Национальный корпус русского языка как источник сведений об устной речи[1]

Е.А. Гришина,
кандидат филологических наук

В статье описаны три подкорпуса Национального корпуса русского языка, из которых пользователь может получать сведения об устной русской речи, — устный корпус, акцентологический корпус и мультимедийный русский корпус (МУРКО). На конкретных примерах показано, какие задачи пользователь может ставить и решать при обращении к каждому из этих корпусов.

Abstract

The paper describes three sub-corpora of the Russian National Corpus (RNC) which are the sources of information concerning the spoken Russian. The sub-corpora are as follows: 1) the Corpus of Spoken Russian, 2) the Accentological Corpus, 3) the Multimodal Russian Corpus (MURCO). The author, using the specific data, gives the detailed description of the problems which may be raised and solved by means of these three sub-corpora of the RNC.

1. Общие сведения о Национальном корпусе русского языка (НКРЯ)

НКРЯ — это достаточно большое по объёму (порядка 160 млн словоупотреблений) собрание аннотированных русских текстов, сбалансированное по времени, по сферам функционирования, жанрам и типам текстов; каждое словоупотребление в НКРЯ размечено — как минимум — морфологически и семантически, а каждый текст в целом имеет метатекстовое описание (дата создания, автор, название, жанр и мн. др.). В соответствии с этой разметкой в НКРЯ можно вести поиск и отбор информации, для чего создан соответствующий интерфейс. НКРЯ с 2003 г. находится в открытом доступе в Интернете по адресу www.ruscorpora.ru. Создают и развивают Корпус группа исследователей, сосредоточенная в основном вокруг Института русского языка РАН, и компания «Яндекс»[2].

[1] Статья написана при поддержке гранта РФФИ 08-06-00371а и программы ОИФН РАН 2009–2011 гг. «Генезис и взаимодействие социальных, культурных и языковых общностей».

[2] Подробнее о НКРЯ в целом и, среди прочего, об истории его создания, можно прочитать на сайте Корпуса, а также в сборниках [НКРЯ2005] и [НКРЯ2009]. На сайте также расположены необходимые инструкции по пользованию Корпусом.

Как и большинство национальных корпусов мира, НКРЯ включает в себя в основном письменные тексты. Однако, ориентируясь на Британский и Чешский национальные корпуса, создатели НКРЯ решили ввести в состав Корпуса также и устные тексты. В настоящий момент ситуация с устной речью в НКРЯ такова:

Таблица 1

Тип подкорпуса	Актуальное состояние	Объём (в словоупотреблениях)	Тип включённых текстов
Корпус устной речи	функционируют, будут пополняться новыми материалами	8 млн	транскрипты публичной и непубличной устной речи, кинотранскрипты
Акцентологический корпус		4,5 млн	стихотворные тексты с размеченными сильными долями (3,2 млн), акцентуированные кинотранскрипты (1,3 млн)
Мультимедийный русский корпус (МУРКО)	находится в стадии разработки	предположительно — от 2 до 5 млн[3]	аудиозаписи и кинофильмы, разрезанные на небольшие фрагменты и выровненные с соответствующими участками аннотированных транскриптов

Далее мы покажем, какие именно задачи можно решать с помощью этих трёх подкорпусов.

2. Корпус устной речи

Корпус устной речи включает транскрипты[4], полученные создателями НКРЯ из разных источников, — как из крупных центров исследования устной речи (например, Саратовский, Петербургский университеты, Институт русского языка РАН[5]), так и от волонтёров, которые самостоятельно записывали и расшифровывали устную речь[6]. Естественно, эти тексты — очень разного уровня по качеству расшифровки (от блестящих образцов, принадлежащих лингвистам-профессионалам, до очень посредственных). Такая неравноценность исходных текстов, а также отсутствие возможности обратиться к звуковому первоисточнику приводят к естественному ограничению на качество тех задач, которые исследователь может ставить при использовании корпуса. Естественно, не предусмотрена постановка задач, каким бы то ни было образом связанных со звучащей речью (чисто фонетических или орфоэпических). Более того, пользователь всегда должен учитывать, что в *сложных* случаях при грамматических или словообразовательных запросах он с осторожностью должен относиться к полученным из корпуса данным. Так, например, при запросах **ишко* или **ишка* (т.е. «найти все слова, которые в исходной форме оканчиваются на *-ишко* или *-ишка*») пользователь должен учитывать, что качество гласного в заударных слогах в расшифровках могли определять непрофессионалы и, следовательно, цитаты типа

[3] Объём МУРКО будет зависеть от масштабов его финансирования, поскольку этот проект требует достаточно больших трудозатрат.

[4] Транскрипты — это расшифровки устных текстов, сделанные в русской орфографии, с заменой пунктуационных знаков на слэши. Примерами могут послужить все цитаты в данном обзоре, кроме (5), (7), (8). Подробнее о форме подачи материала в Устном корпусе см. [Гришина 2005] и [Савчук, Гришина 2009].

