А. Мустайоки из университета Хельсинки. Они тоже создают небольшого объема корпус, тщательно размеченный синтаксически и морфологически; предлагаемая ими система разметки опирается на теорию функционального синтаксиса проф. А. Мустайоки. В заключение в статье М. Копотева дается подробное описание основных корпусных проектов, которые ведутся в настоящее время в Финляндии — одного из форпостов корпусной лингвистики в мире.

Е. В. Рахилина, С. А. Шаров

УДК 811.161.1'1:001.103

Л. А. Вербицкая, Н. Н. Казанский, В. Б. Касевич

Некоторые проблемы создания национального корпуса русского языка*

Описываются подходы к созданию Национального корпуса русского языка как репрезентативного многофункционального корпуса, который должен послужить базой для реализации, как минимум, следующих задач: подготовка новых академических словаря и грамматики русского языка; создание семейства частотных словарей; моделирование восприятия и порождения речи; разработка пособий и справочников по русскому языку. Единицы хранения корпуса — от лексемы до полных собраний сочинений избранных авторов. Система управления корпусом как базой данных и программа конкордансов позволят извлекать нужную информацию по любому набору параметров. Источники, прежде чем стать единицами хранения, будут подвергаться тщательной филологической экспертизе.

О КОРПУСНОЙ ЛИНГВИСТИКЕ

Э. Брукхейзен в рецензии на труды международной конференции "Корпусная лингвистика и лингвистическая теория" пишет: "Лингвистический анализ компьютеризированных текстовых корпусов, который [еще недавно] был занятием маргинальных (и обычно именно так воспринимаемых) исследовательских групп, передвинулся [ныне] в центр исследований в области английского языка. В ходе этого процесса получены впечатляющие результаты, которые, помимо и "сверх" их интереса для англистов, вынуждают нас переосмыслить, причем кардинально и систематически, проблемы лингвистической теории" [1].

Действительно, корпусная лингвистика переживает своего рода "бум". Начало было положено еще в 60-е гг., когда был создан Корпус английского языка Университета Брауна (Brown Corpus of the English Language). Сейчас к нему прибавилось целое семейство английских корпусов: The International Computer Archive of Modern/Medieval English, и British National Corpus, и National American Corpus, и др., объемом от 1 млн до 7,5 млн словоупотреблений. В Германии известен LIMAS-Korpus (1 млн единиц), во Франции с начала 60-х гг. создавался корпус Trésor de la langue française (90 млн словоупотреблений). Достаточно активны и ученые Нидерландов, Японии, Румынии, Индии и ряда других стран. В Чехии силами специально учрежденного для этой цели государственного института под руководством проф. Фр. Чермака создан Корпус чешского письменного языка объемом в 100 млн словоформ (создание корпуса рассматривалось *как национальный проект*). В ноябре 2000 г. этот корпус был представлен научной общественности.

Институализации этой отрасли деятельности содействует целенаправленная подготовка специалистов в области корпусной лингвистики. Так, в Бирмингемском университете, одном из ведущих центров корпусной лингвистики, в 2001 г. открылась кафедра корпусной лингвистики и объявлена программа подготовки по данной специализации. Издается специальный журнал "International Journal of Corpus Linguistics" (в 2002 г. вышел его 7-й том), создана соответствующая международная ассоциация.

К большому сожалению, в России до сих пор нет крупномасштабных многофункциональных корпусов русского языка. Работа над Машинным фондом русского языка (руководитель В. М. Андрющенко) [2] была прекращена по причинам внешнего порядка; фонетический фонд русского языка (руководитель Л. В. Бондарко) ориентирован преимущественно на решение задач в области фонетики и фонологии. Даже литература по корпусной лингвистике на русском языке крайне немногочисленна (к работам, ссылки на которые содержатся в данной статье, можно добавить публикации [3; 4] и, возможно, еще несколько).

В настоящее время имеется несколько компьютерных массивов русских письменных текстов, разного объема и в разной степени лингвистически обработанных. Это — Корпус СПбГУ (руководитель В. Б. Касевич, более 1 млн словоупотреблений) [5], корпус ИЯ РАН "Русский стандарт" (руководитель В. А. Плунгян, ок. 600 тыс. словоупотреблений [6]), "Динамический корпус текстов по современной публицистике (90-е гг.)" — авторская группа: А. Н. Баранов, М. Н. Михайлов, Г. О. Сидоров

*Грант РГНФ № 03-04-00226а

[7]. Нужно также отметить Уппсальский корпус (около 1 млн словоупотреблений), очень тщательно разработанный, на базе которого создан лучший на сегодня частотный словарь русского языка [8].

Существует **настоятельнейшая потребность** создания полномасштабного многофункционального **Национального корпуса русского языка**. Только такой универсальный корпус может поставить на прочную основу филологические и лингвистические исследования в первую очередь в Российской Федерации, а также и в других странах, где существуют развитые традиции в области русистики.

