

УДК 004.522

РОССИЙСКАЯ АКАДЕМИЯ НАУК

Санкт-Петербургский институт информатики и автоматизации

Санкт-Петербургский государственный университет аэрокосмического приборостроения

Санкт-Петербург, 199178, 14 линия, 39.

<http://www.spiiras.nw.ru/speech>

Труды первого междисциплинарного семинара

«Анализ разговорной русской речи» (АР<sup>3</sup> - 2007). – СПб.: ГУАП, 2007. – 87 с.

ISBN 978-5-8088-0270-4

Издание представляет собой сборник докладов, сделанных на заседаниях первого междисциплинарного семинара «Анализ разговорной русской речи» (АР<sup>3</sup> - 2007), проходившего 29 августа 2007 года в Санкт-Петербургском институте информатики и автоматизации Российской академии наук. Семинар посвящен обсуждению особенностей разговорной речи и возможных подходов к автоматическому анализу русской речи. Междисциплинарный подход к изучению речи позволит скорее продвинуться в моделировании речевой деятельности и решить фундаментальную проблему человеко-машинного диалога.

УДК 004.522

Статьи печатаются в авторской редакции.

Издание осуществлено за счет средств европейского гранта SIMILAR IST-2002-507609.

ISBN 978-5-8088-0270-4

© СПИИРАН, 2007

© Коллектив авторов, 2007

© ГУАП, 2007

*Е.П. Комовкина, Н.А. Слепокурова*

## **Об опыте составления частотного словаря устного спонтанного текста\***

*Санкт-Петербургский государственный университет,  
г. Санкт-Петербург, Россия,  
[nataliars@inbox.ru](mailto:nataliars@inbox.ru), [ekomovkina@yahoo.com](mailto:ekomovkina@yahoo.com)*

В ходе выполнения проекта по исследованию русской спонтанной речи авторами настоящего доклада была выполнена письменная расшифровка видеозаписей телевизионных спонтанных диалогов, прозвучавших в эфире летом-осенью 2005 г. Эти записи составляют достаточно надежную репрезентативную выборку (время звучания каждой – не менее 4 час., расшифровка обеих занимает в сумме около 200 стр. текста), в которой зафиксирован определённого рода срез реальной звуковой «материи» русского языка на сегодняшнем этапе его функционирования.

Указанные записи представляют, по нашему предположению, разные пласты живой устной речи – молодежный, «раскованный» (передача ***Большой Брат*** (канал ТНТ) и относительно более «респектабельный» (***Культурная революция*** (канал Культура)). Об этом свидетельствует как разница в возрасте, социальном статусе, образовании, культуре участников телепередач, так и характер ведения в них диалогов. Поскольку в передаче ***Культурная революция*** обсуждаются некоторые интересные для собеседников темы и вопросы мировоззренческого характера, реплики отдельных участников этой программы часто представляют собой спонтанные мини-монологи, в то время как в реалити-шоу ***Большой Брат***, не обременявшем своих участников никакими серьезными интеллектуальными проблемами, ведется как будто бы обычный бытовой диалог с частой сменой собеседников. Об указанной разнице наглядно говорят цифры: по данным первых 50 стр. письменных расшифровок, средняя длина реплики одного участника диалога в ***Большом Брате*** в 5 раз короче, чем в ***Культурной революции***.

Первый этап исследования полученного речевого материала включал в себя общий качественный обзор и сплошную фиксацию замеченных отклонений от соответствующих фонетических, грамматических и лексических норм кодифицированного литературного языка. Состав этих отклонений в обоих текстах вполне соответствовали той картине, которая нарисована в классических работах московских и саратовских исследователей русской разговорной речи [Горбова, Слепокурова и др. 2006].

Далее, естественно, возник вопрос о выявлении степени «спонтанности», «разговорности» этих текстов и о сравнении их по этому параметру друг с другом. Представляется совершенно нереалистичным осуществлять это сравнение в виде некоего подсчета по текстам предварительно отмеченных «разговорных явлений» и их последующего сопоставления по количеству и составу. Это обусловлено тем, что указанные «разговорные явления», представляющие собой отклонения от норм кодифицированного языка, имеют разную степень очевидности и достоверности как для носителей языка, так и для лингвистов. Критерии их выделения предельно прозрачны в сфере фонетики и морфологии (понятно, например, что зафиксированные нами в текстах выражения *бóшку ломит* или *в московском списком «Родины»* содержат в себе абсолютно недопустимые речевые ошибки), значительно менее очевидны в сфере лексики и словообразования (в самом деле, *к проблеме нечитания книг* или *ты не выпендриваешься перед всеми* – это ошибки или нормативные для разговорной речи явления?) и совершенно размыты и неопределенны в сфере синтаксиса, где разного рода эллипсис, употребление именительного темы, более свободный, чем при письме, порядок слов и т.п. уже давно признаются многими авторитетными исследователями устной речи

