Е. А. Гришина

УСТНАЯ РЕЧЬ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА*

Практика создания национальных корпусов свидетельствует, что в качестве необходимой составляющей корпуса, без которой информация о данном языке не может считаться полной, рассматривается не только письменная, но и устная форма речи. Так, около 10% объема (т. е. примерно 10 млн словоупотреблений) Британского национального корпуса (BNC) составляют записи устной речи [см. BNC]. Чешский национальный корпус (ЧНК) также содержит «устную» составляющую — т. н. Пражский разговорный корпус, — но она невелика (почти 800 тыс. словоупотреблений, 300 записанных на магнитофон разговоров) [см. Čermák]. Создатели обоих названных корпусов целенаправленно проводили работу по набору устного материала. Так, создатели BNC с помощью Британского бюро маркетинговых исследований (British Market Research Bureau) снабдили 124 добровольца диктофонами, на которые в течение определенного периода времени записывались все разговоры, которые вели респонденты и их собеседники. Состав добровольцев был сбалансирован по социологическим характеристикам (пол, возраст, район проживания), что позволило в результате получить социологически сбалансированный устный подкорпус. Создатели ЧНК действовали примерно так же, однако, указывая на финансовые и организационные трудности, связанные с таким способом получения материала[1], ограничились в результате весьма небольшим объемом устного подкорпуса.

Изначально Национальный корпус русского языка рассматривался прежде всего как корпус письменных русских текстов. Однако по мере его расширения появилась возможность обратиться и к

---

* Данная статья представляет собой сокращенный и слегка переработанный вариант статьи, опубликованной в журнале «Научно-техническая информация. Сер. 2: Информационные процессы и системы», № 3, 2005.

[1] «Получение такого корпуса — очень дорогая и долговременная процедура» [Čermák] — имеется в виду, — по сравнению с корпусом письменных текстов.

устной речи, в частности, встал вопрос о размещении на сайте Национального корпуса русского языка диалектных текстов[2], а также появившихся в распоряжении разработчиков корпуса материалов литературной устной речи. Вопрос о какой бы то ни было социологической сбалансированности устного подкорпуса НКРЯ в данный момент не ставится по экстралингвистическим причинам, вполне аналогичным тем трудностям, на которые указывают создатели ЧНК[3], при этом, однако, при конструировании устного подкорпуса мы стараемся ориентироваться, по крайней мере, на жанровую и тематическую сбалансированность. Решение этой задачи в значительной степени может быть обеспечено представительностью корпуса: создатели Национального корпуса в настоящий момент поставили себе задачу в конечном итоге выйти на объем устного корпуса в 10 млн словоупотреблений (т. е. на объем, вполне сопоставимый с объемом аналогичного подкорпуса BNC). Достаточный объем подкорпуса и относительно полный набор жанров, как представляется, позволят усреднить, если не компенсировать заведомое отсутствие социологической сбалансированности.

ФОРМА ПОДАЧИ МАТЕРИАЛА

Как мы уже писали более подробно раньше[4], из трех возможных форм расположения устной речи на сайте Национального корпуса (аудиозапись, транскрипция, стандартная орфография) создатели корпуса по причинам как технического, так и идеологического характера выбрали орфографический принцип подачи устной речи: текст записывается в традиционной русской орфографии (с минимализацией стандартных стянутых форм типа *щас*, *тыща*, *чек* (человек)

---

[2] О принципах подачи диалектных текстов см. статью А. Б. Летучего «Корпус диалектных текстов: задачи и проблемы» в настоящем сборнике.

[3] В настоящий момент над пополнением устного подкорпуса работают студенты-практиканты Московского государственного университета, Российского государственного гуманитарного университета и Государственного педагогического университета. Кроме того, большую помощь в пополнение устного подкорпуса, как планируется, окажут российские центры изучения устной речи, такие, как Саратов, Пермь, Сыктывкар. Интересные материалы уже получены от Петербурского государственного университета и от наших финских коллег из Хельсинкского университета.

[4] «Научно-техническая информация. Серия 2: Информационные процессы и системы», № 3, 2005.

