Е. А. Гришина

ДВА НОВЫХ ПРОЕКТА ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА: МУЛЬТИМЕДИЙНЫЙ ПОДКОРПУС И ПОДКОРПУС НАЗВАНИЙ

I. Мультимедийный подкорпус

Обоснование проекта

На данный момент в Национальном корпусе представлены практически все формы существования русского языка, включая и самые маргинальные, — представлены реально, т. е. включены в Корпус, размечены и доступны для пользователя, или в качестве проекта (например, поэтические и песенные тексты, фольклор и др.[1]).

При этом, однако, — на наш взгляд, — Корпусом не охвачен один пласт бытования русского языка, который можно условно назвать «мультимедиа», т. е. тексты, спаянные со зрительным и звуковым рядом. Насколько нам известно, в «эталонных» национальных корпусах, на которые ориентировались создатели Национального корпуса русского языка, прежде всего — в Британском и Чешском, такого рода тексты также не отсутствуют.

Объяснение этому зиянию находится легко: мультимедиа попадают в «провал» между тремя основными формами бытования языка, в данный момент отражаемыми в корпусах, в том числе и в русском Корпусе: письменная речь, устная речь, «электронная» речь (о последней см. [Капанадзе 2005]). Мультимедийные тексты — это не письменная и не электронная речь, поскольку имеют в основном устную форму бытования, но, с другой стороны, это и не устная речь per se: основные характеристики последней — спонтанность и невоспроизводимость (что обеспечивает ее неповторимое синтаксическое своеобразие), а мультимедийная речь в достаточной степени подготовлена (иногда уровень подготовленности

[1] См. статью Е. А. Гришиной и В. А. Плунгяна «Перспективы развития Национального корпуса русского языка» в настоящем сборнике.

мультимедиа существенно выше, чем уровень подготовленности письменной речи) и, безусловно, не просто воспроизводима, а рассчитана на воспроизводимость. Тем самым, мультимедийные тексты, несомненно существуя (и оказывая влияние на язык, в том числе и на его письменную, устную и электронную форму бытования), одновременно оказываются несуществующими, поскольку взгляд исследователя не настроен на их видение — они «проскальзывают» сквозь слишком крупные ячейки классификационной сетки.

Таким образом, отсутствие мультимедийной составляющей в корпусах вполне объяснимо, но при этом, как представляется, совершенно логически не обосновано. Для того, чтобы прояснить это утверждение, представляется целесообразным перечислить (в первом приближении) те классы текстов, которые должны составить ядро мультимедийного подкорпуса. Это прежде всего текстовая составляющая

1) кино- и мультипликационных фильмов[2], теле- и радиопередач

2) теле- и радиорекламы, теле- и радиообъявлений

а также

3) либретто опер и оперетт[3].

Три названные выше позиции позволяют поставить следующие вопросы.

Существуют ли данные тексты в русском речевом космосе? Ответ: очевидно, да.

Попадают ли они в Национальный корпус при нынешних установках на отбор текстов? Ответ: практически нет.

---

[2] Отметим, что *песни* из кинофильмов уже запланированы для включения в состав Корпуса при создании поэтического подкорпуса.

[3] Объективности ради следует отметить, что либретто — в небольшом количестве — уже планируются к включению в состав Корпуса, однако в отсутствие мультимедийного подкорпуса они рассматриваются как одна из разновидностей т. н. ТОГов (текстов ограниченного обращения, см. статью [Савчук, Соколова, 2005]), т. е. как достаточно маргинальный вид текстов. При включении в Корпус мультимедиа либретто займут в нем свое законное место. Данная ситуация до некоторой степени напоминает то положение вещей, которое было характерно для начальной фазы существования Национального корпуса русского языка, когда в нем еще не был предусмотрен устный подкорпус: ряд текстов, очевидным образом «устных», просто не опознавались как таковые и помещались вместе с остальным массивом.

Достаточно ли эти тексты влиятельны, чтобы требовать включения в состав Корпуса? Для того, чтобы ответить на этот вопрос, обратимся хотя бы к словарям крылатых слов русского языка. В небольшом (примерно на 1000 единиц) словаре крылатых слов из области искусства [Шулежкова 2003] упомянуто 200 кино- и мультфильмов, цитаты из которых прочно вошли в русский язык (нужно специально отметить, что эти цитаты не просто устойчиво воспроизводятся в русской речи, но и служат моделями для построения целой серии фраз (например, фраза из кинофильма 1966 г. «Волшебная лампа Аладдина» «В Багдаде все спокойно» как основа для модели *«В X все спокойно» — В Тавриде все спокойно, В Чечне все спокойно* [Шулежкова 2003, с. 36] и т. д. и т. п.).

После советской экранизации (к сожалению, достаточно неудачной) романа «Трое в лодке, не считая собаки» огромное количество материалов в прессе, посвященных приусадебному хозяйству, стали называться «Все в сад!», что восходит к реплике Джорджа в исполнении А. Ширвиндта (не зная этого факта, невозможно опознать в такого рода заголовках цитату).

