим признаком в этом комплексе. Так, Т. В. Шмелева, опираясь на понимание типологии речи как типологии речевых жанров [Бахтин, 1979], предлагает описывать речевые жанры на основе следующей комбинации параметров:

1) коммуникативная цель (их 4: информирование, оценка, побуждение, поддержка социальных отношений, осуществляемых в ритуализованных формах);
2) образ автора;
3) образ адресата;
4) диктум (предмет речи, событийная основа высказывания);
5) фактор прошлого (различает инициативные РЖ и РЖ-реакции);
6) фактор будущего (прогнозирование ответной реакции адресата);
7) формальная организация [Шмелева, 1990].

Как было показано в работе [Шаров, Савчук, 2004], параметрическая классификация текстов, предложенная Синклером, и стилистическая, использующая традиционную номенклатуру жанров, не противоречат друг другу и оказываются вполне совместимыми в пределах единой системы признаков, используемых для описания текстов. Поэтому окончательный вариант метаразметки, принятый в НКРЯ и ориентированный на традиции изучения текстов, сложившиеся в отечественной лингвистике, включает в себя такие характеристики текстов, как сфера функционирования, тип (речевой жанр), хронотоп текста.

Кроме того, параметрический подход, используемый в рекомендациях EAGLES, которым следуют составители европейских корпусов, состоит в том, что первоначально строится «идеальная мо-

дель» будущего корпуса на основе исчисления всех возможных текстовых вариаций, полученных путем перебора всех сочетаний признаков и исключения маловероятных комбинаций. Эта модель затем заполняется реальными текстами в определенных пропорциях, так чтобы каждое сочетание параметров было представлено несколькими текстами [Butler, 2004, 152].

При создании НКРЯ, который, по замыслу разработчиков, должен отразить картину реального употребления языка определенного периода, принято решение минимального вмешательства в соотношение реальных текстов, функционирующих в различных сферах речевой практики. С этой целью в проекте корпуса изначально задана только пропорциональность представленности различных функциональных сфер, которая определялась на основе предварительного анализа количественных показателей существующих корпусов, социологических исследований, мониторинга книжного рынка, газетно-журнальной продукции, электронных ресурсов и др., прогнозирования исследовательских интересов будущих пользователей.

В процессе реального наполнения корпуса текстами вмешательство разработчиков ограничивается, в основном, отбором изданий. При этом учитываются такие факторы, как общественная значимость произведений, оценка специалистов и критики, читательский спрос; газеты и журналы, в том числе литературные, представленные в широком политическом, тематическом, региональном, читательском спектре, включаются в корпус целиком. Таким образом, можно предположить, что в корпусе будет представлено близкое к реальному тематическое и жанровое соотношение опубликованных письменных текстов, функционирующих в период конца XX — начала XXI века[4].

Всего для описания текстов в базе данных корпуса используется 25 параметров. Из них 9 относятся к характеристике самого текста, 3 параметра характеризуют автора, 3 — возможную аудиторию, 4 параметра содержат библиографические данные о тексте, 5 параметров представляют собой служебную информацию, необходи-

[4] Что касается неопубликованных текстов (устных, машинописных, рукописных), а также периодики до начала 1990-х годов, размещение которых в корпусе связано с большими затратами на начальной стадии подготовки текстов, то в настоящее время отбор их производится с учетом жанрового и тематического разнообразия, сложившегося в пределах каждой из функциональных сфер.

мую для учета и организации текстовых файлов в составе корпуса. Рассмотрим каждую группу параметров подробнее.

I. ИНФОРМАЦИЯ ОБ АВТОРЕ

1. Имя автора

Этот параметр может принимать несколько значений:

– конкретный автор: указываются его фамилия и имя;

– обобщенный автор, если текст создается от лица организации, официального органа власти, печатного органа и пр. Такой тип авторства характерен для официальных документов, рекламных текстов, некоторых публицистических текстов, в частности, редакционных статей, редакционных (издательских) предисловий к книгам и пр.

– коллективный автор, если текст создан двумя и более авторами, известными по имени. Такой тип авторства имеют коллективные монографии, совместные публикации (статьи), некоторые виды интервью («круглые столы», форумы) и пр. Если текст имеет двух авторов, они заносятся в базу данных поименно, в остальных случаях автор маркируется как «коллективный».

– неизвестный автор, если автор не указан, но при этом текст нельзя рассматривать как отражение чьей-то коллективной позиции. Этот тип авторства часто встречается в газетно-журнальных текстах — в небольших заметках, новостных и тематических подборках, обзорах и под., в объявлениях, надписях и пр. Иногда текст может быть подписан инициалами или условными именами («Аноним», «Ворчун», «Иван Иванов» и т. п.). Подобные подписи не дают информации о реальном авторе, поэтому автор данных текстов тоже рассматривается как «неизвестный» и соответствующая ячейка базы данных остается незаполненной.

2. Пол автора

Этот признак имеет 3 значения:

– мужской

– женский

– неизвестен

Пол автора указывается в том случае, когда автор известен и он один. Если текст имеет коллективного автора, то пол не указывает-

ся, как не указывается он по понятным причинам и в случае обобщенного авторства. Следует отметить, что в Британском национальном корпусе характеристика по полу автора применяется и в том случае, если авторов больше одного, при этом если они не одного пола, используется характеристика «mixed».

Особого внимания требуют случаи (правда, немногочисленные), когда автор выступает под псевдонимом, при этом автор-мужчина подписывает произведение женским именем и наоборот (Марко Вовчок, Жорж Санд, Антон Крайний). Здесь, если псевдоним не раскрыт, пол определяется относительно имени, под которым текст опубликован[5]. В случае известных литературных псевдонимов указывается пол реального автора[6].

