ИССЛЕДОВАНИЕ ГОРОДСКОЙ ДИАЛЕКТНОЙ ЛЕКСИКИ С ПОМОЩЬЮ ПОИСКОВЫХ СИСТЕМ

THE STUDY OF LOCAL URBAN DIALECTS VOCABULARY BY MEANS OF SEARCH ENGINES

Богданов А.В. (bidon@inbox.ru), *Московский государственный университет*

В работе рассматривается способ исследования диалектной лексики при помощи поисковых систем в сети Интернет. Приводятся некоторые примеры конкретных исследований, обсуждаются проблемы, встающие перед исследователями, а также способы их решения. Говорится о перспективах использования полученной таким образом информации.

1. Введение

С массовым развитием сети Интернет в России у лингвистов, имеющих дело с русским языком, появляется еще один источник для получения языковых данных. Пользователи Интернета общаются в сети, ищут информацию, публикуют сообщения в блогах и на форумах. Все это они делают в большинстве случаев, используя свой родной язык.

Интернет в качестве инструмента для лингвистических исследований неоднократно обсуждался в научной литературе (ср. хотя бы [Беликов 2004, 2005]). В данной статье речь пойдет о сравнительно новых сетевых сервисах и их применении к решению задач, стоящих перед лингвистом, а также о проблемах, связанных с эффективностью их применения.

Поведение человека в сети и особенно при публикации собственных сообщений в блогах и на форумах практически никак не регламентируется. Один и тот же пользователь при написании электронного письма начальству и сообщения на форуме может использовать совершенно разные регистры своего идиалекта. Это может проявляться на всех уровнях – начиная с правописания и синтаксических ошибок кончая лексическим выбором и ограничениями, связанными с обсценной лексикой. Следует при этом отметить, что если орфографические ошибки «верифицируемы» (то есть в каждом случае можно найти правило, которое нарушил или не нарушил пользователь), то ошибки синтаксические часто не обладают таким свойством, так как правила, регламентирующие, скажем, модель управления той или иной лексемы, могут быть найдены редко. Эта разница неслучайна, ведь так называемые синтаксические ошибки обычно представляют собой не что иное, как проявление диалектных черт носителя. Так, например, в (1) представлена фраза, неграмматичная с точки зрения центрального (московского) диалекта, но не с точки зрения диалектов множества других русскоязычных городов.

(1) Из блога vlasenko.livejournal.com

Я вообще-то человек не очень честный: в детстве я украл две ***булки хлеба*** *из универсама и иногда оставался в кино на второй сеанс.*

То же самое можно сказать и об «ошибках», связанных с лексическим выбором. Ту лексику, которую носитель никогда не употребит в официальных текстах (в деловом письме, заявлении, отчете и т.п.), тот же носитель без колебаний употребит на форуме или в блоге, (невольно) продемонстрировав тем самым особенности своего диалекта.

Таким образом, Интернет представляет собой уникальную среду не только для свободного общения, но и для анализа языковых данных. Если для поиска примеров на употребление той или иной **литературной** конструкции или слова Интернет используется лингвистами уже давно, то для исследования диалектных особенностей Интернет только начинает привлекаться, что связано, очевидно, с ускорившимся в последнее время ростом активности региональных сетевых ресурсов.

2. Проблемы исследования

Рассмотрим подробно некоторые проблемы, которые встают перед лингвистом при использовании Интернета для исследования диалектных черт.

2.1 География пользователя

Первая и, по-видимому, самая главная проблема, связанная с такими исследованиями, это определение местоположения (и/или местожительства) пользователя, составившего то или иное сообщение. Как известно, ресурсы в Интернете редко привязаны географически к тому или иному региону или городу. И даже если ресурс представляет собой, скажем, форум какого-либо конкретного города, то нет гарантии, что пользователь, написавший сообщение на этом форуме, является носителем диалекта данного региона. Он мог написать его из любого другого региона, а также есть вероятность, что он живет в этом регионе, но переехал сюда недавно и является носителем другого диалекта.