[5] В частности, в Устный корпус вошли материалы из книг [РРР] и [Китайгородская, Розанова 1999].

[6] Большую помощь в пополнении корпуса оказали студенты Московского педагогического университета, Ульяновского университета. Очень ценные транскрипты удалось получить из фонда фонодокументов В.Д. Дувакина при МГУ, а также от Фонда «Общественное мнение», который передал НКРЯ расшифровки бесед социологов с респондентами, проходивших в разных городах РФ.

(1) [№ 1, муж, 21] *Она чуть ли не плачет / мол мне нужен тока ты и всё тут!* [№ 2, муж, 19] *Во / блин / **актёришка**.* [Рассказ о личной жизни (2006)]

(2) [Андрей, Георгий Делиев, муж, 44, 1960[7]] *Для ни́х о́н никто́! О́н жа́лкий никчёмный **музыка́нтишка**.* [Кира Муратова и др. Настройщик, к/ф (2004)]

(3) [№ 8, муж, 62] *А ведь у них там / если мелкий **городишка** <...> то / конечно / он тоже на несколько милей вытягивается там.* [Беседа с социологом на общественно-политические темы (Самара) // Фонд «Общественное мнение», 2003]

не гарантируют нам, что мы имеем дело с существительными *а*-склонения, — это вполне могут быть существительные *о*-склонения (*актёришко, музыкантишко, городишко*).

Однако на тех языковых уровнях, на которых влияние фонетики минимизировано, устный корпус может быть чрезвычайно полезным. Прежде всего, это касается **лексических запросов.** Например, если нас интересует использование в устной речи тех или иных интенсификаторов (наречий семантического ряда *очень, абсолютно, чрезвычайно, исключительно* и т.д.), то мы можем обратиться к данным корпуса и получить результаты, упорядоченные в таблице 2.

Таблица 2

	Всего	**Очень**	**Абсолютно**	**Чрезвычайно**	**Исключительно**
Непубличная речь	34%	35%	32%	*11%*	*21%*
Публичная речь	44%	43%	**56%**	**62%**	56%
Речь кино	21%	22%	*12%*	27%	23%

Как видим, судя по отклонениям от средних значений (столбец «всего»), для *абсолютно, чрезвычайно* и *исключительно* характерна повышенная частотность в публичной устной речи и пониженная — в непубличной, частной, а для *абсолютно* — пониженная частотность в речи кино. Что касается интенсификатора *очень*, то его распределение не отличается от среднего, т.е. тип устной речи не влияет на частотность *очень*.

Полезен устный корпус и при **грамматических запросах** (в тех случаях, естественно, когда грамматические различия не базируются на чистой фонетике). Можно, например, проверить, как склоняются в устной речи обе части топонима *Москва-река*. Сравнение данных устного корпуса и корпуса письменных текстов (XVIII–XIX вв. и XX в.) даёт следующее распределение:

Таблица 3

	Всего	**XVIII–XIX вв.**	**XX в.**	**Устная речь**
Москвы-реки (склоняемая первая часть)	82%	100%	71%	*44%*
Москва-реки (несклоняемая первая часть)	18%	*0%*	**29%**	**56%**

[7] Здесь и далее первое число в квадратных скобках обозначает возраст говорящего в момент произнесения данной фразы, второе — год его рождения. Если эти данные недоступны, то они не приводятся.

Мы видим, во-первых, что несклоняемость первой части этого топонима — явление, возникшее в XX в. Во-вторых, очевидно, что несклоняемость — явление, гораздо более характерное для устной речи (отклонение от среднего значения — в 3 раза, 56% против 18%), чем для письменной. Соответственно, можно сделать предварительный вывод, что несклоняемость первой части в этом топониме возникает и широко распространяется в XX в. в устной речи и лишь опосредованно отражается в письменной.

Интересные и надёжные результаты предоставляет устный корпус в **зоне русского словообразования.** Так, например, сравнив употребление выражений *на минуту/минутку/минуточку/секунду/секундочку* в письменной и устной речи, мы увидим практически зеркальную картину:

Таблица 4

	Всего	Письменная речь	Устная речь
На минуту	38%	**55%**	*19%*
На минутку	22%	*12%*	**32%**
На минуточку	16%	*4%*	**31%**
На секунду	18%	**28%**	*7%*
На секундочку	5%	*0%*	**11%**

Очевидно, что чем выше эмоциональная окрашенность существительного (выражается суффиксальным нулём в минимуме и суффиксом *-очк-* в максимуме), тем более оно характерно для устной речи.