Как известно, пьеса Е. Шварца «Обыкновенное чудо» была написана в 1954 г. Первая экранизация — 1965 г. (реж. Х. А. Локшина), культовая экранизация М. Захарова — 1978 г. Самый поверхностный анализ уровня цитирования покажет всплеск обращения к этому произведению, начинающийся с 1979 г., а не с 1955 и не с 1966 годов, что показывает, какой именно текст хранится в культурной памяти носителей русского языка, — это не текст пьесы Е. Шварца и не текст фильма Х. Локшиной, а текст фильма М. Захарова, т. е. именно кинофильм 1978 г. сделал это произведение общеизвестным явлением русской культуры.

Примеры можно множить бесконечно. «Кино входит в жизнь, переплетается с жизнью и запечатлевается в ней, прежде всего — в языке. ... Бытовая цитация кинолент — неотъемлемая часть обиходно-речевого поведения человека. ... Словесный кинофольклор грандиозен по своему объему»[4]. Уже одно это показывает, что от-

[4] [Елистратов 1999, с. 4]. Отметим, что «российское» цитирование захватывает и зарубежный кинематограф. Достаточно вспомнить героя Хамфри Богарта из «Касабланки» («Я думаю, это начало прекрасной дружбы»), «Твин-Пикс» Д. Линча («Совы — не то, чем они кажутся»), диалог из «Криминального чтива» («Who is Zed? — Zed's dead, baby, Zed's dead») и, конечно, «I'll be back» и «Hasta la vista» из «Терминатора».

сутствие текстов такого масштаба влияния в Национальном корпусе русского языка напоминает гипотетическую ситуацию отсутствия в Корпусе «Горя от ума»[5] и басен И. Крылова — при том, что тексты, включающие в себя цитаты из этих произведений, в Корпус входили бы.

На этом фоне действительно несколько необычным кажется отсутствие мультимедийной составляющей по крайней мере в Британском национальном корпусе. В. Елистратов полагает, что «таких **масштабов** цитации нет и не было ни в одном языке. Иначе говоря, в России (Советском Союзе) кино было в первую очередь воспринято народом с точки зрения текста...» [Елистратов 1999, с. 6], что, как полагает автор, было в гораздо меньшей степени характерно для других европейских культур. Представляется, однако, что если это и так, то лишь отчасти[6], т. е. основная причина невключения кинематографических текстов в состав корпуса — та, о которой упоминалось выше: непризнание за мультимедийным текстом права на самостоятельность.

В. Елистратов во введении к своему словарю пишет: «для меня совершенно очевидно, что в 1990-е гг. данная эпоха уже закончена» и «Кинологос» «“качнулся” куда-то в сторону политической риторики, риторики шоу-бизнеса, рекламы, рок-текстов» [Елистратов 1999, с. 5]. По общему ощущению, с этим следует согласиться, но это, следовательно, означает, что нынешнее положение вещей подводит нас к необходимости включить в состав Корпуса не только печатные рекламные тексты, что уже осуществлено, но и мультимедийную рекламу[7]. Укажем для примера, что уже давно рухнула пирамида МММ, но фраза *у МММ нет проблем* (и порож-

[5] «О стихах я не говорю — половина должны войти в пословицу» (А. Пушкин).

[6] Самое поверхностное знакомство с зарубежной литературой — как массовой, так и претендующей на элитарность, — показывает, что уровень цитирования (по крайней мере, голливудского кинематографа) достаточно высок, причем не только в Америке (можно упомянуть хотя бы романы Мураками). А если вспомнить, что, например, слово «спам» и название языка программирования Python восходят к гениальному английскому телевизионному проекту «Воздушный цирк Монти Пайтона», то можно высказать предположение, что уровень «мультимедийного» цитирования достаточно высок не только в СССР и России.

[7] Естественно, говорить можно только о рекламе российских товаров и услуг, а также российской по происхождению политико-экономической и социальной рекламы, — чтобы в Корпус по мере возможности не попадали такие переводные «уродцы», как батарейки, которые *«работают до десяти раз дольше»*.

даемая ею модель) все еще в ходу. Уж сколько лет нет в природе банка «Империал», и бог весть, существует ли то агентство, которое проводило его рекламную кампанию, но фраза *Случилось страшное!,* великолепно сыгранная С. Фурманом, — живее всех живых.

Напомним, что о массированном присутствии мультимедийной составляющей в русском языке можно говорить, по-видимому, начиная с 1930-х гг. («Цирк», «Волга-Волга» и нек. др.). Предполагать, однако, что с начала 19 в. и до 1930-х гг. (т. е. в течение того периода, который предполагает охватить Национальный корпус русского языка) «свято место» пустовало, вероятно, неверно. Скорее всего — выскажем это в качестве предположения, — эту нишу занимал театр. К сожалению, записей театральных и цирковых представлений практически не осталось (насколько нам известно), но, по-видимому, в той или иной степени доступны либретто опер и оперетт. Именно поэтому в качестве своего рода мультимедийной ретроспективы мы предлагаем в некотором количестве включить их в состав Корпуса[8].

