НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА (www.ruscorpora.ru)*

I

Несмотря на то, что компьютерная лингвистика - одно из наиболее динамично развивающихся направлений современной науки о языке, русистика неоправданно долго отставала в этой области. До недавнего времени единственным доступным электронным корпусом русских текстов мог считаться только Уппсальский-Тюбингенский корпус (www.sfb441.uni-tuebingen.de/bl/rus/korpora.html). В настоящее время полностью или частично подготовлены и опубликованы в интернете целый ряд корпусов [1]. Однако появление Национального корпуса русского языка (далее НКРЯ) - это событие, которое по праву можно назвать выполнением отложенных обязательств[2]. Широта проекта, высокая квалификация специалистов, участвующих в создании НКРЯ, государственная поддержка - все это делает НКРЯ действительно национальным проектом, сопоставимым если не по размеру, то по значимости с British National Corpus, American National Corpus, Ćesky Národni Korpus, корпусом немецкого языка COSMAS и другими крупнейшими корпусами различных языков мира.

Рецензия на еще не завершенный проект -дело рискованное и для рецензентов, и для авторов проекта. Однако, по нашему мнению, рецензирование НКРЯ оправдывается, во-первых, тем, что даже в незавершенном виде это готовый для использования инструмент; во-вторых, преимущество рецензирования интернет-публикаций заключается в том, что электронная форма позволяет быстро вносить необходимые исправления, корректировать ход работы с учетом мнения рецензентов и пользователей. Наконец, настоящая рецензия ставит своей целью знакомство лингвистов с этим полезным инструментом.

Как и всякое значительное предприятие, НКРЯ не лишен мелких упущений и ошибок, которые можно оперативно исправлять, если будет налажена надежная обратная связь от пользователей к разработчикам. В этом смысле лингвисты, заинтересованные в развитии НКРЯ, имеют возможность активно участвовать в подготовке корпуса, присылая свои замечания на адрес info@ruscorpora.ru. Исходя из этого предлагаемая рецензия не содержит частных замечаний (список которых уже отправлен разработчикам), а посвящена обсуждению фундаментальных решений, которые заложены в корпус при его создании.

II. ОСНОВНЫЕ ОСОБЕННОСТИ КОРПУСА

НКРЯ - долгосрочный проект, который объединяет ведущих специалистов в области компьютерной лингвистики, грамматики, лексикографии, работающих в рамках Ассоциации "Национальный корпус русского языка". В перспективе планируется подготовить серию подкорпусов, охватывающих максимально широкий временной и функциональный диапазоны текстов (см. [Вербицкая и др. 2003]).

В настоящее время частично подготовлены корпус ранних текстов (начало XIX - середина XX века) и корпус современных текстов (середина XX - начало XXI века), общий объем которых на момент написания рецензии превышал 65 миллионов единиц[3].

Опубликованная часть НКРЯ распадается на два больших подкорпуса: тексты со снятой неоднозначностью (омонимией) и тексты без снятой неоднозначности[4]. Это разделение поз-

* Авторы рецензии благодарят за помощь и отдельные комментарии М. Бринге (Северная Каролина, США), Д. Дивьяк (Лейвен, Бельгия), М. Михайлова (Тампере, Финляндия).

[1] Компьютерный корпус текстов русских газет конца XX-го века (www.philol.msu.ru/~lex/corpus); Регенсбургский диахронический корпус русского языка (www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/Corpus/kiss/index-ru.htm); Хельсинский аннотированный корпус ХАНКО (www.slav.helsinki.fi/hanco) и др.

[2] Вынужденное отставание, впрочем, имело и свои преимущества: создатели НКРЯ могли в полной мере учитывать накопленный опыт, опираться на разработанные международные стандарты описания (тегирования), решать многие технические проблемы (поддержка Кириллицы, XML-формат и др.). Это, конечно, не могло помочь в решении всех специфических проблем, связанных с русской грамматикой, однако помогло устранить многие ошибки, неизбежные при работе с нулевого цикла.

[3] В настоящее время (середина 2006 года) объем корпуса уже достиг 120 миллионов единиц *(примеч. ред.)*.