3. Возраст автора

В отличие от Британского и Чешского корпусов, в которых используется относительная характеристика возраста автора в момент создания произведения[7], в НКРЯ указывается абсолютный возраст автора, т. е. год его рождения, точный или приблизительный (в интервале 5-10 лет). В случае если возраст установить не удается, он отмечается как «неизвестный». Что касается относительного возраста автора в момент создания произведения, то он вычисляется при сравнении двух дат — года рождения автора и года написания произведения.

Возраст автора не может быть определен для текстов, имеющих обобщенного, коллективного или неизвестного автора. Особо следует оговорить случаи, когда точное имя автора текста, известное, возможно, разработчикам корпуса, не может быть указано в базе данных по этическим соображениям: такая ситуация возникает при размещении в корпусе личных писем, неопубликованных дневников, записных книжек, записей устной речи, авторы которых

---

[5] Аналогичное решение принято разработчиками Словацкого национально корпуса: при аннотировании текстов указывается имя автора (авторов), «как напечатано в произведении» (Гарабик, 2004, 109).

[6] Разумеется, если, например, автор-мужчина пишет от лица женщины («Я шла по улице»), пол автора всё равно обозначается как мужской.

[7] В Британском корпусе используется многоступенчатая шкала возрастов: 0-14, 15-25, 25-34, 35-44, 45-59, 60+ [BNC: The BNC Users Reference Guide, 2000], в Чешском корпусе — двухступенчатая — до 35 (younger) и после 35 (older) [Čermák; 2001; Kopřivová, 2000]. Если возраст выяснить не удается, он условно обозначается как «средний».

не хотят раскрывать своего реального имени. В этом случае имя автора отмечается как неизвестное или используется условное обозначение, а пол и возраст указываются.

По признакам, характеризующим автора, письменные тексты Корпуса в настоящий момент распределяются так, как показано в таблицах на с. 69.

II. Информация о тексте

4. Название текста

Если текст озаглавлен, то его название приводится полностью. Если текст не имеет заголовка, то это поле остается незаполненным. Короткие газетные и журнальные тексты, помещенные в одну рубрику, подаются целым блоком и получают в качестве названия название рубрики, например, «Полезные советы», «Это интересно», «Полезно знать» и пр. Тексты ограниченного употребления [Савчук, Соколова, 2005] помещаются под условным названием, присваиваемым составителями Корпуса, например: «Договор на строительство гаража», «Письмо из армии», «Предложение нового тарифного плана», «Дневник девушки» и под.

5. Дата создания текста

В самом простом случае дата создания текста может быть указана автором: год написания текста (или годы работы над ним) приводятся в конце произведения: «Валентин Катаев. Алмазный мой венец. 1975-1977». Чаще год создания текста выясняется в результате библиографических, биографических, текстологических исследований, причем при отсутствии точной информации дата устанавливается приблизительно, в интервале 5-10 лет. Для нехудожественных текстов (газетных, журнальных, научных) в общем случае дата написания приравнивается к дате публикации текста. Неопубликованные тексты могут содержать точную информацию о дате создания (деловые документы, дневники, большинство личных писем, электронные письма и рассылки), в других случаях в качестве года создания текста указывается год его регистрации в электронном архиве корпуса.

6. Размер текста в словах

Данный параметр является количественной характеристикой и указывает общее количество словоупотреблений в тексте. В отличие

Автор	**Основной корпус текстов XX-XXI в.**			**Корпус текстов XIX в.**		
	Количество текстов	Количество словоупотреблений		Количество текстов	Количество словоупотреблений	
Единичный	17078	59,253,595	85,5%	663	15,273,999	99,5%
Коллективный	1534	3,874,551	5,6%	1	61869	0,5%
Обобщенный	587	1,100,763	1,6%	1	225	
Неизвестный	6679	5021621	7,3%	1	1102	

Пол автора	**Основной корпус текстов XX-XXI в.**			**Корпус текстов XIX в.**		
	Количество текстов	Количество словоупотреблений		Количество текстов	Количество словоупотреблений	
Мужской	11738	46,872,743		583	14528546	
Женский	5682	12,640,792		80	754919	
Муж|жен	402	1,048,561				
Не определен	8056	8,425,728		3	63196	

Год рождения автора	**Количество текстов**	**Количество словоупотреблений**	
до 1900	174	4,518,294	6,5%
1901-1944	1857	23,931,091	34,6%
1945-1969	1093	9,769,508	14,1%
1970-1986	159	693,014	1%
1987-1993	7	2331	
1994-2004	8	2500	
Не определен	22524	30,333,792	43,8%

от BNC, в котором определен верхний предел объема текста (40-45 тысяч словоупотреблений), в результате чего небольшие по объему тексты попадают в корпус целиком, а большие — в виде начальных, срединных и конечных фрагментов или их композитов, в НКРЯ принято решение включать только целые тексты (как Чешском, Польском и др. корпусах). Вследствие этого объемы текстов могут варьировать от нескольких десятков словоупотреблений (в объявлениях, поздравлениях, новостных сообщениях) до нескольких десятков тысяч в романах.

Характеристика объема текста в словах оказывается чрезвычайно важной для НКРЯ, поскольку она служит единственным средством контроля за сбалансированностью корпуса по различным функциональным сферам и тематическим областям: простое количество включенных текстов не может дать объективной картины того, насколько пропорционально представлены в корпусе тексты различной тематической и жанровой принадлежности. Объем текста в словах подсчитывается с помощью специальной программы на стадии предварительной подготовки электронной версии текста после приведения ее к html-формату.

