Е. А. Гришина,
С. О. Савчук

Корпус устных текстов

в Национальном корпусе русского языка: состав и структура

Исследования устной речи ведутся в русистике с 60-х годов прошлого века во многих научных центрах: в Москве, Санкт-Петербурге, Саратове, Перми, Екатеринбурге, Омске, Красноярске, Ульяновске и др. Хорошо известны работы Е. А. Земской, О. А. Лаптевой, М. В. Китайгородской, Н. Н. Розановой, О. Б. Сиротининой и руководимых ими коллективов, В. Е. Гольдина, Г. Г. Инфантовой, Т. И. Ерофеевой, М. Д. Воейковой и др.

Однако следует отметить, что обычно такая работа базируется на ограниченном материале — записях, сделанных одним исследователем или его группой. Так, например, в распоряжении авторов монографии «Лексика разговорной речи в системе функциональных стилей русского литературного языка» (Саратовский университет) была 100-тысячная словарная картотека, составленная по магнитофонным записям разговорной речи, и 15 тысяч карточек-

контекстов, полученных ручным способом. Это много, но несопоставимо с тем, что может предложить корпус текстов.

Созданный в рамках НКРЯ корпус устных текстов значительно расширяет возможности исследователя-русиста.

1) Корпус содержит подлинные целые тексты, а не отдельные выписки, что позволяет обнаружить то, что ускользает от понимания при выборочных записях.

2) Корпус содержит объем текстов, который значительно превосходит то, чем обычно располагает исследователь устной речи. Это позволяет судить о частотности или случайности явления, обнаружить закономерности, которые проявляются только на больших объемах, делать статистически достоверные выводы об обнаруженных закономерностях.

3) Корпус включает тексты, разнородные с точки зрения половозрастного, социального, профессионального состава говорящих, времени и географии записей.

4) Тексты, собранные в корпусе устной речи, охватывают большой временной диапазон — более 70 лет, если начинать отсчет с транскриптов кинофильмов 1930-х годов. Первые записи разговорной речи относятся к 1956 году, последние сделаны весной 2008 года. Это дает возможность прослеживать изменения, которые происходят в устной речи (а они здесь происходят стремительно), отмечать появление новых тенденций и т.д.

Так, проведенный на материале корпуса анализ частицы *вот* и ее вариантов [Гришина 2008] показал, что *от* — это стилистический вариант частицы *вот*, употребляемый либо в диалектных (квазидиалектных), либо в устаревающих контекстах. В частности, в фильмах до 1961 года этот вариант встречается в 2 раза чаще, чем в фильмах последующих лет. В работе [Савчук 2008] отмечено появление с конца 1990-х годов в непринужденной устной речи молодого поколения новой синтаксической конструкции с местоимением *такой*, используемой для передачи чужого высказывания: «*Мне брат* ***такой*** *на следующее утро:* «*Что, смотрела «Ловкие руки?*» (речь студентки 19-ти лет, Разговор студенток, Ульяновск, 4.05.2006).

5) Корпус содержит (в отличие от коллекций, на которых обычно строятся исследования разговорной речи) устные тексты, относя-

щиеся к разным сферам общения, произнесенные в разных условиях. Мы не разделяем мнения некоторых исследователей, согласно которому «живой русской речью» следует считать только «непринужденную речь горожан в условиях непосредственного контакта говорящих»[1]. Устная речь, понимаемая как форма существования языка (в отличие от письменной формы), представлена в разных сферах функционирования: в разговорно-бытовой сфере — как непринужденная разговорная речь, в научной — как устная научная речь, в публицистической — устная публичная речь, телевизионная и радиоречь, в официально-деловой — устная официальная речь, в производственно-технической — устная профессиональная речь, в церковно-богословской — проповедь, в сфере рекламы — теле- и радиореклама, в художественной сфере — речь кино и театра. Поэтому устный текст в корпусе — это не только диалог в магазине или беседа за столом в кругу семьи, но и научная лекция, доклад на семинаре, встреча автора со слушателями, интервью или токшоу по телевидению, спортивный радиорепортаж и многое другое.

Другой критерий, по которому принято разграничивать разновидности устной речи и который учитывается при отборе текстов в корпус, — степень подготовленности или спонтанности. По степени убывания спонтанности можно расположить типы устных текстов на следующей шкале [Галяшина 2002].

