МЕЖДУ СЦИЛЛОЙ ЯЗЫКОЗНАНИЯ И ХАРИБДОЙ ЯЗЫКА: О РУССКОЯЗЫЧНЫХ КОРПУСАХ ТЕКСТОВ

BETWEEN SCYLLA OF LINGUISTICS AND CHARYBDIS OF A LANGUAGE: ABOUT RUSSIAN CORPORA

Михаил Вячеславович Копотев

Хельсинкский университет, Хельсинки, Финляндия

mihail. kopotev@helsinki. fi

В докладе ставится задача выявить соотношение "большой" лингвистики и "машинной" грамматики русского языка. Опираясь на сформулированные Дж. Личем постулаты аннотирования, автор анализирует русские аннотированные корпуса и формулирует проблемные зоны русской корпусной лингвистики.

1993 году один из создателей корпусов LOB и BNC Джеффри Лич сформулировал 7 постулатов аннотирования [1]. Представляется, что эти постулаты в равной степени относятся и к русским корпусам текстов, больше того – они позволяют наметить проблемные зоны в области русской корпусной лингвистики. Настоящий доклад построен в виде анализа соответствия постулатам Лича четырех русских аннотированных корпусов, доступных в интернете[1]. Это Национальный корпус русского языка (далее **НКРЯ**), Хельсинкский аннотированный корпус (далее **ХАНКО**), Тюбингенский аннотированный корпус[2] (далее **ТАК**) и корпус текстов русских газет конца XX-ого века (далее **КОРГАЗ**). В таблице 1 приведены основные сведения о каждом из них.

Назв.	Адрес	Объем (на нач. 2005 г.)	Типы лингвистического аннотирования (на нач. 2005 г.)
НКРЯ	www. ruscorp ora. ru	более 30 млн. текстоформ	Морфологическая, семантическая; словообразовательная, синтаксическая (фрагментарно)
ХАНКО	www. slav. Helsi- nki. fi\ hanco	100 000 текстоформ	Морфологическая
ТАК	heckel.s fb.uni- tuebing en.de/c gi- bin/cqp tag.pl	165 000 текстоформ	Морфологическая; Словообразовательная, синтаксическая (фрагментарно)
КОР- ГАЗ	www. philol. msu. ru/~lex/ corpus/ corp_de scr. html#5	205 000 текстоформ	Морфологическая; словообразовательная, лексическая и синтаксическая (фрагментарно)

Табл. 1

I. *It should <…> made clear how, and by whom, the annotations were applied.*

Этот постулат выполняется всеми исследователями. На сайтах разработчиков можно найти соответствующие сведения о разработчиках, методах работы и т. д. Там же можно найти контактную информацию.

II. *It should always be easy to dispense with annotations, and revert to the raw corpus. The raw corpus should be recoverable.*

III. *The annotations should, correspondingly, be extractable from the raw corpus, to be stored independently, or stored in an interlinear format.*

Очевидно, что эти технические условия могут соблюдаться разработчиками всех корпусов. Независимо от того, как решается эта задача (представлением информации средствами XML-разметки или в виде базы данных), возврат к чистому тексту технически возможен всегда. Однако если для **НКРЯ**, **ХАНКО** и **КОРГАЗ** это действительно так, и доступные версии корпусов позволяют получить “чистый” текст, то в **ТАК** при заявленной возможности добиться этого не удается: в любом случае результат представлен в виде текста с аннотацией (в подстрочном формате).

[1] Порядок следования постулатов Дж. Лича изменен; в целях экономии места близкие постулаты объединены.

[2] Уппсальский-Тюбингенский корпус состоит как из аннотированных, так и неаннотированных текстов. В настоящей статье анализируется только аннотированная часть корпуса.

В таблице 2 представлены фрагменты текстов из всех корпусов.

НКРЯ	**На** ознакомление с материалами уйдет не меньше года [Пунанов Григорий. Обвиняемым показали дело. На ознакомление с материалами уйдет не меньше года // Известия, 2001.07.20]
ХАНКО	Но **на** вторую полноценную гулянку энергии явно не хватает. [контекст]
ТАК	Источник: И вот зима катит в глаза. “Известия”, 87-09-28 (1.214). надои/$_{substantiv\_masc\_pl\_nom\_unb}$ **на/$_{praep\_prp}$** фермах/$_{substantiv\_fem\_pl\_prp\_un}$
КОРГАЗ	Причем 63 процента этого долга приходится **на** государственные унитарные предприятия “Шелонь“ и “Красный двор“.

Табл. 2

Расширяя рамки, заданные постулатами Дж. Лича, ниже я покажу, как соотносятся возможности поиска и представления текста и его аннотации в указанных корпусах.

Все четыре корпуса позволяют производить поиск по единицам текста (то есть по текстоформам), а также по леммам. **НКРЯ**, **ХАНКО** и **КОРГАЗ** позволяют искать и по лингвистическим параметрам. В **ТАК** возможности поиска по грамматическим параметрам отсутствуют.

Вывод результатов поиска на экран реализован в корпусах по-разному.

