Компьютерная лингвистика и интеллектуальные технологии:
По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.).
Вып. 11 (18). — М.: Изд-во РГГУ, 2012. Т. 1, стр. 37—49

Пролегомены к проекту Генерального интернет-корпуса русского языка (ГИКРЯ)

Беликов В. И. (vibelikov@gmail.com) ИРЯ РАН, Москва, Россия
Селегей В. П. (Vladimir_S@abbyy.com) ABBYY; РГГУ, Москва, Россия
Шаров С. А. (s.sharoff@leeds.ac.uk) University of Leeds, UK; РГГУ, Москва, Россия

В докладе рассматривается задача создания Генерального интернет-корпуса русского языка (ГИКРЯ). Анализируются технологические, структурные, функциональные, контентные недостатки современных корпусов русского языка и обосновывается необходимость и технологическая возможность создания нового корпуса на основе русскоязычного Интернета. Определяются принципы сбора, классификации, разметки, необходимые параметры и функциональность ГИКРЯ.

Ключевые слова: сегментно-статистический анализ Интернета, автоматическое создание корпусов, автоматическая классификация текстов, автоматическая разметка текстов

Preliminary considerations towards developing the General Internet Corpus of Russian

This talk presents the project for creating the General Internet Corpus of Russian (GICR). We start with analysing technological, structural, functional and content problems of existing Russian corpora. Then we discuss the need and the possibility for creating a new corpus which is based on the Russian Internet. Finally, we define the principles for text collection, classification and annotation, as well as the necessary parameters and functions of the interface of GICR.

Введение

Современная лингвистика и лексикография являются в значительной степени корпусной: именно корпуса текстов являются сегодня важнейшим источником сведений о живом языке. Это делает особенно важным исследование вопросов, связанных с адекватностью методик корпусных исследований, правильностью устройства самих корпусов и эффективностью инструментов доступа к ним.

Авторы доклада являются сторонниками той точки зрения, что задача получения полноценного инструмента для корпусных лингвистических исследований все еще весьма далека от полноценного решения.

Доклад продолжает дискуссию, начатую на Диалоге-2011 статьей одного из авторов (Belikov 2011). В докладе делается попытка обобщить все известные проблемы с существующими средствами исследования текстов и анонсируется проект, основанный на идее создания Генерального интернет-корпуса русского языка (ГИКРЯ), в котором эти проблемы могли бы быть преодолены.

Существующие корпуса относительно невелики по объему и их содержимое часто отражает случайные особенности их создания. Так, размеры русскоязычного Интернета на несколько порядков больше корпусов типа НКРЯ. Грубая оценка на основе результатов Gulli, Signorini (2005) дает примерно 500 миллиардов слов русского языка индексированного поисковыми машинами в 2005. Это объясняет желание лингвистов проводить исследования на основе запросов к поисковым машинам. Но такие машины не предназначены для лингвистических исследований и частотная информация из них не всегда адекватна. В них также тяжело задавать лингвистически релевантные запросы.

Решением проблемы объем/функциональность явилось бы создание машины для лингвистической обработки русскоязычного Интернета, которая обеспечивала бы сбор достаточного количества страниц (сравнимого с индексом поисковиков), распознавание структуры текста на страницах, качественный морфосинтаксический анализ со снятием неоднозначности, автоматическую классификация текстов по темам и жанрам, а также по другим релевантным критериям метатекстовой разметки (география, возраст, пол автора, время создания). В этом интерфейсе будет возможно также визуализировать примерное положение текстов, интересующих нас, в сравнении с другими текстами Интернета, например, насколько далеко данные тексты отстоят от типичных новостей, либо от художественной литературы (Wilson, et al, 2011).

Содержательные и технологические проблемы существующих корпусов

Формально корпусом может быть названо любое собрание текстов с заявленным создателями принципами отбора, позволяющими оценить соответствие замысла и исполнения (так сказать, по законам, ими самими над собой установленными). Среди корпусов выделяются «универсальные», то есть такие, которые содержат языковой материал, адекватный по мысли создателей для практически любых исследовательских задач. Далее речь пойдет только об условно универсальных корпусах.