Следует отметить, что в BNC присутствует еще один количественный параметр текста — количество предложений, который в НКРЯ пока не используется.

7. Сфера функционирования текста

Функциональная сфера — самая общая типологическая характеристика текста. Этот термин имеет широкое распространение в отечественной функциональной стилистике и типологии текста и обозначает социально значимую область общественно-речевой практики, которая объединяет тексты определенного содержания и целевого назначения и связана вследствие этого с определенными разновидностями языка. Сферы функционирования не следует понимать упрощенно как области жизни и деятельности человека. Такое упрощенное понимание приводит к выделению многочисленных «сфер» (например сферы отдыха и досуга, внутри которой можно было бы выделить еще сферы спорта, туризма, шоу-бизнеса и развлечений и мн. др.), то есть отождествляет сферу функционирования с ситуацией общения. Так, например, А. И. Горшков, критикуя основной принцип функциональной стилистики, исходит именно из такого понимания термина «сфера функционирования» [Горшков, 2001].

Сферы функционирования — это сферы **речевой** деятельности. Они определяют выбор правил речевого поведения, форм речевого взаимодействия автора и адресата, выбор речевых жанров, принципов построения и языкового оформления высказываний.

Различия между функциональными сферами не сводятся и к тематическим различиям. Например, тематика в бытовой сфере может быть бесконечно разнообразной, а объединяет эти тексты то, что они отражают непринужденное общение людей, не связанных официальными отношениями. В официально-деловой сфере, напротив, отношения официальные и коммуниканты выступают не как индивидуальности, а как социальные единицы (граждане государства, члены трудового коллектива, партийного или иного объединения и т. д.), что накладывает общий отпечаток на функционирующие в этой сфере тексты, будь то правовые, дипломатические или административно-канцелярские документы, и несмотря на различия между ними.

Для описания текстов НКРЯ выделено 8 функциональных сфер: учебно-научная, производственно-техническая, официально-деловая, публицистики (и массовой информации), рекламы, церковно-богословская, художественная, бытовая. Для подкорпуса устных текстов предлагается ввести дополнительно сферу устной публичной речи и устной непубличной речи [Гришина, 2005].

Учебно-научная сфера объединяет тексты научного и научно-методического содержания, относящиеся к различным областям науки и образования, целью которых является описание, объяснение и прогнозирование процессов и явлений действительности в логико-понятийной форме. Основная функция языка в учебно-научной сфере — информативная.

Производственно-техническая сфера — это среда функционирования таких текстов, как описания технических устройств и производственных процессов, предписания, регулирующие профессиональные действия человека в среде искусственных объектов. Она смыкается с учебно-научной сферой, с одной стороны (научно-технические тексты), и с деловой сферой, — с другой. В производственно-технической сфере реализуется информативная функция языка и функция воздействия (в инструкциях).

В *официально-деловой* сфере функционируют тексты, основное назначение которых состоит в регламентации отношений между

государством и его гражданами, организациями и другими государствами, между организациями и внутри них, между организациями и частными лицами в процессе производственной, хозяйственной и юридической деятельности. Функция воздействия является основной наряду с информативной.

Сфера *публицистики* объединяет тексты, назначение которых состоит в информировании населения и формировании общественного мнения по вопросам общественной значимости в области политики, экономики, искусства, науки, морали и пр. В этой сфере реализуются функции информативная, воздействия и отчасти эстетическая.

В сфере *рекламы* функционируют тексты, нацеленные на формирование потребностей, главным образом, материальных; их задача — информировать адресата о достоинствах рекламируемых товаров и услуг и в конечном счете побудить его купить товар или воспользоваться услугой. Сфера рекламы смыкается с деловой сферой в области торговой рекламы и с публицистикой в области социальной и политической рекламы. В ней реализуются функции информативная и воздействия.

Церковно-богословская сфера объединяет тексты религиозного содержания, которые представляют религиозную картину мира и обслуживают различные стороны религиозной жизни индивида (молитвы, церковные обряды, исповедь, проповедь и пр.). Функции языка в этой сфере — информативная и воздействия.

Бытовая сфера — это сфера повседневного, непринужденного, неформального общения в кругу лиц, объединенных неофициальными отношениями (родственников, друзей, коллег по работе, учебе и под.). В бытовой сфере реализуется функция общения, которая может сопровождаться другими функциями (фатической, информационной, воздействия и т. д.). Тексты, относящиеся к этой сфере, существуют в основном в устной форме, но возможна и письменная форма — личная переписка, записки, дневники, поздравления и др. В последние годы возникают новые формы спонтанной коммуникации — электронная переписка, чаты, форумы и т. д. Изучение специфики электронной формы непринужденного общения может привести к постановке вопроса о выделении новой сферы функционирования текстов.

Художественная сфера — это сфера словесного художественного творчества, объединяющая тексты, в которых воплощается

созданный воображением автора мир. Главной отличительной особенностью художественного текста является его особая по сравнению со всеми другими разновидностями предназначенность. Организация языковых средств в художественной литературе подчинена не просто передаче содержания, а передаче содержания в эмоциональной, наглядно-образной форме. В художественной сфере язык выступает в эстетической функции и, опосредованно, в функции воздействия.

Этот параметр не используется в BNC и CNC, в которых в качестве основной типологической характеристики текстов рассматривается их тематика[8]. Однако именно распределение корпуса по сферам функционирования является наглядным показателем его представительности. Тексты НКРЯ распределяются по сферам функционирования так, как это указано на с. 76 (подкорпус текстов XIX в.) и с. 75 (основной корпус).