Особенно насущной становится необходимость в Национальном корпусе в условиях, когда российское общество ищет опору в решении проблемы идентичности, что невозможно вне контекста языка — фундамента любой культуры. Чрезвычайно важны и проблемы, связанные с поддержанием русского языка в странах СНГ и в русской диаспоре дальнего зарубежья; решение этих проблем также должно опираться на постоянный мониторинг языкового существования в соответствующих ареалах в соотнесенности с "языковой метрополией", что также невозможно без специальной службы, ориентирующейся на материалы Национального корпуса русского языка.

Даже отвлекаясь от социокультурных проблем, приходится признать, что традиционные способы сбора материала для исследования лингвистических и филологических задач не являются более адекватными. До недавних пор у исследователей реально не было возможности, в силу огромной трудоемкости сбора языковых данных, накапливать достаточно значительные по объему массивы "примеров". Весьма затруднена была и практическая работа с языковым материалом, зафиксированном на традиционных (бумажных) носителях.

Именно эти проблемы и решает прежде всего корпусная лингвистика. Репрезентативность корпуса обеспечивает репрезентативность построенных на основе корпуса словаря и грамматики. Хотя теоретических критериев репрезентативности корпуса на сегодня не существует, эмпирические исследования склоняют к представлениям, согласно которым репрезентативным можно считать корпус в 10–20 млн словоупотреблений (разумеется, при условии определенным образом организованного отбора текстов, см. об этом ниже). Огрубляя, можно утверждать, что если в представительном корпусе некое языковое явление отсутствует, то оно отсутствует в национальном (литературном) языке как таковом[1].

Опора на корпус позволяет избежать традиционных для лингвистических работ формулировок наподобие "у *некоторых* слов гр. 1а в употреблении отмечаются ненормативные формы [род. мн.] на-*ов*..." ([9, с. 499]; выделено нами.— *Авт.*) или "В данном типе употребления глаголов несов. вида выделяются два ряда случаев. ...2) Сообщается о действии (*обычно* отношения и состояния, но *иногда* и действии, связанном с определенным изменением, развитием)..." ([9, с. 611]; выделено нами.— *Авт.*). Информация, содержащаяся в корпусе, даст возможность исчерпывающим образом перечислить все случаи, которые подпадают под действие соответствующего правила, составляют, наоборот, исключение из правила, характеризуются отклонениями в содержательной (семантической) интерпретации и т. п.[2] Особенно велик потенциальный вклад корпуса в исследование лексики, словаря, лексических правил, поскольку эти аспекты в большей степени, нежели грамматические, предполагают максимальный охват текстового материала. Одновременно стоит заметить, что исследование словаря вообще в современной лингвистике выходит на первый план, ибо вокабула ментального лексикона "несет с собой" и характеризующую ее грамматическую информацию, что уместно воспроизводить в лингвистическом описании-модели, так что представление словаря в пределе стремится к представлению языка как такового (см. об этом [10], а также ниже).

Именно и только наличие репрезентативного Национального корпуса сделает задачу составления академического словаря, академической грамматики современного языка, частотного словаря, различных справочников, высококачественных учебников действительно выполнимой. Отдельно отметим важность — точнее, абсолютную необходимость — такого Корпуса для построения моделей речевой деятельности, центральным компонентом которых также выступает словарь.

Говоря об академических словаре и грамматике, мы ставим исследовательскую задачу в рамках *прескриптивного* подхода к лингвистическому описанию, что предполагает филологический и лингвистический отбор материала. Но нельзя отказываться и от *дескриптивного* подхода, когда лингвист покидает почву строгого нормативно-оценочного отбора материала и учитывает фактически все явления, которые соответствующий социум не рассматривает как прямую ошибку[3].

В середине 60-х гг. И. И. Ревзин [12] предлагал некоторую формализацию понятия "отмеченная фраза" (которое можно и нужно расширить до понятия "отмеченного текста"). Здесь не место обсуждать этот формализм. Исходные содержательные предпосылки рассуждения Ревзина просты и самоочевидны. Начало любого лингвистического исследования — это традиционный "сбор материала". Поскольку лингвист осуществляет не просто сбор, но и отбор материала, фактическим фундаментом лингвистического анализа оказывается *стилистика*: именно стилистика определяет приемлемость/неприемлемость языковых объектов, причем скорее как континуум, нежели дихотомию, с пометами относительно контекстной привязанности каждой из градаций в поле "приемлемость/неприемлемость". "Поскольку, однако, функция, не выполняемая современной стилистикой или выполняемая лишь частично единичными лингвистами, занятыми вопросами "культуры

[1] Мы исходим из задачи создания корпуса *литературного* языка (включая и его устное бытование) как задачи первоочередной, хотя, разумеется, возможны и нужны диалектные, иные типы корпуса.

[2] Традиционные формулировки типа *иногда, в некоторых случаях*, "в большинстве случаев", "редко" и т. п. не всегда означают, что автор не обладает полнотой информации: информация может быть и доступной, но перечисление всех реальных случаев просто перегрузило бы текст книги или статьи. Компьютерный корпус может решить эту проблему и тем, что он "раз и навсегда" устраняет трудности, связанные с допустимым объемом описания, и тем, что он дает возможность *иерархизированного* обращения к информации в зависимости от требуемой степени подробности, частотности соответствующих явлений и т. п., что поддается формализации с использованием системы опций.