Несколько огрубляя, можно говорить о трех основных их типах с точки зрения состава и функционала:

1. Классические корпуса XX века, составленные вручную «закрытые» корпуса. Во многих случаях «корпус» рассматривается пользователями и даже авторами как материализация понятия «норма языка». К сожалению, далеко не всегда ясно, на чем основывается претензия на универсальность такого корпуса, и где ее «границы». Закрытые корпуса наиболее функционально развиты: исследователю может предоставляться возможность работать с подкорпусами, на которых произведено ручное снятие омонимии или проведена разметка, включая синтаксическую и семантическую. Такие подкорпуса позволяют искать и собирать статистику не только для отдельных слов, но и для выделенных при разметке лексических значений, синтаксических конструкций, целых семантических полей. Эти же размеченные подкорпуса используются для машинного обучения систем автоматического анализа текстов.

2. «Открытые», стихийно создаваемые без определенного плана массивы текстов, к каковым прежде всего относится интернет в целом. Несмотря на отсутствие плана считается, что в интернете можно найти всё. Это и дает некоторые основания считать его особым универсальным корпусом. Что касается функционала, то он предоставляет весьма ограниченные и, увы, не слишком надежные средства доступа. Индексирование Интернета поисковиками, подсчеты частот, обработка запросов проводится на основаниях, далеких от лингвистических. Перифразируя известную цитату можно сказать «Сеть наша велика и обильна, а порядка в ней нет».

3. Некоторым компромиссом между этими подходами является бурно развивающаяся область исследований, которую можно назвать «Автоматическое получение корпусов из Интернета». Лингвистический статус автоматических корпусов, собранных обычно с помощью запросов по ключевым словам, не вполне ясен. В хороших случаях авторы подвергают их автоматической жанровой классификации, что дает некоторое представление об их структуре, а затем используют как универсальные ресурсы для лингвистических исследований (Sharoff, 2010; Bidhauer, 2012). Именно это направление и развивает предлагаемый проект.

Проблемы закрытых корпусов

Многие проблемы обсуждаются в уже упоминавшейся работе (Беликов, 2011), особенно в ее более полном электронном варианте[1]. Основной ее вывод: «It is a common belief that text corpora provide the best testing ground for solving any kind of linguistic problems. As far as grammar is concerned, this may be true, but if we focus on investigating the lexicon the results often appear to be rather superficial».

Определение «грамматика» в данном случае следует читать как «универсальная грамматика», поскольку в области лексикализованной грамматики мы сталкиваемся с теми же проблемами недостаточности материала.

Лингвистика и лексикография вынуждены обращаться к открытым ресурсам по двум основным причинам: а) из-за колоссальной динамики языковых изменений и б) из-за региональной, социальной и профессиональной «сегментации» языка, требующей дифференциального подхода к анализу языковых явлений и наличия нужного для получения

[1] См. http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/8.pdf

надежных результатов объема языковых фактов.

Таким образом, основными содержательными недостатками закрытых корпусов являются «типологическая» неполнота или несбалансированность, часто случайность в отборе текстов и малый объем (особенно размеченной их части). Не принижая огромной роли замкнутых корпусов в решении многих задач, укажем на те, которые с их помощью решать не следует:

- Лексикосемантические и сильно лексикализованные синтаксические явления на языковой периферии (при этом, увы, с точки зрения надежности применения статистических методов верификации к периферии приходится относить очень многое).
- Исследование динамики языковых изменений.
- Социальная и региональная дифференциальная лексикография.
- Исследование терминологии.
- Сравнительные исследования текстов разных жанров.
- Фиксация и мониторинг языковой нормы.

Множественность нормы. Корпус и норма

До появления источников электронных текстов не было возможности объективно исследовать узус и выявлять на этом основании норму. Любая кодификация была субъективной. Однако, отождествление замкнутых универсальных корпусов (например, НКРЯ) с «нормой языка» требует весьма существенных оговорок.

Текстовое пространство сегментировано, и каждый сегмент, социальный или географический, имеет свою норму.