Нужно отметить, что авторы подобных исследований обычно доверяют данным с такого рода «городских» ресурсов. Например, на довольно известном форуме «Городские диалекты» (проект ***«Языки русских городов»***[1], описанный в [Беликов 2006]) такого рода данные (полученные с местных ресурсов) обычно используются в качестве доказательств в дискуссии о географическом распределении той или иной лексемы и становятся основой для статей в соответствующем ***словаре***[2]. Наряду с ними используются данные местных периодических изданий, также публикуемых в сети. Эти сведения также сложно считать совершенно достоверными, так как гарантии, что один и тот же автор не публикует свои статьи в изданиях разных регионов, нет.

Интересно то, что техническая возможность отслеживать географию пользователя (по IP-адресу и информации о провайдере) имеется. Она используется, например, для показывания определенной рекламы пользователям данного сетевого ресурса из определенного региона. Однако в сети Интернет пока не практикуется, например, разметка сообщений на форуме или в блоге по географическому принципу, по-видимому, потому, что пока, кроме лингвистов, это никому не нужно.

2.2 Усреднение данных с поправкой на количество пользователей и жителей данного региона

Данная проблема связана с тем, что сетевые ресурсы и регионы различаются по количеству пользователей и жителей соответственно.

Предположим, что

(2) лингвистом найдено n *примеров использования данной лексемы в регионе* A *с населением в* a *жителей и* m *примеров использования той же лексемы в регионе* B *с населением в* b *жителей*

Предположим, что n>m. Этот факт еще ничего не говорит об употреблении данной лексемы в данных регионах, так как если в регионе A в два раза больше жителей, чем в регионе B, а n в два раза больше, чем m, то несмотря на разницу в данных, полученных лингвистом, частотность употребления данной лексемы по данным регионам примерно одинакова.

Если лингвист является опытным диалектологом, то он, конечно, будет сравнивать не величины n и m, а, например, величины n/a и m/b. Только в этом случае разница между величинами будет свидетельствовать о большей или меньшей частотности данной лексемы.

Теперь предположим, что

(2') лингвистом ***в сети Интернет*** *найдено* n *примеров использования данной лексемы в регионе* A *с населением в* a *жителей и* m *примеров использования той же лексемы в регионе* B *с населением в* b *жителей*

Если и в данном случае также использовать величины n/a и m/b, то это сравнение не будет достоверным, так как необходимо сделать поправку на развитость сетевых возможностей данного региона – далеко не всегда количество активных пользователей Интернета в данном регионе прямо пропорционально его населению.

Например, ежедневная аудитория поисковой системы Яndex[3] в Новосибирске, по данным самой системы, составляет 56.000 пользователей при населении в 1.425.600 жителей, тогда как аудитория Саратова – примерно 25.000 при населении в 874.000. В процентах доля пользователей в Новосибирске и Саратове составляет соответственно: 3,9% и 2,8%.

Если же сравнить Екатеринбург (1.293.000 жителя) с Волгоградом (1.012.800 жителей) по их недельной аудитории на Яndex (150.000 и 55.000 соответственно), то в процентах получим: 11% и 5%. Как видно, эти соотношения далеки от прямо пропорциональной зависимости (в случае такой зависимости процентные значения были бы близки к равным).

---

[1] http://www.lingvo.ru/goroda

[2] http://www.lingvo.ru/goroda/dictionary.asp

[3] http://www.yandex.ru

Чтобы учесть эти особенности, лингвист в идеале должен иметь в своем распоряжении информацию о количестве пользователей сети в данном регионе. И если вернуться к ситуации (2') и предположить, что известно количество пользователей в регионе A и B (скажем, i и j соответственно), то величины для сравнения могут иметь, например, такой вид: n/(a*i) и m/(b*j). В таком случае частотность примеров будет вычисляться с поправкой на численность региона и на количество активных пользователей.

2.3 Шум в поисковой выдаче

Еще одна проблема, связанная с использованием поисковых систем для анализа диалектных черт, заключается в том, что, как правило, не все результаты поиска содержат именно те объекты, которые планировал найти исследователь, другими словами, в поисковой выдаче имеется шум.

Предположим, что исследователь ищет употребления диалектной лексемы *бадлон*[4]. В данном случае ему не следует беспокоиться по поводу шума в поисковой выдаче, так как любой случай употребления цепочки символов «бадлон» должен быть учтен, то есть в качестве результата можно просто брать размер выдачи, не просматривая самих результатов.