и проч.). При этом знаки пунктуации внутри предложения вообще снимаются (заменяются слэшами), и в тексте остаются только знаки, функционально равные точке. При этом, конечно, следует все время помнить, что слэши в стенограммах, размеченных подобным образом, *не имеют никакой смысловой нагрузки* (т. е. не обязательно соответствуют, напр., паузам), а расставлены только для удобства чтения[5] (именно поэтому было принято решение для предложения, которое оканчивается точкой, не вводить вместо нее двойной слэш, — чтобы не создавать у пользователя иллюзию правильной фонетической записи). Этим слэши в корпусе отличаются от аналогичного способа нотации, широко принятой в разных фонетический традициях, где одинарной косой обозначается малая пауза, а двойной косой — большая пауза или ее эквиваленты [СРЯ, с. 46].

Потеря «звучащей» компоненты, конечно, лишает исследователя возможности работать с устным текстом на фонетическом уровне (включая суперсегментный), кроме того, отсутствие возможности обратиться к звуковому и/или транскрипционному первоисточнику оставляет неснятыми некоторое число двусмысленностей и неясностей, неизбежно возникающих при расшифровке аудиоматериалов. При этом, однако, стенограмма позволяет продуктивно работать со всеми остальными уровнями языка — морфологией, словообразованием, синтаксисом, семантикой, а также с риторическими аспектами устных выступлений[6].

Заметим, что принятое решение накладывает на создателей корпуса существенное ограничение в использовании в качестве исто-

---

[5] Впрочем, некоторое непрямое отношение к способу произнесения слэши имеют: а) в случае хорошей пунктуации исходной стенограммы — в той мере, в которой русская пунктуация в принципе способна отражать паузировку и эмфазу живой устной речи; б) в случае плохой пунктуации — в той мере, в какой плохо знающий пунктуацию дешифровщик самими своими ошибками отражает отсутствие/наличие пауз и эмфазы в той или иной точке текста.

[6] Специально подчеркнем, что принятое решение о включении в корпус прежде всего стенограмм полностью соотносится с решением, принятым в BNC. Напомним, — там устные «исходники» были получены целенаправленно в качестве аудиозаписей, а затем транскрибированы, и тем не менее в BNC «для устных текстов не предпринималось никаких попыток обеспечить отражение фонетических или просодических характеристик затранскрибированной речи: слова переданы в стандартной английской орфографии (исключение составляют т. н. вокальные паузы, которые специально зафиксированы в соответствующем контрольном списке, а также регионализмы и диалектизмы)» [Burnage, Dunlop].

чников устного подкорпуса НКРЯ тех или иных уже созданных устных корпусов. Так, если корпус существует только в виде аудиозаписей или только в виде транскрипции, то реальное использование его в рамках нашего проекта практически невозможно. Затруднено также использование «псевдофонетических» транскрипций, которые построены на русской орфографии, однако включают в себя обозначения значительного числа редукций, например, «ща», «тыща», «здрасти», «чё», «кшно» и пр. — вместо «сейчас», «тысяча», «здравствуйте», «чего», «конечно»[7]. Таким образом, легко и естественно включаются в НКРЯ, помимо стенограмм, транскрипты, полностью построенные на стандартной русской орфографии с включением тех или иных дополнительных обозначений (последние при помещение материалов в корпус либо просто снимаются, либо автоматически транслируются в стандартную корпусную разметку)[8].

Таким образом, образец устной речи — так, как он выдается на сайте Национального корпуса русского языка, — выглядит, например, таким образом:

I.

Интервьюер. А что у вас было? Какой скот был?

Респондент. А?

Интервьюер. Какой скот был? Сколько у вас было там / много было скота / гектаров?

Респондент. Ну / я скота не... / я помню только / когда бабушка была пока жива / вот я до 32-го года... Так вот знаю / что было две коровы / лошадь была и молодая лошадь / но ее не запрягали / вот. А потом я не знаю. Я по нянькам уходила. Я зиму и лето жила в няньках.

Интервьюер. В деревне?

Респондент. В деревне / в деревне. Вот в 31-ом году... / а в 26-ом году я жила вот на хуторе здесь. В 24-ом я жила вот у Сосны. Это мамина мать была. Они там жили три брата / бабушка и две невестки. У бабушки было три сына (два / женатые) / две невестки. И они меня в 24-ом году / вот сколько мне было? Шесть лет / седьмой. Они меня и взяли вот в няньки / вот где Сосны / они там жили.

Интервьюер. А вас детей было двое?

Респондент. А?