Спонтанная речь	• Спонтанный диалог • Спонтанный монолог
Квазиспонтанная речь	• Интервью (ответы на вопросы) • Монологический рассказ на заранее известную тему • Репродуцирование вслух чужой речи • Обдуманная речь по заранее составленному плану • Стереотипная речь по шаблонному тексту • Речь за суфлером

[1] Живая речь уральского города. Тексты. Екатеринбург, 1995. С. 4.

Заранее подготовленная речь	• Пересказ вслух с опорой на письменный текст • Изложение вслух письменного текста • Воспроизведение вслух выученного наизусть текста • Чтение вслух заранее известного текста • Чтение вслух заранее неизвестного текста

В корпусе устной речи нет текстов, представляющих собой заранее подготовленную речь[2]. Но зато в большом объеме представлены тексты, относимые на этой схеме к квазиспонтанным, — прежде всего это записи публичной речи и подкорпус кино.

6) Подкорпус кино, включающий транскрипты речевой составляющей игровых и мультипликационных фильмов (а в проекте — и документальных фильмов и игровой рекламы[3]) — уникальный компонент корпуса устной речи в составе НКРЯ. Эта сфера существования языка почему-то ускользала от внимания исследователей устной речи и составителей больших корпусов[4]. Между тем влиятельность этих текстов в русском (и не только в русском) речевом узусе чрезвычайно велика, как было показано в работе [Гришина 2005*б*].

В настоящее время общий объем корпуса устной речи составляет более 7,5 млн словоупотреблений, и его можно считать предста-

[2] Записи заранее подготовленной речи являются важной составляющей корпусов звучащей речи.

[3] Первые опыты подготовки текстов теле- и радиорекламы показали, что эти рекламные ролики представляют собой «воспроизведение вслух выученного наизусть текста» и потому не соответствуют критериям отбора текстов для устного корпуса. Все они были включены в состав рекламных текстов корпуса письменной речи.

[4] Область, в которой широко используются корпуса, создаваемые на базе фрагментов игровых фильмов, видеоклипов и видеозаписей телепередач, – психолингвистическое изучение эмоционального поведения человека. Кроме того, на базе киноклипов создаются мультимедийные корпуса; о проекте создания такого корпуса в составе НКРЯ см. статью Е.А. Гришиной «Мультимедийный русский корпус (МУРКО): проблемы аннотации» в наст. сборнике.

вительной коллекцией текстов, отражающей функционирование современного русского языка в его устной форме. Покажем, как это отражается в составе и структуре корпуса[5].

Состав и структура корпуса устной речи

Подобно всем другим текстам, вошедшим в состав Национального корпуса русского языка, устные тексты имеют метатекстовую разметку, позволяющую отбирать из всего массива пользовательский подкорпус, а также анализировать состав корпуса и корректировать его в процессе наполнения. К основным метатекстовым признакам относятся:

- **сфера функционирования**: публичная, непубличная, кино
- **тип** текста: беседа, интервью, микродиалог и пр.
- **тематика** текста: частная жизнь, медицина и здоровье, политика и общественная жизнь и пр.
- **время создания** текста
- **место записи** текста
- **стиль** текста: нейтральный, сниженный, официальный
- **характеристики аудитории**: размер, возраст, уровень подготовки

Приведем количественные показатели корпуса по некоторым метапризнакам.

Тексты распределяются по ***сферам устной коммуникации*** следующим образом:

Сфера функционирования	**Количество словоупотреблений**	**Соотношение в %**
Устная публичная речь	3930076	52 %
Устная непубличная речь	761966	10 %
Речь кино	2819394	38 %

[5] Данные о составе и структуре корпуса приводятся по состоянию на январь 2009 года.

В пределах каждой сферы тексты распределяются по основным ***типам***[6]**.**

Сфера функционирования	**Тип текста**	**Количество словоупотреблений**	**Соотношение в %**
Устная публичная речь	беседа	1064750	27,1%
	интервью	305775	7,8%
	дискуссия	1920306	48,9%
	лекция	116636	3%
	парламентские слушания	86640	2,2%
	конференция	48972	1,2%
	круглый стол	49177	1,3%
	рассказ	75585	1,9%
	прочие	181547	6,1%
Устная непубличная речь	разговор	583752	76,6%
	разговор телефонный	79990	10,%
	рассказ	47340	6,2%
	пересказ	12533	1,6%
	микродиалог	25435	3,3%
	прочие	12916	1,8%

Речь кино распределяется по киножанрам.

