ВЕБ-ПРОСТРАНСТВО КАК ЯЗЫКОВОЙ КОРПУС

THE WORLD WIDE WEB AS LINGUISTIC CORPUS

В. П. Захаров

vz1311@yandex.ru

Кафедра математической лингвистики, Филологический факультет, Санкт-Петербургский государственный университет, г. Санкт-Петербург, Россия

В докладе анализируются поисковые системы Интернета как инструменты лигвистических исследований. Веб-пространство при этом рассматривается как большой корпус. Рассматриваются языки запросов, грамматические средства, выходные интерфейсы существующих поисковых систем. Приводятся данные экспериментов.

Введение

Во всех лингвистических исследованиях существенное значение имеет проблема выборки и репрезентативности анализируемого материала. В последнее время в научный оборот лингвистов все шире входит понятие «корпусов текстов» (КТ), на базе которых развивается корпусная лингвистика [1]. Это понятие сосуществует и нередко сливается в научной литературе с такими понятиями, как «коллекция текстов», «полнотекстовая база данных», «электронный архив», «электронная библиотека». Под *корпусом текстов* в узком смысле обычно понимают **унифицированный, структурированный и размеченный** массив языковых (речевых) данных в электронном виде, предназначенный для определенных филологических и, более широко, гуманитарных исследований. Целесообразность создания КТ определяется двумя предпосылками:

1) данные разного типа находятся в КТ в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;

2) исторически и стилистически репрезентативный объем КТ гарантирует типичность представления языковой информации.

Информационное наполнение сети Интернет (веб-пространство) может рассматриваться как огромный многоязычный корпус. Главный материал лингвистического анализа – язык, зафиксированный в виде речевых произведений – в Интернете представлен в огромном объеме и разнообразии и непосредственно доступен для машинной обработки. Этот факт представляет для лингвистов большую ценность, так как на перевод текстов в машинную форму и на создание корпусов приходилось и приходится тратить много времени, сил и денег.

Следует заметить, что текстовые массивы Интернета широко используются как источник данных для формирования корпусов. Это отдельная тема, которая здесь не обсуждается. Так же широко тексты, представленные в Интернете, используются как тестовый материал для различных программ анализа и обработки текстовой информации (особенно тех, которые базируются на статистических и вероятностных методах).

В то же время веб-пространство может рассматриваться и непосредственно как корпус. Особенно активно эта проблема стала обсуждаться после доклада А.Килгариффа в 2001 г. [2]. Очевидно, что ни один корпус не может сравниться по репрезентативности языкового материала с вебом, куда включаются материалы и других Интернет-сервисов (например, электронной почты). При этом, однако, встает вопрос о сбалансированности веб-корпуса. Очевидно, что в Интернете определенные типы речевых произведений представлены относительно чаще, чем это было в языке до сих пор. Однако и в настоящих корпусах вопрос о сбалансированности решается очень приближенно. И поскольку при формировании корпуса осуществляется отбор текстов, то следует ожидать, что и языковой материал корпусов нередко оказывается необъективным. Это тем более справедливо, если учесть, что при отборе текстов, как правило, во внимание берутся экстралингвистические критерии.

Поисковые системы как корпус-менеджеры

При использовании веб-пространства как корпуса роль корпус-менеджеров (corpus manager) выполняют поисковые системы. В Интернете имеются системы классификационного типа, напоминающие библиотечные каталоги (directories, русское распространенное название каталоги-справочники). Базы данных этих систем в некотором смысле могут рассматриваться как корпуса семантического типа.

Однако основным средством поиска информации в сети сегодня следует считать глобальные ИПС вербального типа (поисковые машины – search engines), индексирующие (по крайней мере, претендующие на это) все Интернет-пространство. Индексы вербальных систем – это как бы конкордансы к текстам. При этом

полезно представлять, как эти индексы строятся и, соответственно, учитывать эти особенности при использовании баз данных поисковых систем как материала для лингвистических исследований.