[4] Доля единиц русского текста, квалификация которых морфологически неоднозначна, может составлять более половины от общего количества текстоформ. Сюда попадают не только омонимы, но и разного рода внесистемные случаи неоднозначности, как в предложении *"...мы **должны будем идти** на непозволительные политические уступки"* в котором случайное сочетание *будем идти* не является омонимичным формам аналитического буду щего, но представляет определенную слож ность для автоматического аннотирования (см. подробнее [Копотев 2004]).

воляет исследователю в зависимости от конкретных задач получать точные результаты, но в меньшей выборке или большее количество примеров, но содержащих какое-то количество шума. Алгоритм поиска дает возможность искать несколько разных единиц с указанием максимального и минимального расстояния между ними, предусмотрен поиск по текстоформам ("словоформам") и леммам ("лексемам")[5]. Собственно языковое аннотирование заключается в определении основных морфологических параметров (разработанных под руководством В.А. Плунгяна рабочей группой в составе Г.И. Кустовой, О.Н. Ляшевской, А.Е. Полякова, Д.В. Сичинавы), семантических параметров (основанных на системе "Лексикограф" (см. [Кустова, Падучева 1994] и адаптированных рабочей группой в составе Г.И. Кустовой, О.Н. Ляшевской, В.А. Плунгяна, Е.В. Падучевой, Е.В. Рахилиной), небольшого количества словообразовательных, синтаксических дескрипторов, расстановки ударений в части текстов. Результат поиска представляет предложения, содержащие единицы, удовлетворяющие условию поиска; при необходимости контекст может быть расширен до семи предложений.

Ниже кратко описаны основные особенности интерфейса корпуса и возможности поисковых запросов.

На первой странице сайта www.ruscorpora.ru приведена основная информация о НКРЯ и новости от его создателей. Удобное меню слева позволяет быстро перемещаться на нужную страницу. Меню состоит из четырех разделов, в первом из которых содержится отсылка к первой странице, второй блок объединяет поисковые ресурсы сайта, третий содержит ссылки на метаинформацию о составе текстов, принципах аннотирования и др. Наконец, последний блок объединяет информацию о создателях НКРЯ, использованных программах, авторских правах на тексты.

Авторы корпуса предусмотрели справочный раздел ("Как искать"), поэтому в настоящей рецензии, по-видимому, нет необходимости подробно останавливаться на всех тонкостях поиска, однако отметить основные все же необходимо.

Ссылка "Мой корпус" позволяет определять подвыборку, группируя художественные и нехудожественные тексты; задавать жанр и тип текста, место и время описываемых событий (для художественных текстов); сферу функционирования, тип и тематику текста (для нехудожественных текстов).

Ссылка "Поиск в корпусе" открывает страницу, в которой можно указать основные поисковые требования. Эти требования (запросы) могут представлять собой произвольную комбинацию трех типов условий: буквенный состав (с возможностью использования подстановочных знаков (*) и логических операторов: "*ть", "по*ся", "недо*" и т. п.), морфологические ("грамм. признаки") или лексико-семантические ("семант. признаки") условия. Примером простейшего запроса может являться напечатанная в строке "Поиск точных форм" словоформа "Отечеством": в результате будут найдены все контексты, содержащие этот набор букв. Графа "Слово 1" позволяет усложнить запрос и искать все формы лексемы ОТЕЧЕСТВО. Добавив к этому грамматические показатели ("Грамм, признаки - выбрать"), можно искать только определенные формы лексемы (например, "Родительный падеж"). Понятно, что морфологические и лексико-семантические признаки можно выбирать произвольно, без какой-либо привязки к конкретной лексеме. И это существенно расширяет возможности поиска. Так, можно искать "существительные со значением 'инструмент' в творительном падеже"; "глаголы со значением 'природное явление' в форме третьего лица" и т. п. Еще одним мощным инструментом является возможность задавать условия для поиска сочетаний неопределенно большого количества единиц ("Слово 2" + "Слово 3" и т. д.), произвольно комбинируя все три типа запросов и расстояние между единицами. Отметим при этом, что запрос в строке "Поиск точных форм" чувствителен к порядку слов, тогда как "Лексико-грамматический поиск" нечувствителен к порядку следования заданных единиц. Все это делает поисковые возможности НКРЯ чрезвычайно богатыми и предоставляет лингвисту тонко настраиваемый инструмент, переводящий исследовательскую и преподавательскую работу на совершенно новый уровень. Так, например, авторы рецензии успешно собирали материал для исследования фразем типа "инфинитив ТАК инфинитив", выяснения прототипичного значения переходности глагола. На основе НКРЯ ведутся исследования сочетаний ХОТЬ с формами императива, составляется вопросник для исследования сочетаний финитных форм глагола с инфинитивом, исследуются сложные предложения с разными типами связи между клаузами. Авторы рецензии уверены, что этот список может быть многократно увеличен. В

