Л.Г. Митюшин

Интернет как корпус лингвистических примеров

Учебное пособие
для изучающих иностранные языки

МОСКВА – 2011

УДК 002.54(075.8)
ББК 81.1я73
М66

Митюшин Л.Г.

М66 **Интернет как корпус лингвистических примеров:** Учебное пособие для изучающих иностранные языки. – М.: МАКС Пресс, 2011. – 28 с.

ISBN 978-5-317-03546-4

Данное пособие, написанное весной 2010 года, основано на измерении частоты появления слов и выражений в Интернете с помощью поисковой системы Yahoo!. В течение нескольких лет эта система работала весьма стабильно. Однако в июле 2010 года исходная поисковая система Yahoo! была заменена другой системой, дающей неправильные показатели частоты. Это относится к версии, вызываемой по адресу www.yahoo.com; пока – в январе 2011 года – еще можно получать правильные результаты, обращаясь к старой версии по адресу uk.search.yahoo.com или ru.search.yahoo.com. По всей вероятности, эта возможность сохранится лишь ограниченное время, после чего для измерения частоты выражений рекомендуется обращаться к системе Yandex. Возникшая ситуация обсуждается в последнем разделе пособия.

Education is not the filling of a pail,
but the lighting of a fire.

W.B. Yeats [1]

1. Введение

Задача этого пособия – показать, как можно использовать глобальную компьютерную сеть Интернет для того, чтобы находить нужные выражения на иностранном языке и оценивать степень их употребительности. Мы будем работать с английским языком, так как он наболее распространен в Интернете и в той или иной степени известен большинству наших потенциальных читателей. Однако многое из сказанного далее применимо к другим языкам, в том числе и к русскому.

Суть предлагаемого подхода – измерение частоты (то есть количества случаев) появления выражений в Интернете. Сегодня англоязычный Интернет охватывает более 30 миллиардов веб-страниц, содержащих триллионы слов. В таком огромном корпусе текстов "хорошие" английские выражения должны встречаться многие тысячи раз, а менее употребительные варианты будут иметь существенно меньшую частоту по сравнению с основными.

Приведем простой пример. Пусть нас интересует, как сказать по-английски *большие деньги, большая сумма денег*. Русское прилагательное *большой* имеет два основных английских эквивалента – *big* и *large*. Посмотрим, сколько раз в британском Интернете упоребляются слова *big* и *large*

[1] Цель образования – не наполнить ведро, а зажечь огонь. (Уильям Батлер Йейтс)

в сочетании с *money*. Для этого зададим в поисковой системе Yahoo! запросы ["big money" site:uk] и ["large money" site:uk]. [2] Мы получим, что число страниц британского Интернета, содержащих выражение *big money*, равно 1 780 000 [3]. Для *large money* это число равно 12 900, то есть почти в 140 раз меньше.

А как обстоит дело с выражением *большая сумма денег*? Задаем поисковые запросы ["big sum of money" site:uk] и ["large sum of money" site:uk]. Для первого получаем число страниц 1540, для второго 127 000 – в 80 раз больше.

Прямой вывод из этих статистических данных такой: *big money* и *large sum of money* – "хорошие" английские сочетания, достаточно часто встречающиеся в Интернете; их альтернативные варианты появляются гораздо реже и, возможно, являются неправильными с точки зрения носителей языка.

В поисковых запросах можно использовать звездочку * в качестве так называемого символа-джокера. В Yahoo! этот символ заменяет одно неизвестное слово, в Google – одно или несколько неизвестных слов. Пусть, например, мы хотим узнать, какой предлог надо употребить в английском переводе выражения *посмотрел в Интернете*. Задаем в Yahoo! поисковый запрос ["looked * the Internet" site:uk] и, просматривая показанные на экране цитаты из найденных 42 900 страниц (конечно, не все, а первые несколько десятков), видим, что чаще всего попадается *on*, но бывает

---

[2] Запросом является текст внутри квадратных скобок. Сами скобки в поле запроса не набираются и служат здесь для отделения запроса от остального текста. Почему Yahoo!, почему в кавычках и почему site:uk объясняется ниже в разделах 2, 3 и 4.