8. Тема текста, или предметная область

Определение тематики текста имеет субъективный характер, поскольку в больших по объему текстах чаще всего обязательно представлено несколько тем. Даже небольшой по объему текст может быть отнесен к разным предметным областям, например, коммерческое предложение интернет-услуг можно отнести к области бизнеса, коммерции или компьютерных технологий; заметку о выставке военной техники — к области техники или военного дела; советы огороднику — могут получить характеристики «дом и домашнее хозяйство» или «сельское хозяйство». Поэтому трудно или даже невозможно составить идеальный перечень тематических областей; необходимо учитывать, что во многих случаях однозначное отнесение текста к определенной предметной области достаточно условно. При характеристике тематики текстов НКРЯ в случае неоднозначности указываются обе «конкурирующие» предметные области.

Тем не менее параметр «тематика текста» используется во всех известных корпусах. В Британском, Чешском, Польском, Американском корпусах классификация предметных областей строится на рекомендациях EAGLES. Набор предметных областей, ис-

[8] Некоторое подобие использования идеи функциональных сфер в BNC и CNC можно увидеть в разграничении письменных текстов на художественно-литературные, специальные и публицистические.

пользуемых при классификации текстов НКРЯ, в основном также совпадает с рекомендациями EAGLES, незначительно различаясь в деталях. Различия касаются членения некоторых областей, что можно представить в таблице на с. 77.

Как видно из таблицы, в перечне тематических областей, предлагаемых EAGLES, в одном ряду встречаются области, находящиеся в отношении соподчинения, например, «естественные науки» и «физика», «биология» и пр., «досуг» и «мода», «путешествия», «спорт». Поэтому Синклер предлагает выстроить их в виде раскрывающегося списка (2 и 2.1, 2.2, 2.3..., 3 и 3.1, 3.2, 3.3...). Предметные области, используемые для описания текстов НКРЯ, образуют линейный список, при этом для уточнения тематики научных текстов применяется двойная характеристика «наука и технологии|конкретная наука»: «наука и технологии|физика», «наука и технологии|социология» и т. д.

В целом в НКРЯ используется более обобщенное представление предметных областей. Прежде всего это касается таких областей, как «Политика», «Искусство», «Досуг». Тематическая область «Компьютеры» не выделяется, в отличие от EAGLES, в качестве самостоятельной, а соответствующие тексты распределяются, в зависимости от содержания, по областям «Информатика», «Техника» или «Наука и технологии».

Не совсем ясно в классификации Синклера наличие трех характеристик, связанных с литературой и чтением: fiction — художественная литература, arts/literature (по-видимому, речь идет о литературной критике и литературоведении) и leisure/reading.

Наконец, в НКРЯ введен параметр «природа» для характеристики текстов, связанных с описанием растительного и животного мира — очерков, зарисовок, записок фенолога и пр., довольно часто встречающихся в прессе. В этом случае термин «экология», предлагаемый EAGLES, не подходит, так как он больше ассоциируется с охраной окружающей среды. Для текстов газетно-публицистических на морально-нравственные, житейские темы и для частных писем введена тематическая характеристика «частная жизнь» (в отличие от тем, касающихся общественно-значимых проблем, которым соответствует признак «политика и общественная жизнь»).

Основной корпус

Сфера функционирования	**Предполагаемое распределение в 100 млн. корпусе, %**	**Текущее состояние в основном корпусе**		
		кол-во текстов	**кол-во словоупотреблений**	**%**
Художественная	34%	1868	27,758,035	40%
Учебно-научная, в т. ч. научно-популярная	15%	733	4,065,636	5,9%
Производственно-техническая	0,5%	39	112,485	0,16%
Официально-деловая	1,5%	275	1,071,432	1,5%
Публицистическая,	40%	21653	32,242,368	46,6%
в т. ч. мемуары	12%	352	11,208,544	16,3%
Церковно-богословская	2%	461	1,324,287	1,9%
Реклама	0,5%	528	250,360	0,4%
Бытовая,	1,3%	57	441,093	0,64%
в т. ч. устная непубличная речь				
	0,3%			
Интернет-коммуникация	1,5%	61	1,132,669	1,6%
Устная публичная речь	3,7%	203	852,165	1,2%
Всего	100%	25878	69,250,530	100%

Подкорпус текстов XIX в.

Сфера функционирования	**Текущее состояние**		
	кол-во текстов	**кол-во словоупотреблений**	**%**
Художественная	495	9,871,849	64,4%
Учебно-научная, в т. ч. научно-популярная	36	931,664	6,1%
Официально-деловая	1	1246	
Публицистическая, в т. ч. мемуары	97	2,771,315	18,1%
	40	1,687,254	11,0%
Церковно-богословская	22	572,244	3,7%
Бытовая	15	380,228	2,5%
Всего	666	15,337,195	100%

Следует отметить, что в BNC тематика определяется только для текстов нехудожественной прозы (вся художественная литература имеет одинаковую тему «life»), в CNC по тематическим областям распределены только специальные тексты, публицистика представлена как целое (journalism); в НКРЯ же тематическую характеристику имеют все тексты (за исключением художественной литературы и мемуаров, для которых разработана особая классификация, описанная в следующем разделе).

Этот факт, наряду с высказанными выше соображениями об относительности и субъективности тематических характеристик, необходимо учитывать при сравнении распределения текстов по предметным областям в разных корпусах (см. также [Čermák, 2001]).