Существует большое количество таких систем, отличающихся друг от друга языком запросов, дизайном, сервисом и другими особенностями. К числу главных поисковых систем вербального типа (в первую очередь, по объему базы данных) следует отнести следующие: Google, Fast Search (AllTheWeb), AltaVista, WiseNut, HotBot, MSN Search, Teoma [3]. Среди российских систем главными являются три: Яндекс (Яndex, Yandex), Рамблер (Rambler), Апорт! (Aport).

В составе любой поисковой системы можно выделить три основные части:

Робот – подсистема, обеспечивающая просмотр (сканирование) Интернета и поддержание инвертированного файла (индексной базы данных) в актуальном состоянии. Этот программный комплекс является основным средством сбора информации о наличии и состоянии информационных ресурсов сети.

Поисковая база данных – так называемый *индекс* – специальным образом организованная структура данных (англ. index database), включающая, прежде всего, инвертированный файл, состоящий из лексических единиц, взятых из проиндексированных веб-документов, и содержащий разнообразную информацию об этих единицах (в частности, их позиции в документах), а также о самих документах и сайтах в целом.

Поисковая система – подсистема поиска, обеспечивающая обработку запроса (поискового предписания) пользователя, поиск в базе данных и выдачу результатов поиска пользователю. Поисковая система общается с пользователем через пользовательские интерфейсы – экранные формы программ-браузеров: интерфейс формирования запросов и интерфейс просмотра результатов поиска.

Индексный файл (или просто индекс) представляет собой набор связанных между собой файлов, ориентированных на быстрый поиск данных по запросу. В основе индекса всегда лежит инвертированный файл. *Инвертированная схема* организации поискового массива основана на принципе обеспечения доступа к документам через их идентификаторы содержания. Такую схему получают путем обработки последовательного массива документов с целью создания специальных вспомогательных инвертированных файлов – точек доступа. Эта структура данных известна давно из "бумажных" поисковых систем, называемых конкордансами – алфавитно упорядоченными списками слов из одного или нескольких текстов с указанием их ближайших контекстов.

Каждая запись такого вспомогательного массива идентифицирована соответствующим идентификатором содержания (дескриптор, ключевое слово, термин, имя автора, название организации и т.п.) и содержит имена (адреса хранения) всех документов, в которых он содержится. Для каждого идентификатора содержания (поискового элемента данных) в инвертированном массиве вместе с адресом (именем) документа может храниться дополнительная информация, как-то: имя поля, номер предложения, в составе которых этот элемент встретился в данном документе, номер слова в предложении и т.д. Фиксация положения слова в тексте с точностью до номера предложения и номера слова в предложении дает возможность построить гибкий язык запросов, позволяющий задавать расстояние между словами в документе. Позиционные характеристики также используются при вычислении коэффициента релевантности и ранжировании документов в выдаче.

Как уже говорилось, индексы (инвертированные файлы) поисковых систем – это, по сути, не что иное, как виртуальные конкордансы к текстам. Более того, результаты поиска в ИПС в виде кратких описаний документов, как правило, содержат контексты, в которых искомые слова встретились в найденных документах. Отличие лишь в том, что конкордансы обычно составляются к конкретному произведению или группе произведений (например, все тексты одного и того же автора), в то время как ИПС Интернета индексируют все доступное множество электронных документов.

Главная содержательная проблема при индексировании веб-сайтов заключается в том, какие термины приписываются документам, откуда они берутся. Не все термины из документов и не всегда попадают в индексы. Активно применяются списки запрещенных слов (stop-words), которые в индекс не попадают – это общая, служебная лексика (предлоги, союзы и т.п.) и незначащие слова. Многие системы индексируют лишь часть документа (обычно начальную), есть роботы, которые обрабатывают только часть веб-страниц с одного и того же сайта. Знание того, как работают роботы, каковы их технические характеристики, полезно и для создателей веб-документов, и для составителей запросов при поиске. Подробное описание работы роботов можно найти в Сети[1]. Сведения о большом количестве роботов (в количестве 298) можно почерпнуть из базы данных The Web Robots Database[2].