[3] Числовые данные в пособии приведены по состоянию на январь 2011 года. Число страниц округляется до трех старших цифр.

также *across, around, at, in, into, onto, through, to, up, upon.*

Для более точной оценки задаем отдельно запросы, где звездочка заменена на каждый из этих предлогов, например ["looked on the Internet" site:uk]. Получаем для *on*, *at*, *around*, *to* и *up* соответственно частоты 34 900, 837, 741, 120 и 105; остальные имеют частоту меньше 100.[4] Вывод: наиболее употребительно (с большим отрывом) выражение *looked on the Internet*.

Эти примеры достаточно ясно демонстрируют суть нашего подхода. В следующих разделах сообщаются основные факты о поисковых системах Google и Yahoo!, описываются особенности языка запросов и приводятся дальнейшие примеры работы с реальным лингвистическим материалом.

2. Интернет и поисковые системы

Интернет – это совокупность компьютерных страниц ("веб-страниц" [5]), размещенных на миллионах компьютеров и доступных для просмотра с помощью специальных программ – так называемых браузеров. Веб-страницы могут содержать информацию в самых разных формах: текст, таблицы, графики, рисунки и фотографии, аудио- и

[4] Частотой выражения мы называем число страниц, на которых оно встречается (при этом не учитывается, сколько раз оно встретилось на каждой странице). Более естественно было бы понимать под частотой общее количество появлений выражения в текстах, но поисковые системы не сообщают этих данных.

[5] Веб-страница (web-page) – то же, что страница Интернета, поскольку Интернет называют также "всемирной паутиной" (World Wide Web).

видеофайлы. Их функция может быть не чисто информационной: через веб-страницы можно совершать покупки в интернет-магазине, играть в компьютерные игры, делать переводы с помощью автоматического переводчика, разговаривать с другими пользователями Интернета и т.д. и т.п.

Интернет возник совсем недавно (в его современном виде – где-то в начале 1990-х годов), и почти сразу стали появляться программы, помогающие искать в Интернете нужную информацию. Сегодня наиболее популярна поисковая система Google (www.google.com или www.google.ru). Ее название представляет собой искаженное английское слово *googol*, которое означает десятичное число, записываемое как единица и сто нулей (то есть 10^{100}). У слова *googol* есть "автор": его лет 70 назад придумал девятилетний мальчик, племянник американского математика Эдварда Казнера, когда дядя попросил его предложить название для очень большого числа.

Второе место по популярности занимает система Yahoo! (uk.search.yahoo.com или ru.search.yahoo.com). Слово *Yahoo* тоже придуманное, оно появляется в "Путешествиях Гулливера", где *Yahoos* – примитивные человекоподобные существа, живущие в стране разумных и добродетельных лошадей. Кроме того, есть междометие *yahoo* – крик радости (что-то вроде "ура"), и восклицательный знак в названии системы, по-видимому, указывает именно на это значение.

Существуют и другие поисковые системы, в том числе специально приспособленные для работы с русской частью Интернета. Наиболее известна среди них система Yandex (www.yandex.ru); мы вернемся к ней в разделе 8.

Работа с поисковой системой выглядит так: на экране есть окно, в котором печатается запрос, например одно или несколько английских слов; затем щелкается кнопка "Search". Система показывает число веб-страниц, соответствующих запросу, и выдает список цитат из этих

страниц. Если цитата содержит слова из запроса, они выделяются жирным шрифтом. Щелкнув по цитате, точнее, по ее верхней строчке, напечатанной синим цветом, можно вызвать на экран всю страницу, что обычно и делается при традиционном информационном поиске.

Системы Google и Yahoo! существенно отличаются друг от друга в одном очень важном для нас отношении: они по-разному "понимают", какие страницы соответствуют запросу. Google включает в подсчет не только страницы, которые содержат слова запроса (что предполагается при простом поиске), но и такие страницы, которые сами этих слов не содержат, но упоминаются в ссылках на страницах первого типа.