9. Хронотоп, или место и время описываемых событий

Наиболее спорной в списке тем, предложенных Синклером, является область «Life», которая используется исключительно для характеристики тематики художественных текстов, мемуаров, дневников. В НКРЯ эта характеристика не применяется, а для уточнения тематики художественных, мемуарных, биографических текстов введен параметр «место и время описываемых событий», который условно определяет хронотоп содержания текста. Напр., мемуары Амосова имеют хронотоп «Россия/СССР: советский период», повесть Б. Васильева «А зори здесь тихие» — «Россия/СССР: 1941-1945», и т. д. Значения параметра образуют открытый список,

Тематическая классификация НКРЯ в сопоставлении с рекомендациями EAGLES

EAGLES	НКРЯ
1. Life	частная жизнь
2. Естественные науки Natsci	наука и технологии|естествознание
	наука и технологии|астрономия
2.1 математика	наука и технологии|математика
2.2 физика	наука и технологии|физика
2.3 химия	наука и технологии|химия
2.4 биология	наука и технологии|биология
2.5 геология, география	наука и технологии|геология
	наука и технологии|география
3. Гуманитарные науки Socsci	наука и технологии|конкретная наука
3.1 юриспруденция	наука и технологии|право
3.2 история, археология	наука и технологии|история
3.3 философия	философия
3.4 психология	наука и технологии|психология
3.5 социология	наука и технологии|социология
	наука и технологии|статистика
3.6 антропология	
3.7 лингвистика	наука и технологии|филология
3.8 образование	наука и технологии|образование
	наука и технологии|культурология
	наука и технологии|искусствоведение
4. Прикладные науки Appsci	
4.1 сельское хозяйство	(наука и технологии |) сельское хозяйство
4.2 медицина	(наука и технологии |) здоровье и медицина
4.3 экология и окружающая среда	(наука и технологии |) природа
4.4 техника и технологии (engineering)	(наука и технологии|) техника /производство
4.5 компьютеры	наука и технологии |информатика / техника
4.6 военное дело	наука и технологии|армия и вооруженные конфликты
4.7 транспорт	(наука и технологии |) транспорт
5. Политика	политика и общественная жизнь
5.1 внешняя политика	
5.2 внутренняя политика	
6. Экономика	бизнес, коммерция, экономика, финансы
6.1 финансы	
6.2 промышленное производство (industry)	производство
	администрация и управление
7. Искусство	искусство и культура
7.1 изобразительное искусство	
7.2 литература	
7.3 архитектура	
7.4 театр, кино, танец (performance)	
8. Досуг (leisure)	досуг, зрелища и развлечения
8.1 чтение	
8.2 спорт	спорт
8.3 путешествия	путешествия
8.4 мода	
9. Религия	религия

который пополняется по мере включения новых текстов. Набор значений этого параметра, актуальный для настоящего состояния корпуса, приведен в таблице ниже. Возможность различных комбинаций значений делает схему достаточно гибкой и позволяет характеризовать содержание большинства текстов без необходимости введения новых характеристик.

Значение признака	**Жанры художественной литературы**
доисторический период античность Средние века Новое время	историческая проза
Россия: Средние века Россия: 15 век Россия: 17 век Россия: 18 век Россия: 19 век	историческая проза, мемуары
Россия: 1900-1914 Россия: 1914-1920 Россия/СССР: 1920-е Россия/СССР: 1930-е Россия/СССР: 1940-1945 Россия/СССР: 1950-е Россия/СССР: 1946-1952 Россия/СССР: 1960-1980 Россия/СССР: советский период Россия/СССР: перестройка Россия: постсоветский период Белоруссия: постсоветский период Средняя Азия: постсоветский период Закавказье: постсоветский период	современная художественная проза автобиографическая проза мемуары
Европа: 19 век Европа: 1-я пол. 20 в. Европа: 2-я пол. 20 в. Америка: 1-я пол. 20 в. Америка: 2-я пол. 20 в. Ближний Восток: античность Ближний Восток: 2-я пол. 20 в. Дальний Восток: 2-я пол. 20 в.	
Америка: современность Европа: современность Австралия: современность Ближний Восток: современность Дальний Восток: современность	современная художественная проза, мемуары
ирреальный мир	фантастика и фэнтези, мифология

10. Тип текста

Параметр «тип текста» определяет принадлежность текста к определенному речевому жанру. Понятие речевого жанра как типической воспроизводимой формы высказывания, характеризующейся триединством тематического содержания, композиции и стиля, было разработано М. М. Бахтиным [Бахтин, 1979] и в настоящее время относится к числу фундаментальных представлений стилистики, лингвистики текста, социолингвистики (см. [Дементьев, 1997; Шмелева, 1990]). Согласно такому пониманию жанра, которое можно назвать лингвистическим, каждая речевая сфера вырабатывает свой репертуар речевых жанров: в учебно-научной сфере специфическими жанрами являются научная статья, монография, учебник, реферат и т.д ., в публицистике — заметка, репортаж, интервью и т. д., в официально-деловой сфере — закон, постановление, приказ, акт и пр., в художественной — роман, повесть, рассказ.

Однако недостатком термина жанр является его многозначность: наряду с рассмотренным выше лингвистическим пониманием термина существует литературоведческая традиция выделения и описания жанров художественной литературы (см. ниже). Поэтому, чтобы избежать смешения терминов, в базе данных корпуса для описания жанровой формы текста используется нейтральный термин «тип текста».

Значения этого параметра представляют собой список, в настоящее время включающий около 100 позиций[9], который принципиально открыт в силу двух обстоятельств: во-первых, в него включены названия типов, которые уже представлены в корпусе, и он будет пополняться по мере появления новых текстовых типов; во-вторых, в науке к настоящему времени не разработано единой типологии текстов [Дементьев, 1997], и полное описание всей системы текстовых типов рассматривается как первоочередная задача лингвистики речи [Шмелева, 1993].