Таким образом, необходимость создания мультимедийной составляющей Национального корпуса русского языка, как кажется, представляется достаточно обоснованной. Встает вопрос, как организовать работу по отбору и обработке соответствующего материала.

Технология подготовки мультимедийного подкорпуса

1. Отбор материала

Наименьшие трудности представляет, по-видимому, отбор кино- и мультфильмов для включения в Корпус. Словари цитат и крылатых слов помогут отобрать наиболее цитируемые картины, а энциклопедии и справочники по русскому кинематографу, в т. ч. и онлайновые, позволят отобрать кино слабо цитируемое, но, безус-

[8] Здесь, конечно, возникнет некоторое количество вопросов, за разрешением которых, возможно, придется обращаться к специалистам (например, включать ли в Корпус как либретто Е. И. Розена «Жизнь за царя», так и либретто С. М. Городецкого (отредактированное М. А. Булгаковым) «Иван Сусанин», или только одно из них, и считать ли эти два либретто одним произведением или разными). Множество аналогичных трудностей, впрочем, возникало в самом начале работы над Национальным корпусом русского языка, и все они были так или иначе преодолены.

ловно, входящее в базовый фонд русской (и советской) культуры[9]. Современный рынок видео и DVD и опыт работы конструкторов Корпуса над стенограммами устной речи позволит решить задачу накопления этого материала достаточно легко[10].

Бóльшие трудности вызовет отбор театрального материала (телеспектакли, например, «Ханума» товстоноговского БДТ, «Необыкновенный концерт» С. Образцова), а также образцов эстрадного искусства (миниатюры в исполнении Райкина, Жванецкого, Хазанова, Карцева и Ильченко), но поскольку число такого рода текстов достаточно невелико по сравнению с фильмофондом, то эта проблема, по-видимому, преодолимая.

Задача собирания современной теле- и радиорекламы легко решается сплошным мониторингом телеканалов и радиостанций (включая тех, которые вещают на УКВ). Желательно при этом, конечно, получить соответствующие материалы, хотя бы в небольшом объеме, не только центральных, но и региональных телеканалов и радиостанций. Здесь, однако, возникает следующая трудность: мультимедийная реклама имеет довольно ограниченный срок жизни, в связи с этим невозможен легкий доступ к рекламным материалам, относящимся к предшествующим периодам истории русского языка, прежде всего, — к 1990-м годам, когда, собственно, и началось победоносное рекламное шествие по россий-

[9] В частности, освоенность на уровне цитат и устойчивых выражений фильмов А. Тарковского, по-видимому, близка к нулю, что, однако, не является аргументом для их невключения в состав Корпуса. Интересно, кстати, будет проверить: слово «сталкер», достаточно активно используемое носителями современного русского языка, восходит скорее к роману братьев Стругацких «Пикник на обочине» или к «Сталкеру» Тарковского?

[10] Отметим, что при включении в мультимедийный подкорпус фильмов следует отказаться от соблазнительно легкого решения проблемы — от включения сценариев, а не результирующих текстов. Известно, что съемочная площадка меняет изначальный сценарий, иногда радикально (например, очень любил импровизацию на съемке Л. Гайдай, и довольно значительное число любимых народом выражений в изначальном сценарии просто отсутствует). Кроме того, сравнение итоговых текстов даже тех фильмов, авторы которых старались бережно относиться к оригиналу (например, «Обыкновенное чудо» М. Захарова), с их первоисточником («Обыкновенным чудом» Е. Шварца) демонстрирует существенную разницу между ними (в частности, Волшебник Шварца и Волшебник Янковского-Захарова расходятся достаточно далеко, не говоря уже о том, что «Бабочка крылышками бяк-бяк-бяк-бяк» у Захарова-Кима-Миронова превратилось в песню, до сих пор имеющую высокую популярность, а у Шварца это — достаточно проходная фраза Администратора).

ским просторам (кажется, вся советская реклама исчерпывалась классическими «Храните деньги в сберегательной кассе», — которое в качестве крылатого слова вошло уже в «Иван Васильевич меняет профессию» Л. Гайдая, — и «Летайте самолетами “Аэрофлота”»). Здесь, по-видимому, будет необходимо обращение к телевизионным и радиоархивам, а также к архивам рекламных агентств, значительное время работающих на российском рынке (если таковые архивы существуют).