Речь кино	кинодрама	661963	23,5%
	кинокомедия	1049043	37,2%
	кинодетектив	256423	9,1%
	киноповесть	239922	8,5%
	кинофантастика	83812	3%
	кино детское	233797	8,3%
	прочие	294427	10,4%

[6] Поскольку допускается отнесение текста одновременно к нескольким типам, например, для речи кино—к нескольким киножанрам (кинодетектив|кинокомедия, кинокомедия|кино детское|киносказка), то сумма долей разных значений этого признака может превышать 100%.

В корпусе представлены тексты разнообразной ***тематики***. Наиболее частотны тексты, имеющие помету «частная жизнь» (более 50% всех текстов), затем по степени убывания идут тексты на темы политики и общественной жизни, искусства и культуры, науки, досуга и развлечений, спорта.

По ***времени записи*** бóльшая часть текстов относится к современному периоду — 2003–2006 годы, немалая часть — больше 400 тысяч словоупотреблений — относится к периоду 1990-х годов, период 1970-х годов — 260 тысяч, 1980-х годов — 160 тысяч, до 1970 — 160 тысяч.

География Корпуса живой русской речи достаточно широка. В Корпусе представлены тексты, записанные в Москве и Московской области (их большинство), в Санкт-Петербурге, Саратове, Самаре, Таганроге, Воронеже, Новосибирске, Ульяновске, Екатеринбурге, Кировской области.

Источниками текстов для корпуса послужили:

- записи устной речи, опубликованные в хрестоматиях и сборниках, составленных специалистами в области разговорной речи: под редакцией Е. А. Земской, О. А. Лаптевой, Н. Н. Розановой и М. В. Китайгородской, А. С. Герда и др.;
- ранее не публиковавшиеся коллекции записей устной речи, собранные в различных исследовательских центрах: ИРЯ им. В. В. Виноградова, МГУ (Москва), СПбГУ, Саратовском, Ульяновском университетах;
- стенограммы бесед социологов в фокус-группах на различные общественно-значимые темы, предоставленные Фондом «Общественное мнение»;
- записи устных текстов, выполненные сотрудниками корпуса или под их руководством.

Лингвистическая аннотация

Для корпуса устной речи характерны те же виды разметок, что и для всего НКРЯ, — метатекстовая, морфологическая и семантическая, т.е. в устном корпусе возможны те же типы формирования подкорпусов и типы поиска, что и в «письменном» корпусе. Однако в лингвистической разметке устного корпуса есть и некоторые особенности, из которых следует упомянуть две.

1. Сохраняющая разметка. В устной речи, как известно, употребляется большое количество стяжек (самые стандартные — *тыща, здрасти, щас* и проч.), растяжек (*нууу, вооот*), игровых форм (*зерба, ды* — название буквы «д», *вурдулак*), диалектизмов (*кажный, дак*), искажений иностранцами (*слюшай*) и под. Нам чрезвычайно не хотелось включать эти искажения в основной словарь НКРЯ, поскольку за исключением очень небольшого количества стандартных стяженных форм или фразеологизованных игровых форм (например, *хоккей* в значении *о'кей*), все остальные представляют собой случайные осцилляции и часто не имеют лингвистического значения сами по себе, а лишь как манифестации некоторых общих особенностей устной речи. Но поскольку этих форм нет в словаре НКРЯ, постольку морфологический парсер, который размечает грамматику и семантику в корпусе автоматически, оставляет такие искаженные формы вообще без разметки или приписывает им неправильную разметку (например, варианты частицы *вот*, весьма частотные в устной речи,— *во, от, о* — распознаются как соответствующие предлоги). Такого рода ошибки морфологической (и, соответственно, семантической) разметки некритичны для НКРЯ в целом, ввиду большого объема последнего, но весьма неприятны в небольшом устном корпусе.