Интервьюер. Вас было детей двое?

Респондент. А мы были оставши / вот сестра старше меня на 4 года. А она тоже / все время она по нянькам жила / и я / по нянькам. Дома находились очень / очень...

---

[7] Заметим, что такого рода включения спорадически встречаются и в письменной речи, а именно, — в художественной литературе, но там они в общем и целом не влияют на общую литературность и легко нивелируются, — в отличие от устных текстов, где при такого рода транскрибировании количество редуцированных включений очень велико.

[8] Такова, например, транскрипция, предложенная А. А. Кибриком и В. И. Подлесской в рамках проекта создания устного корпуса русского языка [см. Кибрик, Подлесская].

/ только когда из нянек пустят там / в воскресенье прибежишь и опять в няньки. Вот так. В школу я ходила / в школу я ходила с заговен. Мы не знали / какого числа в школу / а вот заговены / когда уже все поля уберут / лен там / все. Уже в няньках делать нечего / хозяйки уже дома. Тогда я шла домой / ходила я в школу. Я ходила / считай три зимы / но я начинала / вот когда заговены / а заговены / это в октябре месяце. Вот из нянек приду / в школу ходила. Пасха. Пасха / не в числах. До Пасхи меня уже брали в няньки / то / что дом моют там да готовятся к празднику / меня брали в няньки / я уходила в няньки на все лето. Домой я приходила в воскресенье там / или в праздник какой прибегала. Мне была бабушка дорогая. Мне ни мать / ничего... У матери была другая семья / другая семья. Мы были чужие. Прибежишь к бабушке...

II.

Модератор: Ну хорошо. Мы об этом поговорим еще. Кто-то еще. Вы там просто были в Белоруссии. А у кого-то вот / какие еще / просто мнения есть? Что вы думаете о современной Белоруссии?

БОРИС: Ну / мне кажется / что это передовая республика была / ну / бывшая республика / нерядовая. Во-первых / у нас здесь в Воронеже минская вся продукция / хорошая / мы даже покупаем покупаем / хорошие холодильники / товары / опыт хороший / то есть там поставлено дело / мне кажется / нормально.

НАДЕЖДА: Даже с точки зрения славянских народов / даже с этой точки зрения / понимаете.

Модератор: Ну / а вот сама Белоруссия / какое ваше мнение / что это за государство такое?

НАДЕЖДА: Трудяги / замечательные люди. Трудяги / белорусы всю жизнь были трудягами великими. И / как люди / они хорошие. Вот. А то / что политики там меняются... ну / Лукашенко / он немного своеобразный человек / но тем не менее у нас вот / очевидец говорит / конечно / мы в Белоруссии не были / если там такой порядок в сельском хозяйстве / а сельское хозяйство / это житница. Это житница / если такой порядок в сельском хозяйстве / то / наверное / и в промышленности. И промышленность / хотя бы легкая промышленность / мы можем судить о легкой промышленности / правильно товарищ сказал.

МЕТАРАЗМЕТКА УСТНЫХ ТЕКСТОВ

Следующий блок вопросов связан с метаразметкой устных текстов, т. е. с приписыванием каждой стенограмме некоторого набора признаков, которые позволяют пользователю определенным образом сортировать устные тексты и формировать из них те или иные подкорпуса[9]. Описываться будут только те параметры метаразметки, где устная речь имеет отклонения или определенные нюансы по сравнению с письменной.

1. «Нулевая» зона. Как мы уже говорили, основной массив текстов в корпусе — это письменные прозаические тексты, написанные на литературном русском языке в широком его понимании, следовательно, при метаразметке немаркированным по умолчанию считаются именно а) письменные б) прозаические (т. е. недраматургические и не-поэтические) и в) литературные (не в смы-

[9] Об общих пинципах метаразметки см. статью С. О. Савчук в настоящем сборнике.

сле «беллетристические», а в смысле «относящиеся к русскому литературному языку») тексты. Как только хотя бы по одному из этих параметров используется маркированный признак, возникает необходимость его отметить. Таким образом, при метаразметке маркированных текстов должны быть введены следующие «нулевые» (т. е. имеющие «ноль» в качестве соответствия в основном массиве корпуса) графы метаразметки: устная речь, поэзия, драматургия, диалектная речь.