[3] Различение указанных подходов, по-видимому, примирит тенденции, которые в литературе противополагают как corpus-based и corpus-driven ориентации [11].

речи", необходима не только для развития науки, но и для функционирования языка в обществе, то она выполняется художественной литературой, вернее — некоторое группой писателей, которые становятся своего рода жрецами, хранителями понятия языковой нормы, книги которых повсеместно признаются собранием правильно построенных, образцовых фраз" ([12, с. 5]).

Метафора Ревзина относительно "жреческой" роли писателей на самом деле должна быть скорректирована. Писатель не "хранит" норму (не говоря уже о "понятии о норме") — он ее *создает*. Поэтому функция писателя скорее сближается с таковой "демиурга", "жрецом" же выступает *лингвист*: филолог, социолингвист, специалист в области стилистики. Именно эти специалисты-профессионалы выделяют круг текстов, которые они (специалисты) расценивают как воплощение нормы; обычно уже другие специалисты (грамматисты, лексикологи) используют полученные таким образом тексты для построения равно инвентаря и системы лексических/грамматических правил языка.

Применительно к задаче создания Корпуса, из сказанного выше следует, что его формированию должна предшествовать *филологическая экспертиза* текстов, которые могли бы войти в состав Корпуса. Прежде всего, необходима сплошная паспортизация всех текстов/фрагментов текстов, вошедших в Корпус. В паспорте текста/фрагмента текста следует приводить реквизиты издания (место, издательство, год, какое издание — первое, второе и т. д., прижизненное или нет, наличие/отсутствие редактора и имя последнего — возможно, также и корректора, страницы и т. д.; полный список требований к паспорту текста еще предстоит выработать). Если источник текста — сетевые ресурсы, прежде всего Интернет, необходимо указать адрес соответствующего сайта, дату помещения на сайт, информацию о том, имеет ли электронная версия прототип на бумажном или каком-либо ином носителе и, если да, информацию о прототипе, о его характеристиках и о соотношении разных версий, дату последней редакции электронной версии текста и т. п.

Если корпус содержит фрагмент текста, а не весь текст, то такой фрагмент должен обладать формально-семантической целостностью. Как верно отмечает А. Н. Баранов, такой фрагмент "не должен содержать неоднозначности любых типов, в частности, местоимений, для которых невозможно установить антецедент и пр. В тех случаях, когда единицы хранения включают случаи языковой игры, связанной с неоднозначностью, рамки контекста должны быть таковы, чтобы пользователь мог легко определить, что речь идет о языковой игре, а не об ошибке в вычленении единицы хранения" [13, с. 119].

Важнейшее требование репрезентативности текстов предполагает, что состав корпуса должен быть каким-то образом сбалансирован по жанрам: в нем должна быть представлена художественная проза, драма (в известной степени моделирующая разговорную речь), научная проза, деловая проза, общественно-публицистическая проза, эпистолярная проза, записи устной речи, поэзия; вероятно, это не полный перечень, не говоря уже о возможности выделения релевантных подтипов, ср., например, язык научно-популярной литературы, который находится "на пересечении" художественной, научной и публицистической литературы, или языковую специфику детской литературы. Что касается самих по себе жанров, то здесь, опять-таки, задача стилистики — предложить систему жанров и определить критерии отнесенности конкретных текстов к тому или иному из жанров. Что касается количественной представленности в Корпусе текстов, принадлежащих разным жанрам, то приходится признать, что эта проблема плохо поддается решению. Ясно, что осмысленно ее можно ставить только применительно к условно выделенным группам носителей языка; например, тексты, в изобилии содержащие терминологию конкретных наук, не являются ни продуктом порождения, ни, нормально, объектом восприятия абсолютного большинства носителей языка, их лексические и, отчасти, грамматические особенности не соотносимы с языковыми системами "средних" говорящих/слушающих. Вероятно, не столь уж малочисленна группа носителей языка, которые практически не сталкиваются с поэтическими текстами, особенно если речь идет о современном верлибре, о поэзии концептуалистов и т. д. и т. п. В существующем на сегодня Корпусе СПбГУ количественно в равной степени представлены тексты художественной литературы, драмы и публицистические тексты (всего 181 автор при общем объеме корпуса в 1.031.920 словоупотреблений). Разумеется, это сугубо временное, условное решение — над проблемой жанровой сбалансированности Корпуса нужно работать отдельно; насколько нам известно, для существующих корпусов универсального характера эта проблема не только не решена, но даже и не поставлена в сколько-нибудь явном виде[4].