[5] Текстоформа - "аналог" словоформы в компьютерной лингвистике; на практике часто понимается как единица от пробела до пробела. Лемма - "аналог" лексемы, результат автоматического сведения текстоформ к начальной форме.

этой связи не лишним будет напомнить, что возникновение корпусной лингвистики означает если не революцию, то существенное изменение исследовательских практик (см. [Fillmore 1992]).

III. ПРОБЛЕМЫ И РЕШЕНИЯ

III. 1. Теоретическая эклектичность

Один из "постулатов аннотирования", сформулированных Дж. Личем, определяет теоретические основания аннотирования так: "Annotation schemes should preferably be based as far as possible on 'consensual', theory-neutral analyses of the data" [Leech 1993: 275]. Несмотря на это, существуют языковые базы данных, которые связаны с определенными теоретическими установками [например, проект FrameNet (www.icsi.berkeley.edu/~framenet), основанный на жестких принципах семантического описания]. В этих случаях разметка единиц более или менее точно соответствует исходным теоретическим положениям разработчиков.

Однако очевидно, что проект общедоступного многоярусного корпуса должен следовать постулату Дж. Лича. И в этой сфере разработчики русскоязычного корпуса сталкиваются с первой серьезной проблемой.

В НКРЯ, как было указано выше, представлена морфологическая, лексико-семантическая информация, а также частично словообразовательная и синтаксическая. Однако ясно, что степень полноты и общепризнанность классификаций языковых уровней существенно различается. Так например, в научной литературе по морфологии могут дискутироваться вопросы о количестве русских падежей, но, как кажется, не вызывает сомнения сам факт существования категории падежа. В области синтаксиса, как известно, такого единства нет. Широко распространенная в практике преподавания классификация, опирающаяся на представление о главных и второстепенных членах предложения, не может считаться общепризнанной; современные синтаксические теории, описывающие синтаксические отношения в виде дерева зависимостей, не имеют прямой корреляции с теорией главных/второстепенных членов предложения, метаязык "грамматики конструкций" Ч. Филлмора не может быть согласован с положениями "Русской грамматики" 1980-го года и т.д. (Грамматика-80).

Трудно представить корпус, который смог бы объединить все теории, поэтому многоярусный корпус неизбежно (во всяком случае, в обозримом будущем) оказывается или эклектичным, или узконаправленным. К этому добавляется и техническая проблема: если большинство программ автоматического морфологического аннотирования русского языка базируется на "Грамматическом словаре" А.А. Зализняка, то в основе алгоритмов синтаксических парсеров часто лежат разные синтаксические теории, среди которых синтаксис в духе Грамматики-80 не является самым популярным. Еще больше сложностей возникает при описании семантического компонента языковых единиц.

Осознавая проблему теоретической эклектичности, создатели НКРЯ выбирают разные подходы для разных языковых уровней. Если морфология в НКРЯ представлена в достаточно традиционном виде (о деталях см. ниже) и не апеллирует к специальным сведениям, выходящим за базовые знания выпускника филологического факультета, то семантическая разметка представляет собой воплощение на широком языковом материале оригинальной системы семантических дескрипторов, которая требует предварительного знакомства со справкой на сайте или - в идеале - с соответствующими работами авторов [Кустова, Падучева 1994; Рахилина 2000]. Конечно, такую систему семантической разметки ни в коем случае нельзя назвать недостатком, поскольку она оказывается гораздо более полной и строгой, чем существующие "традиционные" классификации лексики. С другой стороны, синтаксическая разметка (на сегодняшний день еще не представленная в корпусе) будет гораздо менее подробной и ограничится лишь "малым" синтаксисом (на уровне словосочетания).