Комбинация этих признаков задаст тот или иной подкорпус (напр., комбинацией «устная речь» + «поэзия» + «диалектная речь» задается подкорпус поэтического диалектного фольклора, комбинация «поэзия» + «драматургия» задаст подкорпус стихотворных пьес (напр., «Горе от ума»), и т. д.)[10]. Набор комбинационных возможностей здесь достаточно велик, перечислять и моделировать их в данных заметках нет смысла, важно, однако, что все эти комбинации в той или иной степени осмысленны, а следовательно, при метаразметке таких маркированных типов текста в «нулевой» зоне следует предусматривать четыре столбца — 1) устная речь, 2) поэтическая речь, 3) драматическое произведение, 4) диалектная речь. В тех случаях, когда произведение по данному признаку не маркировано, соответствующую ячейку в соответствующем столбце следует заполнять так, как мы заполняем ее в случаях, напр., когда неизвестен (или неважен) автор произведения.

2. Автор текста. Автором устного текста считается автор монолога или участники диа- или полилога. Естественно, имя автора текста не указывается, когда автор неизвестен (так мы поступаем и в остальном массиве корпуса).

В случае, когда имя автора / имена авторов известны, принимается следующее ограничение — считается желательным, чтобы у устного текста *не было больше двух авторов*.

В случае монологического текста проблем не возникает.

[10] Теоретически рассуждая, все комбинации этих признаков осмысленны, но, конечно, далеко не все из них будут реализованы — даже в далекой перспективе: напр., комбинация «устная речь» + «драматургия» в принципе дает нам возможность получить стенограммы того или иного спектакля, но вряд ли такого рода тексты будут представлять первоочередной интерес для лингвиста и, следовательно, вряд ли в скором времени войдут в корпус.

В случае диалога и полилога может возникнуть проблема выбора автора, и здесь лучше не полагаться на филологическое чутье разметчика, а максимально возможным способом формализовать процедуру выбора. Для этого предлагается использовать несколько терминов:

1) ***равноправный диалог/полилог:*** когда для обоих участников априори предполагается попеременная свободная мена ролей в разговоре — спрашивающего/отвечающего, задающего направление беседы, и под.;

2) ***неравноправный диалог/полилог:*** когда один из участников беседы задает тему и провоцирует остальных на ее обсуждение, а также задает другим участникам вопросы (т. е. является модератором или интервьюером, см. ниже);

3) ***модератор:*** участник неравноправного полилога, задающий вопросы, провоцирующий на высказывания и в целом направляющий беседу;

4) ***интервьюер:*** модератор в диалоге;

5) ***главные участники:*** участники неравноправного диалога/полилога, не являющиеся модератором (интервьюером);

6) ***сопоставимость:*** ситуация, когда модератор (интервьюер) сопоставим по общественной значимости с главным участником/главными участниками[11].

Специально следует отметить, что понятия **равноправности/неравноправности** не имеют никаких оценочных характеристик, а также не являются социолингвистическими, в отличие, например, от понятий **симметричности** собеседников и их **паритетности**, — понятий, которые используются рядом исследователей при социолингвистической характеристике устных жанров: «Симметричность/несимметричность отражает тождество-нетождество социальных признаков партнеров по коммуникации вне оценочной их характеристики по шкале «выше-ниже». Соотношение ролей по признаку паритетность/непаритетность строится на принципе равенства-неравенства (оценочная шкала «выше-ниже») в соответст-

[11] Специально подчеркнем, что все предлагаемые термины не являются строго научными, но достаточно ясны интуитивно, чтобы ими можно было пользоваться практически.

вии со стереотипными социальными нормами»[12]. Степень равноправности описывает *стратегическую роль* говорящего в диа- и полилоге. Что касается параметра **сопоставимость**, то он как раз соотносится с понятиями «симметричности» и «паритетности»: сопоставимые участники симметричны, несопоставимые же — характеризуются отсутствием паритетности. Впрочем, очевидно, что равноправие и сопоставимость, тем не менее, связаны: интервьюер (модератор) обычно несопоставим со своими собеседниками, а главные участники — сопоставимы, даже если и неравноправны[13]. Различение этих параметров полезно в тех случаях, когда модератор по своему социальному статусу вполне сопоставим с главными участниками или даже превосходит их (например, при беседе социолога с участниками фокус-группы и заведомо — в беседе диалектолога с информантами), но его участие в беседе по определению неравноправно.