Хорошим примером может служить различие профессиональной и общей нормы, ярко проявляющееся, например, в именовании общеизвестных биологических объектов. Так, данные, собранные в рамках проекта «Языки русских городов» показывают, что денотатом известной строки «Шумел камыш…» обычно признаются растения, известные биологам как рогоз и тростник. В то же время толковые словари чаще закрепляют именно профессиональную биологическую норму, которая практически всегда имела определенные (но в разных случаях разные) лингвогеографические обоснования.

Охота на корпусных снарков

В лингвистике и лексикографии как объекты описания, так и их свойства делятся де-факто на основные (как бы важные) и периферийные (как бы не очень важные). Важные являются относительно хорошо документированными, периферийные описаны не всегда верно, неполно и непоследовательно.

Замкнутые условно универсальные корпуса отличаются следующей особенностью: они хорошо подходят для описания основных объектов, и плохо — для периферийных.

Состав явлений, попадающих в корпусную «периферию», весьма неоднороден. Очень плохо, что по большому счету никто ее структуру не изучал. Речь не всегда идет об абсолютно редких объектах. Часто периферией оказываются объекты, быстро меняющие свое поведение (в том числе — появляющиеся и исчезающие), склонные к вариативности или очень неравномерно распределенные.

Не претендуя на создание термина, назовем для удобства ad hoc эти трудноуловимые объекты корпусными снарками[2]. Охота на снарков в замкнутом корпусе является по определению делом достаточно безнадежным.

Типичными представителями корпусных снарков являются:

- единицы региональная лексики;
- общая лексика за пределами частотного словаря в 30–40 тыс. слов;
- новые значения и новые модели управления;

---

[2] *They sought it with thimbles, they sought it with care,*
They pursued it with forks and hope,
They threatened its life with a railway share,
They charmed it with smiles and soap.
The Hunting of the Snark (An Agony in Eight Fits by Lewis Carroll).

• объекты актуального паремийного фонда;
• социолекты;
• распределение конкурирующих способов выражения (включая явления в грамматике);
• etc.

Так, лингвист, пытающийся на основании встреченного газетного примера «Мэрия согласовала нам маршрут движения…» провести корпусное исследование необычной модели с помощью НКРЯ, найдет лишь 2 примера разных лет. Никакого исследовательского вывода из неудачной охоты на снарка сделать, разумеется, нельзя.

Обращение к Интернету, даже без возможности задать грамматическое значение в запросе, немедленно показывает, что у «согласовать» быстро появляется новое значение ‘разрешить’ с похожей моделью управления.

Конкурирующие способы выражения это не только лексические синонимы (у которых реальное использование и стилистическая квалификация часто зависит от возрастных и социальных характеристик говорящего, ср. *захворать/заболеть/приболеть*). Фактическая (некорпусная) норма отражает «разноспрягаемость» синонимичных глаголов лазать/лазить: статистическое большинство предпочитает вариант *лазит*, но избегает формы *лажу*. Вариативность формообразования ряда глаголов хорошо известна, но здесь также проявляется неописанная «разноспрягаемость» (преобладает *плескаюсь в*, но *плещется в, плещутся в*).

Очевидно, что корпусные снарки являются одновременно и словарными: относительно них нет и не может быть и надежных словарных фиксаций.

В области хуже в целом описанной фразеологии соотношение идиом-конкурентов практически не исследовано. Так, словари отражают лишь *телячий восторг*, но в узусе даже старших возрастов он уже заметно уступает *щенячьему*.

Анализ языковой «периферии» вообще является делом чрезвычайно сложным. В замкнутых корпусах снарки легко путаются с сопоставимыми по частоте шумами и девиациями. При этом явления, плохо отслеживаемые корпусной статистикой, составляют в целом заметную часть текстов. Тут на место поиска «with forks and hope», примеров которого в лингвистических работах имеется достаточно много, должны прийти надежные корпусные методы, возможные только с использованием открытых корпусов.