Эта трудность могла бы быть преодолена, если бы было принято решение принудительно трансформировать искаженные формы в правильные. Однако такое снятие проблемы существенно обедняет наши перспективы в изучении устной речи — мы теряем возможность анализировать именно и только искаженные формы (в частности, в их соотнесении с неискаженными, словарными). А в ряде случаев такая нормализация и вовсе невозможна, например, контексты с несловарным вариантом *щаз* не могут быть приведены к контекстам с *сейчас*, поскольку *сейчас* и *щаз* имеют существенно разные значения (в частности, в высказывании *Щаз!, Бегу!* есть некоторые компоненты значения — сарказм, ирония,— которые отсутствуют или ослаблены в *Сейчас! Бегу!*), или, например, некоторые контексты с *о* (вариантом частицы *вот*) не могут быть заменены аналогичными контекстами со стандартным *вот* (см. об этом [Гришина 2008]). В связи с этим было принято решение в случае искаженных форм применять так называемую

сохраняющую разметку, суть которой можно выразить следующей схемой:

$$\text{Incorrectness}\left\{\text{Correct}_{\text{Spelling}} + \text{Grammatical}_{\text{Characteristics}} + \text{Semantic}_{\text{Characteristics}}\right\}$$

Согласно этой схеме, каждая Inc (неправильность) сохраняется в тексте, при этом ей приписывается правильная, словарная форма (Cor), которая, в свою очередь, традиционными для НКРЯ способами, с помощью грамматического парсера, получает свою грамматическую (Gram) и семантическую (Sem) разметку.

Сохраняющая разметка предоставляет пользователю устного корпуса возможность произвести следующие действия:

1) Найти все случаи вхождения данной Cor в виде Inc (например, искаженные формы *здравствуй(те) — здрасьте, издраствуй, здряствуй, здрааасьте, драствуй, здрассте*).

2) Найти все случаи вхождения данной Cor в виде Cor, без Inc (например, все контексты, где слово *тысяча* используется в полной форме, а не в форме *тыща*).

3) Найти все контексты с Cor, включая Inc (например, все случаи употребления местоимения это, включая апокопированный вариант *эт* (*Эт что такое?*) и безударный *йто* (*Что йто случилось?*)).

4) Найти все ответы на запрос от определенного Gram и Sem, включая или исключая искаженные формы (например, на запрос «наречия направления» будет получен результат, включающий в себя апокопированный вариант *прям* < *прямо*, хотя формально вариант *прям* совпадает с краткой формой мужского рода прилагательного *прямой*, а не с наречием *прямо,* и при отсутствии сохраняющей разметки именно так и был бы размечен; при этом же запросе, но исключающем искаженные формы, будут получены только контексты с наречием *прямо* — разумеется, среди прочих наречий направления).

2. Социологическая разметка. Помимо морфологической и семантической разметки, в корпусе устных текстов используется так называемая социологическая разметка — характеристика словоупотребления с точки зрения пола и возраста употребившего его говорящего (если эта информация, естественно, доступна создателям корпуса).

Социологическая разметка позволяет пользователю создать свои подкорпуса:

- по полу говорящего (т.е. пользователь может сформировать подкорпуса женской или мужской устной речи);
- по возрасту говорящего (например, пользователь может сформировать подкорпус реплик подростков);
- по году рождения говорящего (доступно только для кинотранскриптов — можно, например, отобрать реплики актеров, родившихся в XIX в.);
- по имени актера (например, можно сформировать подкорпус кинореплик Евгения Леонова).

Очевидно, что социологическая разметка может быть дополнена метатекстовой — позволяющей отобрать тексты, созданные одним говорящим, что предоставляет возможность вынести его имя и год рождения в описание текста как целого (понятно, что в случае, если а) говорящих в тексте больше одного, б) говорящие по этическим причинам безымянны, в) их возраст либо неизвестен, либо слишком разнообразен, — эти параметры не могут быть вынесены в описание целого текста и приходится обращаться исключительно к социологической разметке).

Возможности и перспективы использования социологической разметки довольно широки. Проиллюстрируем это следующим примером: проверим, есть ли какие-нибудь статистически значимые различия между мужчинами и женщинами в использовании прилагательных формы с уменьшительно-ласкательным суффиксом *-еньк-*. Наиболее частотными в этой зоне являются прилагательные *кругленький* и *пухленький*. Распределения здесь таковы:

	Всего	**Говорящий — женщина**	**Говорящий — мужчина**
кругленький (о вещи)	29%	50%	8%
кругленький (о человеке)	25%	0%	50%
пухленький (о человеке)	29%	42%	17%

Как видим, по отношению к вещи женщины употребляют слово *кругленький*, а мужчины избегают такого определения (при этом, надо заметить, слово *круглый* по отношению к вещам и мужчинами, и женщинами употребляется в равной степени). Что касается определения человека, то здесь между мужчинами и женщинами наблюдается существенное различие — женщины предпочитают прилагательное *пухленький*, а мужчины в этом же значении употребляют слово *кругленький*. Таким образом, определение *пухленький* — в значительной степени «женское» слово, а слово *кругленький* свойственно и мужчинам, и женщинам, но по отношению к разным классам предметов.