В BNC при разметке текстов их жанровая принадлежность не учитывалась и балансировка корпуса строилась на трех показателях — «предметная область», «время создания текста» и «источник»

[9] О составе речевых типов в подкорпусе устной речи см [Гришина, 2005] и статью Е. А. Гришиной в настоящем сборнике.

Тип текста	**Функциональная сфера**
заметка	публицистика
информационное сообщение	публицистика
интервью	публицистика
комментарий	публицистика
обзор	публицистика
отзыв	публицистика, учебно-научная
отчет	публицистика, учебно-научная, официально-деловая
очерк	публицистика, художественная
рецензия	публицистика, учебно-научная
статья	публицистика, учебно-научная
хроника	публицистика
объявление	реклама
анонс	реклама
монография	учебно-научная, публицистика
справочник	учебно-научная, производственно-техническая
учебник	учебно-научная
учебное пособие	учебно-научная
закон	официально-деловая
постановление	официально-деловая
кодекс	официально-деловая
автобиография	официально-деловая
акт	официально-деловая
договор	официально-деловая
заявление	официально-деловая
инструкция	официально-деловая, производственно-техническая
письмо служебное	официально-деловая
резюме	официально-деловая
рекомендация	официально-деловая
характеристика	официально-деловая
путеводитель	публицистика, реклама
рецепт	бытовая
гороскоп	бытовая
дневник, записные книжки	бытовая, публицистика
письмо личное	бытовая
сочинение	учебно-научная
проповедь	церковно-богословская
молитва	церковно-богословская
роман	художественная
рассказ	художественная
повесть	художественная
сказка	художественная
миниатюра	художественная
мемуары	публицистика
ассоциативная проза	художественная
письмо литературное	художественная
эссе	художественная
пьеса	художественная

[BNC, 2000]. Недостатки такого решения разработчиков BNC подробно рассматриваются в обширной статье Д. Ли, который изучил жанровый состав корпуса и составил новую специальную базу данных (The BNC Index), включающую и жанровую характеристику текстов [Lee, 2001]. Для BNC Д. Ли выделил 70 жанров (и субжанров): 24 — для записей устной речи, 46 — для письменных текстов. Однако то, что Д. Ли предлагает рассматривать как жанры, является по большей части типологическими образованиями, лежащими на пересечении тематики и функциональных сфер. Так, среди 46 «жанров», выделенных для письменных текстов, наряду с реальными жанрами («биографии/автобиографии», «школьные сочинения», «личные письма», «деловые письма») встречаются следующие типы: «академическая проза: естественные науки», «академическая проза: политика, законодательство, образование», «академическая проза: технология, компьютеры, инженерное дело», «печатная реклама», «художественная литература: драма», «художественная литература: поэзия», «художественная литература: проза», «центральные газеты: искусство/культура», «центральные газеты: разное», «региональные и местные газеты: политика», «региональные и местные газеты: наука» и т. д.

В CNC используется несколько типологических характеристик текста:

- тип текста: художественный, информативный, переходные типы;
- тип жанра: список жанров различен для специализированных и неспециализированных текстов и включает около 60 типов (например, драма, роман..., музыка, философия,.. промышленность, спорт, ... религия и т. д.);
- тип субжанра (жанровой разновидности), например, учебник, критическая статья, энциклопедия;
- текстовый тип, т. е. стихи или проза [Čermák, 2001].

Здесь видим то же смешение тематических и жанровых характеристик, что и при описании жанрового состава BNC. Таким образом, простое количественное сравнение распределения текстов по типам в разных корпусах не даст ожидаемого результата, поскольку типы выделяются по разным принципам и представляют собой образования разных уровней.

11. Жанр художественной литературы

Данный параметр используется только для описания художественных текстов. В традиции литературоведения и теории словесности произведения художественной словесности принято делить на роды, виды и жанры. В соответствии с самым общим делением — на роды — разграничивают эпос, лирику и драму. Внутри родов выделяются виды: роман, повесть — виды эпоса; элегия, мадригал — виды лирики; комедия, трагедия — виды драмы. Частное проявление вида, определяемое тематикой произведения, называется жанром художественной литературы [Горшков, 2001]. Именно в таком значении используется параметр «жанр» при описании текстов Национального корпуса. Поскольку в настоящее время художественные тексты представлены в основном прозой, список жанров невелик:

внежанровая проза	признак используется для характеристики основного потока «серьезной» художественной литературы
историческая проза	основное содержание — изображение конфликтов иной эпохи
приключения	основное внимание на перипетиях сюжета
фантастика	изображение фантастических миров — гипотетического будущего или параллельных мифических миров (фэнтези)
детектив, боевик	внимание на описании преступления и процесса его раскрытия
юмор и сатира	тексты комического содержания: скетчи, юморески, миниатюры
любовная история	жанр массовой литературы, в центре сюжета — история любви
детская	литература для детей и подростков
автобиографическая проза	художественное произведение на документальной основе

Следует отметить, что в Британском и Чешском корпусах для художественной литературы используется самое общее разграничение поэзии, прозы и драмы.

12. Стиль текста

С помощью этого параметра описывается языковая форма текста, прежде всего его лексический состав. Система кодирования стилистических особенностей разработана отдельно для нехудожественных текстов и художественной литературы. В центре стилис-

тических оппозиций находится нейтральный стиль. Нейтральный стиль отражает стилистическую норму данной функциональной сферы. Естественно, что в разных функциональных сферах эта норма будет разной. Так, основу стилистической нормы текстов научных, публицистических, официально-деловых составляют книжно-письменные языковые средства. В бытовой сфере стилистическая норма формируется устно-разговорными средствами (о фундаментальном членении литературного языка на книжно-письменную и устно-разговорную разновидности см. [Лаптева, 2003]).