Таким образом, эклектичность и неравномерная представленность разных языковых уровней в НКРЯ выявляет две существенные проблемы современной лингвистики: отсутствие полных теоретически обоснованных и общепринятых классификаций, с одной стороны, и сложность (граничащая с невозможностью) автоматического аннотирования на основе этих классификаций, - с другой. Всякий языковой корпус в силу необходимости тотального описания материала кристаллизует проблемные области в описании того или иного языка. И в этом смысле НКРЯ является не только инструментом для быстрого поиска примеров, но и бесконечным источником совершенствования теоретических и чисто дескриптивных подходов к русскому языку.

III.2. "Машинная грамматика" русского языка и ее отношение к "большой" лингвистике

Вторая проблема, встающая перед создателем аннотированного корпуса, связана с дилеммой объем материала vs. точность обра-

ботки. Одним из промежуточных итогов компьютерной лингвистики стало признание факта невозможности точного автоматического анализа текста. Для русского языка, богатого флективным словоизменением и омонимичностью грамматических показателей, создание анализатора, безошибочно производящего дизамбигуацию (снятие неоднозначности), практически невозможно. На сегодняшний день качественное аннотирование русского текста всегда связано с существенной ручной "постобработкой", проводимой квалифицированными специалистами. В этом смысле при относительной ограниченности организационных возможностей перед создателями любого корпуса всегда стоит выбор: сравнительно небольшой, но выверенный корпус или объемный, но аннотированный автоматически.

Разрешая эту дилемму, создатели языковых корпусов вынуждены идти на более или менее серьезные упрощения лингвистических классификаций, выбирая между скоростью обработки материала и точностью интерпретаций. Фактически речь идет о том, что разработчики принимают те или иные решения, которые противоречат языковой реальности или лингвистическим представлениям; принятие этих решений часто связано не с теоретическими установками создателей, а с задачей облегчения автоматической обработки. Создатели НКРЯ тоже должны были идти на подобные компромиссы, вызванные и пробелами в описании системы русского языка, и требованиями автоматической обработки. Не обсуждая мотивировок, заставивших разработчиков прийти к тому или иному решению, приведем список таких допущений.

a. Не разводятся лексические омонимы. Пользователя НКРЯ не должно вводить в заблуждение то, что авторы корпуса предлагают поиск в подкорпусе со снятой омонимией. Речь идет только об устранении грамматической неоднозначности (например, существительное и глагол ПЕЧЬ). Лексические омонимы в корпусе считаются одной леммой, в силу этого поиск только одного члена полной омонимичной пары невозможен. Так, например, запрос "ЛУК: существительное: 'оружие'" выдает и контексты такого рода: *Золотистые связки лука над крыльцом* [Сергей Довлатов. Заповедник (1983)].

b. Не учитываются (или недостаточно учитываются) следующие многокомпонентные единицы.

- Формы сослагательного наклонения глагола: *прочитал бы, сходил бы* и более сложные случаи, как в предложении *Я хочу, чтобы студенты прочи-тали эту книгу*, в котором слившиеся частица и союз не отменяют аналитизма сослагательной формы *прочитали [бы]*.
- Формы сложного будущего времени: *буду читать*.
- Аналитические формы прилагательных и наречий: *более быстрый*, *более быстро* (оставляем в стороне вопрос о спорности их выделения в "большой" лингвистике).
- Аналитические формы местоимений: *ни от кого*.
- Составные и дробные числительные: *сто сорок восемь*, *две третьих*. Записи "двадцать три" и "23" считаются разными единицами и состоят из разного количества лемм ("лексем"). Добавим, что традиционные разряды числительных заменены основанными на морфологических критериях выделениями и интуитивно понятными "числительными" и "числительными прилагательными".
- Служебные фраземы (так называемые "эквивалентны слова"): *потому что*, *в течение* и т.д.). Список из 180 служебных фразем ("сложных лексических единиц", по терминологии авторов) далек от полноты. Оставляя в стороне полнозначные фразеологизмы типа *сесть в калошу*, которые тоже должны выделяться как отдельные единицы (см., например [ТКС 1984]), отметим, что по данным словарей [Рогожникова 2003; Ефремова 2004] список "сложных лексических единиц" должен быть увеличен примерно в 10 раз (например, отсутствуют такие очевидные кандидаты, как *в числе*, *в ногу*, *на износ*, *на зависть* и др.).

c. Не учитываются однословные морфологические признаки, сложные для автоматического анализа.