Разные возможности комбинации предложенных выше параметров дают нам следующий алгоритм:

1) монолог = один автор;
2) равноправный диалог = два автора;
3) неравноправный диалог с несопоставимым интервьюером = один автор (главный участник диалога[14]);
4) неравноправный диалог с сопоставимым интервьюером = два автора (интервьюер и главный участник[15]);

[12] [Китайгородская, Розанова 2003, с. 106-107].

[13] Например, если в беседе В. Познера и В. Соловьева последний будет отмалчиваться, а первый — непрерывно говорить, равноправность В. Соловьева пострадает, хотя его сопоставимость останется неприкосновенной. С другой стороны, если в дружеской застольной беседе кто-то из участников начнет вести себя как модератор, т. е. настойчиво выстраивать стратегию беседы, то через некоторое время у его собеседников вполне возможно возникновение следующей, например, реакции: «Это что, допрос?», т. е. ситуативно сопоставимые собеседники должны камуфлировать свою неравноправность, даже если она есть.

[14] Напр., интервью, которое Иван Иванов берет у Владимира Познера, — автор В. Познер.

[15] Напр., интервью, которое Владимир Познер берет у Владимира Путина, — авторы В. Познер и В.Путин.

5) неравноправный полилог с участием трех человек (в т. ч. с участием несопоставимого модератора) = два автора (главные участники[16]);

6) неравноправный полилог с участием трех человек (в т. ч. с участием сопоставимого модератора)[17], равноправный полилог с любым числом участников, неравноправный полилог с числом участников больше трех — автор не указывается, поскольку авторов во всех таких случаях получается больше двух.

Все вышеизложенное звучит достаточно сложно, но на самом деле, как ясно из примеров в сносках, всего лишь формализует естественную трактовку авторства при анализе ситуаций разных типов.

3. Название текста. У устных текстов в нормальном случае нет названия. Его следует искусственно генерировать из полей: 1) автор текста, 2) тип текста (для художественных устных текстов — жанр текста[18]) — см. пункт 7, 3) тема текста (если она одна и вычленяется достаточно легко), 4) дата записи текста, 5) место записи текста, — напр., «Беседа Г. Ярцева и В. Колоскова о российском футболе, Колыма, 13.10.2004». Если значение одного из полей неизвестно, то оно, соответственно, не заполняется.

Отдельно следует отметить ситуацию, когда у текста по тем или иным причинам не отмечается автор (см. выше). В таких случаях разной стратегии следует придерживаться при описании устной публичной и устной непубличной речи (см. п. 6). Для устной публичной речи (при наличии модератора/интервьюера) следует, по-видимому, в названии текста в ряде случаев указывать *социальный статус* модератора/интервьюера — например, дискуссия на фокус-группе, происходившая в Москве 12.12.2002 по изложенным выше правилам должна была бы называться: «Беседа на общественно-политические темы, Москва, 12.12.2002», но, как представляется, более верным было бы назвать ее «Беседа с социологом на общественно-политические темы, Москва, 2002». Аналогичным

[16] Иван Иванов беседует с Светланой Сорокиной и Владимиром Познером, — авторы С. Сорокина и В. Познер.

[17] Светлана Сорокина ведет беседу с Савиком Шустером и Владимиром Познером.

[18] Список возможностей здесь такой же — или, по крайней мере, не шире, — как и у письменных текстов.

образом, для ситуативно-обусловленных микродиалогов и микромонологов разумно указывать в названии локус текста, например, «Разговор на рынке, Норильск, 1984» (см. п. 7).

4. Место записи текста. В корпусе письменных текстов такого параметра нет вообще — местоположение указывается только для издательства, опубликовавшего текст. Очевидно, что для значительного числа устных нелитературных текстов — и для ста процентов диалектных текстов — эта характеристика является чрезвычайно важной.

5. Время записи текста. Для письменных текстов аналогом этого является год издания произведения.

