Невод, закинутый в Сеть

**Л.А.Ашкинази,
М.Л.Гайнер,
А.В.Кузнецова**

Авторы статьи применили Интернет как инструмент для исследования: 1) грамотности, 2) интереса к другим странам и народам, 3) распространения материалов, 4) отличия письменного языка от «языка Интернета», более близкого к устному и 5) «притяжения» слов друг к другу.

В Интернете можно найти любую информации, используя поисковую систему. Но мы просто извлекаем информацию, которую кто-то ранее добыл и выложил в Сеть. Однако Интернет можно применить и для получения новой информации, которой не существовало ранее, то есть он может быть инструментом исследования. Самое простое — определение числа ссылок, полученных в ответ на запрос, как характеристики интереса, проявленного людьми к данному предмету. При этом возможно измерение собственно числа ссылок, их динамики, распределения по языкам, странам, типам сайтов и др. Исследователи еще не распробовали как следует этот метод изучения общества, хотя подобные данные и начали появляться в статьях и книгах.

Мы задали Сети пять простых вопросов. Первый: насколько неграмотны люди и как эта неграмотность изменяется со временем? Второй: насколько велик у людей интерес к другим странам и народам? Третий: часто ли в Сети таскают друг у друга материалы? Четвертый: отличается ли язык Сети от русского письменного языка? Пятый: как определить, насколько два конкретных слова тяготеют друг к другу?

Грамотность

Что ответила Сеть на первый вопрос, показано в таблице — мы для примера исследовали несколько слов. Показаны неправильное написание, правильное и отношение частоты неправильного написания к правильному. Yandex позволяет для каждого слова установить количество страниц, на которых нашлось это слово, количество сайтов, на которых находятся эти страницы, и количество запросов в течение данного месяца. Поэтому доля неправильных написаний может вычисляться по запросам, по страницам и по сайтам (левое, среднее и правое число в таблице). Расчет по запросам дает наиболее чистые данные, не искаженные влиянием редакторов, но их недостаток в том, что они доступны только за месяц, а это уменьшает статистику. Доля неправильных написаний, рассчитанная по страницам, существенно меньше, поскольку тексты на некоторых страницах правят. Доля неправильных написаний, рассчитанная по сайтам, лежит посередине, поскольку, если на сайте встретилось одно неправильное написание, он попадает и в «правильные» и в «неправильные».

Из таблицы видно, что за полтора года грамотность пользователей Интернета (графа «Запросы») несколько возросла, в то время как грамотность текстов на сайтах несколько уменьшилась. По-видимому, это означает ослабление редактирования. Полтора года назад «страницы» были грамотнее «запросов» в шесть раз, сегодня — в три раза, а еще года через три они сравняются — то есть редактирование отомрет, как явление. Разумеется, этот вывод ввиду малого интервала наблюдений и малой статистики носит лишь демонстрационный характер.

Логично было бы исследовать зависимость частоты ошибок от типа ошибки и от контекста. Одним из результатов такого исследования мог бы стать вывод — как надо учить, чтобы уменьшить частоту ошибок.

Классификация ошибок может быть двух типов. Во-первых, по формальной стороне: что именно происходит с буквами при ошибке. Таким образом можно выделить четыре типа ошибок: вставка, удаление, замещение и перестановка. Вторым же подходом к классификации может быть «сущностный». Почему делается та или иная ошибка? Например, «карова» человек может писать потому, что плохо усвоил орфограмму о безударных гласных, а в соответствующем классе плохо учил «словарные слова». Такая классификация может быть подробной, включающей в себя указания на все орфограммы русского языка. Интересно было бы сопоставить относительные частоты ошибок в Интернете и, скажем, во вступительных сочинениях в вуз, то есть в набранных на компьютере текстах и в текстах, написанных от руки.