Каждый текст (из отобранных) будет представлен в Корпусе в нескольких версиях. Во-первых, это исходная версия, максимально полно воспроизводящая оригинал — вероятно, даже с ошибками и опечатками, которые тоже представляют определенный интерес. Во-вторых, это рабочая версия, в которой произведены следующие модификации: устранены опечатки и очевидные ошибки[5]; последовательно введена буква *ё*, которая в абсолютном

[4] Здесь можно видеть параллель к проблеме отбора множества *языков* для получения типологически релевантных выводов. Как известно, для типологических нужд используются достаточно определенные критерии, применение которых приводит к представлениям о том, что 50-ти языков (из общей численности в 6–7 тыс. языков мира) достаточно для выведения лингвистических генерализаций, если ни одна произвольная пара языков множества не связана ни генетически, ни ареально [14, 15]. Парадоксальным образом оказывается, что установить критерии отбора материала для всех языков относительно проще, чем решить сходную задачу для одного-единственного языка.

[5] Нужно отдавать себе отчет в том, что определение "очевидности" ошибки — не вполне тривиальная задача. Например, Ю. Д. Апресян справедливо указывает, что фраза из газетного репортажа *Преступники угнали несколько государственных и собственных машин* семантически аномальна, поскольку получается, что "преступники обокрали себя угнав свои собственные машины" [16, с. 11]. Высказывание аномально с точки зрения литературного языка (прескриптивно-аномально) — но нельзя исключить наличие некоторого идиома, в котором у прилагательного *собственный* есть значение 'частный', и в этом случае фраза будет признана для данного идиома дескриптивно адекватной, не аномальной. Полностью отбрасывая такие факты, мы рискуем утратить представления о развитии языка, пример — признание еще недавно ненормативного употребления слова *эпицентр* (*в эпицентре внимания* и т. п.) в последних словарях русского языка.

большинстве публикаций не используется; во всех словах проставлены ударения, которые нормально присутствуют лишь в некоторых учебных текстах; знаки переносов повсеместно удалены; все сокращения развернуты и заменены их полными исходными прототипами; то же действительно для встречающихся в текстах цифр и, вероятно, формул. Не очень понятно, что делать с передачей на письме попыток использования графических соответствий фонетических особенностей произнесения, ср. *у-у-у как хо-о-о-лодно* или *р-р-революционный!* Равным образом нет универсального решения для слов и выражений иностранного происхождения (не ассимилированных). Например, целые абзацы французской речи в текстах русских классиков XIX в., скорее всего, должны просто опускаться (хотя в исходной версии они должны сохраняться; возможно, для определенных задач типологии текстов, изучения переключения кодов варваризмы должны быть представлены и в рабочей версии). В то же время использование выражений наподобие *in situ* и даже *deadline* есть основания признавать узусом для "интеллигентской" речи и, стало быть, нужно найти способ передачи этого в текстах Корпуса.

Придется признать права гражданства для ненормативной (табуированной) лексики (которая, увы, изобильно представлена в современных публикациях) — по крайней мере, для дескриптивно ориентированных задач; в том числе, придется развертывать до полных исходных вариантов те табуированные слова, где вместо части букв используются многоточия. Тем более сохраняются в Корпусе просторечия и вульгаризмы.

Наряду с указанными, должны существовать и другие версии — прежде всего, разные варианты транскрибированных текстов, где будут использоваться разные типы транскрипций в зависимости от решаемых задач (подробнее см. статью А. В. Венцова, В. Б. Касевича, Е. В. Ягуновой в данном номере).

ЛИНГВИСТИЧЕСКИЕ И ФИЛОЛОГИЧЕСКИЕ ПРИМЕНЕНИЯ КОРПУСА

В данном разделе речь пойдет о возможных расширениях сферы корпусной лингвистики — о ее включении в более широкую филологическую сферу. Мы имеем в виду прежде всего возможность использования Корпуса для литературоведческой работы и издания литературных текстов, в частности, академического издания классических текстов. Для корпуса как такового, предназначенного для лингвистических нужд, конечно, не имеет особого смысла включать *полные* тексты, например, Л. Толстого или М. Булгакова — из них будут представлены извлечения, отдельные рассказы. Но кажется очевидным, что современное академическое издание писателя, поэта, философа практически невозможно без введения в компьютер всех массивов текстов, релевантных для создания академического собрания сочинений соответствующего автора. Именно компьютерная версия позволит сделать максимально эффективным сравнительное изучение редакций, версий и т. п. Чрезвычайно ценно использование на материале Корпуса существующих программ построения конкордансов, особенно программ, позволяющих получать окружение заданного элемента, обычно слова, с заданными же параметрами — левое, правое, левое и правое, однословное, двусловное окружение и т. д. Разумеется, столь же ценна эта возможность и для лингвистики.

Обладая компьютерными версиями текстов, литературоведу, текстологу гораздо легче снабжать сочинения соответствующих авторов указателями, комментариями и иным справочным и служебным аппаратом. Точно так же становятся весьма простыми такие задачи, как составление словарей рифм того или иного поэта, получение всевозможных статистических данных (например, относительно тех же рифм), изучение математических закономерностей текстов, их метрики и ритмики и т. д., и т. п. Существенно облегчается задача атрибуции анонимных и псевдонимных текстов.