6. Сфера функционирования. В связи внесением в корпус устных текстов некоторые дополнения придется внести и в эту графу метаразметки. В основном корпусе под сферами функционирования текстов подразумеваются те сферы речевой деятельности, которые «определяют выбор правил речевого поведения, форм речевого взаимодействия автора и адресата, выбор речевых жанров, принципов построения и языкового оформления высказываний»[19]. Можно было бы воспользоваться для описания устной речи теми сферами функционирования, которые выделены для письменных текстов, но мы, имея в виду тот факт, что на степень самоконтроля говорящего при произнесении устного текста, на степень подготовленности устного текста прежде всего влияет степень «публичности» высказывания, рассчитанности на постороннюю аудиторию, сочли более адекватным для материала ввести здесь иное подразделение. В качестве сфер функционирования устных текстов предлагается выделить следующие: 1) ***устная публичная речь***, которая либо априори предполагает наличие слушателей и запись на носителях, либо просто допускает их как естественные, и 2) ***устная непубличная речь,*** не предполагающая наличия посторонних слушателей и фиксации на носителях[20].

---

[19] См. статью С. О. Савчук в настоящем сборнике, с. 71.

[20] Обратим внимание, что в BNC устные тексты разбиваются на две большие группы, в принципе, соотносимые с предлагаемым нами разбиением: 1) **контекстно-обусловленная устная речь** (context-governed part of the Spoken Corpus) — соответствует устной публичной речи и включает в себя следующие подразделения: а) обучающие и информационные речевые события (лекции, выпуски новостей, обсуждения в классе и под.), б) деловая речь (демонстрации товаров, консультации, интервью при приеме на работу и под.); в) институциональные и пуб-

7. Тип текста. Здесь на данном этапе развития корпуса предлагаются следующие подразделения.

Для устной публичной речи:

Монологические тексты:

лекция
речь
комментарий (напр., футбольный, новостной и пр.)
рассказ
презентация
проповедь
сообщение...[21]

Диа-/полилогические тексты:

беседа
комментарий (ведущийся несколькими персонажами)
творческая встреча (с читателями, со слушателями и проч.)
интервью
дискуссия (например, политическая или научная)
конференция
круглый стол
парламентские слушания
пресс-конференция
семинар
совещание

Для устной непубличной речи:

Монологические тексты:

пересказ (сна, фильма, книги...)
рассказ

Диа-/полилогические тексты:

домашний разговор
микродиалоги (в больнице, на почте, в аптеке, в библиотеке, в доме отдыха, в кассе, в магазине, в лифте, в

личные речевые события (проповеди, политические речи, парламентские дискуссии и под.); г) беседы на досуге (спортивный комментарий, клубные беседы, звонки на радио) и 2) **демографические тексты** (the demographic part of the spoken corpus) — соответствуют устной непубличной речи и включают записи ситуативно-обусловленных текстов, которые уже упоминались выше [см. BNC].

[21] Многоточие обозначает, что этот список пока открыт.

транспорте, дома, информационные, на кухне, на пороге, на работе...)
праздный разговор
разговор-воспоминание
разговор (в магазине, в парикмахерской, при встрече...)
спор
телефонный разговор[22].

Как видим, типы устных текстов не складываются в строгую классификацию, что, по-видимому, естественно. При формировании списка типов устной речи (на данный момент список открыт, но по мере пополнения устного подкорпуса мы будем стремиться к его закрытию) составителями Корпуса учитывались следующие факторы:

1) ***Устная публичная речь*** имеет развитую систему жанровых самоназваний (см. выше). Задача составителей корпуса в этой зоне разметки заключалась в том, чтобы преодолеть предлагаемое языком (и жизнью) богатство самоназваний и максимально укрупнить классификацию, огрубив жанровую систему. Так, например, в отличие от BNC, мы в данный момент отказались отдельно выделять такой жанр, как *ток-шоу*, поскольку этот тип публичной речи без остатка распределяется между такими базовыми публичными жанрами, как *беседа* и *дискуссия*. Совокупность жанров в ***устной непубличной речи*** устроена существенно иным способом: непубличная речь чаще всего не содержит жанровых самоназваний, представляя собой нерасчлененную стихию *разговора*. Здесь задача составителя корпуса состоит в том, чтобы неким единообразным и понятным для пользователя способом назвать разные «участки» этого разговора, помещаемые в корпус как отдельные единицы описания.
2) Для выделения отдельного типа устного текста использовалась некоторая совокупность ***доминант***, свойственных прежде всего данному типу и ослабленных в других типах. Так, в публичных монологических текстах можно отметить доминанты *обучения* (лекция, проповедь), *описания текущего состояния* (комментарий, презентация), *описания прошлого* (рассказ), *воздействия на аудиторию* (речь), *информирования* (со-

[22] См. также [Китайгородская, Розанова 1999, с. 225 и далее].