Исследования устной речи на основе корпуса

Приведем пример использования корпуса устной речи, который касается вопроса о различии устной и письменной речи и предлагает образец его решения на основе количественных данных, предоставляемых корпусом.

Этому вопросу посвящена обширная литература, описывающая как экстралингвистические факторы, обусловливающие разграничение устной и письменной речи, так и собственно лингвистические признаки. Выявлен ряд статистических показателей, релевантных для дифференциации устной и письменной речи, спонтанной и подготовленной, монологической и диалогической [Галяшина 2002].

Исследование, выполненное на материале Национального корпуса русского языка [Гришина 2007*а*,*б*], показало значительное расхождение по ряду показателей между текстами устного и письменного корпуса. Эти показатели были названы **маркерами устной речи.** К числу признаков, обнаруживших в ходе сплошного обследования самые существенные расхождения между устными и письменными текстами, были отнесены следующие:

I. Средства, позволяющие говорящему **ориентировать слушающего** в логическом и прагматическом устройстве своей речи в отсутствие знаков препинания (наряду с интонацией).

1. Межфразовые скрепы, прежде всего *ну, а, да*.
2. Метатекстовые вставки: *вот, вот так, так вот, вот что, значит*.
3. Перформативные лексемы (*считаю, обещаю, спрошу* и пр.), эксплицитно выражающие речевое намерение говорящего, тип речевого акта — для этой цели используются глаголы речи и ментальной сферы.
4. Личные местоимения 1 и 2 лица, подчеркивающие роли участников речевого акта.
5. Контактные слова, привлекающие внимание слушающего к речи:
   а) глаголы восприятия и ментальной сферы в форме 2 лица (*понимаешь/понимаете, знаешь/знаете, видишь/видите, (по)смотри/(по)смотрите* и др.);
   б) обращения к слушающему;
   в) частицы-обращения, формально совпадающие со скрепами *ну, а, да*, но произносимые с вопросительной интонацией.

II. Эгоцентрические элементы, проявляющие говорящего в его речи (наряду с местоимениями 1-го лица)

6. Глаголы в форме 1 лица, описывающие действия говорящего.
7. Слова *да, нет*, служащие для выражения согласия — несогласия.
8. Междометия и оценочные слова, прежде всего слово-интенсификатор *очень*.
9. Слова, выражающие ближайшие намерения говорящего и слушающего — глаголы движения.

III. Дейктические элементы

10. Наречия, привязывающие высказывания к настоящему моменту — *здесь, сейчас, сегодня*;
11. Указательные местоимения и наречия *тут, там, тогда, такой, так*.

По всем этим параметрам корпус устных текстов, как было показано в статьях [Гришина 2007а,б], существенно отличается от корпу-

са письменных текстов: разница составляет от 2 (параметр *сегодня*) до 10 раз (междометия).

Ниже приведены результаты более детального исследования устной речи с использованием перечисленных маркеров. Оно выявило особенности представленных в корпусе разновидностей устной речи в сравнении с типами письменных текстов, обнаруживших разную степень «устности»[7]. Результаты в чем-то подтвердили наши интуитивные представления о характере устной и письменной речи, а в чем-то и удивили.

Сопоставлялись следующие подкорпуса текстов.