Для нехудожественных текстов отмечаются отклонения от нейтрального стиля в сторону большей официальности (в официально-деловой сфере и в некоторых жанрах публицистики) и академичности (в учебно-научной сфере). Помету «официальный» получают тексты, в которых преобладают книжные средства и конструкции официально-деловой речи: такой стиль характерен для законов, юридических документов, официальных сообщений в прессе и т. д. Помету «специальный» имеют научные тексты, рассчитанные на специалистов: в них преобладает терминология, отсутствуют эксплицитные объяснения, и т. д.

Для художественной прозы принята следующая система помет: нейтральный — региональный — сниженный — индивидуально-авторский стиль. Если в тексте преобладают общелитературные средства, то его стиль характеризуется как «нейтральный». Если в нем велика доля средств, выходящих за пределы литературной нормы, и это расширение происходит за счет диалектизмов и регионализмов, то стиль текста получает помету «региональный» (некоторые рассказы В. Шукшина). Если текст насыщен элементами просторечия, жаргонизмами или обсценной лексикой, то его стиль характеризуется как «сниженный» (Ю. Алешковский, Н. Медведева). Помету «индивидуально-авторский» получают тексты, носящие следы языкового эксперимента, которые отличает специфическое словоупотребление, смещение значений, словотворчество, особый синтаксис и т. д. (тексты С. Соколова, В. Нарбиковой и др.).

К сожалению, в стилистике не выработаны критерии количественного измерения стилистических свойств текста — степени его «сниженности», «официальности» или «индивидуальности». Пока эти признаки определяются субъективно, на основе экспертных оценок, общего впечатления от языковых особенностей текста.

Так, например, помету «сниженный» получают художественные тексты, в которых сниженные языковые средства присутствуют не только в речи персонажей, но и в авторской речи. Но, вероятно, именно стилистическая разметка текстов корпуса будет способствовать изучению этого вопроса и позволит в дальнейшем выявить объективные количественные показатели стилистических характеристик.

III. Информация об аудитории

13. Возраст аудитории

Ориентация на возраст предполагаемой аудитории во многом определяет содержание текста и его языковое оформление. Этот параметр позволяет разграничить детскую литературу в составе художественной, специфические «молодежные» издания в составе публицистики и учебную литературу для разных категорий учащихся. Выделяется несколько значений этого признака, которые характеризуют аудиторию как детскую (0-10 лет), подростковую (11-17 лет), молодежную (18-34 года). В остальных случаях тексты получают помету «н-возраст», что означает, что признак возраста не оказывает существенного влияния на свойства текста. Таким образом, помета «взрослая аудитория», которая первоначально использовалась по умолчанию для характеристики текстов, немаркированных по признаку возраста, оказывается избыточной, не говоря уже о ее амбивалентности: взрослую аудиторию можно трактовать как «любую, смешанную» и как «только для взрослых».

14. Уровень образования аудитории

При оценке уровня образования аудитории учитываются два показателя: 1) знание о конкретном предмете (общее и специальное) и 2) уровень образования (высокий и низкий). Эти параметры взаимно дополняют друг друга: тексты могут быть предназначены для аудитории без специальных знаний о предмете, но предполагают общий высокий уровень образования. Это означает, что такие тексты используют минимум специальной терминологии, но могут апеллировать к абстрактным категориям. Напротив, другие тексты могут быть предназначены специалистам с невысоким уровнем общего образования и использовать большое количество специальной терминологии, но не абстрактных рассуждений по данной теме.

Для описания текстов корпуса используются четыре значения этого параметра: а) «высокий» уровень образования, если текст рассчитан на читателя с высоким уровнем общего образования и с общим знанием о предмете; б) «профессиональный», если текст рассчитан на специалистов с различным уровнем общего образования; в) «низкий», если текст предназначен для нетребовательного читателя с низким уровнем общего образования и отсутствием специальных знаний о предмете (например, публикации в «желтой прессе»); г) в остальных случаях, если признак нерелевантен, используется помета «н-уровень.

15. Размер аудитории

Этот количественный параметр позволяет разграничить большие классы текстов: тексты, предназначенные для публичной аудитории, которая может быть малой (до 1 тысячи человек), средней (1-50 тысяч человек), большой (до 1 млн человек) и очень большой (свыше 1 млн человек), и частной аудитории, в свою очередь подразделяемой на личную (1 человек) и групповую (от 5 до 30 человек). Публичная аудитория характерна для печатных изданий, электронной коммуникации; групповой аудитории, как правило, адресованы канцелярские документы, учебные лекции, личную аудиторию имеет личная переписка.

IV. БИБЛИОГРАФИЧЕСКОЕ ОПИСАНИЕ ТЕКСТА

16. Источник текста

Тексты поступают в корпус из разных источников: электронные версии выпущенных книг, газет и журналов могут быть предоставлены издательствами и информационными агентствами, тексты могут быть взяты из общедоступных электронных библиотек и сверены с оригиналом, могут быть получены путем сканирования и ручного набора с печатных или рукописных оригиналов. При метаразметке издательских версий в качестве источника указываются выходные данные книги, название и дата выхода газеты или журнала. Тексту, взятому из электронной библиотеки и сверенному с печатным оригиналом, приписываются выходные данные печатной версии. Аналогично описывается и отсканированный изданный текст. Для неизданных текстов источник определяется как рукопись. Для текстов, полученных из Интернета, указывается адрес сайта.