---

\* Работа выполнялась в рамках проекта «Русский язык и современная Россия» при поддержке федеральной целевой программы «Русский язык» и Благотворительного фонда В.Потанина.

«сущностными» и потому, очевидно, нормативными атрибутами русской спонтанной речи. Эти соображения, основанные на предварительном, но в то же время далеко не поверхностном знакомстве с имеющимся в нашем распоряжении речевым материалом, заставляют предполагать, что единственным по-настоящему продуктивным подходом к сформулированной задаче является **сплошной** просмотр и анализ текстов с целью выявления в них некоторых строго определенных лексико-грамматических явлений. Последнее было осуществлено нами в виде попытки создания **частотного словаря** устного спонтанного текста.

Основная трудность, с которой пришлось столкнуться в процессе выполнения работы, – отсутствие прецедентов. Хотя в литературе и были обнаружены некоторые данные о попытках составления словарей устной русской речи, относящиеся к 60-ым гг. прошлого века, не удалось найти сведений о конкретных процедурах. Нельзя было ориентироваться и на опыт создания немногих имеющихся частотных словарей русского языка, основанных на письменных текстах, в которых единицей текста считается любое слово «от пробела до пробела» и каждая входящая в грамматическую парадигму словоформа возводится к основной словарной форме, – в результате же получается точная количественная информация о представленности в исследованных текстах неких «инвентарных» лексико-грамматических единиц непонятной психолингвистической природы.

Поэтому в данной работе мы опирались, прежде всего, на те принципы, которыми руководствуются разработчики и создатели существующих в настоящее время независимых отечественных корпусов русского языка – Национального корпуса русского литературного языка (НКРЛЯ, Санкт-Петербург, руководитель – проф. В.Б. Касевич) и Национального корпуса русского языка (НКРЯ, Москва, руководитель – проф. В.А. Плунгян). В обоих этих корпусах, создающихся на основе огромных массивов письменных текстов, не предусмотрена процедура лемматизации и отдельной словарной единицей считается **словоформа**. Это решение соответствует последним экспериментальным психолингвистическим данным, согласно которым единицей перцептивного лексикона (а равным образом и единицей текста) у носителей языков с развитым словоизменением является не лексема, а словоформа, и, возможно, – фонетическая словоформа, представляющая собой, в общем виде, просодически единый комплекс из знаменательного слова и клитик [Венцов, Касевич 2003]. Совершенно очевидно, что и любые современные попытки создания словарей **устных** разговорных текстов должны учитывать реальную психолингвистическую природу входящих в словник единиц. Поэтому в настоящей работе в качестве базовой словарной единицы рассматривалась отдельная словоформа.

Две других весьма серьезных трудности, с которыми пришлось столкнуться при подходе к практическому решению поставленной задачи, связаны с проблемами лексико-грамматической омонимии и так называемых «составных слов». Эти трудности, являющиеся главным камнем преткновения и при разработке корпусов русского языка, на наш взгляд, вполне адекватно решаются создателями НКРЛЯ. Проблема омонимии преодолевается в нем путем приписывания **каждой** словарной единице так называемого дескриптора, в котором указываются все лексико-грамматические признаки этой единицы; что же касается «составных слов», под которыми понимаются единицы, представляющие собой сочетания формальных слов, но не образующиеся по правилам, а воспроизводимые в тексте в «готовом» виде (например, *друг друга, в целом, в самом деле, может быть, не по себе* и т.д.), то они учитываются как отдельные текстовые и словарные единицы и их список по мере вхождения в корпус новых текстов постоянно пополняется.