Открывая нажатием на кнопку "Cached" в конце цитаты сохраненные копии страниц, которые находит Google, мы нередко видим вверху такое сообщение, касающееся заданных слов: "these terms only appear in links pointing to this page" (эти слова присутствуют только в ссылках на данную страницу). С Yahoo! такого не бывает, показанные страницы всегда сами содержат слова запроса. И хотя данные о числе страниц часто являются приблизительными (поскольку это лишь некоторые оценки истинного числа страниц, соответствующих запросу), способ подсчета Yahoo! лучше отражает реальную частоту появления слов и выражений в Интернете. По этой причине автор рекомендует находить частоты с помощью Yahoo!, и именно так получены все данные в этом пособии. [6]

[6] Сказанное вовсе не означает, что система Google "вообще плохая", просто Yahoo! лучше приспособлена для наших специальных задач.

3. Как строятся запросы

Ниже описываются формы запросов в системах Google и Yahoo!, достаточные для наших целей. Запрос представляет собой цепочку элементов, разделенных пробелами. Есть следующие три базовых типа элементов:

1. Слово – цепочка букв и цифр без пробелов.
   Примеры: book, tomorrow, didn't, USA, 2010.

2. Заключенная в кавычки цепочка слов с пробелами между ними. Разрешаются также знаки препинания и символы *. В Yahoo! каждый символ * должен отделяться от слов и других символов * пробелами.
   Примеры: "tomorrow morning", "didn't * properly", "thank * * much", "I am * forward to * from you".

3. Префикс "site:", за которым следует двухбуквенный код страны.
   Примеры: site:ru, site:uk (uk = United Kingdom).

Будем говорить, что веб-страница соответствует элементу запроса типа 1, если она содержит указанное слово. Страница соответствует элементу типа 2, если она содержит указанную цепочку слов без учета знаков препинания (можно представлять себе, что при сравнении текста на странице и текста в запросе знаки препинания, в том числе дефис, заменяются пробелами). При этом вместо звездочки * в цепочке слов на странице должно стоять одно произвольное реальное слово (при поиске в Yahoo!) или одно или несколько слов (при поиске в Google). Наконец, страница соответствует элементу типа 3, если она находится на компьютере в указанной стране.

Страница соответствует запросу, составленному из элементов типа 1, 2 и 3, если она соответствует каждому из элементов запроса в отдельности. Следует учитывать, что

при сравнении элементов запроса с текстом на странице заглавные и строчные буквы не различаются.

Пример 3.1. Запрос [yellow submarine]. Ему соответствуют страницы, содержащие одновременно слово *yellow* и слово *submarine*, при этом они могут быть расположены в любых местах текста и не иметь никакого отношения друг к другу.

Пример 3.2. Запрос ["yellow submarine"]. Ему соответствуют страницы, содержащие в точности эту цепочку слов, возможно, со знаками препинания между ними.

Пример 3.3. Запрос ["yellow submarine" site:uk]. Результатом являются такие же страницы, как в примере 3.2, с дополнительным условием, что они находятся на компьютерах в Великобритании.

Перед элементом запроса можно поставить знак минус, означающий отрицание. Пробел после минуса не ставится. В этом случае при проверке соответствия требуется, чтобы страница не содержала данного элемента, а в случае "–site:..." – чтобы страница находилась вне указанной страны.

Пример 3.4. Запрос [yellow submarine –"yellow submarine"]. Ему соответствуют страницы, содержащие слова *yellow* и *submarine*, но не в комбинации *yellow submarine*.

В заключение этого раздела отметим одну техническую особенность системы Yahoo!: она иногда "устает". Если задать ей в быстром темпе несколько десятков запросов, в какой-то момент на экране вместо ответа появляется сообщение "Sorry, Unable to process request at this time – error 999." и список рекомендуемых действий: проверить компьютер на вирусы, обратиться к провайдеру и т.п. Ничего этого делать не надо, следует просто подождать минут 10–15, и система "отдохнет" и заработает снова.

4. Запросы без звездочки

Часто приходится сравнивать варианты, которые заранее известны. Пусть, например, мы хотим проверить, нужен ли в некотором выражении артикль, или выбрать один из двух-трех возможных предлогов. Или, скажем, мы не знаем, какие из заданного списка прилагательных или глаголов лучше сочетаются с данным существительным. Во всех этих случаях может помочь сравнение частот альтернативных вариантов.