17. Название издания

Этот признак релевантен только для печатных изданий (книг, газет, журналов). В данном поле указывается название тома, в составе которого опубликован помещенный в корпус рассказ, повесть, статья и т. д. Для газетных и журнальных статей в этой указывается только название печатного органа.

18. Название издательства

Этот признак используется только для характеристики текстов, опубликованных в книгах.

19. Год издания

Этот признаки используются для характеристики текстов, опубликованных в книгах и в периодических изданиях.

20. Тип носителя

Указывается, в какой материальной форме существовал текст до его включения в корпус: рукописной, машинописной (деловая документация), печатной (в виде книги, газеты. журнала, брошюры, листовки и т. д.), электронной и пр.

В полях 17-19 фактически дублируется информация об источнике текста, однако это позволяют организовать поиск по изданию (например, отобрать тексты из газеты «Известия», журнала «Новый мир» за 1997 год и под.).

V. СЛУЖЕБНАЯ ИНФОРМАЦИЯ

В эту группу объединяются несколько параметров:

21. Качество электронной версии текста

22. Условное название подкорпуса, в состав которого включается текст, например, «корпус со снятой вручную омонимией», «корпус с неснятой омонимией», «устный корпус», «диалектный корпус» и др.

23. Комментарии. Сюда заносится любая дополнительная информация о тексте, которая не нашла отражения в основных полях базы данных.

24. Спонсор. Здесь указывается название организации или имя конкретного лица, предоставившего электронную версию текста в распоряжение разработчиков корпуса.

25. Ответственный. В этом поле указываются имена сотрудников, ответственных за подготовку текста к помещению в корпус.

* * *

Описанная система метаразметки позволяет пользователю отбирать тексты по любому из признаков или их комбинациям и формировать свой подкорпус текстов для решения конкретных лингвистических задач.

Литература

Бахтин М. М. Проблема речевых жанров // М. М. Бахтин. Эстетика словесного творчества. М., 1979

Васильева А. Н. Курс лекций по стилистике русского языка. М., 1976

Гарабик Р. Словацкий национальный корпус // Труды международной конференции «Корпусная лингвистика-2004». — СПб., 2004. С. 99-121

Гольдин В. Е., Сиротинина О. Б., Ягубова М. А. Русский язык и культура речи: Учебник для студентов-нефилологов. М., 2002

Горшков А. И. Русская стилистика. М., 2001

Горшков А. И. Русская словесность: Учебник для общеобразовательных учреждений. М., 2000. С. 221-238

Гришина Е. А. О принципах размещения устных текстов в Национальном корпусе русского языка // НТИ. Сер. 2. № 3. 2005. С. 31-38

Дементьев В. В. Изучение речевых жанров: обзор работ в современной русистике // ВЯ, 1997, № 1. С. 109-121

Кожина М. Н. Стилистика русского языка: Учебник для студентов пед. ин-тов. М., 1993

Лаптева О. А. Теория современного русского литературного языка. М.: Высшая школа, 2003

Савчук С. О., Соколова Е. Г. Тексты ограниченного обращения в составе Национального корпуса русского языка // НТИ. Сер. 2. № 3. 2005. С. 13-23

Современный русский язык: Социальная и функциональная дифференциация / Отв. ред. Л. П. Крысин. М.: Языки славянской культуры, 2003

Чебанов С. В., Мартыненко Т. Я. Семиотика описательных текстов: Типологический аспект. СПб.: Изд-во СПб ун-та, 1999

Шаров С. А. Представительный корпус русского языка в контексте мирового опыта // НТИ, Сер. 2, № 6. 2003. С. 8-18.

Шаров С. А., Савчук С. О. Типология текстов для представительного корпуса. // Труды международной конференции «Корпусная лингвистика-2004». СПб., 2004. С. 352-362

Швейцер А. Д., Никольский Л. Б. Введение в социолингвистику. М., 1978

Шимкова М. Репрезентативность корпуса как лингвистическая проблема // Международная конференция MegaLing'2005. Прикладная лингвистика в поиске новых путей. 27июня — 2 июля 2005. Украина, Крым. Материалы конференции.

Шмелева Т. В. Речевой жанр: Возможности описания и использования в преподавании языка // Русистика, 1990, № 2. С. 20-32

Шмелева Т. В. Повседневная речь как лингвистический объект // Русистика сегодня: Функционирование языка: лексика и грамматика. М., 1993. С. 8-15

BNC: The BNC Users Reference Guide, 2000. http://www.natcorp.ox.ac.uk/World/HTML.

Čermák Fr. Language Corpora: The Czech Case // Text, Speech and Dialogue, TSD 2001, eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer. Springer Berlin etc. 2001, 21-30

Christopher S. Butler. Corpus studies and functional linguistic theories // Functions of language 11:2 (2004), p. 152

Kopřivová M. Český národní korpus na přelomu tisíciletí. // Český národní korpus/ Praha 2000 <http://ucnk.ff.cuni.cz>

Lee, D. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* Vol. 5, No. 3, September 2001, pp. 37-72. http://llt.msu.edu/vol5num3/pdf/lee.pdf

Sinclair, J. Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P, 1996. http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Sharoff, S. Towards basic categories for describing properties of texts in a corpus. In *Proc. of Language Resources and Evaluation Conference (LREC04)*. May, 2004, Lisbon, Portugal, http://www.comp.leeds.ac.uk/ssharoff/texts/lrec-04.pdf

Sharoff, S. Open-sourse Corpora: using the net to fish for linguistic data // http://corpus.leeds.ac.uk/serge/internet-corpora.pdf