Проблемы открытых ресурсов

Примеров корректно проведенных лингвистических исследований на открытых корпусах пока не так много. Назовем из последних лексикографические исследования на основании анализа блогов (Б. Л. Иомдин 2012, В. И. Беликов 2012)

В некоторых случаях исследователям удается, опираясь на недокументированные возможности поисковых систем и/или во взаимодействии с их разработчиками, решать некоторые из указанных проблем, но систематического их решения ожидать вряд ли стоит.

Проблемы поиска стандартными поисковиками

Проблемы с использованием стандартных поисковиков широко известны, см. примеры в (Belikov, 2011), а также классический пример Veronis[3]:

Кратко эти проблемы таковы:

Надежность выдаваемой статистики

- Надежность выдаваемых цифр тем больше, чем цифры — меньше. А полностью доверять можно только тем цифрам, которые можно проверить полным просмотром выдачи.
- К числовым результатам не применимы аксиомы классической арифметики. Так, здесь часть вполне может быть больше целого (уточняя запрос получаем больше

[3] http://blog.veronis.fr/2005/01/web-googlean-logic-en.html.

«результатов», чем для грубого запроса).

- Имеет место нестабильность: результат меняется во времени в произвольную сторону (не связанную с реальным изменением числа релевантных объектов) (Rayson, et al, 2012).

Очевидно, что информация о частоте запроса предназначена для совершенно других целей, чем лингвистические исследования. Результат зависит от множества разных факторов, не имеющих отношения к частоте употребления слов и конструкций (в настоящее время Яндекс стал говорить «Нашлось 42 тыс. ответов», а Гугл «Результатов: примерно 143 000 000» без указания единицы измерения).

Ненадежность в интерпретации результатов поиска хорошо известна, но лингвисты все равно пользуются ими, поскольку традиционные корпуса не дают возможности ответить на большинство запросов о частоте употребления даже среднеупотребимых конструкций (см. Крылов 2007). Отчасти помогают N-граммы (но такой базы на основе русскоязычного Интернета пока нет). Однако, в них, разумеется, не снята омонимия, что в отсутствии исходных контекстов ограничивает их применение как корпуса. Еще большая проблема таких списков — полная неясность с тем, как в них отражается дублетность, а это может исказить реальную картину на порядок и больше.

Снятие омонимии, синтаксическая разметка, язык запросов

Для поисковых машин не очень важна проблема снятия даже морфологической неоднозначности в самом запросе. Задавая вопросы поисковой системе, пользователи быстро приучаются вводить необходимые фильтрующие шум модификаторы. С другой стороны, сами системы поиска адаптируются к статистике запросов массовых пользователей. И вольно или невольно оказываются на поводу у ловких промоутеров и оптимизаторов, чем безусловно искажают реальность.

Совершенная иная ситуация возникает при использовании поисковиков в качестве устройств для лингвистического исследования Интернета (в частности, Рунета).

Здесь при исходном запросе ‘*помятые брюки*’ полное равноправие вариантов «*не помяв новые брюки*» и «*подшивая брюки, помните, что сзади они должны доходить до каблука обуви*» вызывает серьезные проблемы для интерпретации статистики..

К сожалению, проблема в целом решается только параллельным снятием омонимии в запросе и в найденном тексте.

Для работы лингвистов корпус должен очевидным образом давать возможность задавать в запросе грамматические значения, пунктуацию, учет капитализации и прочее, к чему привыкли пользователи закрытых корпусов. Соответствующее направление развития на наш взгляд не совпадает со стратегией развития универсальных поисковиков, ориентированных в бизнес-процессе даже не столько на информационный, сколько на потребительский поиск товаров и услуг.

История с географией

Различие в подходах хорошо заметно на примере с географической атрибуцией интернет-страниц. В недавней работе (Volkov, Serdyukov, 2012) описывается алгоритм, применяемый в системе Яндекс.