- Не выделены отдельно, а включены в категорию множественного числа формы Pluralia tantum.
- Текстоформы типы *красивее* (КРАСИВО, КРАСИВЫЙ) считаются одной формой. Так, запрос "прилагательное в форме компаратива" возвращает среди результатов предложение *Моторы ревели теперь* ***ровнее,*** *и ящики успокоились* [И. Грекова. На испытаниях (1967)].
- Отсутствуют собственные/нарицательные существительные. Параметра "ан-

тропонимы" в семантической разметке, очевидно, недостаточно, поскольку в русском языке есть, например, топонимы, омонимичные именам нарицательным.

- В списке частей речи без достаточного теоретического обоснования указана такая часть речи, как "вводное слово". Надо отметить, что эту категорию-"призрак" можно найти во многих словарях русского языка, в том числе и в "Грамматическом словаре" А.А. Зализняка. Однако нам неизвестны классификации, в которых бы фигурировала такая часть речи.
- Без достаточных объяснений местоимения разделены на морфологически мотивированные местоимения-существительные, местоимения-прилагательные, местоимения-предикативы, местоименные наречия (традиционные разряды местоимений отнесены к семантическим признакам). И хотя эта классификация выглядит обоснованной, поиски некоторых местоимений оказываются непростой задачей (а поиск реципрока *друг друга* по морфологическим показателям вообще невозможен).

Подводя итог этой части рецензии, необходимо признать, что в настоящее время (как кажется, и в обозримом будущем) могут существовать два разнонаправленных принципа создания русскоязычных корпусов, в силу того, что объем материала и точность аннотирования представляются на практике взаимоисключающими критериями. В то же время оба подхода имеют свое право на существование, и оба вполне оправданы при подготовке современных корпусов. Так, относительно небольшой корпус ХАНКО позволяет искать формы сослагательного наклонения, аналитического будущего, существительные Pluralia tantum и другие единицы, о которых шла речь выше, однако следствием ограниченного объема является, например, то, что в нем нет ни одной звательной формы.

III. 3. Состав корпуса (текстовая сбалансированность)

Еще одна серьезная и крайне болезненная проблема при создании любого корпуса - сбалансированность состава источников. Эта проблема широко обсуждается создателями корпусов (см. [Sinclair 1991; D. Biber 1993] и др.) и, тем не менее, не может считаться окончательно решенной. Типология текстов Синклера [Sinclair 1996], адаптированная С.А. Шаровым, кажется серьезной базой для этой работы, но и она не дает четких критериев для определения пропорциональной представленности разных типов текстов в корпусе. Авторы НКРЯ утверждают, что "Корпус характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленных в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Следует иметь в виду, что хорошая представительность достигается только при значительном объеме корпуса (десятки и сотни миллионов словоупотреблений). Планируемый составителями объем Национального корпуса русского языка - 200 млн. слов" (www.ruscorpora.ru/corpoa-intro.html).

Необходимо признать, что в настоящее время это, безусловно, верное утверждение является скорее заявкой планов на будущее, чем констатацией существующего положения дел. На момент подготовки рецензии объем представленного материала был в почти четыре раза меньше запланированного. В этой связи нецелесообразно обсуждать сбалансированность существующих материалов, поскольку авторы и сами прекрасно понимают незавершенность проекта. Однако некоторые замечания общего характера все же необходимо сделать.