Если же исследователь ищет слово *баллон* в значении 'стеклянная банка объемом 2, 3 или 5 литров', то недостаточно просто зафиксировать размер выдачи по запросу «баллон», так как в ней будет довольно весомая часть поискового шума, образованного контекстами со словом *баллон* в значении, скажем, 'газовый баллон'.

В случае таких (неоднозначных) запросов очевидным выходом является просматривание результатов поиска и подсчет в ручную искомых контекстов. Однако более эффективным выходом является расширение поискового запроса в целях сужения выдачи до нужных контекстов. Так, например, для получения контекстов со словом *баллон* в значении 'банка' можно использовать запросы «"баллон молока"» или «баллон /1 молоко», первый из которых (в поисковой системе Яndex) задаст все контексты, в которых встречается точное словосочетание «баллон молока», а второй – все контексты, в которых слова *баллон* и *молоко* расположены рядом и стоят в любой форме. К слову, размер выдачи по первому запросу (в поисковой системе Яndex) почти втрое меньше размера выдачи по второму запросу (36 результатов против 95).

Сразу очевидны как плюсы, так и минусы такого метода. Безусловным плюсом является тот факт, что исследователю не приходится просматривать все результаты поиска и выбирать нужные контексты в ручную, однако минусом является то, что чем у́же задан контекст, тем меньше (по абсолютному количеству) будет найдено употреблений, а это может быть существенным фактором особенно в случае поиска и без того редких диалектных значений лексемы или же в случае ее малой распространенности, то есть тогда, когда каждый найденный случай употребления представляет высокую ценность.

3. Примеры исследований

Приведем теперь конкретные примеры подобных исследований и проиллюстрируем на них, как и насколько успешно авторы решали описанные выше проблемы.

3.1 Исследование городской диалектной лексики с помощью сервиса Яндекс.Города[5]

Данное исследование было проведено нами для того, чтобы показать аудитории форума «Городские диалекты» (проект «Языки русских городов»), как можно использовать в диалектологических целях сервис Яндекс.Города.

В основе этого метода лежат данные, предоставляемые сервисом Яндекс.Города, а именно возможность поиска по определенному городу (региону) и так называемый интернет-индекс каждого города. Интернет-индекс города – это численный показатель, в котором учитывается (помимо прочего):

- количество и авторитетность городских сайтов
- число сайтов на душу интернет-пользователя
- тИЦ[6] «топовых» сайтов данного города

[4] Бадлон (Петербург) тонкое облегающее трикотажное изделие на верхнюю часть туловища, без застёжки, с длинными рукавами и высоким воротом.

[5] http://goroda.yandex.ru

[6] Тематический индекс цитирования (тИЦ) — технология поисковой машины «Яндекс», заключающаяся в определении «авторитетности» интернет-ресурсов с учетом качественной характеристики ссылок на них с других сайтов. тИЦ рассчитывается по специально разработанному алгоритму, в котором особое значение придается тематической близости ресурса и ссылающихся на него сайтов.

• число запросов с именем города к Яндексу

• население города (по данным последней переписи)[7]

Как видно, все эти показатели, кроме последнего, таковы, что чем больше их абсолютное значение, тем бóльшую сетевую активность проявляет данный город. Последний же показатель вносит необходимую нам поправку на численность населения данного региона. Таким образом, интернет-индекс сочетает в себе всю информацию, о которой говорилось в разделе 2.2.

Собственно метод заключался в том, чтобы задавать один и тот же поисковый запрос по разным городам и делить результаты (количество найденных сайтов) на интернет-индекс данного города. Таким образом, если запрос задает некоторую диалектную черту, то получившийся в результате деления показатель определяет ее относительную частотность в данном регионе (в котором уже учтена сетевая активность данного региона и численность его населения).

Посмотрим на результаты такого анализа для запроса «плойка»[8] (по выборочному списку городов, в котором представлены все регионы) в таблице 1.