Интерес к другим странам и народам

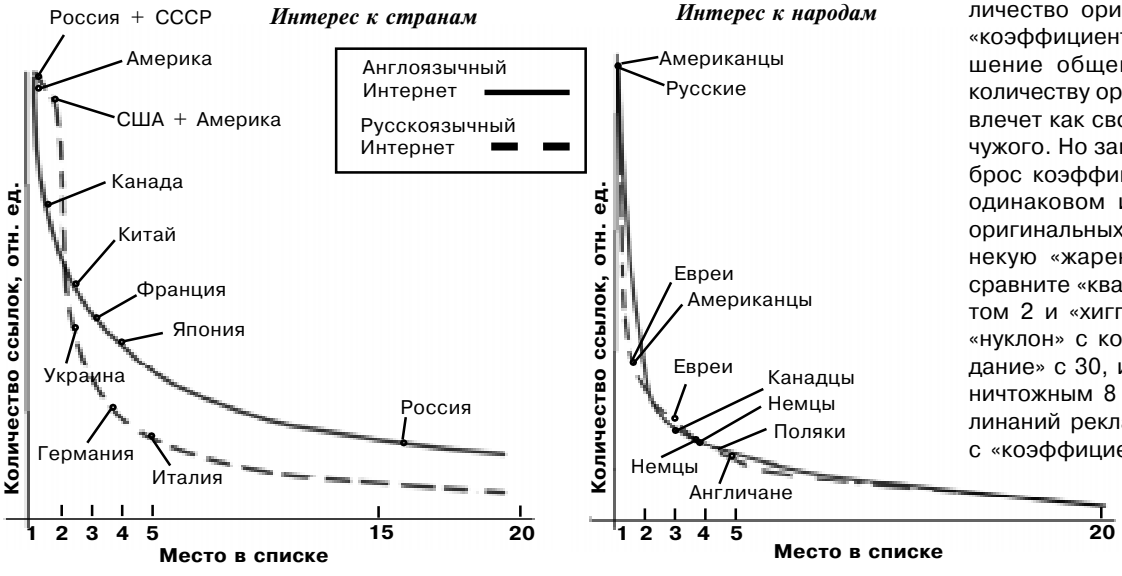
Второе маленькое исследование — это частота употребления в русскоязычном и англоязычном Интернете названий стран и народов. Если расположить страны в порядке убывания частот упоминания в Интернете и нарисовать зависимость частоты от места в списке (ранга), получатся вот такие кривые (за 1 принята частота упоминания «России» и «русских» в русскоязычном Интернете и «США» и «американцев» — в англоязычном). Глазом, вооруженным Google'ом, видно, что англоязычному населению Интернета (3/4 его — американцы) окружающие страны более интересны, нежели русскоязычному, а народы — интересны в той же мере. Так что тупо повторяемая фраза, что американцам ничто не интересно, кроме самих себя, — миф. Но вот в русскоязычном Интернете очень высок интерес к США — почти так же высок, как к себе.

Если же посмотреть на отношение интереса к стране и ее народу, то видно, что есть ситуации, когда страна интереснее народа, а бывает и наоборот. Например, в англоязычном Интернете в сторону страны смещен интерес к Китаю, Франции, Японии и Индии (то есть интерес носит политический характер), а в сторону народа — для Польши, России, Греции (культурный интерес). В русскоязычном Интернете «политичен» (а вернее — «туристичен») интерес к Италии, Кипру, Испании, Турции, Индии, Болгарии, а «культурен» — к Англии, Японии, Китаю, Германии.

Ощущение своего места в мире может проявляться и иначе. В русскоязычной литературе нам дважды встречались утверждения, что в некой

	Запросы			Страницы			Сайты		
Год	2004	2005	2005	2004	2005	2005	2004	2005	2005
Месяц	IV	IV	X	IV	IV	X	IV	IV	X
абревиатура/ аббревиатура	0,25	0,25	0,16	0,02	0,08	0,057	0,42	0,83	0,9
абонимент/ абонемент	0,1	0,11	0,07	0,02	0,04	0,06	0,4	0,68	0,68
абардаж/ абордаж	0,06	0,04	0,02	0,01	0,014	0.024	0,1	0,19	0,18
агенство/ агентство	0,8	0,55	0,54	0,1	0,12	0,1	0,25	0,36	0,43
агригат/ агрегат	0,01	0,003	0,003	0,001	0,002	0,002	0,04	0,27	0,31
аксесуар/ аксессуар	0,3	0,22	0,18	0,02	0,03	0,03	0,05	0,35	0,36
алкаголь/ алкоголь	0,03	0,03	0,018	0,005	0,008	0,008	0,1	0,24	0,3
алитерация/ аллитерация	0,2	0,14	0,12	0,05	0,015	0,027	0,08	0,04	0,1
альботрос/ альбатрос	0,015	0,02	0,014	0,003	0,005	0,027	0,05	0,07	0.08
анотомия/ анатомия	0,04	0,02	0,02	0,02	0,04	0,007	0,1	0,4	0,45
ангажимент/ ангажемент	0,1	0,2	0,08	0,03	0,04	0,2	0,05	0,07	0,14
аникдот/ анекдот	0,003	0,002	0.002	0,0005	0,001	0,001	0,03	0,1	0,15
(аппеляция+ апеляция+ аппелляция)/ апелляция	0,9	0,7	0,9	0,23	0,18	0,22	1,1	1,2	1,8
аперцепция/ апперцепция	0,25	0,15	0,1	0,03	0,05	0,04	0,5	0,1	0,1