Если исходить из уже опубликованных материалов и заданных в поясняющих статьях пропорций, то соотношение художественных/нехудожественных текстов представляется вполне сбалансированным (в настоящее время в НКРЯ не представлена поэзия и разговорная речь, но это, насколько нам известно, дело ближайшего будущего). Вызывает вопросы соотношение подтипов внутри двух указанных групп. Прежде всего, бросается в глаза большое количество публицистических текстов. Их объем - 15 328 251 текстоформ (74.4 процента от всех нехудожественных текстов). На этом фоне непропорционально мало́ количество "учебно-научных" текстов (1 685 877 текстоформ, 8.2 процента от общего количества нехудожественных текстов) или официально-деловых текстов (942 767 текстоформ, 4.6 процента). Однако если иметь в виду, что конечный объем корпуса - около 200 млн. текстоформ, то пропорции еще могут быть выправлены.

Еще одна тенденция, проступающая в НКРЯ в его современном виде, - несбалансированность содержательного состава корпуса. Так, список тем нехудожественных текстов, упорядоченных по объему выборок, вызывает определенные вопросы. Пятипроцентный барьер преодолевают следующие темы.

Тема	Количество текстов	Объем	%
Прочее	582	6484279	31.46
Политика и общественная жизнь	6574	6158456	29.87
Искусство и культура	1364	1633208	7.92
Наука и технологии	267	1511409	7.33
Религия	480	1401076	6.80

Как кажется, это не отражает реального состава текстов современной публицистики, в которых такие темы, как "политика и общественная жизнь", "война и вооруженные конфликты", "криминал" занимают несопоставимо большее место, чем "спорт", "искусство и культура", "наука и технологии". Так например, материалы, посвященные "войне, вооруженным конфликтам" и "криминалу" вместе взятые, занимают в корпусе в 11 раз меньший объем, чем "искусство и культура". Как бы нам ни хотелось, чтобы эта пропорция была верна, она не отражает реального положения дел. С другой стороны, преобладание в НКРЯ текстов, сфера функционирования которых определена как "общественно-политическая" грозит превратить корпус в аналог "Частотного словаря русского языка" [Засорина 1977], составителям которого было высказано немало замечаний как раз в связи с явным уклоном в область общественно-политической лексики.

IV. ЗАМЕЧАНИЯ И ПРЕДЛОЖЕНИЯ

Жанр рецензии предполагает список опечаток и неточностей. Предусмотрен такой список и в настоящей рецензии. Однако существенное ее отличие от рецензирования любого бумажного издания заключается в том, что наши замечания и предложения могут быть внесены в корпус (конечно, при том условии, что разработчики с ними согласны). Развивая эту мысль, можно сказать, что при ответственном отношении k НКРЯ со стороны лингвистического сообщества корпус может постепенно избавляться от опечаток и других ошибок, которых, volens-nolens, не лишен и НКРЯ. Соответствующий список опечаток и других неточностей, обнаруженных рецензентами, был выслан разработчикам, к этому мы призываем и всех пользователей, заинтересованных в дальнейшем развитии НКРЯ.

Более существенные предложения, которые, по нашему мнению, помогут сделать НКРЯ еще более удобным для работы, приведены ниже.

-Полезной кажется функция, позволяющая собрать весь найденный материал (с контекстом или без него) на одну страницу с последующим сохранением в один файл. Такой файл, содержащий выбранные предложения в контексте, существенно облегчил бы работу пользователя.

-По-видимому, можно продумать более удобный способ просмотра информации о выбранной текстоформе (сейчас это можно сделать, только подведя курсор мыши к нужной единице).

-Не лишней была бы возможность сохранения выборки со всеми тегами в каком-нибудь удобном формате (скажем, "текст с разделителями") для последующей обработки в СУБД или в табличных процессорах.

-Представляется, что есть смысл добавить возможность вывода информации в формате стандартного конкорданса (как это реализовано на сайте corpus.leeds.ac.uk/ruscorpora.html) со всеми возможными статистическими сведениями (по жанрам, авторам, грамматическим параметрам и т.д.). Это облегчило бы исследование сочетаемостных свойств лексем и позволило бы подключить к корпусным исследованиям аппарат статистических методов.

-Во многих случаях полезно получать списки коллокаций и коллигаций с указанием частотности. Это позволило бы более эффективно изучать сочетаемость лексем, частотность различных грамматических конструкций и т.п.

-Для решения определенных задач удобным может оказаться генеральный словник, с возможностью поиска "от леммы".

-Удобной была бы возможность ограниченного поиска материала, с возможностью определять объем выборки: ограничение на количество примеров, "шаг" выборки (один пример из ста.