Что касается либретто, то здесь возникает две проблемы. Во-первых, проблема поиска самих текстов. По-видимому, в качестве одной из многих задач этот вопрос следует ставить перед создателями подкорпуса языка XIX века. Можно только высказать надежду, что обращение к театроведам позволит хотя бы частично решить проблему накопления данных. Второй вопрос, скорее, теоретический. Естественно, музыкальный театр обращался не только (и, может быть даже не столько) к исконно русским произведениям, а следовательно, возникает вопрос о включении в Корпус либретто опер и оперетт зарубежных авторов. Напомним, что в общем Корпусе эта проблема решена однозначно: в его состав включаются тексты, которые исконно написаны на русском языке (или существуют в авторизованном переводе на русский, например, «Лолита» В. Набокова)[11]. Казалось бы, этот же принцип должен соблюдаться и для либретто. Обратим, однако, внимание на то, что перевод либретто — это совсем не то же самое, что перевод прозаического текста. Переводчик либретто должен не только следить за тем, чтобы его перевод был полноценным русским текстом, а не близким к оригиналу подстрочником, но и за тем, чтобы полученный в результате текст легко и правильно «ложился» на музыкальную основу произведения, а также содержал такой набор звуков и характеризовался такой системой пауз, чтобы исполнитель мог это спеть, а зритель понять. В результате этой работы, как представляется, русские тексты либретто настолько далеко уходят от немецкого — французского — итальянского оригинала, что их следует

[11] Кажется, исключение сделано только для Библии в синодальном переводе, а также для «Волшебника Изумрудного города» А. Волкова и «Буратино» А. Н. Толстого, которые «переперли нам Шекспира на язык родных осин» с таким успехом, что об исходных текстах, по-видимому, можно забыть, — обе сказки стали фактом русской культуры (уровень и качество переработки здесь таковы, что позволяют сопоставить два этих текста с «Пиром во время чумы» А. Пушкина, о котором также трудно говорить как о переводном произведении).

считать не переводами, а отдельными авторскими произведениями, написанными на русском языке. Впрочем, этот вопрос требует отдельного обсуждения.

2. Метаразметка текстов

1. Автор текста. Для *фильмов и телеспектаклей* авторами текста считаются режиссер и автор сценария (причем если фильм поставлен на литературной основе, то автор литературного произведения также должен указываться (например, в составе авторов фильма «Иван Васильевич меняет профессию» обязательно должен быть упомянут М. Булгаков, невзирая на то, что текст фильма достаточно далеко ушел от текста пьесы). Для *рекламных текстов* указывается коллективный автор, однако совершенно необходимо упоминать среди авторов текста яркие актерские работы (таковых немного), например, С. Фурмана в упомянутой выше рекламе банка «Империал». Для телеминиатюр должен быть упомянут не только автор текста (например, М. Жванецкий), но и исполнитель (например, Р. Карцев — аналогичная проблема возникнет и при внутритекстовой разметке фильмофонда, см. ниже). Как авторы переводных *либретто* должны быть указаны не только авторы исходного текста, но и авторы русского перевода (если таковые известны).

2.-3. Пол и возраст автора. Как видим, автор мультимедийного текста либо неизвестен, либо коллективен, либо далеко не единственный. Следовательно, данные пункты при метаразметке, скорее всего, останутся не заполненными.

4. Сфера функционирования. Здесь мультимедийные тексты не представляют собой единства. Для рекламы и объявлений необходимо будет ввести подраздел *«мультимедийная реклама»* в уже существующей общей сфере «реклама», для остальных же текстов придется вводить новую сферу, которую предварительно можно обозначить как *«художественные мультимедиа»*, причем определение «художественные» будет характеризовать этот вид мультимедиа так же, как определение «художественные» характеризует «художественную литературу», а «мультимедиа» дают пользователю Корпуса понять, что данный текст, помещенный в Корпус, не исчерпывает всех информационных возможностей данного произведения, включающего в себя не только текстовый, но и визуально-звуковой ряд.

5. Жанр/тип текста. *Художественные мультимедиа* (см. п. 4) имеют собственное подразделение на жанры. Некоторые из них уже определены (для фильмов, спектаклей, телеминиатюр — комедия, триллер, боевик, фантастика, комедийная миниатюра, комедийный спектакль, драма, сериал и др.), некоторые же определяются с большими затруднениями (с аналогичными проблемами создатели Корпуса сталкивались при определении жанров художественной литературы, в результате чего пришлось ввести специальное «нулевое», немаркированное подразделение «нежанровая проза»). Вероятно, сходное решение понадобится принять и для художественных мультимедиа, но об этом можно будет судить только по мере накопления материала. Что касается *мультимедийной рекламы*, то нам пока видятся здесь только два возможных типа текста — «рекламный ролик» и «объявление». Так ли это, покажет постепенное пополнение Корпуса.

Остальные разделы стандартной метаразметки, как представляется, не претерпят существенных изменений при метаразметке мультимедийных текстов.

3. Внутренняя разметка текстов

При внутренней разметке мультимедийных текстов возникает по крайней мере две проблемы (вероятно, их будет больше, но предсказать заранее, каковы они будут, не представляется возможным).