- В **НКРЯ** результат представлен в виде контекста, равного одному предложению без возможности расширения, аннотация представлена в виде неудобной для использования всплывающей подсказки.
- В **ХАНКО** начальный контекст тоже равен предложению, однако его можно расширить до целого текста. Вся грамматическая аннотация для всех слов может быть выведена в отдельное окно.
- В **ТАК** пользователь может задать контекст, равный произвольному количеству знаков, слов, предложений до и после искомой единицы. Грамматическая информация может быть выведена на экран в виде набора сокращений в скобках после текстоформы (см. пример в таблице 2).
- В **КОРГАЗ** контекст по умолчанию составляет 30 текстоформ справа и слева, он может быть легко расширен до целого текста. Кроме того, существует возможность получить список единиц без контекста (“словарное” представление). Грамматическая информация не выводится в результатах поиска.

IV. *The scheme of analysis presupposed by the annotations — the annotation scheme — should be based on principles or guidelines accessible to the end-user. <…>*

Этот постулат Дж. Лича кажется очевидным, однако в полной мере не реализуется ни в одном из анализируемых корпусов. Конечно, все четыре корпуса снабжены более или менее подробной справкой, но получение точных сведений о том, что значит, каждый параметр описания, какие теоретические или чисто практические мотивировки стоят за выделением той или иной единицы – все это часто остается неэксплицированным или комментируется в узкоспециализированных сборниках статей.

Приведем доказательства.

В **КОРГАЗ** существует возможность производить поиск по морфологическим, словообразовательным, синтаксическим параметрам, однако список параметров недоступен, и поиск, например, одушевленных существительных превращается в занимательный детектив.

Тэги, выводимые после текстоформы в **ТАК**, почти всегда без труда распознаются (см. примеры в таблице 2), однако их полный список с комментариями существенно облегчил бы работу с корпусом. Например, далеко не сразу становится ясно, что запись “на/praep_prp” значит 'предлог *на*, управляющий предложным падежом'.

Кроме этого, ни один из корпусов не эксплицирует разницу между леммой и лексемой, словоформой и текстоформой. Это приводит к тому, что неискушенный в компьютерных конвенциях исследователь может предположить, что создатели корпусов считают лексемы ЛУК ‘оружие' и ЛУК ‘растение' одной лексической единицей, а текстоформы “читал” и “бы” — двумя словоформами. Надо отметить, что непоследовательность в использовании терминов приводит к определенным недоразумениям: так, составители **НКРЯ** на странице “Статистика” пишут о “словоупотреблениях” (видимо, о текстоформах), а ниже те же цифры относятся уже к “словам”.

V. *Therefore, to avoid misapplication, annotation schemes should preferably be based as far as possible on ‘consensual’, theory-neutral analyses of the data.*

Выполнение этого постулата сталкивается, возможно, с самыми серьезными трудностями. Дело в том, что степень полноты и общепризнанность классификаций языковых уровней существенно различается. Например, в научной литературе по морфологии могут дискутироваться вопросы о количестве русских падежей, но 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 структуры составляющих, не имеют столь же широкого распространения, особенно в учебной

практике; подходы функционального синтаксиса плохо согласуется с положениями "Русской грамматики" 1980-го года и т. д.

Как кажется, трудно создать систему аннотирования, которая бы объединяла все теории, поэтому многоярусный корпус неизбежно оказывается или эклектичным, или узконаправленным. К этому добавляется и техническая проблема: если большинство программ автоматического морфологического аннотирования русского языка базируются на общепринятом стандарте — "Грамматическом словаре русского языка" А. А. Зализняка, то в основе алгоритмов синтаксических парсеров часто лежат совершенно разные синтаксические теории. Не меньше сложностей возникает при описании семантического и словообразовательного компонента языковых единиц. Уже сейчас можно утверждать, что эта работа связана со множеством сложных вопросов, на многие из которых современная лингвистика еще не нашла ответа. Различия в точности описания разных языковых уровней приводит в итоге к тому, что всякий многоуровневый корпус оказывается эклектичным.

Эта эклектичность, например, ярко проявляется в **НКРЯ**. Если морфология в **НКРЯ** представлена в достаточно традиционном виде, то семантическая разметка представляет собой воплощение на широком языковом материале оригинальной системы семантических дескрипторов [3, 4]. 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. Однако дисбаланс авторских и общепринятых подходов к языковой системе сохраняется.

Как кажется, эклектичность и неравномерная представленность разных языковых уровней выявляет две существенные проблемы современной русистики: отсутствие *полных* теоретически обоснованных и общепринятых классификаций, с одной стороны, и сложность (граничащей с невозможностью) автоматического аннотирования на основе этих классификаций — с другой. В этом смысле всякий языковой корпус в силу необходимости тотального описания материала кристаллизует проблемные области в описании того или иного языка. Он оказывается не только инструментом для быстрого поиска примеров, но и источником совершенствования теоретических и чисто дескриптивных подходов к языку.