тысячи), количество примеров одного автора, из одного текста и т.п.

V. ЗАКЛЮЧЕНИЕ

Поскольку работа создателей корпусов сопоставима с работой лексикографов, рассчитанной на самую широкую аудиторию, необходимо учитывать готовность пользователей понимать и принимать те или иные решения. В этом смысле любой корпус - и НКРЯ не исключение - базируется на принципе, сформулированном Дж. Личем: "There can be no claim that the annotation scheme represents 'God's truth'. Rather, the annotation scheme is made available to a research community on a *caveat emptor* principle. It is offered as a matter of convenience only, on the assumption that many users will find it useful to use a corpus with annotations already built in, rather than to devise and apply their own annotation schemes from scratch (a task which could take them years to accomplish)" [Leech 1993: 275]. Принимая его, создатели НКРЯ делают все возможное, чтобы с одной стороны, приблизиться к адекватности лингвистического описания, а с другой, отчетливо понимают, что это невозможно.

Авторы рецензии полагают, что НКРЯ - исключительно полезный инструмент, который даже в сегодняшнем незаконченном виде предоставляет лингвисту богатые возможности для изучения русского языка. Работа, проделанная коллективом авторов, не оставляет сомнений, что по завершении НКРЯ станет вехой в развитии русистики, сопоставимой с изданием Б АСа, публикацией текстов по разговорной речи, подготовкой и изданием Словаря русского языка X-XVII веков. Кажется, именно такая задача способна объединить силы лингвистов, заинтересованных в создании действительно востребованного и современного ресурса. Нам остается только пожелать составителям удачного продолжения, а пользователям - скорейшего включения Национального корпуса русского языка в необходимый инструментарий исследователя.

СПИСОК ЛИТЕРАТУРЫ

Вербицкая и др. 2003 - *Л.А. Вербицкая, Н.Н. Казанский, В.Б. Касевич.* Некоторые проблемы создания национального корпуса русского языка // НТИ. Сер. 2. № 6. 2003.

Ефремова 2004 - *Т.Ф. Ефремова.* Толковый словарь служебных частей речи русского языка. М., 2004.

Засорина 1977 -*Л.Н. Засорина* (ред.). Частотный словарь русского языка. Л., 1977.

Копотев 2004 - *М.В. Копотее.* Неоднозначность и пути ее разрешения в Хельсинкском аннотированном корпусе ХАНКО // Труды международной конференции "Корпусная лингвистика-2004". СПб., 2004.

Кустова, Падучева 1994 - *Г.И. Кустоеа, Е,В. Падучева.* Словарь как лексическая база данных // ВЯ. 1994. № 4.

Рахилина 2000 - *Е.В. Рахилина.* Когнитивный анализ предметных имен: семантика и сочетаемость. М., 2000.

Рогожникова 2003 - *Р.П. Рогожникова.* Толковый словарь сочетаний, эквивалентных слову. М., 2003.

ТКС 1984 - *ИЛ. Мельчук, А.К. Жолковский.* Толково-комбинаторный словарь. Вена, 1984.

Biber 1993 - *D. Biber.* Representativeness in corpus design // Literary and linguistic computing. 1993. 8/4.

Fillmore 1992 - *Ch. Fillmore.* "Corpus linguistics" vs. "Computer-aided armchair linguistics" // Directions in corpus linguistics. Berlin, 1992.

Leech 1993 - *G. Leech.* Corpus annotation schemes // Literary and linguistic computing. 1993. 8/4.

Sinclair 1991 - /. *Sinclair.* Corpus, concordance, collocation. Oxford, 1991.

Sinclair 1996 - J. *Sinclair.* EAGLES preliminary recommendations on text typology // www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html.

М .В. Копотее, Л. Янда

РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВОПРОСЫ ЯЗЫКОЗНАНИЯ

ЖУРНАЛ ОСНОВАН В ЯНВАРЕ 1952 ГОДА

ВЫХОДИТ 6 РАЗ В ГОД

Издается под руководством
Отделения историко-филологических наук РАН

СЕНТЯБРЬ - ОКТЯБРЬ

"НАУКА"

МОСКВА - 2006