Разумеется, лингвисту корпус (и только корпус) дает, прежде всего, надежную основу для выполнения двух своих первоочередных профессиональных задач: составления словаря и составления грамматики соответствующего языка; обладая достаточно богатым набором грамматик и словарей разных языков (50-и?), уместно проверить существующие положения общего и типологического языкознания — и вывести новые.

КОРПУС И СЛОВАРИ

Уже говорилось выше о том, что современный словарь любого языка мыслим лишь на базе репрезентативного корпуса. Это относится, разумеется, к словарям всех типов. Кажется, пока нет еще практики привлечения к составлению переводных словарей корпусов *параллельных текстов*, но эту возможность тоже необходимо иметь в виду (о корпусах параллельных текстов см., например, [7]).

Важно отчетливо сознавать, что адекватному описанию языка отвечает не один словарь, а некоторое *семейство* словарей. Мы имеем в виду далеко не только установленный Л. В. Щербой набор из четырех переводных словарей для произвольной пары языков, где эта четверка определяется на основании двух двухзначных признаков, первый из которых носит семантический характер, а второй — прагматический: "входной/выходной язык" и "пользователь словаря — носитель входного языка/пользователь словаря — носитель выходного языка". Еще одна важнейшая идея Щербы должна быть генерализована, также с получением результата, который заключается в умножении семейства словарей. Мы имеем в виду знаменитое предложение Щербы различать активную грамматику — грамматику порождения речи и пассивную — грамматику восприятия речи. Есть все основания распространить эту оппозицию на словарь, различая, соответственно, словарь порождения и словарь восприятия речи. Главное отличие последнего — это признание вокабулой такого словаря не лексему (вернее, лемму, словарную форму лексемы), а *словоформу* [5; 18].

Один из важнейших типов словарей — это, конечно, частотный словарь[6]. Но и здесь мы видим необходимость различения нескольких подтипов словарей. Если выделять, как предлагалось выше, словарь восприятия речи, или перцептивный словарь, признавая его вокабулой словоформу, то и частотностью для перцептивного словаря будет частотность словоформ. Построение такого словаря традиционно выступает для всех авторов существующих частотных словарей лишь начальным предварительным этапом, неизбежным, поскольку в тексте представлены и доступны статистическим операциям именно словоформы, но, скорее, служебным: после составления словаря словоформ проводится операция лемматизации — сведения словоформ в лексемы, и словарная лемма получает кумулятивную частотность соответствующих словоформ. Однако с точки зрения восприятия речи именно словарь словоформ выступает основным.

Применительно к задачам восприятия речи диверсификация словарей не исчерпывается выделением словаря словоформ. Приходится выделять *поверхностное* восприятие как особый тип восприятия речи, при котором семантика играет ограниченную роль, перцептивная обработка текста носит достаточно механический характер (см. об этом [19]). В целом ряде случаев словарь, обслуживающий поверхностное восприятие речи, будет отличаться от словаря для "полного" восприятия, причем прежде всего именно с точки зрения частотных характеристик. Покажем это на простых примерах (ср. [5]).

В существующих частотных словарях русского языка содержится, естественно, слово *друг*, принадлежащее к высокочастотным слоям лексики. Однако в них отсутствует сочетание *друг друга*, из чего следует с очевидностью, что словоформы *друг* и *друга* из этого сочетания "отдали" свою частотность слову *друг*. В этом можно видеть двойное искажение языковой реальности: с синхронной точки зрения, компоненты указанного сочетания не имеют никакого отношения к слову *друг*, ср. семантически безупречное высказывание *Они ненавидят друг друга;* иначе говоря, частотность слова *друг* выступает **завышенной** — при том, что частотность вполне "легитимного" идиоматического сочетания (аналитического местоимения-анафора) *друг друга* оказывается просто **нулевой**. То же самое можно сказать о всех фразеологизмах, идиоматических сочетаниях и их компонентах, ср. *повесить нос* и т. п.

Вместе с тем подход, при котором раздельно учитывается словоформа *друга* из контекста наподобие *У меня нет друга лучше, чем Иван* и та же словоформа из контекста типа *Иван и Петр ненавидят друг друга*, отражает не всю реальность. Если исходить из описания и моделирования **поверхностного** восприятия речи, то приходится принимать во внимание его известную механичность, уже упоминавшуюся выше, когда, скорее всего, не различаются сочетания *друг друга* из высказываний, например, *Друг друга хорошо знает* и *Друг друга хорошо знают;* для этого требуется анализ согласовательных характеристик глагола, но именно такого рода анализ едва ли представлен в процедурах поверхностного восприятия. Если это так, то приходится допустить два типа словарей: в одном, рассчитанном на поверхностное восприятие речи, сочетание *друг друга* в качестве отдельной вокабулы отсутствует (именно такова ситуация в существующих частотных словарях); в другом, призванном обслуживать "полное" восприятие, включены в качестве отдельных вокабул и словоформы *друг, друга*, и сочетание *друг друга*.