Этот же пример показателен с точки зрения фиксации новых значений в устной речи. Выражения *на минуточку/секундочку* (но не *на минуту/минутку/секунду*) в современной речи приобретают новое значение «между прочим» — говорящий считает, что его собеседник не придаёт значения некоторому важному обстоятельству, и фиксирует внимание собеседника на этом обстоятельстве с помощью вводных слов *на минуточку/секундочку*[8]:

(5) *Сила этого искусства такова, что около 900 препаратов, приготовленных им, до сих пор хранятся в Кунсткамере в превосходном состоянии и не нуждаются в замене. Препаратам* ***на минуточку*** *триста лет.* [Дуня Смирнова. Гробъ богатого китайца (1997) // «Столица», 1997.12.08]

(6) [Чернов, Андрей Панин, муж, 40, 1962] *Ты* ***на секундочку*** *моя дочь. Мне бы вообще-то не хотелось / чтобы ты связалась с каким-нибудь там проходимцем.* [Руслан Бальтцер, Александра Большакова. Даже не думай, к/ф (2002)]

[8] В данной статье нет возможности подробно обсуждать этимологические аспекты такой трансформации значений. По-видимому, эта трансформация связана с использованием выражений *на минуточку/секундочку* в качестве способа привлечь к себе внимание собеседника, например: (4) [Любовь Андреевна, Нина Усатова, жен, 37, 1951] [художнику] *Тсс! ти́ше!* [Художник, муж] [вызывает Любовь Андреевну в коридор] *На мину́точку!* <...> [Любовь Андреевна, Нина Усатова, жен, 37, 1951] [выходит в коридор] *Фу́ / накури́ли!* [Юрий Мамин, Владимир Вардунас. Фонтан, к/ф (1988)].

В варианте *на минуточку* эта семантическая трансформация впервые фиксируется в публицистике (пример (5)), т.е. в письменной речи, и только в варианте *на секундочку* — в устной речи (пример (6)). При этом, однако, «плотность» этого нового явления в устной речи существенно, в 11 раз, превышает его «плотность» в речи письменной. Очевидно, таким образом, что новые явления в русском языке, зарождающиеся обычно в устной речи, в устной же речи легче обнаруживаются.

Устный корпус также предоставляет возможность пользователю формировать подкорпуса в соответствии с ограниченным числом **социологических параметров** (можно сформировать корпус мужской/женской речи, корпус речи какого-либо актёра, корпус говорящих определённого возраста или года/периода рождения, а затем проводить на этих корпусах соответствующие изыскания). Например, с помощью соответствующих запросов по корпусам мужской и женской речи можно определить, что для речи женщин междометие *ах!* характерно в большей степени, чем для речи мужчин, а для речи мужчин, напротив, более характерны междометия *ух!* и *эх!* Что касается *ох!*, то его распределения в мужской и женской речи не отличаются от среднего.

Ещё пример. Запрос по группам говорящих разного возраста показывает, что частота употребления слова *умный* молодыми (10–30 лет) людьми и людьми старшего возраста (31–80 лет) не отличается от средней частоты этого слова по выборке в целом, а вот слово *мудрый* в заметно меньшей степени характерно для речи молодых людей.

Примеры можно множить, но, как представляется, и приведённых иллюстраций достаточно, чтобы понять перспективы использования устного подкорпуса НКРЯ[9].

3. Акцентологический корпус

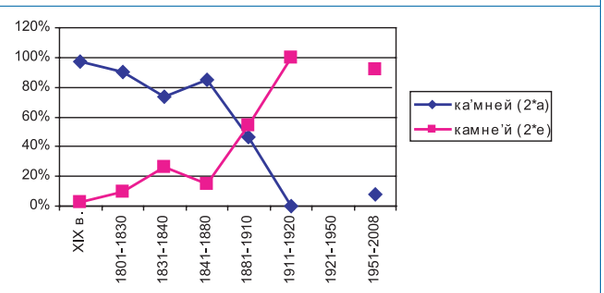
Акцентологический корпус обеспечивает пользователя сведениями о таком немаловажном аспекте устной русской речи, как ударение. Корпус состоит из двух зон: 1) *поэтическая зона* содержит тексты, в которых в соответствии с метром стиха размечены сильные доли (т.е. слоги, на которые потенциально может падать ударение); путём пересчёта по определённым правилам из совокупности сильных долей мы можем получить точные сведения об ударении в том или ином слове[10]; 2) *прозаическая зона* на сегодняшний день включает акцентуированные кинотранскрипты, в которых ударения расставлены не в соответствии с правилами литературного русского языка, зафиксированными в словарях и нормативных справочниках, а в соответствии с реальным ударением в том или ином фильме. Формирование этого корпуса ещё не закончено, будут пополняться как поэтическая, так и прозаическая зона, причём последняя не только новыми кинотранскриптами, но и акцентуированными расшифровками реальных устных текстов.

[9] Подробнее об устном корпусе можно узнать в работах [Гришина 2005], [Савчук, Гришина 2009].