Город	**Кол-во сайтов**	**Интернет-индекс**	**(Кол-во сайтов)/(Интернет-индекс)**
Новосибирск	51	568	0,0898
Ростов-на-Дону	18	202	0,0891
Киев	41	476	0,0861
Челябинск	30	362	0,0829
Томск	12	185	0,0649
Красноярск	20	335	0,0597
Екатеринбург	39	662	0,0589
Омск	12	226	0,0531
Харьков	8	174	0,046
Кемерово	6	131	0,0458
Хабаровск	7	157	0,0446
Пермь	15	356	0,0421
Южно-Сахалинск	6	145	0,0414
Тюмень	12	292	0,0411
Оренбург	6	149	0,0403
Волгоград	9	264	0,0341
Воронеж	9	292	0,0308
Астрахань	3	100	0,03
Нижний Новгород	10	334	0,0299
Вологда	4	151	0,0265
Казань	8	353	0,0227
Саратов	5	245	0,0204
Петропавловск-Камчатский	3	148	0,0203
Мурманск	4	211	0,019
Тверь	3	162	0,0185
Владивосток	6	349	0,0172
Ярославль	4	254	0,0157
Ставрополь	3	224	0,0134
Краснодар	5	376	0,0133
Самара	5	480	0,0104

Таблица 1. Поисковый запрос "плойка"

Строки в таблице 1 упорядочены по убыванию значения правого столбца. Как видно, города, представленные в данной выборке, довольно сильно разнятся по этому показателю и, что характерно для этого запроса, количество найденных сайтов (вторая колонка) во всех случаях не очень велико. Также можно отметить, что если бы были использованы абсолютные результаты – количество найденных сайтов (вторая колонка), то порядок сле-

[7] Часть из этих данных была любезно предоставлена автору директором по специальным проектам компании «Яндекс» Андреем Себрантом.

[8] Плойка – инструмент для завивки волос.

дования городов сильно отличался бы, что видно, например, по Нижнему Новгороду и Кемерово.

Итак, используя данные, представленные в таблице 1, можно утверждать, что:

- *плойка* часто употребляется в Сибири (Новосибирск, Томск, Красноярск, Омск, меньше в Кемерово)
- *плойка* часто употребляется на Урале (Челябинск, Екатеринбург, меньше в Тюмени)
- *плойка* часто употребляется на Украине (Киев, меньше в Харькове)
- *плойка* редко употребляется в Поволжье (Самара, Саратов, Нижний Новгород, Казань)
- *плойка* редко употребляется на Дальнем Востоке (Владивосток, Петропавловск-Камчатский, чаще в Южно-Сахалинске и в Хабаровске)
- *плойка* редко употребляется в центральном регионе и на северо-западе (Ярославль, Тверь, Мурманск, Вологда, Воронеж)
- *плойка* редко употребляется на юге (Краснодар, Ставрополь, Астрахань, Волгоград)

Город	**Кол-во сайтов**	**Интернет-индекс**	**(Кол-во сайтов)/(Интернет-индекс)**
Иркутск	68	321	0,2118
Томск	39	185	0,2108
Одесса	33	166	0,1988
Барнаул	29	147	0,1973
Красноярск	63	335	0,1881
Ростов-на-Дону	35	202	0,1733
Кемерово	22	131	0,1679
Новосибирск	93	568	0,1637
Чита	12	78	0,1538
Челябинск	52	362	0,1436
Омск	32	226	0,1416
Хабаровск	22	157	0,1401
Владивосток	39	349	0,1117
Луганск	8	74	0,1081
Екатеринбург	65	662	0,0982
Тюмень	28	292	0,0959
Пермь	32	356	0,0899
Саратов	21	245	0,0857
Астрахань	8	100	0,08
Ставрополь	16	224	0,0714
Уфа	23	329	0,0699
Краснодар	26	376	0,0691
Южно-Сахалинск	10	145	0,069
Донецк	10	147	0,068
Калининград	11	178	0,0618
Киев	29	476	0,0609
Петропавловск-Камчатский	9	148	0,0608
Волгоград	16	264	0,0606
Нижний Новгород	18	334	0,0539
Оренбург	8	149	0,0537
Ижевск	9	186	0,0484
Архангельск	7	153	0,0458
Рязань	7	158	0,0443
Самара	21	480	0,0438
Сыктывкар	6	145	0,0414
Вологда	6	151	0,0397
Воронеж	11	292	0,0377
Казань	12	353	0,034
Харьков	5	174	0,0287
Псков	4	145	0,0276
Тверь	4	162	0,0247
Ульяновск	3	124	0,0242
Ярославль	6	254	0,0236
Петрозаводск	5	216	0,0231
Белгород	4	174	0,023
Новгород	5	223	0,0224
Мурманск	4	211	0,019
Львов	3	169	0,0178

Таблица 2. Поисковый запрос “булка /1 хлеб”

Эти результаты практически совпадают с высказывавшимися на форуме «Городские диалекты» гипотезами об областях распространения данной диалектной лексемы. Заметим, что если бы в качестве результатов использовались абсолютные значения количества найденных сайтов, то такого распределения по регионам не получилось бы.