VI. *No annotation scheme can claim authority as a standard, although de facto interchange 'standards' may arise, through widening availability of annotated corpora, and perhaps should be encouraged.*

В силу того, что создатели русскоязычных корпусов опирались на существующие традиции создания корпусов (прежде всего англоязычных), а исследовательские коллективы имели возможность общаться друг с другом такие "стандарты описания" de facto существуют. Конечно, определенные внутрикорпоративные конвенции должны существовать, однако необходимо эксплицировать их и – при необходимости – искать способы их устранения. Не обсуждая мотивировок, заставивших разработчиков прийти к тому или иному решению, приведем список таких допущений.

- Обычно не разводятся лексические омонимы, совмещаясь в одной лемме. В силу этого поиск только одного члена полной омонимичной пары невозможен. Особенно остро этот вопрос встает в связи с семантической разметкой корпуса. Так, например, запрос "ЛУК: существительное: 'оружие'" в **НКРЯ** выдает и контексты такого рода: *"Золотистые связки лука над крыльцом". [Сергей Довлатов. Заповедник (1983)]*.
- Недостаточно учитываются аналитические формы (**НКРЯ**, **ТАК, КОРГАЗ**).
- Формы сослагательного наклонения глагола: *сходил бы*.
- Формы сложного будущего времени: *буду ходить*.
- Аналитические формы прилагательных и наречий: *более быстрый, более быстро*.
- Составные и дробные числительные, числовые и буквенные написания числительных: *сто сорок восемь, 148, две третьих*.
- Аналитические формы местоимений: *<u>ни</u> от <u>кого</u>*.
- Служебные фраземы, или так называемые "эквивалентны слова": "потому что", "в течение". Исключениями являются **НКРЯ** и **ХАНКО**: в первом выделены все употребления 180 служебных фразем; в **ХАНКО** на основе списка из [2, 5] выделяется приблизительно 2000 единиц.
- Не учитываются морфологические формы, сложные для автоматического выделения.
- Формы Pluralia tantum (**НКРЯ**, **ТАК, КОРГАЗ**).
- Текстоформы типы *красивее (КРАСИВО, КРАСИВЫЙ)* считаются одной формой (**НКРЯ**, **КОРГАЗ**).
- "Малые" части речи: *это* (местоимение, частица, связка), *что-то* (местоимение, наречие) (**НКРЯ**, **ТАК, КОРГАЗ**).
- Для двувидовых глаголов не приводится указание на вид по контексту (**НКРЯ, ХАНКО, ТАК, КОРГАЗ**).
- Не определяется возвратность/невозвратность глагольных форм (**НКРЯ**, **КОРГАЗ**).
- Вводятся "фантомные" части речи и грамматические признаки. Например, в **НКРЯ** без достаточного теоретического обоснования

указана такая часть речи, как “вводное слово”. В **ХАНКО** и **НКРЯ** в список грамматических параметров попали сокращения, в **ХАНКО** список частей речи пополнился пунктуационными знаками; в **ТАК** можно найти одушевленные и неодушевленные местоимения.

Известно, что серьезная проблема, встающая перед создателями аннотированного корпуса, связана с дилеммой объем материала *vs* точность обработки. Создание анализатора, безошибочно производящего дизамбигуацию для русского языка, по-видимому, невозможно. На сегодняшний день качественное аннотирование русского текста всегда связано с ручной постобработкой. 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. Представляется, что оба принципа имеют право на существование.

Задачей настоящего доклада является анализ русских корпусов текстов с позиций, так сказать, заинтересованного “внешнего наблюдателя”. Это позволяет выявить проблемные зоны, главными из которых я считаю следующие:

- отсутствие доступных описаний;
- эклектичность или необщепринятость схем аннотирования или их отдельных частей;
- наличие допущений, неизвестных обычному пользователю.

Поскольку работа создателей корпусов сопоставима с работой лексикографов, рассчитанной на самую широкую аудиторию, мы должны учитывать готовность пользователей понимать и принимать наши решения. В этом смысле создатели корпуса должны донести до пользователей последний постулат Дж. Лича.

VII. *There can be no claim that the annotation scheme represents ‘God’s truth’. Rather, the annotated corpus is made available to a research community on a caveat emptor principle. <…>*

Принимая его, мы должны сделать все возможное, чтобы, одной стороны, приблизиться к адекватности лингвистического описания, а с другой, – дать ясно понять пользователю, что это невозможно.

Список литературы:

1) Leech G. Corpus annotation schemes // Literary and Linguistic Computing, 1993, 8/4, 275-81.
2) Ефремова Т.Ф.Толковый словарь служебных частей речи русского языка // М.: Астрель-АСТ, 2004.
3) Кустова Г.И., Падучева Е.В. Словарь как лексическая база данных // Вопросы языкознания, 1994, 4, 96-106.
4) Рахилина Е.В. Когнитивный анализ предметных имен: семантика и сочетаемость // М.: Русские словари, 2000.
5) Рогожникова Р.П., 2003: Толковый словарь сочетаний, эквивалентных слову. М.: Астрель-АСТ. 2003.