[10] Правила пересчёта иктов в ударения подробно изложены в [Гришина 2009а], а в более простом варианте — на сайте НКРЯ, в инструкции к Акцентологическому корпусу, http://www.ruscorpora.ru/instruction-accent.pdf. Об использовании поэтического корпуса как источника сведений об акцентологии можно также прочитать в [Корчагин 2008], [Гришина 2009].

Стандартные задачи, которые можно решать на материале акцентологического корпуса, иллюстрирует следующий пример. Как известно, в современном русском языке слово *камень* имеет колеблющуюся парадигму склонения: в основном это схема 2*е (т.е. в ед.ч. и в им./вин. пад. мн.ч. — ударение на основе, а в косвенных падежах мн.ч. — ударение на окончании), однако возможно и постоянное ударение на основе в обоих числах, т.е. схема 2*а (см. [Грамм]). Интересно посмотреть, как устанавливалась эта система на протяжении последних трёх веков[11]. Построим диаграмму по данным акцентологического корпуса.

Диаграмма 1



Как видим, ситуация в начале XXI в. зеркально противоположна ситуации в XVIII — нач. XIX в.: полностью победила новая схема (2*е: *ка'мни, камне'й*), при абсолютном доминировании в XVIII в. старой схемы (2*а: *ка'мни, ка'мней*). Перелом, как видно из диаграммы, произошёл в период 1881–1910 гг., в течение которого схемы ударения были практически равночастотны. Интересен при этом «скачок» в использовании схемы 2*е в промежутке 1831–1840 гг., который связан с творчеством М.Ю. Лермонтова, вообще тяготеющего, как показывает анализ системы ударений в его поэтическом творчестве, к использованию «прогрессистских» акцентологических моделей.

Помимо чисто акцентуационных задач, акцентологический корпус позволяет ставить задачи, связанные с манерой произнесения. Прежде всего это касается проблемы выбора *ё* или *е* при использовании тех или иных словоформ. Все мы со школьных времён помним цитату из Н. Некрасова:

(7) Иди́ в ого́нь за че́сть отчи́зны, За у́бежде́нье, за́ любо́вь... Иди́, и ги́бни ***бе́зупре́чно***. Умре́шь не да́ром, де́ло про́чно, Когда́ под ни́м струи́тся кро́вь ... [Н. А. Некрасов. Поэт и гражданин (1855–1856)]

Предположим, мы хотим узнать, является ли произнесение *безупрёчно* поэтической вольностью Некрасова или «общим местом» для языка 19 в. Запрос в акцентологическом корпусе «*безупречный* в зоне рифмовки» предлагает нам семь примеров *безупречно* в позиции рифмы. И только в одном случае из семи *безупречно* произносится с *е*, а не с *ё* (поскольку рифмуется с *бессердечно*):

(8) Ра́дость сия́ла, чиста́ ***безупре́чно***, В ча́с, как тебя́ оброни́ла неве́ста. Не́т, не поки́ну тебя́ *бессерде́чно*, Зде́сь, у меня́ на груди́ тебе ме́сто. [А.А. Фет. «Окна в решётках, и сумрачны лица...» (1882)]

В остальных случаях *безупречно* рифмуется со словоформами *точно, цветочный, беспорочен, прочной, непорочны*, т.е., как и в примере (7), произносится с *ё*, из чего мы вправе сделать вывод о том, что произнесение этой словоформы с *ё* было для языка этого периода стандартным.

[11] Естественно, пока работа над акцентологическим корпусом не закончена, в нём остаются хронологические провалы — например, недостаточно или вообще нет данных для начала XX в., для XX в. практически нет сведений из поэтических текстов. Однако для понимания общих тенденций материала всё же хватает.

4. Мультимедийный русский корпус (МУРКО)

В отличие от устного и акцентологического корпусов, МУРКО находится пока на стадии проектирования, хотя и достаточно продвинутой. Как следствие, иллюстрировать задачи, которые может решать МУРКО, с помощью конкретных примеров весьма затруднительно. Поэтому в данном разделе обзора мы опишем типы разметки, которые предполагается осуществить в МУРКО, и задачи, на решение которых ориентирован тот или иной тип аннотации.

4.1. Стандартная разметка НКРЯ. В стандартную разметку НКРЯ для устных текстов входят: 1) лемматизация (т.е. сведение всех словоформ данной лексемы к единой лемме), 2) морфологическая разметка, 3) семантическая разметка, 4) социологическая разметка. Все эти типы разметки будут использованы и в МУРКО. Это даёт нам возможность ставить и решать следующие задачи.

1) Поиск звукового материала для данной конкретной словоформы, данной конкретной леммы, данного словообразовательного форманта, а также для словосочетаний с конкретными составляющими. Таким образом нам же удалось в предварительном порядке определить хронологические и гендерные факторы, влияющие на выбор произношения в словах с корнем *бухгалтер-* (заднеязычный фрикатив или заднеязычный смычный), а также в сочетаниях предлога *к* со словами, начинающимися на *к-* (см. [Гришина 2009б]).