Рассмотрим еще один пример такого же анализа по другому поисковому запросу – «булка /1 хлеб»[9]. Такой запрос был сформулирован для того, чтобы отделить значение ‘буханка’ от стандартного ‘булка’. Результаты представлены в таблице 2 (по всем городам, информация о которых доступна в данном сервисе).

Как видно из таблицы 2, внизу таблицы (низкая частотность) расположились:
- Центральный регион (Белгород, Ярославль, Тверь, Воронеж Рязань)
- Северо-западный регион (Мурманск, Новгород, Петрозаводск, Псков, Вологда, Сыктывкар, Архангельск)

Ближе к центру (средняя частотность) расположилось:
- Поволжье (Нижний Новгород, Оренбург, Ижевск, Саратов, Пермь, Самара, Казань, Ульяновск)

В верхней части (высокая частотность) находятся:
- Дальний Восток (Хабаровск, Владивосток)
- Сибирь (Иркутск, Томск, Барнаул, Красноярск, Кемерово, Новосибирск, Чита)
- Урал (Челябинск, Екатеринбург, Тюмень)
- Юг (Ростов-на-Дону, Астрахань, Ставрополь, Краснодар)

Разделение на эти три группы носит условный характер, но, как представляется, этих данных достаточно, чтобы показать, что такой способ позволяет давать регионам характеристику по частотности того или иного явления, выраженного в поисковом запросе.

Интересно, что города Украины распределились практически по всей высоте таблицы (ср. только позиции Львова и Одессы). Особая позиция Львова в этой ситуации не столько удивительна, сколько позиции других городов. Причин такого разброса может быть много, в том числе, например, подсчет интернет-индекса по украинским городам может быть затруднен или вестись не совсем так, как для российских городов. О д н а к о результаты данного исследования по российским городам в целом достаточно точно коррелируют с данными словаря «Языки русских городов» (сноска 9).

3.2 Фрагмент сравнительного анализа аудитории «блогосервисов»

Данное исследование было опубликовано в анонимном блоге sheldon-j.livejournal.com[10]. В нем проводится сравнительный анализ аудитории трех «блогосервисов» – Livejournal[11], Liveinternet[12] и Блоги@Mail.ru[13]. По заявлению автора, целью исследования было «составить примерный портрет аудитории каждого из этих сервисов».

На первом этапе автор определяет так называемые коэффициенты популярности сервисов (величина, по смыслу параллельная интернет-индексу в предыдущем примере исследования). Способ определения этих коэффициентов достаточно прост. Автор использует поисковые запросы с так называемыми «нейтральными» словами – «стол» и «окно». Поиск при этом проводится отдельно по каждому сервису. Его результаты в таблице 3.

Запрос	Сервис	Найдено записей
«окно»	Livejournal	686578
	Liveinternet	372433
	blogs.mail.ru	87350
«стол»	Livejournal	500745
	Liveinternet	229459
	blogs.mail.ru	49423

Таблица 3. Результаты поиска по “нейтральным” словам

[9] Булка III. (Украина Латвия Казахстан Дон Сев. Кавказ Урал Сибирь Дальний Восток) крупное хлебобулочное изделие, буханка (из словаря «Языки русских городов»).
[10] http://sheldon-j.livejournal.com/20437.html
[11] http://www.livejournal.com
[12] http://www.liveinternet.ru
[13] http://blogs.mail.ru

Округленное среднее арифметическое найденных записей по каждому сервису принимается за коэффициент популярности сервиса. Таким образом, коэффициент популярности Livejournal равен 594, Liveinternet – 301 и Блоги@Mail.ru – 68.