Ориентируясь на этот подход и полностью с ним солидаризируясь, мы, тем не менее, не могли его копировать в данной работе – во-первых, из-за ограниченности времени и ресурсов и, во-вторых, из-за того, что, в отличие от составителей НКРЛЯ и НКРЯ, мы имели дело с записями **просодически размеченных устных** разговорных текстов. И хотя эта разметка была достаточно примитивной (отмечались паузы между синтагмами и высказываниями; коммуникативный тип высказывания – утверждение, побуждение, вопрос; ненормативные лексические ударения), опираясь на нее, а также на некоторые литературные данные и собственную интуицию, мы сочли возможным и необходимым использовать в своей работе

более широкую интерпретацию понятия «составное слово», чем это допускается при составлении словарей на основе письменных текстов.

Что касается омонимии, то ее фактический масштаб стал очевидным лишь при нацеленной на выполнение поставленной задачи дальнейшей обработке текста: выяснилось, что количество случаев употребления говорящими в устной речи лексически и грамматически омонимичных словоформ огромно: ср., например, семантику и частеречную принадлежность таких частотных и общеупотребительных лексем, как *это* и *хорошо*, в следующих примерах: *Что это ты сегодня такая грустная?*, *Что это значит?*, *Это дело всем вместе решать*, *Кингстон это такой город на Ямайке*, *Она хорошо поет*, *Здесь хорошо*, *Хорошо / я ему скажу*. См. также примеры грамматической омонимии, т.е. случаи совпадения одинаковых словоформ в разных грамматических значениях: *сигареты на столе*, *возьми там сигареты*; *большой перерыв*, *с большой надеждой*, *к большой радости*, *на большой тарелке* и т.д.). Естественно, каждая из отмеченных выше словоформ должна учитываться в словаре в качестве самостоятельной единицы. Для «разведения» омонимичных словоформ было решено ввести систему цифровых маркеров. Различные маркеры использовались для разных падежей существительных, прилагательных и местоимений, некоторых глагольных форм (например, инфинитива в функции императива – в отличие от инфинитива после модальных и фазовых глаголов). Особыми пометами в случаях необходимости отмечались также род прилагательного, субстантивированные прилагательные, наречия в предикативной функции и т.д.

Возвращаясь к лексической омонимии, следует оговориться, что, поскольку целью работы была попытка составления **частотного словаря**, мы не ставили перед собой задачу проводить тонкую семантическую дифференциацию на основе, скажем, имеющихся толковых словарей русского языка. Кроме того, характер текста позволял предполагать, что в словарях, составленных в основном по письменным источникам и отражающих функционирование письменного русского языка в лучшем случае трех – четырехдесятилетней давности, едва ли могут фиксироваться все лексико-семантические особенности современной, «раскованной», иногда шокирующей речевой «стихии» (весьма показательно, например, что встретившиеся в тексте личные глагольные словоформы *свалит* и *тормозит* употреблены лишь в сленговых значениях, зафиксированных в «Большом словаре русской разговорной речи» В.В. Химика и отсутствующих в академических толковых словарях). К тому же, объем рассматриваемого материала и ограниченные сроки исполнения работы делали такую дифференциацию совершенно нереалистичной. Поэтому сплошное (по тексту) разграничение омонимичных словоформ проводилось в основном путем тщательных повторных просмотров текста и в опоре на лингвистический опыт и интуицию исполнителей работы.

Наиболее трудным, серьезным и, разумеется, дискуссионным стал для данной работы вопрос о выделении и учете в рассматриваемом материале текстовых и словарных единиц, представляющих собой сочетания двух и более отдельных лексических элементов. И собственный речевой опыт лингвистов – авторов данной работы, и давно ставшие классикой наблюдения над русской разговорной речью ее московских и саратовских первооткрывателей, и тонкая интуиция известных исследователей-филологов (см., в частности, понятие коммуникативного фрагмента в [Гаспаров 1996]), и, наконец, вынужденно глубокое погружение исполнителей работы в анализируемый текст – все говорило о высокой степени клишированности разговорной речи, из чего с неизбежностью вытекала необходимость ее отражения в задуманном частотном словаре.