стране карты мира нарисованы так, что данная страна всегда находится в центре картинки. Мы встретили такие утверждения про США и Новую Зеландию. Простейший анализ не только показывает, что это — вранье, но и позволяет обнаружить интересные различия между странами и, возможно, между типами сайтов. Карта мира на сайтах Новой Зеландии расположена Новой Зеландией вперед в 30 % случаев, на сайтах Австралии — в 40 % случаев (эта разница при объеме выборки 100 сайтов несущественна), а на сайтах Японии — в 80 % случаев. В остальных случаях Земля изображена Европой вперед. Поэтому можно предположить, что 40 процентных пунктов разницы — эффект ментальности, а общие 40% составлены из ментальности и удобства пользования. Что же касается США, то на сайтах com, org, edu и gov Земля изображена Америкой вперед в 2, 3, 5 и 8% случаев соответственно (в целом — около 2% случаев), в остальных случаях — Европой вперед.



Авторское бесправие

Третий вопрос, который мы задали Сети, — часто ли в ней «таскают» материалы. Возможность для такого исследования дает, например, Google, поскольку он делит ссылки на «наиболее значимые» и «очень похожие на них». Оказалось, что отношение общего количества ссылок (обе категории вместе) к количеству оригинальных ссылок (таковыми мы считали «наиболее значимые») изменяется в широких пределах, по крайней мере, — от 1 до 230. Вот некоторые примеры, причем первое число — количество оригинальных ссылок, второе — общее количество. «Квазигруппа» — 20/40, «хиггсовский бозон» — 30/230, «асимптотическая свобода» — 50/150, «поле температур» — 100/140, «тахион» — 250/1000, «фуллерит» — 250/800, «поле скоростей» — 300/800, «липосакция» — 500/6000, «нуклон» — 500/3000, «похудание» — 500/16000, «гравитационное поле» — 600/7000, «майкрософт» — 700/70000, «кристаллография» — 700/5000, «ожирение» — 700/46000, «магнитное поле» — 750/29000, «электрическое поле» — 800/10000, «кристалл» — 800/190000.

Понятно, что с увеличением общего интереса к теме должны расти как количество оригинальных ссылок, так и «коэффициент размножения» — отношение общего количества ссылок к количеству оригинальных, ведь интерес влечет как свою работу, так и таскание чужого. Но заметен существенный разброс коэффициента размножения при одинаковом или близком количестве оригинальных ссылок, указывающий на некую «жареность» темы. Например, сравните «квазигруппу» с коэффициентом 2 и «хиггсовский бозон» с 8, или «нуклон» с коэффициентом 6 и «похудание» с 30, или «кристаллографию» с ничтожным 8 и одно из любимых заклинаний рекламщиков — «кристалл» — с «коэффициентом жарености» 230.

Более подробное рассмотрение самих ссылок показало, что лишь треть из них является ссылками на тот же документ, но лежащий на другом сайте. В эту группу входят как законное выкладывание документа второй раз, например на сайте-зеркале, так и пиратское. Однако провести этот анализ вручную можно только при относительно небольшом количестве ссылок.