2) Поиск звукового материала для того или иного морфологического значения. Например, можно подобрать звуковой материал для анализа произношения безударного окончания в 3 л. мн.ч. настоящего времени глаголов 2 спряжения, т.е. отслеживать произношение *любют* vs *любят*, *ходют* vs *ходят* и под.

3) Отбор материала на тех или иных социологически ограниченных подкорпусах, т.е. осуществление фонетико-орфоэпических исследований с социологической (гендерной, возрастной, личностной) специализацией. Например, возможно будет исследовать историю произносительных норм, касающихся слов *когда* [*када/кагда*] и *тогда* [*тада/тагда*], в разные хронологические периоды, а также у говорящих определённого года рождения[12].

4.2. Фонетико-орфоэпическая разметка. Мы предполагаем ввести в МУРКО вид разметки, который не используется больше ни в какой зоне НКРЯ, а именно — фонетико-орфоэпическую разметку, основанную на орфографии. Каждому текстовому фрагменту, который образует отдельный документ в МУРКО[13], будут приписаны аннотационные цепочки, включающие два типа информации.

[12] В МУРКО, естественно, будет предусмотрена и стандартная для НКРЯ семантическая разметка, но поскольку, по первому впечатлению, семантика напрямую с фонетикой/орфоэпией не связана, семантическая разметка будет подключаться, скорее всего, как удобный ограничитель материала при использовании других видов разметки, хотя не исключено, что со временем проявятся возможности использовать семантическую разметку в МУРКО по прямому назначению.

[13] Эти текстовые фрагменты будут весьма невелики по объёму, иначе пользователю будет трудно ориентироваться в «прикреплённом» к тексту аудио- и/или видеоматериале.

1) *Сведения о звукосочетаниях*. В каждом текстовом фрагменте будут размечаться границы слов структуры C#C[14], V#V и C#V. Внутри слова будут размечаться буквосочетания структуры CC, CCC (сочетания двух и трёх согласных) и VV (сочетания двух гласных). Например, для следующего фрагмента:

(9) [Макарыч/ Алексей Смирнов] *Кузнечик.* [Кузнечик/ Сергей Иванов] *Я.* [Макарыч/ Алексей Смирнов] *Иди к командиру. Настроение / о!* [Л. Быков и др. В бой идут одни «старики», к/ф (1973)]

будут размечены ЗН (*кузнечик*), K#K, НД (*к командиру*), СТР, ОЕ, ИЕ (*настроение*), E#O (*настроение о*). Соответственно, по всей этой разметке будет возможен поиск, так что пользователь без малейшего напряжения сможет получить, среди прочего, звуковые цепочки, которые содержат сочетание ЗН или НД (и определить, для каких из них перед гласными переднего ряда зафиксировано произношение [з'н], [н'д], а для каких [зн], [нд]), сочетание K#K (и определить, как именно оно реализуется — как одиночное [к], как гемината [кк] или как сочетание фрикатива и смычного [хк]).

2) *Сведения об акцентологической структуре*. Поскольку в качестве первоначальной базы для МУРКО (по экстралингвистическим соображениям) были выбраны кинотранскрипты, все текстовые расшифровки фильмов будут включать в себя сведения об ударении (т.к. в перспективе все фильмы, включённые в устный корпус, будут акцентуированы). Следовательно, мы имеем возможность учитывать при разметке структуру слова с точки зрения ударения. Предполагается, что каждому слову будут приписываться сведения о качестве и номере ударного слога, о качестве и номере предударных и заударных гласных, о количестве слогов в слове. Таким образом, текстовому фрагменту (9) пословно будет приписана следующая информация:

Таблица 5

Слово	Ударный гласный и номер ударного слога	Предударный гласный	Заударный гласный	Количество слогов
кузнечик	E2[15]	у1	и1	3
я	А1	—	—	1
иди	И2	И1	—	2
к	—	—	—	—
командиру	И3	а1, о2	у1	4
настроение	Е3	о1, а2	и1, е2	5
о	О1	—	—	1

Представляется, что фонетико-орфоэпическая разметка такого рода существенно облегчит сбор материала и постановку задач для тех исследователей, которые занимаются русской фонетикой и орфоэпией.

4.3. Разметка речевых действий. Этот тип разметки создаётся в помощь исследователям русской интонации и системы русских речевых актов. Совершенно новые перспективы открывает

[14] C = согласный, V = гласный, # = пробел между словами.

[15] Использование «е», а не «э» обозначает «э» после мягкого гласного.

МУРКО также для исследователей русских вокальных жестов, междометий, физиологических действий, включённых в речь, поскольку доступными становятся большие массивы ситуационно разнообразного материала.