Поэтому в рамках выполнения поставленной задачи встал вопрос о получении некоего списка лексически объединенных форм – аналогов того, что в начале данного раздела было названо «составными словами», а далее будет именоваться рабочим термином «идиоматизмы». Мы сочли необходимым прибегнуть к такому переименованию потому, что, судя по описаниям подходов к созданию НКРЛЯ (см., например, [Баскулина 2006]), понятие «составного слова», возникшее из потребности адекватной лексической интерпретации прежде всего письменных текстов, является, несмотря на отсутствие полного списка таких слов, гораздо более строгим и отрефлексированным, чем то условное и недостаточно определенное с точки зрения наличного

лингвистического инструментария содержание, которое вкладывается нами в термин «идиоматизм». И «составные слова», и «идиоматизмы» – это предположительно единицы ментального лексикона, функционально равные слову и всякий раз воспроизводимые (и воспринимаемые) в речи в виде неких цельных комплексов. Разумеется, в состав «идиоматизмов» входят и «составные слова», но круг первых, на наш взгляд, может и должен быть гораздо более широким – хотя бы потому, что они принадлежат разговорной речи с ее высокой долей автоматизма, следствием чего является стремление говорящих к использованию «готовых» единиц. Ср.: «Говорящий, находясь в условиях непринужденного неподготовленного общения, стремится упростить и облегчить свое «речевое поведение», поэтому он легко и часто прибегает к готовым языковым формулам, в том числе всякого рода клише, шаблонам и стереотипам» [Земская, Китайгородская, Ширяев 1981: 6].

Априорно можно было предполагать, что выделенные нами «идиоматизмы» будут обладать разным языковым статусом и разной функциональной нагруженностью и, скорее всего, не будут жестко отграничиваться друг от друга, образуя своеобразный континуум. На одном полюсе этого континуума – грамматически не изменяемые единицы с внутренними связями такой силы, которая не позволяет вносить в эти единицы какие-либо изменения, на другом – единицы с возможностью грамматического изменения и с ослабленными внутренними связями, что выражается в спонтанной речи как возможностью перестановок образующих их элементов, так и инкорпорированием «посторонних» элементов.

Приступив к задаче непосредственного выделения «идиоматизмов» в анализируемом тексте, мы пытались использовать для формального обоснования этой процедуры все известные и доступные нам правила, критерии и средства. В их число вошли: 1) критерии разделения слова и словосочетания, предложенные В.Б. Касевичем (см. [Касевич 1988] и [Касевич 2006]), критерии разделения слова (=словоформы) и морфемы, предложенные В.А. Плунгяном [Плунгян 2003], критерии И.А. Мельчука, «противопоставляющие словоформу, с одной стороны, частям словоформы (морфам и цепочкам морф), а с другой стороны – группам словоформ (=словосочетаниям)» [Мельчук 1997], и 2) критерий просодической слитности. Из первой группы наиболее полезным и имеющим почти универсальную силу действия оказался для наших целей лишь критерий идиоматичности (неаддитивности); соответствие вычлененных «идиоматизмов» второму критерию (отчасти параллельному выделенному И.А. Мельчуком такому свойству словоформы, как звуковая связанность) было полным. Кроме того, при выполнении данной работы практически неизбежной оказалась сильная опора на субъективный фактор – лингвистическую интуицию и семантическую интроспекцию исполнителей.

В общей сложности в анализируемом тексте было выделено более 900 «идиоматизмов». По критерию идиоматичности полученные нами объединенные формы можно разделить на две группы.

Первая – аддитивные **аналитические единицы**, образующиеся в речи с большой степенью регулярности. В эту группу вошли следующие аналитические формы глагола (следует оговориться, что приводимый далее перечень далеко не полностью соответствует принятым в русистике взглядам, отраженным, в частности, в Грамматике-80): отрицательная форма; темпоральные формы (глагол *быть* + инфинитив; в основном так образуется будущее время); сослагательное наклонение (*б(ы)* + глагольная форма на *-л*); повелительное наклонение (*давай(те)* + инфинитив); желательное наклонение (*пусть/пускай* + личная форма глагола). В эту же группу вошли формы превосходной степени прилагательного (*самый/ая/ое* + прилагательное), формы превосходной степени наречия (типа *больше всего, быстрее всех, меньше всех*) и формы сравнительной степени прилагательного и наречия (*более* + прилагательное или наречие).

Единицы этой группы характеризуются большой регулярностью образования и силой внутренних связей. Хорошим примером этого может служить отрицательная форма глагола, которая образуется регулярно и при этом не допускает ни вставок, ни перестановок (в сочетаниях со вставкой типа *не очень хотел, не он хотел* и т.д. отрицание относится не к глаголу, а к «вставке»).