В этом и следующем разделах мы всегда включаем в запросы элемент site:uk. Это значит, что при подсчете частот мы ограничиваемся британской частью Интернета. Логика здесь простая: предполагается, что в британском Интернете качество английского языка в среднем выше, чем в мировом Интернете, поскольку английский язык является родным для большей доли авторов (хотя заведомо не для 100%). Соответственно, статистические данные будут более достоверными.

Пример 4.1. Слово высокий (в физическом смысле) может переводиться как *high* или *tall*. Известно, что когда речь идет о человеке, надо говорить *tall*. Действительно:

" a tall man " site:uk	42 100
" a high man " site:uk	62

Кстати, зачем здесь артикль? Почему просто не задать запросы ["tall man" site:uk] и ["high man" site:uk]? Ответ: присутствие артикля сильно увеличивает вероятность того, что слово *tall* или *high* в найденных текстах будет относиться к следующему за ним слову *man*. Без артикля это становится менее обязательным. Например, для запроса ["high man" site:uk] на одной из страниц получаем следующий текст: *I think you hold man too* ***high. Man*** *is a very simple*

creature... Формально здесь все соответствует запросу, поскольку знаки препинания при поиске не учитываются, а заглавные и строчные буквы не различаются, но на самом деле это не то, что мы ищем.

Посмотрим, как *high* и *tall* сочетаются с другими существительными, например *mountain, tree* и *tower*. Частоты демонстрируют явную "избирательность" для *mountain*, менее сильную (и в обратную сторону) для *tree*, и еще меньшую для *tower*:

" a high mountain "	site:uk	20 500
" a tall mountain "	site:uk	263
" a tall tree "	site:uk	11 500
" a high tree "	site:uk	961
" a tall tower "	site:uk	9 380
" a high tower "	site:uk	2 940

Пример 4.2. Пусть мы хотим перевести выражение *сильная простуда*. *Простуда* по-английски *cold,* и в "Longman Dictionary of Contemporary English" мы находим пример *I've got a bad cold.* Нельзя ли, однако, употребить слово *strong* – основной английский эквивалент для слова *сильный*? Сравниваем два варианта:

" had a bad cold "	site:uk	7 710
" had a strong cold "	site:uk	15

и получаем ответ: лучше не надо.

Пример 4.3. Допустим, мы хотим сказать, что кто-то поступил правильно, используя для этого выражение *right thing to do*. Нужен ли здесь артикль, и если да, то какой? Смотрим частоты:

" it was the right thing to do "	site:uk	81 300
" it was a right thing to do "	site:uk	20
" it was right thing to do "	site:uk	58

Различие больше чем в 1000 раз дает основание считать, что

второй и третий варианты практически наверняка неправильны.

Могут возразить: "Зачем в этом случае вообще смотреть частоты? Здесь ситуация такая же, как с выражениями *to tell the truth* и *to tell a lie*: поскольку есть единственный правильный способ поведения, нужен артикль *the*." Это звучит убедительно, но ... Померим частоты, заменив *right* на *wrong*:

" it was the wrong thing to do " site:uk	2 430
" it was a wrong thing to do " site:uk	51
" it was wrong thing to do " site:uk	17

Частоты отличаются меньше, примерно в 50 раз, но этого достаточно, чтобы и здесь решительно предпочесть артикль *the*, хотя с точки зрения приведенного выше аргумента надо было бы выбрать *a*. Язык, как живое существо, не подчиняется слишком жесткой логике.

Пример 4.4. Рассмотрим конструкцию *kind of* плюс существительное, которой по-русски соответствует слово *вид* или *тип* в сочетании с существительным в родительном падеже: *this kind of literature – этот вид литературы*. Если это существительное исчисляемое, то по-русски оно должно иметь форму множественного числа: *этот вид книг, такие типы вопросов*. Посмотрим, что происходит в английском:

" this kind of questions " site:uk	442
" these kinds of questions " site:uk	11 000
" this kind of question " site:uk	12 700
" these kinds of question " site:uk	200

Таким образом, по-английски правило другое: существительное после *of* должно иметь такую же форму числа, как слово *kind*. Хотя остальные две комбинации тоже встречаются, их частоты намного меньше, поэтому их лучше избегать.