Во-первых, разметка автора реплики в фильме. Как уже было принято нами для внутренней разметки драматических произведений и устных текстов, автор реплики должен будет, по-видимому, отмечаться таким образом, чтобы пользователь Корпуса видел на экране, кому эта реплика принадлежит, но при этом сам текст этой ремарки должен будет «выводиться» на другой уровень текста — так, чтобы неизбежный повтор ремарки не влиял на статистику выдачи по пользовательским запросам. «Выведение» текста ремарки за пределы текста фильма, кроме того, указывает на то, что это, собственно, не текст самого фильма, а некоторая служебная информация, добавленная составителями Корпуса для удобства пользователя. Таким образом, при внутренней разметке, например, «Джентльменов удачи» Г. Данелии классическая реплика «Чуть что — сразу Косой» получает ремарку «**Косой:** Чуть что — сразу Косой». Однако сразу бросается в глаза недостаточность такой пода-

чи. Как пишет В. Елистратов, «именно Слово, звучащее из уст любимого актера и включающее в себя его специфическую интонацию, мимику и жест — вот главный герой советского кинематографа ... 30-80-х гг. XX века» [Елистратов 1999, с. 5]. Представляется, что это совершенно верно — во всяком случае, для огромного большинства фильмов, заслуживающих быть помещенными в Национальный корпус русского языка. Следовательно, более полной и правильной подачей материала будет указание имени актера, произносящего данную реплику: «**Косой/С. Крамаров:** Чуть что — сразу Косой»[12]. Возможно, со временем создателям корпуса удастся наладить сортировку и выдачу текстов по авторам реплик (такая задача давно уже стоит для драмы и для устной речи), и тогда, например, появится возможность получить все тексты, произнесенные в русском кинематографе, например, Е. Леоновым, что, как представляется, может иметь если не лингвистический, то уж точно — культурологический и исторический интерес[13].

Вторая проблема является, по сути, логическим продолжением первой, а именно: как далеко должны заходить создатели Корпуса по пути наполнения текста фильма служебным материалом, прежде всего, ремарками, обозначающими действия героев фильма и вообще событийный ряд. Представляется, что если не поставить на этом пути довольно жестких ограничений, то в результате мы придем к тому, что служебный текст, принадлежащий создателям корпуса, по объему будет превосходить собственный текст фильма

[12] Требуют отдельного обдумывания и обсуждения проблемы, связанные с минимальной социологической информацией в мультимедийном, прежде всего в кинематографическом подкорпусе (о социологической разметке см. с. 108 настоящего сборника). Очевидно, что авторов реплик достаточно легко разметить по признаку пола. Что касается возраста и профессии, то здесь могут возникнуть проблемы.

[13] Заметим, что аналогичная проблема — введение в Корпус информации о самом значимом (или единственном) исполнителе — неизбежно возникнет и при создании песенного подкорпуса. В частности, как известно, одна из самых известных песен советского репертуара, «Течет Волга», была исполнена впервые М. Бернесом, затем ее пел В. Трошин, однако самую широкую популярность она получила в исполнении Л. Зыкиной, что, как представляется, обязательно должно быть каким-то образом отмечено (например, посредством указания Л. Зыкиной среди авторов этой песни). С другой стороны, как известно, автором текста к знаменитой песне группы «Наутилус» (как и многих других) был поэт Илья Кормильцев, но исчерпать его фамилией авторский коллектив этой песни, не упомянув группу «Наутилус» и В. Бутусова, как представляется, будет верным лишь формально, но не содержательно.

(особенно это верно для фильмов со скупым текстовым рядом, например, для некоторых фрагментов фильмов Тарковского). По-видимому, в качестве исходной установки следует поставить задачу вообще не использовать вспомогательной информации, а ограничиваться только собственно текстом. Так, например, классический диалог: «**Трус/Г. Вицин:** Жить, как говорится, хорошо. **Балбес/Ю.Никулин:** А хорошо жить — еще лучше!» в таком виде и должен войти в состав Корпуса, без указаний типа: *«Пьют пиво на кавказском курорте»*. В случае, если включение ремарки представляется неизбежным (например, если без нее фраза двусмысленна или непонятна), то она должна размечаться так же, как ремарки в драматических произведениях и устных текстах, т. е. выводиться за пределы речевой ткани фильма.

На этом предварительное описание мультимедийной составляющей Национального корпуса русского языка, вероятно, можно закончить. В заключение отметим только, что количественные показатели этого подкорпуса обозначить на данном этапе довольно трудно, — и не только потому, что аналогов такого подкорпуса нет, но и потому, что сейчас совершенно непонятен средний объем текстового ряда самого обыкновенного фильма, длящегося 1,5-2 часа, и, соответственно, нет ни малейшей возможности даже очень приблизительно сказать, на какой объем выйдет самая крупная составляющая мультимедийного подкорпуса — кинематографическая.

II. Подкорпус названий

Обоснование проекта

Предположим, пользователь Корпуса поставил перед собой задачу исследовать систему газетных заголовков и ее развитие в один из исторических периодов, «покрываемых» материалами Корпуса (что такая задача вполне осмысленна, по-видимому, очевидно). На данный момент вся необходимая информация в том или ином объеме (в зависимости от выбранного периода) уже в Корпусе содержится. Для того, чтобы ее получить, пользователю необходимо воспользоваться зоной «Мой корпус», задать там искомый период и тип текстов. Тем самым поставленная задача вродк бы будет выполнена.