Подкорпус текстов	**Количество текстов**	**Объем в с/у**
Устная непубличная речь	522	486788
Устная публичная речь	660	3827200
Речь кино	185	1195671
Драматургия (1950–2006)	53	541618
Художественная проза (1950–2006)	2249	33016014
Нехудожественные тексты (интервью)	2057	2810521
Нехудожественные тексты (статьи)	18011	23647354
Электронная коммуникация	89	1192121

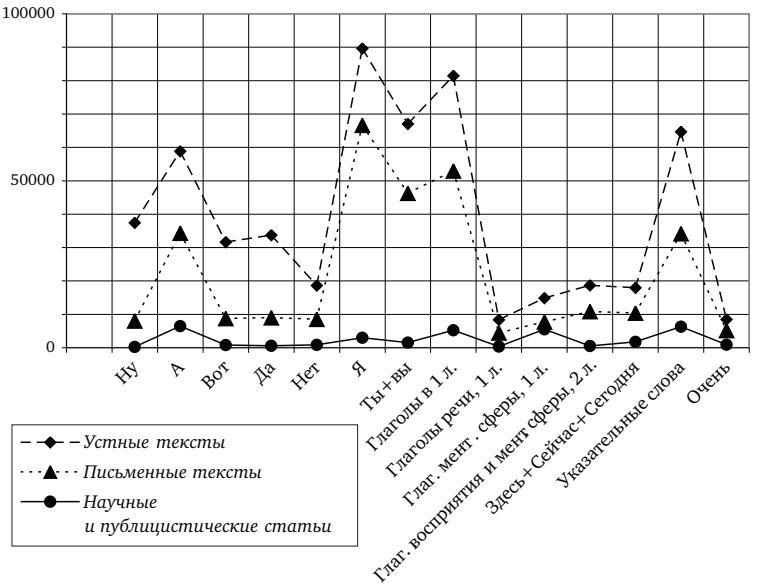
Для каждого подкорпуса были вычислены **абсолютные частоты** встречаемости маркеров — как отношение количества контекстов к количеству словоупотреблений в подкорпусе (для удобства вычислений эти величины пересчитаны на миллион словоупотреблений). Результаты представлены в таблице на следующей странице.

[7] Следует отметить, что исследование проводилось дважды, с разницей в один год, на корпусах разного объема. При этом значения параметров, различаясь в абсолютных цифрах, сохранили свое соотношение в текстах разных типов.

Частота встречаемости дискурсивных маркеров в текстах разных типов (ipm)[8]

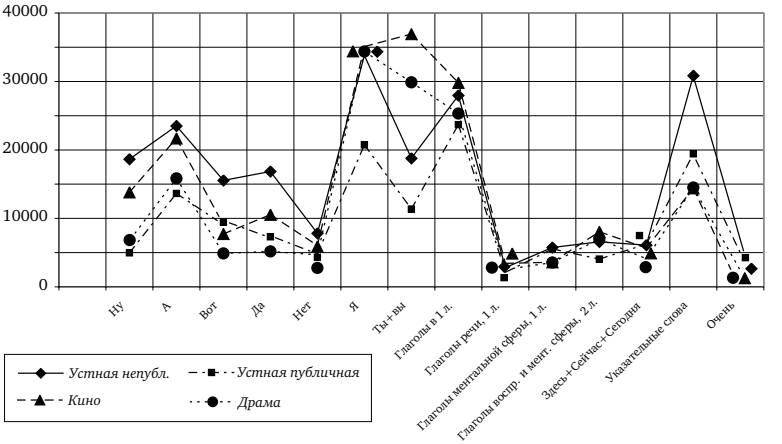
Маркер	Устная непубличная	Устная публичная	Кино	Драма	Худож. проза	Нехудож. (интервью)	Нехудож. (статьи)
Ну	18645	4970	13799	5628	1724	671	215
А	23500	13641	21679	15836	10318	8160	6463
Вот	15545	8363	7721	4882	2313	1527	761
Да	16849	6331	10534	5173	2492	1319	534
Нет	7800	4790	6047	4667	2031	1835	830
Я	33790	20722	35092	34365	17648	14606	2979
Ты+вы	18764	11348	36909	29877	10135	6281	1490
Глаголы в 1 л.	27953	23686	29809	25319	12060	15557	5228
Глаголы речи, 1 л.	2798	2112	3425	2517	976	950	311
Глаголы ментальной сферы, 1 л.	5748	5523	3582	3543	1643	2535	585
Глаголы восприятия и ментальной сферы, 2 л.	6557	4048	8045	6974	2496	1380	513
Здесь+Сейчас +Сегодня	6085	6217	5653	4097	2480	3807	1728
Указательные слова	30835	19419	14400	14497	10573	9073	6299
Очень	2911	3600	1931	1416	1186	2580	880

[8] ipm (instances per million words)—общая частота, или число употреблений на миллион слов корпуса.