КОРПУС И ГРАММАТИКА

В современной бурно развивающейся корпусной лингвистике огромное внимание уделяется грамматическим характеристикам слов (иногда и иных единиц), входящих в корпус и в словарь, построенный на базе корпуса. Для эффективного использования и корпуса, и словаря — в частности, в моделях речевой деятельности — соответствующие единицы должны быть грамматически интерпретированы[7]. Обычно в качестве первого шага избирается присваивание этим единицам помет частеречной принадлежности. Появляющиеся время от времени (главным образом в Интернете) сообщения о разработках универсальных программ автоматического определения частей речи в тексте (Language-Independent Taggers) вызывают серьезные сомнения. Хотя в языках типа русского, казалось бы, можно с достаточной степенью надежности установить частеречную принадлежность слов по их формальным признакам (ср. знаменитую *глокую куздру* Щербы), эти возможности не следует переоценивать. Даже отвлекаясь от того, что формальные признаки данной словоформы конкретного высказывания нередко дают лишь вероятностную характеристику[8], мы должны констатировать, что в реальном тексте встречается огромное множество слов, частеречная принадлежность которых вовсе не очевидна даже в условиях исчерпывающего знания их семантики,

[6] Строго говоря, лишь с некоторой прагматической точки зрения частотный словарь является самостоятельным типом словаря. Частотность — неотъемлемая характеристика слова, словоформы и других объектов лингвистического анализа, она должна быть частью, соответственно, присловной информации в словаре, наряду с информацией о частеречной принадлежности и т. д., так что вынесение этой характеристики в особый словарь выступает, скорее, как технический прием (равным образом можно было бы говорить о словаре существительных, глаголов и т. п. — просто это менее оправданно с прагматической точки зрения).

[7] В последнее время сами средства аннотирования текстов (как обычно называют присваивание словам и иным единицам индексов грамматической информации) принято организовывать в виде особых баз данных (treebanks).

[8] Даже в хрестоматийной *глокой куздре*, если отвлечься от интонации и пунктуации, вполне возможна интерпретация слова *глокая* как деепричастия. Ср. другой пример на материале "нормального" высказывания *Шаланды, полные кефали...*: если воспринимающий это высказывание не знает слова *кефаль*, ничто не мешает ему интерпретировать это слово как глагол прош. вр. (ср. *шаланды полные сверкали, бежали* и т. п.). Можно вспомнить и эпизод из истории работы над машинным переводом, когда в одной из программ слово *порося* "для простоты и удобства" трактовалось как деепричастие.

рфологических свойств и синтаксических функ-
й. Это относится к многочисленным случаям не-
ределенности решения в пользу то ли существи-
льного, то ли прилагательного (ср. *Да она про-
го сумасшедшая!*), к вводным словам наподобие
нечно, которые чисто механически относят к на-
чиям, хотя они не выполняют каких бы то ни бы-
наречных функций, и т. д. (Нельзя не отметить
есь особо, что работа с корпусами поистине бес-
нна для лингвиста уже потому, что она просто
нуждает его искать объяснение **каждому** упо-
еблению слова, конструкции и т. д., поскольку
опуск любой характеристики из предусмотрен-
х автоматически приведет к сбою в работе со-
ветствующих программ. Во многом именно это
стоятельство и заставляет нас "переосмыслить,
ичем кардинально и систематически, проблемы
нгвистической теории", по словам Брокхойзена,
иведенным в начале статьи.) Неопределенности
определении частеречной принадлежности, упо-
нутые выше на примере проблемы "существи-
льное или прилагательное?", как можно наде-
ься, удастся снять путем привлечения множе-
ва релевантных контекстов, "поставляемых" ре-
езентативным корпусом — хотя, вероятно, нель-
отрицать и своего рода прагматического реше-
я, когда все решает эффективность обращения
системе аннотирования при моделировании то-
или иного аспекта речевой деятельности. Ведь
сти речи (как и прочие виды классификации), с
нгвистической точки зрения, не самоценны, это
его лишь способ краткого указания на системы
авил, ограничений, функций, присущих данной
ксеме при ее использовании в речевой деятель-
сти. В этом смысле соотношение признаков при-
длежности к части речи и самих по себе частей
чи обратимо: признаки служат основанием клас-
фикации, но именно к ним и обращаются соот-
тствующие программы, когда алгоритм предпо-
гает использование информации о частеречной
инадлежности слова, см. [20].

На примере определения частеречной принад-
жности мы можем видеть принципиально **ре-
рсивный** принцип работы лингвиста в обла-
и корпусной лингвистики. Нет сомнения, что уже
иступая к этой работе, лингвист должен обла-
ть некоторой концепцией — вернее даже, некото-
м набором концепций — относительно тех лекси-
ских и грамматических явлений, которые подле-
ат категоризации в рамках корпуса и основанных
корпусе моделей лингвистического описания. В
мках этих концепций, в терминах соответству-
щих категорий должен быть оценен, помечен ка-
ый из элементов корпуса, словаря и т. д., если
п этого элемента принят в качестве структурной
иницы корпуса и моделей языка, текста, речевой
ятельности. Но в процессе этой работы, прежде
его под "давлением" материала, предоставленно-
корпусом, с неизбежностью выявятся недочеты,
обелы предварительно принятых концепций, ко-
рые должны привести к их модификации или за-
не, за чем последует новый раунд категоризации
ементов словаря — уже в терминах, отвечаю-
их пересмотренным концепциям. Ничто не дает
рантии, что число таких раундов не окажется
сконечным...