Особенности построения и структура индекса напрямую связаны с языком запросов и возможностями поисковых систем. Наиболее важными с точки зрения лингвистического анализа текстового материала представляются следующие особенности ИПС:

- "грамотная" работа со словоформами – способность ИПС отождествлять разные словоформы одной и той же лексемы, по-другому, порождать каноническую форму – лемму, и возможность выделять среди множества словоформ конкретную форму;

[1] См., в частности, http://www.searchengineworld.com/robots/norobots.htm

[2] http://www.robotstxt.org/wc/active/html/

- поиск слов с заданным или произвольным усечением, как правым, так и левым;
- индексирование полных текстов в полном объеме без исключения. Многие системы, как уже говорилось, не включают в индекс служебную и незначимую лексику;
- работа со словосочетаниями – учет расстояния между элементами словосочетаний и порядка их следования;
- различение больших и малых букв.

Также важно, какую информацию и в каком виде можно извлечь из выходных интерфейсов ИПС. Интерфейс выдачи (форма представления результатов) у разных систем включает такие параметры, как статистика слов из запроса, количество найденных документов, количество найденных сайтов, количество документов на странице с результатами поиска, средства управления сортировкой документов в выдаче, описание сайта, с которого взят соответствующий документ, описание документа. Последнее, в свою очередь, содержит в своем составе заглавие документа, URL (адрес в сети), размер документа (объем), дата создания, кодировка, аннотация (краткое содержание), визуальное выделение в аннотации слов из запроса, указание на другие релевантные веб-страницы того же сайта, ссылка на рубрику каталога, к которой относится найденный документ или сайт, коэффициент релевантности, ссылки на другие возможности поиска (поиск похожих документов, поиск в найденном).

Из всех этих реквизитов для задач лингвистического исследования наибольший интерес представляют частотные характеристики и выдача контекста. Следует различать два типа частот, учитываемых и выдаваемых системами, пословную и подокументную. Сведения о количестве языковых единиц в разных системам и разных режимах поиска могут относиться как к словоформам, так к лексемам. Некоторые системы ведут журнал запросов с возможностью повторных поисков и выдачей статистики по запросам. Полезной и интересной возможностью является также отнесение документов к тематическим классам.

Экспериментальная база

На нескольких примерах покажем возможности поисковых систем для получения лингвостатистических данных о частоте использования тех или иных слов или словосочетаний. В принципе, нас, как правило, интересуют относительные частоты, а для этого достаточно проведения сравнительных поисков в рамках одной ИПС. Однако для того, чтобы убедиться в достоверности данных и показать особенности разных систем мы выбрали для эксперимента пять ИПС, наиболее популярных и обладающих наиболее развитым лингвистическим обеспечением. В первую очередь, это российские ИПС Яндекс, Рамблер и Апорт. Возможно, наиболее мощный лингвистический аппарат имеет ИПС "Артефакт" (фирма "Интегрум-Техно, г. Москва), однако ее наполнение по составу документов заметно отличается от других и, главное, эта система является коммерческой и ее базы данных массовому пользователю недоступны. Из западных систем (к сожалению, в большинстве своем не обладающих развитыми лингвистическими средствами анализа текстового материала), мы взяли хорошо известные ИПС Google и AltaVista.

Кратко охарактеризуем особенности этих систем применительно к нашим задачам (наличие или отсутствие соответствующих возможностей помечено знаками "+" и "–")

	Яндекс	Рамблер	Апорт	Google	AltaVista
Поиск по лемме	+	+	+	–	–
Поиск по словоформам	+	+	+	+	+
Учет синтагм [3]	+	+	+	+	+
Учет больших и малых букв	+ (в синтагмах)	+ (в синтагмах)	–	–	–[4]
Частота пословная	+	–	–	–	–
Частота подокументная	+	+	+	+	+

Табл. 1. Характеристика поисковых систем

[3] неразрывныя словосочетания

[4] Ранее большие и малые буквы различались; в ныне работающей версии эта возможность отсутствует.