Во вторую группу вошла большая группа неаддитивных «идиоматизмов» (*грубо говоря, голову морочить, дело за вами, чё-то не так, вроде как, какого чёрта, всё что угодно, по этой причине, таким макаром, ни фига себе, задавай вопрос, вся в меня, ничего/ничё страшного, водишь машину, без проблем, в первых рядах, внештатные ситуации, ничего не поделаешь, чувство юмора, горячие точки, ничё/ничего себе* и т.д.). В отличие от единиц первой группы, они не образуются регулярно и каждый из них уникален по своей структуре. С точки зрения как действия критериев переставимости и вставимости, так характеризующих возможности изменения линейного порядка группы единиц, как и возможности грамматического словоизменения компонентов, эта группа представляет собой пеструю неструктурированную картину. Мы отдаем себе отчет в том, что выделение по меньшей мере некоторых «идиоматизмов», отнесенных к этой группе, может стать объектом дискуссии, в рамках которой главным аргументом авторов в пользу данного решения может служить лишь интуиция, а оправданием в случае доказанной неудачи – абсолютно пионерский характер работы и необходимость «с чего-то начинать».

Основным результатом проделанной работы является пробная версия частотного словаря, полученная в результате компьютерной обработки размеченного текста по программе, созданной старшим научным сотрудником лаборатории моделирования речевой деятельности СПбГУ (руководитель – проф. В.Б. Касевич) канд. биол. наук А.В. Венцовым. Этот словарь включает в себя всего 4778 расположенных в алфавитном порядке и снабженных индексом частотности единиц.

Полученная версия частотного словаря, разумеется, не может претендовать на какую-либо даже весьма условную полноту охвата реального разговорного словника современной русской речи – как из-за ограниченного объема анализируемого материала, так и из-за достаточно сильной тематической привязки рассмотренного текста к событиям, происходившим в реалити-шоу *Большой Брат*. Однако мы надеемся, что в качестве некоего первого опыта, словарь может оказаться полезным прежде всего с методической точки зрения, а принципы, положенные в основу его составления, могут стать подготовительным материалом для дальнейших исследований в этой области. Ближайшей задачей для нас является составление подобного же словаря (а затем и получение «объединенного» частотного словаря) и для записей телепрограммы *Культурная революция*.

Ниже в качестве образца приведен небольшой фрагмент полученного частотного словаря: в столбцах цифр направо от словарных единиц указаны индексы частотности, межсловное нижнее подчеркивание соединяет в единое целое «идиоматизмы», цифра, соединенная со словарной единицей нижним подчеркиванием, соответствует грамматическому / лексическому маркеру, маркирующему, падеж (в случае омонимии падежных форм) а также некоторые особенности глагольных форм. Знаком + обозначен ударный гласный.

вошли	1	всё_ещё	1
вперёд	4	все_за_одного	1
впереди	1	всё_необходимое_3	3
впрочем	1	всё_нормально	2
вращаться	1	всё_равно	11
вре+менном_5	1	всё_хорошо	1
времени_1	7	всё_что_угодно	1
время	13	всё-всё-всё	1
время_3	8	всегда	12
врёшь	5	всего_1	1
вроде_как	1	всего_2	1
вру	1	всего_лишь	1
вряд_ли	1	всего_хорошего	1
все	35	всём	2
всё	78	всем_02	17
все_03	5	всем_4	1
всё_1	20	всеми	4
всё_2	27	всё-таки	5
всё_3	1	всё-то	1
всё_3	1	всех	9
всё_время	1	всех_01	6

## Литература

1. Баскулина Ю.Н. «Составные слова» в современном русском языке (на материале Национального корпуса русского литературного языка // Материалы XXXV Международной филологической конференции. 13-18 марта 2006 г. Вып. 22. Секция общего языкознания. Ч. 2. С.17-23.
2. Венцов А.В., Касевич В.Б. Проблемы восприятия речи. М., 2003.
3. Гаспаров Б.М. Язык. Память. Образ: Лингвистика языкового существования. М., 1996.
4. Горбова Е.В., Слепокурова Н.А. и др. Предварительные результаты мониторинга современной русской устной спонтанной речи // Современная русская речь: состояние и перспективы. Часть II. СПб, 2006.
5. Земская Е.А., Китайгородская М.В., Ширяев Е.Н. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. М., 1981.
6. Касевич В.Б. Семантика. Синтаксис. Морфология. М., 1988.
7. Мельчук И.А. Курс общей морфологии. Т. 1. М.; Вена, 1997.
8. Плунгян В.А. Общая морфология. Введение в проблематику. М., 2003.