Далее автор использует этот коэффициент для подсчета разных характеристик данных сервисов. При этом методика подсчета совпадает с методикой, продемонстрированной в предыдущих примерах исследований – формулируется поисковый запрос, задающий какую-либо характеристику, результаты (количество найденных вхождений) по данному запросу фиксируются и делятся на коэффициент популярности данного ресурса (интернет-индекс в предыдущих примерах), а затем сравниваются.

Покажем лишь некоторые из полученных автором результатов в таблице 4.

Запрос	**Сервис**	**Коэффициент популярности сервиса**	**Найдено записей**	**Частота**
«прив»	Livejournal	594	2937	4,94444444
	Liveinternet	301	15464	51,3754153
	blogs.mail.ru	68	2317	34,0735294
«пасиб»	Livejournal	594	14420	24,2760943
	Liveinternet	301	16306	54,1727575
	blogs.mail.ru	68	1079	15,8676471
«кажеться»	Livejournal	594	13306	22,4006734
	Liveinternet	301	17612	58,5116279
	blogs.mail.ru	68	2963	43,5735294
«пробывал»	Livejournal	594	3333	5,61111111
	Liveinternet	301	2949	9,79734219
	blogs.mail.ru	68	516	7,58823529

Таблица 4. Результаты анализа по словам сетевого сленга и популярным орфографическим ошибкам

В таблице 4 показаны результаты анализа по запросам, соответствующим словам современного сетевого сленга («прив» ‘привет’ и «пасиб» ‘спасибо’), а также популярным орфографическим ошибкам («кажеться» и «пробывал»). Автор делает на основе этих данных вывод об отличии сервиса Livejournal от остальных двух участников анализа, утверждая, что пользователи Livejournal в целом более интеллигентны. Нас же интересует не столько интерпретация этих результатов, сколько сама методика. Нетрудно видеть, что методика в целом параллельна той, что была использована в продемонстрированных выше примерах исследований (которые, к слову, проводились нами независимо от данного).

Можно выделить следующие ключевые особенности продемонстрированных нами примеров исследований:

- **Подбор поискового запроса.** Запрос должен однозначно задавать некоторую характеристику содержания сетевого ресурса (будь то диалектные черты, сленговая лексика или же наличие орфографических ошибок).
- **Вычисление индекса (коэффициента) популярности.** Данный показатель может вычисляться по-разному в зависимости от целей конкретного эксперимента. Он должен отражать популярность ресурса (или же множества ресурсов, как в случае с интернет-индексом), а также (если это необходимо для конкретного анализа) количество ресурсов в данном множестве и количество потенциальных пользователей (как в случае с поправкой на население региона).
- **Возможность сужения области поиска.** Важным фактором является возможность сужать область поиска (будь то регион поиска или же отдельный ресурс, как в случае с «блогосервисами»).

4. Заключение и перспективы

В заключении отметим, что продемонстрированные выше примеры конкретных исследований говорят о том, что сеть Интернет на сегодняшний день предоставляет достаточное количество средств для разнообразных статистических исследований, в том числе и диалектологических. Интересен также тот факт, что авторы представленных здесь исследований действовали независимо друг от друга и выработали при этом сходные по своей структуре методы исследований.

Говоря о перспективах, хотелось бы отметить тот факт, что подобные исследования могут производиться автоматически – исследователю в идеале нужно только в явном виде указать поисковые запросы, области поиска и предоставить системе данные для подсчета индекса популярности. Эта возможность позволяет, например, «размечать» различные ресурсы (в том числе и отдельные сайты) по различным характеристикам – от уровня грамотности пользователей до их диалектных особенностей. Возможны применения подобных методик и для тематического ранжирования сетевых ресурсов.

Список литературы

1. Беликов В.И. Yandex как лексикографический инструмент // Компьютерная лингвистика и интеллектуальные технологии». Труды Международной конференции Диалог 2004. М.: Наука, 2004.
2. Беликов В.И. Верификация словарных помет Интернетом // «Компьютерная лингвистика и интеллектуальные технологии». Труды Международной конференции Диалог 2005. М: Наука, 2005.
3. Беликов В.И. Словарь «Языки русских городов»: подбор примеров и Интернет // «Компьютерная лингвистика и интеллектуальные технологии». Труды Международной конференции Диалог 2006. М.: Ин-т проблем информатики РАН, 2006.