"Поиск по лексемам" означает, что результат сравнения слов документов и запросов признается положительным при наличии в документе любой формы слова из запроса, что обеспечивается механизмом автоматической лемматизации.

"Поиск по словоформам" означает, что результат сравнения документов и запросов признается положительным при наличии в документе словоформы, точно совпадающей со словоформой из запроса, что происходит при отсутствии автоматической лемматизации или обеспечивается особым механизмом учета словоформ.

"Частота подокументная" означает, что в результате поиска выдается сообщение о количестве релевантных документов, т. е. документов, содержащих данное слово (словоформу) или словосочетание.

"Частота пословная" означает, что в результате поиска дополнительно выдаются сведения об общем количестве словоупотреблений данной лексемы (независимо от формы) или конкретной словоформы в поисковой базе данных (индексе).

Примеры экспериментальных исследований

Начнем наши лингвистические изыскания с вопроса, как следует называть программу просмотра веб-страниц (англ. browser): "броузер" или "браузер". В нормативных словарях русского языка это слово отсутствует. Поиск в "Яндексе" дает следующие результаты: статистика слов: *броузер*: 2737640, *браузер*: 10349916; запросов за месяц: *броузер*: 4351, *браузер*: 23102[5]. Из этих данных можно сделать вывод, что написание "браузер" в настоящее время утверждается как языковая норма.

[5] Два с половиной года назад эти данные выглядели следующим образом: статистика слов: *броузер*: 472847, *браузер*: 997666; запросов за месяц: *броузер*: 2150, *браузер*: 5335

Поиск со словами "пергамент" и "пергамен" показывает: статистика слов: *пергамен*: 635, *пергамент*: 59585; запросов за месяц: *пергамен*: 4, *пергамент*: 240.

Еще один пример. Анализ поисковой базы "Яндекса" показывает, что наряду с написанием "офсайд" (122030 словоупотреблений, 84652 веб-страницы) в русском языке также достаточно широко используется написание "оффсайд" (41942 словоупотребления, 27392 веб-страницы). Синонимичное словосочетание "вне игры" по данным "Яндекса" (142107 веб-страниц) используется чаще, чем "офсайд", что можно объяснить, по-видимому, его частым использованием в переносном смысле[6].

Еще одно наблюдение. Данные таблицы 2 показывают, что неологизм Н.В. Гоголя "*мартобря 86 числа*" (. "Записки сумасшедшего") зажил в русском языке вполне самостоятельной жизнью.

	Ян-декс	Рам-блер	Goo-gle	Alta Vista	Апорт
"марто-бря 86 числа"	31	17	38	19	7
Марто-бря	915	499	676	359	190

Табл. 2. Частота употребления слова "мартобря" (количество найденных докум

[6] Два с половиной года назад показатели были следующие: "*офсайд*" (27168 словоупотреблений, 19106 веб-страниц) по сравнению с 34217 веб-страницами для "*вне игры*". По этим цифрам и данным предыдущего примера можно судить о росте объемов документальной информации в Интернете.

ентов (веб-страниц))

Приведем частоту употребления в русскоязычном Интернете некоторых редких слов и словосочетаний.

	Яндекс	Рамблер	Google	Alta Vista	Апорт	Примечания
	док.	док.	док.	док.	док.	
Кейф	936	402	536	181	154	
Кайф	21994 9	97904	94000	43496	1983	
Пампа	4070	1619	1500	757	558	
Пампасы	24103	6972	6150	3467	1610	
Падекатр	84	58	122		29	
па-де-катр	3518	280	257	145	100	
Падетруа	875	54	26	13	8	
па-де-труа		668	709	388	235	
"терновый венок"	537	331	380	234-	287	синтагма
"терновый венец"	4229	2469	3550	1825-	1860	синтагма

Табл. 3. Частота употребления отдельных лексических единиц

Другие примеры лексикологических изысканий приводятся в докладе.

В ряде случаев базы данных ИПС Интернета можно использовать для изучения и грамматических явлений. Например, предлог "*согласно*" по нормам русского языка требует после себя дательного падежа (*согласно предписанию, согласно уставу*) [4, с.178], однако в последнее время нередко можно услышать или прочитать конструкции, когда этот предлог управляет родительным падежом. Проверим, насколько сильна эта тенденция.