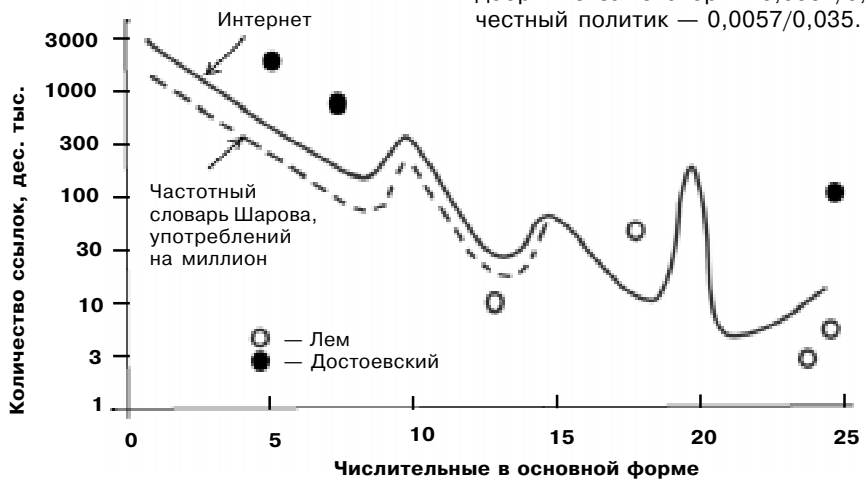
Язык и сетевой язык

Четвертый вопрос Сети: далек или близок ее язык к русскому письменному. Языки можно сравнивать многими способами, для примера мы сравнили частоты числительных. На рисунке показаны частоты первых 25 числительных (в основной форме — один, два, три...). Видно, что по этому параметру язык Сети близок к письменному русскому. Но одно отличие заметно: в Сети реже употребляются большие числа и чаще — числа менее 15. Что касается пана Станислава, то он почему-то любит число 18 (что мы отмечали и ранее).

Можно обсудить и сам вид частотных зависимостей. Где реальная частота объектов (один любимый, два сапога, три стакана, четыре конечности...), а где психология людей, влияние которой на частоты употребления числительных отмечал еще Перельман? Насколько различаются частоты у разных авторов (скажем сразу, что различия есть) и почему? Как проявляется психология в разных «слоях» языка? Например, частота в пословицах — совпадает она или нет с частотой в языке в целом? Различны ли эти частоты в разные эпохи, в разных языках, у разных народов?

«Притяжение» слов

Наконец, пятый вопрос — «притяжение» слов, то есть вероятность для двух слов оказаться рядом или в одной фразе по отношению к вероятности оказаться на одной странице (разумеется, возможны и другие нормировки, и смысл получаемых в этих случаях результатов достоин обсуждения). Ниже приведены относительные частоты для некоторых пар существительное — прилагательное. Большинство результатов довольно понятно, но нам показалось удивительным, что слова «умная» и «женщина» притягиваются в языке (и, стало быть, в сознании) намного сильнее, чем «глупая» и «баба» и чем «глупая» и «женщина». Пары слов упорядочены по убыванию притяжения (числитель — отношение вероятности оказаться рядом к вероятности оказаться на одной странице, знаменатель — отношение вероятности оказаться в одной фразе к вероятности оказаться на одной странице): голая баба — 0,21/0,28, рыжий кот — 0,19/0,27, пьяный мужик — 0,095/0,15, умная женщина — 0,047/0,15, глупая женщина — 0,017/0,10, глупая баба — 0,016/0,040, умная баба — 0,014/0,032, трудолюбивый студент — 0,0070/0,022, добрый экзаменатор — 0,0067/0,018, честный политик — 0,0057/0,035.



Другой вариант пар — это употребление имени собственного вместе с названием класса (как пояснение). В качестве объекта мы взяли названия гор-восьмитысячников и посмотрели, часто ли употребляются вместе с этими словами названия классов — «пик» и «вершина». И тут обнаружилось нечто неожиданное. Понятно, что Джомолунгма употребляется как «вершина Джомолунгма» и «пик Джомолунгма» редко — в 2% случаев, ибо само это название употребляется часто (14 тыс. ссылок) и хорошо известно. Но остальные названия, встречающиеся все с примерно одинаковой частотой (от 540 до 970 ссылок), разделились на две группы. Аннапурна, Лхоцзе и Чо-Ойю употребляются вместе со словами «пик» или «вершина» в 31–46 % случаев, а все остальные — в 6,0–12 % случаев. Причины такого странного распределения непонятны.

С помощью Интернета можно изучать и явные связи между словами, например устойчивые словосочетания, отношения сравнения (быстрый, как свет), уточнения (красный, точнее, малиновый), приближения (худой, почти тощий). Например, «не промахнись» в 10% случаев входит в «не промахнись, Ассунта», «промахнулся» — в 8% случаев входит в «Акела промахнулся», причем «Акела» входит в «Акела промахнулся» в 40% случаев. «Быстрый» — это в 9% случаев «быстрый, как ветер», в 2% — «как свет», в 1% — «как олень».

Словоупотребление вообще (и связи между словами в частности) может зависеть от характера сайта. Например, data — это данные, причем формально это множественное число, и для него есть архаичное единственное — datum. Запросы «data is» и «data ar» дают примерно одинаковое количество ссылок. Но оказалось, что «data is» сосредоточено на сайтах .com и меньше на .org, а «data ar» на .edu, .uk, .org и немного на .com. Нормативная ситуация, согласно Оксфордскому словарю, сегодня такова — это множественное число, но поскольку в компьютинге data повсеместно употребляется как единственное, то в этой области «data is» считается допустимым.