Видно, что выбор параметров машинного обучения ориентирован на событийную, а не языковую специфику. В этом случае текст новосибирца о поездке в Москву будет атрибутирован скорее всего как московский. Сказать, правильно это или нет, не имея определения региональности и списка задач, ради которых она определяется, достаточно затруднительно[4]. Ясно, что для исследований регионального языка следует опираться на

[4] Другой вопрос, которого мы здесь не касаемся, это насколько последовательно реализовано то, что можно назвать информационно-региональной привязкой. Желающих адресуем к статье [В. И. Беликов. «Яндекс-регионы: найдется всё, выдастся без разбору»], являющейся дополнением к данной, в электронных материалах Диалога-2012.

надежно установленные языковые региональные особенности, а не (только) на реалии, топонимы и проч.

Например, ФНС тщательно контролирует использование кассовых аппаратов. Запрос о *непробитых / невыбитых чеках* по делам, рассмотренным в 17 Арбитражном апелляционном суде (обслуживает Пермский край, Свердловскую область и Удмуртию) даст не более четверти релевантных документов, поскольку в остальных речь будет идти о *неотбитых чеках*.

Проблемы дублирования и скрытого цитирования

Это одна из самых неприятных проблем, требующая колоссальных усилий от добросовестного исследователя. Тут есть вещи относительно очевидные — многократная перепечатка текстов, не очень очевидные — цитаты, и совсем не очевидные — паремии.

Исследователю, работающему с открытым корпусом, должны быть ясны принципы, на которых поисковик относит тексты к дублетам. Проблему, какие дубли возможны и важны, а какие должны удаляться, мы полагаем совершенно нерешенной.

Неполнота и непоследовательность метатекстовой разметки

Печальным недостатком Интернета как открытого корпуса является отсутствие метатекстовой разметки. Даже дата создания является крайне ненадежным параметром, поскольку время публикации только для определенных сегментов Интернета совпадает с датой написания.

При этом жанрово однородных сегментов в Интернете очень мало. Имеющаяся в отдельных сегментах Интернета метатекстовая разметка непоследовательна. Скажем, даже в Классике библиотеки Мошкова беллетристика, поэзия и эпистолярий не разделены. В Журнальном Зале тексты упорядочены по изданиям, но жанровой классификации нет. Блоги наиболее насыщены метатекстовой информацией, но и здесь сложно говорить о полноте и последовательности ее приписывания. Так, в некоторых регионах блоги mail.ru не размечаются по возрасту, что делает практически невозможным проведение исследований по этом параметру. Кажется, что метатекстовая разметка в блоге имеет характер бонуса, а не обязательного параметра, за который отвечает разработчик поисковика.

Структурная неоднородность страниц

Страница Интернета в общем случае является сложно устроенным документом, с большим многообразием типов структур. Имеются два типа неоднородностей:

- Информацию на странице можно разделить на существенную для анализа постоянную «ингерентную» часть и динамическую составляющую: рекламу, новости, общую для нескольких страниц служебную информацию. К сожалению, страницы часто индексируются целиком.
- Часто структурно неоднородна и «ингерентная часть» страницы. Так, в Классике библиотеки Мошкова тексты авторов и комментарии к ним (обычно второй половины XX века) хранятся одним документом[5]. Очевидно, что языковые свойства этих разделов сильно различаются.

Программа проекта

Перечисленные выше проблемы можно решить только с помощью корпусов нового типа, сочетающих необходимую полноту исследовательского материала с наличием релевантной лингвистической разметки и основанной на ней надежной статистики.

Процитируем еще раз работу (Belikov, 2011): «WWW contains some relatively homogeneous arrays of texts formed independently of linguists, in some cases emerging quite spontaneously…Frequencies of the same lexical items differ greatly from one segment to another,

[5] На ФЭБе это разные документы, хотя все «пушкинское» лежит вместе и никаким сужением адреса эти части не разделить.

and this statistics is very significant for sociolinguistics. The main problem in applying the method of segmental statistics is the lack of a suitable instrument for automatic data processing". То есть, возможность детального исследования многих типов узуса есть, но реализовать эту возможность без специального инструментария можно только «точечно», лишь в отношении конкретных частных языковых фактов, при этом такое исследование сегодня является весьма трудоемким.

До недавнего времени такая задача казалась неразрешимой, однако, мы полагаем, что сегодня имеются уже все основания для успешной реализации проекта, отвечающего нуждам лингвистов, и такой проект для русского языка (Генеральный интернет-корпус РЯ) должен быть обязательно запущен.