Обратим, однако, внимание на то, что полученный таким образом материал можно использовать очень ограниченно, прежде

всего потому, что современными средствами в подкорпусе названий, собранных таким образом, невозможно осуществлять поиск — ни лексический, ни морфологический.

Этот довольно простой пример порождает достаточно серьезные следствия, которые в конечном итоге можно сформулировать как необходимость сконструировать в Корпусе то, что условно можно назвать **подкорпусом названий**.

Под ***названиями*** в дальнейшем мы будет понимать следующие группы лингвистических объектов: 1) по отношению к артефактам (включая тексты) — то, что в русской орфографической традиции обычно пишется в кавычках, 2) по отношению к природным объектам[14] — то, что обычно пишется с прописной буквы[15].

1) ***Названия артефактов*** в первом приближении, по-видимому, включают в себя следующие группы:

1.1) *Заголовки*, т. е. названия текстов (включая названия мультимедийных текстов, прежде всего фильмов, см. часть I настоящей статьи)

1.2) *Ярлыки* — а) названия учреждений (например, магазинов и фирм), б) названия объектов культуры (например, картин, скульптур, а также тексты на плакатах).

2) ***Названия природных объектов*** включают в себя

2.1) *антропонимы* (имена, фамилии, прозвища)

2.2) *топонимы*.

Формирование самостоятельного подкорпуса названий позволит на этом материале решать как стандартные задачи, которые обычно ставятся на материале Корпуса (использование того или иного слова или формы слова в названии, морфология и синтаксис, а также словообразовательные характеристики названий), и, кроме того, исследовать историко-хронологические и культурологические проблемы функционирования разных классов названий в разные периоды развития русского языка.

---

[14] В число природных объектов, таким образом, согласно правилам русской орфографии, входят и артефакты «большого объема», например, города.

[15] Это техническое определение порождает не такой уж большое количество неоднозначностей и ошибок, а безусловно более предпочтительное онтологическое определение понятия «название» увело бы нас слишком далеко от задач настоящей статьи, и даже если бы его удалось найти, то оно было бы полно столь глубокого философского смысла, что использовать его при решении конкретных вопросов, по-видимому, было бы просто невозможно.

ТЕХНОЛОГИЯ ПОДГОТОВКИ ПОДКОРПУСА НАЗВАНИЙ

1. Отбор материала

Как уже было сказано выше, существенная часть предлагаемого подкорпуса названий уже содержится в Корпусе. Это, прежде всего, названия текстов, уже вошедших в Корпус. Кроме того, в Корпусе в составе текстов уже содержатся антропонимы и топонимы, которые, в случае, если они отмечаются как таковые при морфологической разметке текстов, могут быть вычленены и по определенным правилам скопированы в подкорпус названий. Задача для этой группы названий, таким образом, ставится следующим образом: 1) названия текстов, включенных в Корпус при метаразметке этих текстов должны автоматически по некоторому алгоритму получать свою собственную метаразметку (как отдельные тексты) и вместе с этой метаразметкой копироваться в подкорпус названий; 2) антропонимы и топонимы должны при морфологической разметке текстов отмечаться как таковые, что позволит выгрузить их в подкорпус названий и оттуда получить по соответствующему запросу исследователя.

Однако часть названий не входит в корпус по определению (например, ярлыки типа «Родина-мать зовет!», «Всё в прошлом», какой-нибудь «Мир окон», «Елки-палки» или «Рога и копыта»), а что касается антропонимов и топонимов, то они фиксируются текстами, вошедшими в корпус, весьма опосредовано, не прямо. Следовательно, включение таких ярлыков, антропонимов и топонимов в подкорпус названий должно происходить и из внетекстовых источников, например, из телефонных справочников, записей актов гражданского состояния, приходских и писцовых книг, некрополей, каталогов, энциклопедий, словарей и любых иных аналогичных списочных и справочных источников, относящихся к тому или иному периоду в развитии русского языка.

2. Метаразметка текстов

Для того, чтобы понять, какой метаинформацией должны быть снабжены элементы[16], входящие в подкорпус названий, следует хотя бы в первом приближении сформулировать те типы запросов, которые могут быть обращены пользователем к такому подкорпусу.

[16] Текстами эти элементы можно назвать со слишком большой степенью условности.

Для этого выделим следующие четыре группы:

1) заголовки
2) имена (антропонимы/топонимы) текстового происхождения
3) имена нетекстового происхождения
4) ярлыки.