1. Частота встречаемости маркеров в устных и письменных текстах

На диаграмме 1 показано соотношение суммарных частот маркеров в устных текстах и в письменных текстах, отличающихся повышенной степенью диалогичности (драма, художественная проза, газетно-журнальные интервью). Как видим, значения частот маркеров в устных текстах выше, чем в письменных (в некоторых точках в 2–3 раза), причем пропорционально выше, что хорошо видно на графике. Это, несомненно, свидетельствует о том, что маркеры устной речи выбраны точно и отражают именно существенные ее особенности, прежде всего диалогичность. Для сравнения на этом же рисунке графически представлено поведение маркеров в текстах современных научных и публицистических статей, из которого видно, что для данного типа текстов, в отличие от текстов с повышенной степенью устности, эти маркеры не являются значимыми.



На диаграмме 2 наглядно представлено соотношение значений маркеров в разных типах устных текстов в сравнении с драмой. Здесь обращают на себя внимание следующие моменты:

1. Речь кино можно рассматривать как точную имитацию устной речи. По отдельным показателям она ближе к публичной речи (*вот*, указательные слова), по каким-то — к непубличной устной речи (глаголы в форме 1 лица, *а, ну*), а по каким-то показателям даже превосходит естественную речь и приближается к драме (местоимения 1 и 2 лица, глаголы в 1 лице, глаголы речи, 1 л.). Это оправдывает включение кино в корпус устных текстов.

2. Драматические тексты обнаружили практически полное совпадение по данным параметрам с текстами кино (а ведь драма относится к письменному корпусу!). По некоторым показателям они даже превосходят естественную устную речь (так имитация акцентирует наиболее характерные особенности имитируемого)[9].

---

[9] Особенно обращает на себя внимание высокая частота местоимений 2 л. в кино и драме, даже по сравнению с непубличной речью. Это может говорить о том, что модель общения, воссоздаваемая в литературных диалогах, коммуникативно более правильная, в ней ярче выражена установка на собеседника, что отражается в экспликации местоимений 2 л. Это обстоятельство еще требует уточнения и может быть проверено при пополнении корпуса новыми записями непубличной речи.

О чем это говорит? Во-первых, это те особенности, которые сразу опознаются на слух и «бросаются в глаза» в письменном тексте. Они воспроизводятся и используются авторами — драматургами и сценаристами — для имитации устной речи персонажей в пьесе и в кино, воссоздающих на сцене и экране модель реальной жизни[10].

Во-вторых, исследователи русской разговорной речи в 1950–60-е годы были не так уж далеки от истины, когда изучали особенности разговорной речи на материале текстов пьес (в частности, данные о разговорной речи в частотном словаре Засориной получены на таком материале). Но поскольку язык драмы все-таки нельзя считать спонтанной устной речью, наглядные количественные показатели принадлежности драмы к письменной речи, вероятно, нужно искать в области синтаксиса, строения текста, лексического разнообразия.

Перспективы развития корпуса устной речи

В ближайших планах развития НКРЯ — создание Акцентологического корпуса (см. статью Е. А. Гришиной «Корпус “История русского ударения”» в наст. сборнике) и Мультимедийного корпуса устной речи (см. в наст. сборнике статью Е. А. Гришиной «Мультимедийный русский корпус (МУРКО): проблемы аннотации»). Есть ли на фоне этих проектов перспективы развития у корпуса устной речи — ведь он явно проигрывает в полноте представления материала и акцентологическому, поскольку не содержит информации об ударении, и уж тем более мультимедийному, дающему живой портрет высказывания?

Ответ на этот вопрос можно дать только положительный по нескольким причинам.

Во-первых, корпус устной речи отличается от акцентологического и мультимедийного корпусов прежде всего составом текстов. Как уже говорилось, в устном корпусе собраны образцы устной речи, записанные в разных регионах России и в широком временном диапазоне. В принципе при наличии аудиозаписи, материальных и че-

[10] По терминологии В.Д. Левина, такие признаки живой речи являются «сильными», в отличие от «слабых», которые не выходят за пределы устной коммуникации [Лаптева 2003, 272].

ловеческих ресурсов нет никаких препятствий к тому, чтобы привести расшифровки в соответствие с реальным звучанием, оформить тексты так, как это делается для акцентологического корпуса.