Нам уже приходилось подробно описывать теоретические основы, на которых базируется эта разметка (см. [Гришина 2009б]). В данной статье мы хотели бы перечислить базовые параметры, которые учитывает аннотация речевых действий в том виде, в каком она используется в Marker — рабочем месте разметчика, созданном специально для аннотирования МУРКО (см. [Кудинов, Гришина 2009]):

количество и пол говорящих в данном клипе (т.е. можно отобрать клипы с *монологом, диалогом, полилогом, внутригендерные* и *кросс-гендерные*)

типичные социальные ситуации, в которых функционирует говорящий: *неспецифическая ситуация, разговор с представителем власти, покупка в магазине, заказ в ресторане* и под.

язык речевого действия: *русский, русский с акцентом, английский, французский, немецкий и т.д., тайный язык, абракадабра*

собственно типы речевых действий: *вопрос*[16]*, согласие, отрицание, этикетные высказывания, утверждения, императивы, модальные высказывания, шутливые/насмешливые высказывания, чужая речь, торговля, пейоративные высказывания, похвала, апеллятивы*

полнота речевого действия: *полное, автопрерывание, вопрос без ответа, наложение реплик, незаконченное, прерванное, продолженное, жест вместо слова*

типы повторов (если повторы есть): *однократный/многократный, однословный/неоднословный, переспрос, повтор с интенсификатором, повтор с разной интонацией, эхо, передразнивание*

манера говорения: *нормальная, быстрая, невнятная речь, голос за кадром, декламация, дефекты дикции, диктовка, дубляж, крик, напевать, оговорка, говорение «про себя», пьяный разговор, говорение на слезе, на смехе, скандирование, чревовещание, чтение*

вокальные жесты: *звонок, дразнить, издавая звуки, заполнители паузы, изобразительные звуки, обращение к животному, хмыканье, холодно, цокать языком* и др.

междометия: *аккомпанементы* (*да, ну, э*), *боль, волнение, вопрос, восхищение, жалость, испуг, насмешка, недоверие, недовольство, недоумение* и мн. др.

физиологические действия: *вздох, втянуть носом воздух, выдох, зевок, кашель, плач, плевок, поцелуй, принюхиваться, причмокиванье* и др.

[16] Каждый из типов имеет разновидности: например, вопрос может быть частным, общим, косвенным, обратной связи, нечленораздельным, переспросом, контактным; согласие может быть подтверждением, пониманием, признанием, разрешением, согласием, подчинением, но в данной статье из-за недостатка места мы приводим только основные типы.

Перечисленные параметры приписываются не отдельному слову или отдельной реплике, а клипу как целому. Таким образом, исследователь, которого, к примеру, интересуют типы интонации, характерные для общего вопроса, в качестве ответа на соответствующий запрос («общий вопрос») получит всё множество клипов с прикреплёнными текстовыми расшифровками, в каждом из которых — среди прочего — содержатся и общие вопросы. Сочетание запросов — например, «императивы» + «повторы разных типов» — выдаст пользователю материал, предоставляющий возможность анализировать использование повторов в высказываниях, содержащих в себе императивную составляющую.

4.4. Разметка жестов. Впервые при создании МУРКО ставится задача сплошной разметки жестового ряда, сопровождающего речь. Уже первые обращения к результатам этой разметки оказались, как представляется, чрезвычайно продуктивными. В частности, учёт жестикуляции при исследовании русских вокальных жестов оказывается в высшей степени полезным (см. [Гришина 2009в]). Недостаток места не позволяет нам подробно изложить ни теоретические, ни практические основы, на которых базируется разметка жестов (отошлём ещё раз к статье [Гришина 2009б], в которой они описаны достаточно подробно), поэтому, как и в предыдущем разделе, мы просто перечислим параметры разметки, которые были учтены при создании рабочего места разметчика жестов GesturesMarker, созданного специально для аннотации МУРКО в диалоговом полуавтоматическом режиме (см. [Кудинов, Гришина 2009]):

Социальные параметры жеста: *имя актёра, пол актёра/персонажа, возраст актёра/персонажа, социальная ситуация*

Параметры описания жеста:

кратность жеста: *однократный/многократный*

основной орган (зона тела жестикулирующего, в которой локализован данный жест): *голова, корпус, рука, руки, нога, ноги*

активный орган[17] (список зависит от выбора основного органа): *брови, бровь, глаз, глаза, голова, губы, зубы верхние, лицо, подбородок, рот, язык, нос, ухо, губа нижняя* (для основного органа *голова*), *корпус, плечи, спина, плечо* (для основного органа *корпус*) и т.п.

пассивный орган[18] (зависит от выбора активного органа): *грудь/живот, кисть, лицо, подбородок, рот, корпус, голова, нет пассивного органа* (для активного органа *кисть* < основного органа *рука* < многократного жеста) и т.п.

адаптор[19] (зависит от выбора пассивного органа): *внешний объект, собеседник, тяжёлый предмет, нет адаптора, поверхность, одежда* (для позиции *нет пассивного органа* < активного органа *кисть* < основного органа *рука* < многократного жеста), и т.п.