Заголовки — это тексты второго порядка: метатекст по отношению к основному тексту и одновременно просто текст как таковой, хотя и небольшого объема. Из этого следует, что практически все основные позиции метаразметки, характерные для собственно текстов, следует сохранить и для заголовков: автор, пол автора, возраст автора, дата создания, объем в словах, сфера функционирования, жанр/тип текста, тематика, хронотоп, стиль и т. д. Единственный традиционный пункт метаразметки, который заведомо будет отсутствовать при метаразметке заголовков, — это «заглавие», поскольку очевидным образом заголовки его не имеют. Этот пункт метаразметки должен быть заменен на «стратификационный», например, *название* (или *заголовок*). Таким образом, к заголовкам могут быть применены практически все те типы сортировок, которые традиционны для обычных текстов: можно отобрать все заголовки одного автора, все заголовки авторов мужского/женского пола, автора того или иного возраста, заголовки, характерные для той или иной сферы функционирования языка (для художественной, научной, публицистической и проч. литературы), для того или иного жанра/типа текстов, для текстов определенной тематики или определенного хронотопа, стиля, времени создания, места фиксации и проч. Кроме того, выбрав заголовки в качестве «Моего корпуса», на этом материале можно осуществлять все стандартные типы морфологического поиска (по конкретному слову/словам[17], по определенной морфологической форме[18], по той или иной конструкции[19]).

---

[17] Например, узнать, какие произведения, кроме романа Н. Чернышевского и труда В. И. Ленина, имели название «Что делать?».

[18] Например, узнать, насколько традиционно для русской культуры включение в название инфинитивов и в каких типичных конструкциях (в частности, любопытно было бы выяснить, характерны ли для множества русских заголовков конструкции типа «Убить пересмешника» или «Убить Билла»).

[19] Например, попытаться определить, прервалась или продолжается традиция присвоения заголовков конструкции «X и Y» — «Отцы и дети», «Преступление и наказание», «Война и мир», «Баргамот и Гараська», «Чук и Гек», «Мастер и Мар-

Отдельно следует отметить, что для заголовков необходимо предусмотреть функцию расширения контекста, которая сейчас вполне успешно работает и в стандартном Корпусе. Как зону расширения контекста здесь следует, по-видимому, рассматривать первую фразу текста. Таким образом, по желанию пользователя, Корпус сможет предоставить информацию о том, определяет ли данный тип заголовка то или иное начало текста, как связаны между собой два этих метатекстовых события — заголовок и первая фраза, — и т. д.

Замечание. Проблема отражения в корпусе элементов метатекстовой организации текста, в принципе, требует обсуждения и, по-видимому, должна быть в будущем так или иначе решена. Метатекстовая организация текста в минимальном варианте предполагает вычленение 1) названия текста, 2) первой фразы, 3) последней фразы, 4) метатекстовых ремарок внутри текста. На данном этапе развития Корпуса размечены только ремарки внутри драматических произведений — таким образом, чтобы они из текста при выдаче его на экран не исчезали, но одновременно не влияли бы на статистические параметры поиска. Пока, впрочем, нельзя задать опцию «выдать все драматические ремарки». Проблема первой и последней фразы текста пока даже не ставилась (очевидно, она возникнет при подготовке для размещения в Корпусе стихотворных материалов, хотя бы потому, что огромное количество поэтических текстов пользователь будет пытаться найти или опознать по первой строчке, а не по официальному названию — многие ли помнят, что песня «Славное море — священный Байкал» имеет официальное название «Думы беглеца на Байкале»? Кроме того, как показывают проведенные исследования, прагматическая характеристика зачина и концовки поэтического текста имеет достаточно существенное значения для типологии поэтической композиции [см. Гришина 1989]). Между тем, представить себе лингвистические, литературоведческие или культурологические исследования, основанные на работе с метатекстовой разметкой, не только легко, но и достаточно заманчиво. Пока, к сожалению, Корпус не предоставляет такой возможности, хотя эта задача, будучи поставленной, может быть решена, и в достаточно обозримые сроки.

Отдельного обсуждения требует проблема включения в Корпус переводных названий. Как уже упоминалось выше, общий принцип Корпуса на данный момент — включать только русские непе-

---

гарита», «Живые и мертвые», «Жизнь и судьба» и т. д., — и каковы смысловые отношения элементов X и Y в заголовках такого типа.

реводные тексты. Отметим, однако, что название текста, которое вполне можно рассматривать как очень маленький самостоятельный текст, очень часто именно в таком виде и остаются в языке. Например, название фильма 1966 г. «Никто не хотел умирать» — единственный текстовый элемент этого фильма, который прочно сохранился в русском языке[20]. Аналогичным образом в языке функционируют и названия зарубежных фильмов. Например, «Плата за страх» (фильм А. Ж. Клузо 1953 г. «Le salaire de la peur»), «Анатомия любви» (польский фильм Р. Залуского 1974 г.), название французского балета Л. Герольда «Тщетная предосторожность» (премьера в России 1885 г.) и мн. др. Классический случай — «They Like It Hot», фильм, название которого прочно вошло в русский язык сразу в двух вариантах, российском («В джазе только девушки») и американском («Некоторые предпочитают погорячее»). Как представляется, обеднять Корпус отсутствием такого рода заголовков, мотивируя это тем, что они (как и озаглавленные ими тексты) являются переводными, было бы ненужным формализмом. Кажется, разумнее снабжать такого рода заголовки иноязычной параллелью (например, Нет мира под оливами» | «Non s'e pace fra gli ultivi»), в графе «Дата создания» давать год создания исконного текста, а в графе, аналогичной «Году издания» для книг — год его появления в русском языковом космосе (для фильмов это год выхода на советские (российские) экраны).