	Яндекс		Google	Примечания
	док.	сайтов	док.	
"согласно постановления"	30534	1642	4540	Синтагма
"согласно постановлению"	138950	1554	61700	Синтагма
"согласно приказа"	9969	1672	4190	Синтагма
"согласно приказу"	46783	1622	25800	Синтагма

Табл. 4. Частота употребления падежных конструкций с предлогом "согласно"

Как видим, норма еще держится, однако число неправильных употреблений превышает уже 20%.

Подобные изыскания каждый лингвист может провести в большом количестве и с высокой степенью достоверности, не тратя времени на сбор текстового материала.

Кроме того, Интернет – источник данных для так называемых параллельных корпусов. Упомянем здесь хотя бы систему STRAND (Structural Translation Recognition Acquiring Natural Data) [5].

Выводы

Данные экспериментов косвенно позволяют судить и об объеме баз данных различных систем. Можно сделать вывод, что наиболее полно русскоязычный Интернет представлен в системе "Яндекс". Наибольшие нарекания в последнее время вызывает "Апорт", по причине как малых объемов выдачи, так и необъяснимости отдельных результатов. В то же время следует обратить внимание, что сравнительно небольшие объемы выдачи в Google и AltaVista объясняются еще и тем, что сравнение терминов запроса и документа ведется по словоформам, а не по лексемам. Если в запросах Google и AltaVista через оператор "ИЛИ" задать все словоформы той или другой лексемы (а для глаголов это еще и причастия, и деепричастия), то результаты будут другими. Далее, если ограничить область поиска временны**м** интервалом, а именно, отбирать документы только последнего периода, например, за последние 6 месяцев, то на первом месте по объему выдачи даже при поиске по словоформам иногда оказывается Google, что говорит о том, что эта система индексирует информационное пространство Интернета наиболее оперативно и полно. С точки зрения частотных характеристик и возможностей языка запросов наиболее полезной для лингвостатистических исследований, безусловно, является система "Яндекс". К слову

сказать, именно на поисковом механизме "Яндекса" базируется первый общедоступный экспериментальный корпус русского языка[7].

Кроме поиска во всем Интернете (точнее, в той его части, которая индексируется соответствующей поисковой системой), поисковые системы позволяют формулировать запросы, ограничивающие область поиска определенными формальными признаками (например, заглавие, сайт, дата, домен, регион и т.п.).

И все же основные режимы использования Интернета как корпуса ограничены изучением лексического материала. И здесь возможности очень велики. Что же касается грамматических исследований на базе Интернета, то без предварительной металингвистической разметки они сводятся к минимуму.

Также и выходные интерфейсы поисковых систем заметно уступают программам-конкордансерам и корпус-менеджерам. Следует, однако, заметить, что ничто не стоит на месте, и в последнее время появились специальные системы интерфейсного типа: корпус-менеджеры по форме запросов, служащие посредниками между лингвистом и глобальными ИПС. В качестве примера можно привести систему WebCorp.[8]

Список литературы

1. Leech G. The State of Art in Corpus Linguistics // English Corpus Linguistics / Aimer K., Altenberg K.(eds.). London: 1991. P. 8-29.
2. Kilgariff A. Web as corpus // Proc. of Corpus Linguistics 2001 conference (Lancaster University). Lancaster: 2001. P. 342-344.
3. Greg R. Notess. Search Engine Statistics: Relative Size Showdown // http://www.searchengineshowdown.com/stats/size.shtml
4. Словарь русского язвка в четырех томах. Изд. 3-е / Ред. А.П. Евгеньева. М.: Русский язык, 1988. Т.3.
5. Resnik, P., Smith, N.A. The Web as a Parallel Corpus // Computational Linguistics , 2003. Vol. 23(N3). P.349-380.

[7] http://www.ruscorpora.ru

[8] http://www.webcorp.org.uk/