Что же касается технологии аннотирования
рпуса, словаря, осуществляемого в терминах ча-
ей речи и иных категорий, то реален полуав-
матический метод (который также нормально
является рекурсивным). При таком подходе, как
и при акцентуации текстов, пометы частей речи
проставляются вручную в словаре, а затем уже
программным способом распространяются на все
тексты корпуса. В принципе **предварительную**
частеречную разметку словаря можно осуществить
и по формальным признакам слов (для разных язы-
ков это задача разной степени выполнимости), но
пока таких программ в нашем распоряжении нет.

Помимо присвоения словам частеречных помет,
помет подклассов слов (валентностные классы гла-
голов и т. п.), необходимо и аннотирование в тер-
минах формообразовательных категорий. Обладая та-
кой разметкой, мы можем получить явно не лишен-
ную интереса информацию о, например, частотно-
сти отдельных падежей (кажется, в литературе нет
соответствующих сведений) — не говоря уже о том,
что формообразовательная информация абсолютно
необходима для моделирования порождения и, в не-
сколько меньшей степени, восприятия речи.

Любопытно, что применительно к рассматрива-
емым задачам может оказаться в известном смысле
иррелевантным выбор между решениями, предла-
гаемыми разными авторами. Примером может слу-
жить проблема причастий и деепричастий. Хотя
сейчас, кажется, абсолютное большинство русис-
тов считает причастия и деепричастия глагольны-
ми формами, а не словами особых частей ре-
чи (альтернативный вариант решения), в Корпусе
СПбГУ причастие и деепричастие выступают "на-
равне" с существительным, глаголом и т. д. Дело
в том, что при принятой установке на **комплекс-
ное** использование классификационной (части речи
и пр.) и формообразовательной информации поме-
ты "причастие" и "деепричастие" нетрудно интер-
претировать как означающие "причастная форма
глагола" и "деепричастная форма *глагола*", соот-
ветственно. Иначе говоря, данные пометы, являясь
формообразовательными, поглощают информа-
цию о классификационной отнесенности, что дела-
ет систему помет-дескрипторов более экономной.
Внешне такая система помет может выглядеть как
возвращение к позиции вычленения причастия и
деепричастия в качестве самостоятельных частей
речи, но в действительности мы здесь имеем дело,
как уже отмечено выше, с принципом комплексного
и иерархического использования информации, пе-
редаваемой системой дескрипторов.

Само собой разумеется, в данном разделе затро-
нуты лишь некоторые из великого множества про-
блем грамматики, которые возникают перед разра-
ботчиками-конструкторами корпуса. Словарь, по-
строенный на основании корпуса, безусловно дол-
жен быть толково-комбинаторным и включать все
зоны, предусмотренные словарем этого типа в мо-
дели "Смысл ⇔ Текст" [21; 22]. Соединение идей
этой модели и возможностей, которые предста-
вит репрезентативный корпус, должны дать каче-
ственно новые результаты.

КОНСОРЦИУМ ПО СОЗДАНИЮ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

Представляется уместным в данной публика-
ции кратко изложить конкретные планы по созда-
нию полномасштабного репрезентативного много-
функционального корпуса — Национального кор-
пуса русского языка.

Для реализации этих планов в настоящее время (2003 г.) создан Консорциум, в который вошли Санкт-Петербургский государственный университет, Институт лингвистических исследований РАН, Институт языкознания РАН, Институт русского языка РАН и Всероссийский институт научно-технической информации РАН. Координатор проекта, для выполнения которого сформирован Консорциум — Л. А. Вербицкая.

В рамках Консорциума запланирована работа пяти исследовательских групп, которые совместными усилиями будут вести работу по созданию Корпуса, отвечая каждая за один из аспектов общего задания.

Группа 1. Группа общих проблем лингвистики корпуса (руководители Н. Н. Казанский, В. А. Плунгян) должна определить параметры Корпуса: типы единиц Корпуса (соотношение текстов и фрагментов текстов); структуру и объем Корпуса (деление Корпуса на подкорпусы, структуру такого деления, определение критериев включения/невключения единиц хранения в Корпус, разработку схемы соотношения единиц хранения по жанрам); определение уровня открытости/закрытости Корпуса; структуру и формат единиц Корпуса (определение глубины и тип разметки, соглашение о формате представления).

Группа 2. Группа филологических проблем формирования Корпуса (руководитель Н. Н. Казанский) отвечает за общий список включаемых в Корпус авторов, произведений и филологическую экспертизу текстов.