Если обобщить все претензии и пожелания к имеющимся ресурсам и инструментам доступа к ним, то получится следующая программа для проекта.

- Генеральный корпус должен быть настолько большим, чтобы быть релевантным для решения задач дифференциальной лингвистики и лексикографии (лексикографии жанровых, социальных и региональных различий). Поскольку речь не идет о задачах информационного поиска, объем корпуса не должен быть равным объему всего русскоязычного Интернета. Тем не менее, он должен быть на несколько порядков больше современных условно универсальных корпусов РЯ (10–100 миллиардов словоупотреблений).
- Этот корпус должен представлять все существенные социальные, жанровые, тематические сегменты Интернета и давать статистически достоверную картину относительного распространения текстов данного сегмента в сети.
- Генеральный корпус должен обновляться синхронно с обновлением Интернета (постоянно). Для целей обучения могут фиксироваться некоторые его версии.
- Интерфейс к корпусу должен обеспечивать поиск и подсчет частот с учетом любых параметров метатекстовой разметки.
- Поиск должен обеспечиваться технологиями статической (на уровне индекса) и динамической (на уровне обработки выдачи по покрывающему запросу) автоматической лингвистической разметки, позволяющей искать и статистически оценивать любые параметры языковых структур.
- Корпус должен предоставлять возможность использовать для динамической разметки альтернативные лингвистические модели.

Рассмотрим, насколько достижимы эти цели на современном уровне технологий, и что должно быть сделано в рамках проекта по созданию подобного корпуса.

Получение корпуса нужного объема и состава

Генеральный корпус должен быть открытым подмножеством Рунета, постоянно подпутывающимся новыми образцами. Получение сбалансированного корпуса нужного объема потребует многих итераций работы краулера на основании параллельно модифицируемого сегментного классификатора и сегментной карты Интернета.

Используемый нами термин *сегмент* применяется для указания на хорошо выделяемые и однородные в отношении некоторых параметров метатекстовой разметки подмножества текстов в Интернете. Сегменты могут быть физически компактными, например, электронные библиотеки или блоги, или виртуальными, распределенными по интернету. Разумеется, наиболее эффективна работа с компактными сегментами, которые систематически пополняются релевантными текстами.

Одним из наиболее ценных физических сегментов является блогосфера. Это, безусловно, наиболее динамичный, актуальный и размеченный массив текстов. Кроме того, только здесь и на форумах есть представительная информация о диалоговом общении.

Метатекстовая классификация и разметка

Частично классификация может вестись по материалам метаданных в отдельных ресурсах (пол, регион, возраст автора, иногда и его образовательный уровень в блогах, время

создания/перевода текста в литературных коллекциях), но основу метатекстовой разметки в корпусе такого объема может составить только автоматическая классификация страниц.

В настоящее время достаточно хорошо себя зарекомендовали методы классификации по темам и жанрам, использующие самые простые и доступные признаки, такие как частота морфологических тегов или символьных N-грамм (Sharoff, 2010; Sharoff, et al, 2010).

В этом вопросе, однако, у авторов нет полного согласия. Необходимо проводить серьезные исследования с различными априорными классификаторами, выбор которых отнюдь не однозначен. В частности необходимы исследования по параметрам автоматической региональной привязки.

В интерфейсе также предполагается визуализация различий между исследуемыми группами текстов (см. Biber, 1989). Например, на основе частотности поверхностных признаков текста в сравнении с аннотированной нормой можно провести многомерное шкалирование (multidimensional scaling) либо анализ методом главных компонент (principal component analysis), чтобы определить позицию данного текста по шкале аргументативности или нарративности и сравнить его со средним значением для данного сегмента Интернета (Wilson, et al, 2011).

Автоматическая лингвистическая разметка

Полная ручная разметка Генерального корпуса заведомо невозможна, а частичная — бессмысленна, если не говорить о каких-то отдельных узких сегментах Интернета, позволяющих сделать компактный, но достаточно представительный подкорпус.