Для ***имен текстового происхождения***, т. е. для антропонимов и топонимов, встретившихся в текстах, уже вошедших в состав Корпуса, разумно предусмотреть следующие элементы метаразметки: автор, название, дата создания, сфера функционирования, жанр/тип, тематика/хронотоп, стиль *того текста, в котором данное имя зафиксировано*. Кроме того, для онимов этого типа следует предусмотреть две статистические характеристики: 1) частота встречаемости в данном конкретном тексте и 2) количество текстов, в которых данный оним зафиксирован. Кроме того, оним должен иметь следующие характеристики: 1) для антропонимов — стратифицирующая характеристика (имя — прозвище — фамилия), пол чело-

[20] По данным словаря С. Г. Шулежковой, эта фраза до сих пор не только широко используется в русских текстах сама по себе, но и служит моделью для образования тех или иных выражений: *Никто не хотел отвечать, Никто не хотел рисковать, Никто не хотел отступать, Никто не хотел заполнять* (*декларацию о доходах*) и т. п. (см. [Шулежкова 2003, с. 219-220].

века, к которому относится данное имя (если это возможно определить), 2) для топонимов — стратифицирующая характеристика (название города, реки и под.), если это поддается определению.

Такая метаразметка позволит узнать, например, действительно ли А. Пушкин «впервые» именем Татьяна «страницы своего романа ... самовольно освятил», т. е. действительно ли имя Татьяна до Пушкина не встречалось в литературных произведениях? И если это так, то встречалось ли это имя в текстах других сфер функционирования, типов, тематики и проч.? Какой частотой характеризовалось использование этого имени на страницах художественных произведений по мере развития русской литературы и в каких отношениях эта частота находилась, например, с частотой употребления этого имени в периодике? С какого времени появились в художественных произведениях герои с украинскими фамилиями на *енко? Какое наиболее любимое мужское имя русского романа? И так далее. Конечно, такого рода исследования можно проводить и сейчас, но при «затерянности» имен в просторах большого Корпуса получение необходимой информации такого рода требует от исследователя большой ручной работы.

Имена нетекстового происхождения должны иметь минимум характеристик: стратифицирующие (см. выше), частотные (т. е. частота фиксирования на тот или иной период) и год фиксации в том или ином списке или справочнике. Это позволит не только получать корпус зафиксированных русских имен и фамилий, названий географических объектов, но и изучать их динамику с точки зрения хронологии.

Для ***ярлыков*** следует предусмотреть стратифицирующую характеристику (например, для «Не ждали» — «картина», «Мир стульев» — «мебельный магазин» и под.). Уникальные объекты (в основном это объекты культуры) должны снабжаться сведениями об авторе и о дате создания, неуникальные — датой фиксации в том или ином справочнике и статистической характеристикой, т. е. количеством фиксаций в определенный период, например, в определенный год. Тем самым можно будет получить сведения о том, когда и в каком количестве появились в России после 1991 г. первые «Миры» («Мир окон», «Мир стульев», «Мир дверей» и мн. под.), каковы стандартные названия ресторанов в России и Советском Союзе разных периодов, какой временной зазор разделяет написание И. Репиным картины «Бурлаки на Волге» и активное использова-

ние названия этой картины в качестве крылатого слова в русских текстах.

Что касается **морфологической разметки**, то ею должны быть снабжены все названия; при этом морфологическая информация об именах по определению будет редуцирована до рода и падежа, а ярлыки и заголовки будут иметь более полный набор морфологических признаков.

* * *

На этом можно закончить предварительное описание потенциальных подкорпусов. Они, безусловно, выбиваются из стандартных способов наполнения любого корпуса, но, как представляются, эта необычность недостаточна для того, чтобы априори вывести вышеописанные способы пополнения корпуса из рассмотрения. Основное, что позволяет рассматривать эти проекты всерьез, это не только филологическая и культурологическая осмысленность полученной с их помощью информации, но и техническая возможность воплощения этих проектов в жизнь.

Литература

Гришина 1989 — *Е. А. Гришина*. Структура поэтического текста с точки зрения теории речевых актов (на материале русского восьмистишия начала 20 в.). АКД, М., 1989.

Елистратов 1999 — *В. С. Елистратов*. Русский кинемалогос (о целях и структуре словаря). // Он же. Словарь крылатых слов (русский кинематограф). М., 1999.

Капанадзе 2005 — *Л. А. Капанадзе*. На границе письменного и устного текста: структура и тенденции развития электронных жанров // Она же. Голоса и смыслы. Избранные работы по русскому языку. М., 2005. С. 305-320.

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

Шулежкова 2003 — *С. Г. Шулежкова*. «Словарь крылатых выражений из области искусства». М., 2003.