Группа 3. Группа лексико-грамматических проблем (руководители В. А. Виноградов, В. Б. Касевич, В. А. Плунгян, Е. В. Рахилина) должна предложить систему лексических и грамматических категорий, используемых для разметки текста (части речи, валентностные классы, члены предложения/синтаксемы и т. д.), а также наметить общие контуры концепции создания семейства словарей и грамматик на базе Национального корпуса русского языка.

Группа 4. Группа алгоритмов речевой деятельности (руководители А. В. Венцов, В. Б. Касевич) должна разработать представления об основных процедурах речевой деятельности, используемых человеком с опорой на словарь (лексикон).

Группа 5. Группа учебников и учебных пособий (руководители С. И. Богданов, А. М. Молдован) начинает работу над созданием (на базе Корпуса) нового семейства учебников и пособий по русскому языку, базирующихся на надежной фактической основе и созданных при участии профессиональных лингвистов (специалистов по грамматике, стилистике, этимологии, социолингвистике и пр.).

Понятно, что в настоящей статье затронуты "грубыми мазками" лишь некоторые — как и значится в названии — проблемы, возникающие при решении такой сложнейшей проблемы, как создание Национального корпуса русского языка. Мы оставили в стороне все, что выходит за пределы собственно лингвистики и филологии, в том числе важные вопросы программного обеспечения (см. об этом, например, [23]). Полностью исключены из рассмотрения проблемы фонетических и полимодальных корпусов — эта область достаточно специфична, и ее нужно обсуждать отдельно. Почти не затронуты вопросы типологии в обоих своих аспектах — как с точки зрения использования корпусов для типологических исследований, так и с точки зрения типологических особенностей разных языков, которые диктуют свои подходы к созданию корпусов. Нетрудно предсказать с высокой вероятностью "успеха", что в ходе дальнейшей работы появятся новые проблемы, о существовании которых мы, возможно, даже не подозреваем.

СПИСОК ЛИТЕРАТУРЫ

1. Broekhuizen E., van. [Rec. ad op.:] Mair Ch. & M. Hundt (eds.) Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20). Amsterdam; Atlanta, 1999 // LINGUIST List.— 2001.— Vol. 12.— 272.

2. Андрющенко В. М. Концепция и архитектура машинного фонда русского языка.— М., 1987.

3. Фрэнсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике. Вып. XIV: Проблемы и методы лексикографии.— М., 1983.

4. Баранов А. Н. Автоматизация лингвистических исследований: Корпус текстов как лингвистическая проблема // Русистика сегодня.— 1998.— № 1–2.

5. Венцов А. В., Касевич В. Б. Словарь для модели восприятия речи // Вестн. СПб ун-та.— 1998. Сер. 2.— Вып. 3.

6. Сичинава Д. В. К задаче создания корпусов русского языка // НТИ.— 2002.— Сер. 2, № 11.

7. Баранов А. Н., Михайлов М. Н., Сидоров Г. О. "Динамический корпус текстов" как новая технология прикладной лингвистики // Тр. междунар. семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Т. 2. 1998.

8. Lönngren L. et al. Частотный словарь современного русского языка (Studia Slavica Uppsaliensia, N 32).— Uppsala, 1993.

9. Русская грамматика. Т. 1.— М., 1980.

10. Касевич В. Б. Лексикон и лексикология (в печати).

11. Tognini-Bonelli E. Corpus Linguistics at Work.— Amsterdam, 2001.

12. Ревзин И. И. Отмеченные фразы, алгебра фрагментов, стилистика: К лингвистическому обоснованию теории моделей языка // Лингвистические исследования по общей и славянской типологии.— М., 1966.

13. Баранов А. Н. Введение в прикладную лингвистику.— М., 2001.

14. Perkins R. D. Deixis, Grammar, and Culture.— Amsterdam/Philadelphia, 1992.

15. Bybee J. L. Morphology: A Study of the Relation between Meaning and Form.— Amsterdam/Philadelphia, 1985.

16. Апресян Ю. Д. Лексическая семантика.— М.: Наука, 1974.

17. Melamed I. D. Empirical Methods for Exploiting Parallel Texts.— Boston, 2001.

18. Kassevitch V. B., Ventsov A. V., Yagounova E. V. The simulation of continuous text perceptual segmentation: A model for automatic segmentation of written text // Язык и речевая деятельность.— 2000.— Т. 3, ч. 2.

19. Венцов А. В., Касевич В. Б. Проблемы восприятия речи.— СПб, 1994.

20. Касевич В. Б. Семантика. Синтаксис. Морфология.— М., 1986.

21. Мельчук И. А., Жолковский А. К. Толково-комбинаторный словарь русского языка: Опыт семантико-синтаксического описания русской лексики.— Вена, 1984.

22. Мельчук И. А. Русский язык в модели "Смысл ⇔ Текст".— М., 1995.

23. Исаев И. А. Опыт автоматизации лексикографических исследований: Система DIALEX // Слово Достоевского.— М., 1996.