Нужны новые инструменты, позволяющие производить автоматическую лингвистическую разметку (лемматизация, аннотирование частеречных признаков и синтаксический анализ) текстов, не связанную уже с ограничением на объем.

Было показано, что современные технологии морфологического и синтаксического анализа (Sharoff, Nivre, 2011) могут успешно работать на корпусе в 1–2 миллиарда словоформ.

В лингвистической разметке полезно различать условно физический и логический уровни. Генеральный корпус должен быть физически размечен признаками, с которыми готовы работать большинство исследователей. Что касается логического уровня, то он обеспечивается динамической постобработкой базового запроса на основании лингвистической модели разметки, которая может иметь альтернативы. Так, вряд ли можно говорить о едином синтаксическом представлении.

Системы семантической разметки еще сложнее унифицировать. Назовем для примера разметки на основе PropNet, систему разметки, основанную на модели Смысл-Текст (Apresjan et al. 2006), автоматическую семантическую разметку в системе Compreno (Selegey, 2012).

Очистка и структурирование страниц

Очистка страницы имеет насколько уровней: от избавления от форматирующей информации до определения подструктур документа (см. выше про неоднородность).

Это очень важная тема, на которой мы не можем здесь останавливаться подробно. В целом, имеется определенная аналогия между задачами распознавания структуры документа в оптическом распознавании и для структурирования интернет-страниц. В нашем случае помогает доступная HTML-разметка.

Нужно иметь представления о формате основных коммуникативных форм: форумы, блоги, чаты.

Новые типы информации, представляемые корпусом

Лингвистическая и метатекстовая разметка позволяют решать задачи, которые сегодня еще не доступны исследователям, работающим с корпусами. Приведем только 2 примера:

- Переход в оценке от отдельных вхождений к числу страниц и далее — документов, был очень важен. Но все же наиболее надежная из всех видов статистики — авторская.

Особенно эффективно инкорпорированность лексики в идиолекты можно оценивать в блогосфере. Легко выяснить, сколько блоггеров употребили за свою историю конкретное слово (словосочетание и т. п.). Для новых выражений при погодовых срезах отчетливо виден, например, переход от *конкретных* (*чуваков*) к *реальным* (*пацанам*), от *тусоваться* к *тусить*. Но в отсутствии разметки приходится «изворачиваться». Так, лингвист, интересующийся распределением авторов между *лажу/лазию*, должен смотреть *не лажу/ не лазию*. (без отрицания на *лажу* получишь в основном *лажу*). И можно, наконец, выяснить, кого больше: тех, кто говорит *лингвистика* или тех, кто говорит *языкознание.*

- Возможности выявления фонетической специфики. Использование поэтических и песенных сайтов (*stihija.ru, stihi-rus.ru, rupoem.ru, pesni.net* и т. п.) позволяет по размеру делать наблюдения над ударением (можно доказать, например, широкую распространенность «запрещаемого» ударения *пико́вый*, — оно обычно для всех контекстов, кроме пиковой дамы, но и *пико́вая дама* встречается), по рифмам выявлять относительную встречаемость [чн] / [шн] в словах типа *скучно*, *булочная*, *конечно*. Много *булошных* и не ожидалось, но торжество *коне*[ч]*но* удивило. Некоторые ресурсы (*stihija.ru*, например) производят сугубо графоманское впечатление. Но для лингвистики это скорее хорошо: человек «фонетизирует» без оглядки на традицию.

Открытые вопросы для изучения

Переход к открытым исследовательским корпусам с автоматической разметкой порождает новые важные вопросы :