общение). Это не значит, что данные параметры не существенны для других типов речи, — просто в них они отступают (в классическом варианте) на второй план[23]. Для устной непубличной речи существенна, прежде всего, доминанта *стандартизованности*: до какой степени данный текст имеет стандартную структуру, которой говорящий обязан следовать, если он хочет быть правильно понятым адресатом. Именно эта доминанта является решающей для классификации непубличных устных жанров, а не противопоставление монолога диалогу, поскольку чистый монолог в реальной непубличной речи практически не встречается (если не считать монологов с самим собой и разговоров с вещами и животными, — да и те в большинстве своем построены как диалоги). Именно по параметру стандартизованности различаются, с одной стороны, микродиалоги и телефонные разговоры, а с другой, — все остальные типы устной непубличной речи. Для стандартизованных диа- и полилогов определяющей является доминанта локуса, т. е. где именно, в какой жизненной ситуации происходит данное стандартизованное общение[24]. Как только доминанта стандартизации ослабляется, вступают в игру другие доминанты: ориентирован ли данный текст на опровержение собеседника (спор), на общение per se (праздный разговор, домашний разговор), ориентирован ли он на прошлое (разговор-воспоминание, рассказ) и т. д.

8. Аудитория. Характеристики аудитории (ее возраст, уровень развития, объем) по-видимому, должны сохраняться только для публичной устной речи (см. пункт 6), которая априори предполагает наличие слушателей. Для непубличной речи дается характеристика аудитории только с точки зрения ее размера — *личная* или *групповая*.

Подготовка устных текстов

Подготовка устных текстов для размещения в корпусе включает в себя все те действия, которые предусмотрены для подготовки

[23] Отметим при этом, что в реальности базовая доминанта может противоречить заявленному жанру устного публичного текста, и тогда у составителя Корпуса возникает дилемма — следовать ли за самоидентификацией текста, или ориентироваться на его реальную доминанту.

[24] См. [Китайгородская, Розанова 1999, с. 264 и далее].

любого письменного текста (в частности, — проверку орфографии, разметку иноязычных, стихотворных и проч. включений, и т. д., и т. п.). В дополнение к этому стандартному набору действий в устном тексте снимаются знаки препинания — в соответствии с изложенными выше принципами.

Разметка метатекста. Отметим, однако, что устный текст, точнее, стенограмма устного диалога/полилога, содержит некоторые элементы (аналоги которых включает в себя из всех типов речи еще только драматургия), а именно, ***метатекстовые вставки***, т. е. элементы текста, заведомо не принадлежащие его авторам[25]. Не отличать текст от метатекста методологически неверно: в частности, в случае такого неразличения существенно меняются частотные характеристики текста. Так, например, если в интервью, имевших место на радиостанции «Эхо Москвы», при размещении в Корпусе не разметить соответствующим образом авторов реплик, то частота фамилии «Бунтман» будет существенно выше той, которая в действительности эту фамилию характеризует[26].

Следующий вопрос, который здесь возникает, — как именно размечать эти метатекстовые включения. Здесь существенно различение, с одной стороны, обозначений авторов реплик, а с другой, — ремарок (типа «Все молчат», «Смеется» и под.). Последние не таят за собой никакой интриги: при разметке они помещаются между двумя уникальными (т. е. не используемыми в данном тексте ни для каких иных целей) знаками, например, #Все молчат.#, и при стандартной обработке специальным набором программ (автор — А. Поляков), обязательным для всех текстов, располагаемых на сайте корпуса, автоматически переводятся на «другой уровень»

[25] Для драматургии метатекстом являются элементы текста, которые, принадлежа, как и остальной текст, автору произведения, тем не менее, при реальной постановке пьесы на сцене не получают словесного воплощения, — это ремарки и имена действующих лиц, вводящие их реплики.

[26] Точно так же и при подготовке к размещению в корпусе драматических произведений пренебрегать различием текста и метатекста не только культурологически, но и лингвистически некорректно, в частности, например, статистические результаты при анализе пьесы с неразмеченным и размеченным метатекстом будут существенно различными (так, например, в «Горе от ума» слово «внучка» при отсутствии разметки метатекста имеет частоту 21, а при различении текста и метатекста (т. е. при отделении ремарки Графиня-внучка: от самих реплик графини-внучки) — частоту 0).