Однако это не всегда возможно. Значительная часть текстов устного корпуса (прежде всего ранние записи, а также переданные в корпус коллекции из региональных центров изучения устной речи) существует только в виде транскриптов: магнитофонные записи либо не сохранились, либо вообще не делались (в случае ручной записи микроситуаций). Это относится прежде всего к текстам, изданным в составе хрестоматий (РРР 1978; Китайгородская, Розанова 1999; РРР-СВ 1998, Живая речь 1995 и др.). Эти записи могут быть представлены только в составе корпуса устной речи.

Несмотря на усовершенствование звукозаписывающих устройств расшифровки аудиозаписей и в наши дни остаются наиболее распространенным (и наиболее надежным) способом фиксации устного материала, и этот источник пополнения корпуса устной речи нельзя недооценивать. Как показал опыт проведения практики по сбору устной речи студентами московских вузов, транскрипты не всегда сопровождаются полноценными аудиофайлами, пригодными для использования в корпусе. Причины могут быть разными — техническими и случайными: низкое качество записи, редкий формат файлов записывающего устройства, ошибки при конвертации и др. Такие записи не могут быть использованы в акцентологическом корпусе, но могут занять достойное место в устном корпусе. Таким образом, по объему и составу текстов устный корпус превосходит и акцентологический, и планируемый мультимедийный.

Вторая причина, по которой следует продолжать развитие устного корпуса, — характер лингвистической разметки и поиска в нем. Корпус только тогда становится эффективным инструментом исследования, когда разметка в нем соответствует тем лингвистическим задачам, которые ставит исследователь при обращении к данному ресурсу. Так, анализ большинства морфолого-синтаксических и лексико-семантических особенностей устной речи удобнее проводить на материале устного корпуса: его достоинства — это большой объем и разнообразие текстов, разметка, сопоставимая с разметкой в корпусе письменных текстов, которая позволяет легко сравнивать результаты, полученные на материале текстов разных

типов. Если же речь идет об изучении фонетических, акцентологических, просодических, паралингвистических характеристик устных высказываний, то следует обратиться к акцентологическому или мультимедийному корпусам.

Таким образом, ближайшей задачей развития корпуса устных текстов можно считать наращивание объема корпуса до 10 млн словоупотреблений за счет текстов, пока недостаточно в нем представленных, прежде всего записей непубличной речи, и обеспечение сбалансированности корпуса.

Другой задачей является расширение географии корпуса за счет включения записей русской устной речи, сделанных в различных регионах России, в странах ближнего и дальнего зарубежья, что позволит изучать состояние русского языка в контакте с другими близкородственными и неродственными языками, в иноязычном окружении.

Литература

Галяшина 2002 — Е. И. Галяшина. Проблема дифференциации спонтанной и подготовленной речи. // Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям *http://www.dialog-21.ru/materials/archive.asp?id=7287&y=2002&vol=6077*

Гришина 2005а — Е. А. Гришина. Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. — М.: Индрик, 2005. — С. 94–110.

Гришина 2005б — Е. А. Гришина. Два новых проекта для Национального корпуса: мультимедийный подкорпус и подкорпус названий. — Там же. С. 233–250.

Гришина 2007а — Е. А. Гришина. О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.). — М.: Издательский центр РГГУ, 2007. — С. 147–156.

Гришина 2007б — E. Grishina. Text Navigators in Spoken Russian. // Proceedings of the workshop "Representation of Semantic Structure of Spoken Speech" (CAEPIA'2007, Spain, 2007, 12–16.11.07, Salamanca). — Salamanca, 2007. — P. 39–50.

Гришина 2008 — Е. А. Гришина. Варианты частицы *вот* в непринужденной речи // Инструментарий русистики: корпусные подходы (Slavica helsingiensia, 34). — Хельсинки, 2008. — Р. 63–91.

Живая речь 1995 — Живая речь уральского города. Екатеринбург, 1995.

Лаптева 2003 — О. А. Лаптева. Теория современного русского литературного языка. — М., 2003.

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

РРР 1978 — Русская разговорная речь: Тексты/ Отв. ред. Е. А. Земская, Л. А. Капанадзе. М., 1978.

РРР–СВ 1998 — Русская разговорная речь европейского северо-востока России / Под ред. Н. С. Сергиевой и А. С. Герда. Сыктывкар, 1998.

Савчук 2008 — С. О. Савчук. Местоимение *такой* в функции маркера чужой речи в устном высказывании // В печати.