1. В какой степени корпуса, размеченные автоматически, пригодны для целей лингвистики и лексикографии, и как следует оценивать качество их разметки. Несомненно, что интегральная оценка качества более или менее адекватна для компьютерно-лингвистических приложений (для информационного поиска), но в случае исследовательских задач нельзя полагаться на средние данные (что-то вроде средней температуры по больнице) и нужно применять дифференцированные подходы к оценкам качества разметки с учетом различных языковых явлений.
2. Обоснование выбора языков разметки. Чем дальше мы продвигаемся от морфологии и частных задач разрешения лексической неоднозначности (основанных на каких-то канонических системах лексических значений типа WordNet), тем больше возникает проблем, связанных с различием подходов к описанию синтаксиса и семантики. Именно поэтому в современной корпусной лингвистике давно уже принят подход, различающий физическую разметку корпуса и его логические разметки. Важно, чтобы проект Генерального интернет-корпуса предусматривал возможность добавления альтернативных разметок.
3. Большой размер универсального корпуса не во всех случаях является благом. Необходимо изучать вопросы определения выделения исследовательского подкорпуса — такого, на котором наиболее эффективно проводить конкретное исследование, включая и машинное обучение (Gasco, et al. 2012).

Заключение

Проект ГИКРЯ ни в коей мере не является альтернативой проектам создания образцовых замкнутых условно универсальных корпусов русского языка (например, НКРЯ) или специализированных корпусов, ориентированных на исследование определенных явлений (референции, афазии, etc). Важно понимать, для решения каких задач какой тип корпуса эффективнее.

Реализация проекта ГИКРЯ невозможна без широких обсуждений и участия заинтересованных исследователей. У него есть все шансы стать «рамочным» проектом — настоящим полигоном для апробации новых методов компьютерной лингвистики и лексикографии, прежде всего — методов автоматической классификации и разметки.

References

1. *Apresjan J., Boguslavsky I., Iomdin B., Iomdin L., Sannikov A., Sizov V._* A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, 2006. pp. 1378–1381.
2. *Belikov V._* 2011. What are sociolinguists and lexicographers lacking in a digitized world? Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011". Moscow. pp. 60–67.
3. *Belikov V._* 2012. Segment-statistical approach to Internet as a corpus (on the example of blogosphere's analysis) Available at: http://www.abbyy.ru/science/ seminars/archive/
4 *Bidhauer F., Schafer R._* COW und texrex: Gigatoken Webkorpora and Tools zur Ad-hoc-Webkorpuserstellung. 34. Jahrestagung der Deutschen Gesellschaft fur Sprachwissenschaft, Frankfurt am Main, 2012.
5. *Gasco, G., Rocha, M., Sanchis-Trilles, G., Andres-Ferrer, J., Casacuberta, F,_* 2012. Does more data always yield better translations? EACL2012, Avignon, France.
6. *Gulli, Signorini,_* 2005. The Indexable Web is More than 11.5 billion pages. WWW05, Chiba, Japan.
7. *Iomdin B., Lopukhina A., Piperski A._* et al. Thesaurus of Russian everyday life terminology: new problems and new techniques. Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2012". Vol.1 Moscow. pp. 213–226.
8. *Rayson P., Charles O., Auty I.,_* Can Google count? Estimating search engine result consistency. Proceedings WAC'7, April 17, 2012, Lyon, France.
9. *Selegey V._* 2012. On automated semantic and syntactic annotation of texts for lexicographic purposes. Proceedings of the International Conference "Euralex 2012"
10. *Sharoff S._* In the garden and in the jungle: Comparing genres in the BNC and Internet. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, Genres on the Web: Computational Models and Empirical Studies, pages 149–166. Springer, Berlin/New York, 2010.
11. *Serge Sharoff, Zhili Wu, and Katja Markert._* The Web library of Babel: evaluating genre collections. In Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010, Malta, 2010.
12. *Sharoff S., Nivre J.,_* 2011. The proper place of men and machines in language technology. Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011". Moscow. pp. 591–604.
13. *Volkov A., Serdyukov P._*, Unified Classification Model for Geotagging Websites. Proc. of WWW 2012 April 16–20, 2012, Lyon
14. *Wilson, J., Hartley, A., Sharoff S., Stephenson, P._* (2011). Advanced corpus solutions for humanities researchers. Proceedings of the Pacific Asia Conference on Language, Information and Computation 2010.
15. *Krylov S. A._*, 2007. On the frequency of invective statements with universal reference and their negative correlates ("Yandex" as assistance to WCIOM: an attempt of quantitative analysis). Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2007". Moscow.