ориентация ладони: *вверх, вниз, перпендикулярно телу говорящего, к телу говорящего, наружу, ориентация не важна*

ориентация кисти: *вперёд, внутрь, вверх, кулак* и др.

[17] Активная часть основного органа, выполняющая жест.

[18] Орган человеческого тела, пассивно участвующий в осуществлении жеста.

[19] Необходимый участник жестикуляции, не являющийся частью тела жестикулирующего (своего рода объективированный вне говорящего пассивный орган), например, *внешний объект* для указательного жеста, *собеседник* для пейоративных жестов, *небо* для жеста *воздеть руки* и под.

направление движения: *вперёд-назад, вперёд, из стороны в сторону, вбок, изнутри наружу, снаружи в центр, на себя, снизу вверх, круг вертикальный, круг горизонтальный* и др. (набор значений зависит от выбора активного органа)

Тип и значение жеста:

тип жеста: дейктические, декоративные жесты, жесты — речевые действия, жесты внутреннего состояния, изобразительные, корпоративные, пейоративные, поисковые, регулирующие, риторические, условные, физиологические, этикетные жесты

значение жеста — см. таблица 6

название жеста: стандартное название жеста, имеющего данное значение (например, для значения *прощание* [этикетный жест] зафиксированы жесты, имеющие названия *воздушный поцелуй, вскинуть руку, вскинуть руку к голове, кивнуть, махать/махнуть рукой, поцеловать кого-л., рукопожатие, троекратный поцелуй;* для значения *призыв к порядку* [регулирующий жест] зафиксированы жесты, имеющие названия *дёрнуть кого-л. за какую-л. часть тела, качать пальцем, коснуться кого-л., поднять палец, посмотреть строго, стукнуть/стучать по чему-л., ткнуть пальцем, толкнуть кого-л. локтем* и т.д.). Таких пар (значение жеста — название жеста) на сегодняшний день зафиксировано порядка тысячи.

Дополнительные характеристики жеста:

полнота: *полный, автопрерывание, прерванный, редуцированный, трансформированный*

аутентичность: *аутентичный, притворный, игровой, зеркальный, показывать на себе*

сопровождающие эмоции: *улыбка, смех, слёзы, нет*

Все перечисленные параметры аннотации жестов являются одновременно поисковыми параметрами, т.е. пользователь может запросить жесты с данным значением, данного типа, данного стандартного названия (например, подобрать все клипы, где говорящий машет рукой), основанные на том или ином активном органе (например, все жесты, которые осуществляются с использованием в качестве активного органа указательного пальца), на той или иной ориентации ладони (например, все жесты с ориентацией ладони *наружу*); кроме того, естественно, пользователь может комбинировать эти параметры (например, подобрать все клипы, в которых *отказ* [жест — речевое действие] осуществляется с помощью основного органа *рука*). Наконец, немаловажной является возможность комбинировать «жестовые» запросы с запросами лексическими, грамматическими, акцентуационными, семантическими, а также с запросами, касающимися речевых действий.

5. Заключение

Как видно из вышеизложенного, информация об устной речи, которую можно получить в НКРЯ, достаточно разнообразна, причём каждая из трёх составляющих (собственно устный корпус, акцентологический корпус и МУРКО) подстраховывает и дополняет две другие. От пользователя требуется лишь чёткое понимание того, какие задачи следует ставить на данном корпусе, а какие не могут получить

Таблица 6

Тип жеста	Значение жеста
Дейктические жесты	демонстрация, идентификация собеседника, общеуказательный, самоидентификация, фиксация объекта
Декоративные жесты	общедекоративный
Жесты — речевые действия	возражение, договор, клятва, назидание, намёк, насмешка, одобрение, отказ, отрицание, подтверждение, подчинение, поздравление, понимание, похвала, предложение, просьба, секретничать, согласие, требование, угроза, упрёк, я так думаю
Жесты внутреннего состояния	беспокойство, благоговение, вожделение, возмущение, вызов, высокая оценка, горе, готовность, дистанцирование, догадаться, дружелюбие, задуматься, игривость, интерес, испуг, кокетство, лукавство, манерность, негодование, недоверие, недовольство, недоумение, неожиданность, нервозность, нетерпение, облегчение, огорчение, ожидание, озабоченность, опасение, осторожность, отвращение, отчаяние, печаль, подобострастие, потрясение, предвкушение, пренебрежение, привязанность, радость, раздражение, разочарование, расслабленность, растерянность, решимость, скука, смущение, солидарность, сосредоточенность, сочувствие, спохватиться, стыд, уверенность, удивление, удовлетворение, что тут поделаешь!, ясно без слов
Изобразительные жесты	действие, животное, качество, количество, направление, объект, топология
Корпоративные жесты	молитва, пароль, пионерское приветствие, воинское приветствие
Пейоративные жесты	дразнить, дурак!, критические замечания, передразнивание, подумаешь!, пошёл вон!, сумасшедший!
Поисковые жесты	искать что-л., обратить внимание, оценивать обстановку, оценка веса, оценка внешнего вида, оценка температуры, поиск слова, узнавание
Регулирующие жесты	достаточно!, задавать ритм, замолчи!, запрет, иди!, начали!, не трогай!, остановить кого-л., остановить машину, подбодрить, прекрати!, привлекать/привлечь внимание, призыв к порядку, самоуспокоение, торопить, успокаивать, утешение
Риторические жесты	интенсификация действия, материализация речи (аргумент, брань, возражение, вопрос, отрицание, перечисление, побуждение, просьба, требование, убеждение, утверждение, финал, чужая речь), новая тема, потребность в поддержке, предвосхищение отрицания, предвосхищение согласия
Условные жесты	пьяный, сдаваться, тост
Физиологические жесты	больно, горько, делать что-л. с усилием, душно, жажда, кричать, не видно, не слышно, недосып, неприятно, нервно, стереть чужое прикосновение, ударить, усталость, холодно, чешется, чисто
Этикетные жесты	благодарность, вы правы!, зевать, извинение, не стоит благодарности, помощь, при кашле, при смехе, приветствие, приглашение, прощание, спросить разрешение, я слушаю