внутритекстового существования (в частности, позволяющий не учитывать их при статистическом анализе).

Что касается обозначений авторов реплик, то, в принципе, в этом случае можно было бы поступать так же, как и в случае ремарок. Однако разработчики корпуса сочли это нерациональным: при таком решении пропадает существенная информация, которой на данном этапе воспользоваться из-за программных ограничений, к сожалению, невозможно, но в будущем снятие этих ограничений весьма вероятно. Имеется в виду социологическая информация — характеристики говорящих с точки зрения имени (если оно известно), пола, возраста и профессии, — которые могут иметь существенное значение для социолингвистических исследований[27]. При наличии социолингвистической разметки в будущем появится возможность, например, сформировать подкорпус высказываний программистов, слесарей, женщин после 55 лет и под.[28] Представляется, что с лингвистической точки зрения постановка такой задачи вполне осмысленна.

Таким образом, учитывая все вышесказанное, стратегия социологической метаразметки устного текста выглядит следующим образом:

	Текст считается имеющим одного автора[29]	Текст считается имеющим двух авторов	Текст считается не имеющим автора, поскольку реальных участников беседы слишком много
Социологическая информация находится в таблице метаразметки текста	+	+	–
Социологическая информация помещается при каждой реплике соответствующего автора	–	+	+

[27] Ср., например, работу [Rayson, Leech, Hodges].

[28] Для драматургии это дает возможность, в частности, при анализе «Горе от ума» образовать подмножество «высказывания Молчалина, Фамусова, Лизы» и т. п., при анализе советской драматургии — подмножество высказываний председателей парткомов, при анализе драматургии 19 века — подмножество купеческих высказываний, и т. п.

[29] В соответствии с принципами, изложенными выше.

Таким образом, мы видим, что при этой системе подачи материала социолингвистическая информация о говорящих остается доступной для исследователя даже в том случае, если текст считается не имеющим автора.

* * *

Таковы в общем принципы подготовки устных текстов для размещения в Национальном корпусе русского языка. Очевидно, однако, что по мере расширения корпуса устной речи (как мы уже говорили, сейчас разработчики ставят перед собой задачу довести объем устного корпуса до 10 млн словоупотреблений; на данный момент обработано в соответствии с вышеизложенным порядка 4 млн) возможна некоторая корректировка описанных в настоящих заметках решений. Задача разработчиков на данный момент, соответственно, состоит в том, чтобы при разметке и метаразметке устного корпуса не предпринимать необратимых действий, которые в дальнейшем ни при каких условиях не могут быть исправлены.

Литература

[Кибрик, Подлесская] — А. А. Кибрик, В. И. Подлесская. К созданию корпусов устной речи: принципы транскрибирования. — НТИ. Сер. 2. Информационные процессы и системы, 2003, № 10, с. 5-12

[Китайгородская, Розанова 1999] — М. В. Китайгородская, Н. Н. Розанова. Речь москвичей. Коммуникативно-культурологический аспект. М., 1999

[Китайгородская, Розанова 2003] — М. В. Китайгородская, Н. Н. Розанова. Современное городское общение: типы коммуникативных ситуаций и их жанровая реализация (на примере Москвы). — Современный русский язык. Социальная и функциональная дифференциация. М., 2003

[СРЯ] — Современный русский язык, М., 2002

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

[Burnage, Dunlop] — G. Burnage, D. Dunlop. Encoding the British National Corpus. — Oxford University Computing Services Published in English Language Corpora: Design, Analysis and Exploitation, Papers from the 13th international conference on English Language research on computerized corpora, Nijmegen 1992, edited Jan Aarts, Pieter de Haan and Nelleke Oostdijk.

[Čermák] — Frantiek Čermák. Language Corpora: The Czech Case In Text, Speech and Dialogue, TSD 2001, eds. V. Matouek, P. Mautner, R. Mouček, K. Taušer Springer Berlin etc. 2001, 21-30

[Rayson, Leech, Hodges] — P. Rayson, G. Leech, M. Hodges. Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. UCREL (University Centre for Computer Corpus Research on Language) Lancaster University, Lancaster LA1 4YT, United Kingdom