здесь правильного решения. Например, старая проблема устного корпуса — проблема расставленных в устных транскриптах слэшей, которые замещают знаки пунктуации и имеют лишь опосредованное отношение к реальной расстановке пауз. Очевидно, что эта проблема может быть решена только на материале МУРКО, когда пользователь получает возможность проанализировать реальную паузацию данного текстового фрагмента. Аналогично, только с помощью МУРКО могут быть решены проблемы акцентуации словосочетаний, состоящих

из (полу)служебных слов (например, ответить на вопрос, есть ли ударение на слове *ты* во фразе *Что ты делаешь?*, можно только с привлечением соответствующего звукового фрагмента); проблемы значения междометий и вокальных жестов (которые правильно истолковываются только при использовании соответствующего жестового ряда — одного аудиоряда в данном случае часто бывает недостаточно). Однако большое количество проблем, связанных с устной речью, можно решить, не привлекая аудио- и видеоряд, — и для этого будет совершенно достаточно материалов устного и акцентологического корпусов.

В заключение отметим, что основной задачей описанных корпусов, как и НКРЯ в целом, является обеспечение академических исследований и нужд преподавания (именно поэтому НКРЯ является абсолютно открытым и доступным ресурсом и таковым, мы надеемся, и останется). Это имеет своей оборотной стороной недостаточное представление в НКРЯ узкоспециализированных материалов, которые чаще всего интересуют инженерные и коммерческие структуры. Но это естественное и непреодолимое ограничение всех национальных корпусов, и для того, чтобы его компенсировать, необходимо параллельное создание узкоспециализированных и специфических корпусов, в том числе и устных.

Литература

1. Грамм — А.А. Зализняк. Грамматический словарь русского языка. М., 2003.
2. Гришина 2005 — Е.А. Гришина. Устный корпус в Национальном корпусе русского языка // НКРЯ2005, с. 94-110.
3. Гришина 2009а — Е.А. Гришина. Корпус «История русского ударения». Проект // НКРЯ2009 (в печати).
4. Гришина 2009б — Е.А. Гришина. Мультимедийный русский корпус (МУРКО): проблемы аннотации // НКРЯ2009 (в печати).
5. Гришина 2009в — Е.А. Гришина. К вопросу о соотношении слова и жеста (вокальный жест О в устной речи) (в печати).
6. Китайгородская, Розанова 1999 — М.В. Китайгородская, Н.Н. Розанова. Речь москвичей: Коммуникативно-культурологический аспект. М., 1999.
7. Корчагин 2008 — К.М. Корчагин. Поэтический подкорпус Национального корпуса русского языка как акцентологический источник // http://www.dialog-21.ru/dialog2008/materials/html/Korchagin.htm.
8. Кудинов, Гришина 2009 — М.С. Кудинов, Е.А. Гришина. Инструменты полуавтоматической разметки для Мультимедийного русского корпуса (МУРКО) (в печати).
9. НКРЯ2005 — Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.
10. НКРЯ2009 — Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. М., 2009.
11. РРР — Русская разговорная речь: Тексты / Академия наук СССР, Институт русского языка // Отв. ред. Е. А. Земская, Л. А. Капанадзе. М.: Наука, 1978.
12. Савчук, Гришина 2009 — С.О. Савчук, Е.А. Гришина. Устный корпус в Национальном корпусе русского языка: состав и структура // НКРЯ2009 (в печати).

Е.А. Гришина —

кандидат филологических наук (1989 г.), старший научный сотрудник Института русского языка им. В.В. Виноградова РАН, активный участник проекта «Национальный корпус русского языка».