5. Запросы со звездочкой

В примерах предыдущего раздела мы сравнивали между собой несколько известных вариантов. Пользуясь символом-джокером *, можно находить заранее не известные слова, которые хорошо вписываются в заданный контекст. При этом среди найденных слов может не быть таких явных лидеров по частоте, как в предыдущем разделе, распределение частот может быть более "пологим". Ценность представляет сам спектр значений этих слов в заданном лексическом окружении. Впрочем, иногда некоторые варианты оказываются намного более частыми, чем другие, как в следующем примере.

Пример 5.1. Как сказать по-английски *большой промежуток времени*? Предположим, что искомое выражение имеет такой вид: *a* (или *an*) – прилагательное – существительное – *of time*. Сначала пробуем несколько прилагательных, имеющих значение "большой": *large*, *big*, *great*, *huge* и (в применении ко времени) *long*, оставляя место существительного свободным. Получаем такие частоты:

" a long * of time " site:uk	991 000
" a great * of time " site:uk	663 000
" a huge * of time " site:uk	98 700
" a large * of time " site:uk	60 900
" a big * of time " site:uk	12 000

Считаем, что остаются два кандидата: *long* и *great*. Чтобы найти самые частые существительные, просматриваем цитаты из страниц, найденных для первых двух запросов.[7] Видим, что после *long* в большинстве цитат идет

[7] Число цитат, показанных на странице поисковой системы Yahoo!, может быть задано в интервале от 10 до 100.

существительное *period*, но бывает также *space* и *span*. После *great* преобладает *deal*, другие возможности – *length* и *amount*.

Таким образом, для каждого прилагательного получаем три возможных существительных – всего шесть комбинаций. Выражения *a great deal of time* и *a great amount of time* по смыслу не соответствуют выражению *большой промежуток времени*, так как означают большое количество времени, а не единый непрерывный промежуток. Частоты оставшихся четырех вариантов выявляют бесспорного лидера:

" a long period of time " site:uk	965 000
" a great length of time " site:uk	2 140
" a long span of time " site:uk	764
" a long space of time " site:uk	270

Обратимся к ситуации другого типа, с гораздо большим разнообразием возможностей.

Пример 5.2. *Произвести впечатление* по-английски – *to make an impression*. По-русски мы говорим: *произвел большое/огромное/сильное/глубокое/неизгладимое/хорошее/плохое/положительное/отрицательное впечатление*. А что говорят по-английски? Задаем запрос

" made a * impression " site:uk 373 000

а на случай прилагательного, начинающегося с гласной, – запрос

" made an * impression " site:uk 33 200

Удобно задать 100, чтобы делать меньше переходов при просмотре. Для этого надо щелкнуть кнопку "Options", затем "Advanced Search", затем внизу открывшейся страницы установить "Number of Results" равным 100 и щелкнуть "Yahoo! Search" справа.

Смотрим цитаты и записываем "заполнители" для *. Сразу видно, что много *big*, *huge*, *good*, *great* и *lasting*. Записав эти слова, добавляем их в конце запроса после site:uk со знаком минус, "чтобы не мешали", и запускаем поиск снова. Тогда в цитатах эти слова уже не появляются и идут следующие по частоте: *deep*, *strong*, *positive*...

Набрав 20 прилагательных, прекращаем просмотр цитат (можно было бы еще долго увеличивать список, но надо где-то остановиться). Получать числовые значения частот при поиске со звездочкой не обязательно – обычно и так видно, какие слова попадаются чаще всего. Впрочем, ради любопытства измерим в этом примере частоты для всех 20 слов. Ниже мы пишем только сами прилагательные, остальная часть запроса подразумевается – например, значение частоты 62 900 соответствует запросу ["made a big impression" site:uk].

big	62 900
lasting	41 200
good	28 400
great	19 100
huge	18 400
deep	13 900
strong	9 590
immediate	8 840
instant	4 260
positive	2 740
real	2 000
profound	1 130
favourable	794
indelible	778
excellent	738
significant	702
powerful	660
early	627

considerable	559
bad	541

Обращает на себя внимание низкая частота слова *bad* по сравнению с прилагательными положительной оценки. Посмотрим аналогичные показатели в русском Интернете:

" произвел хорошее впечатление " site:ru	7 420
" произвел плохое впечатление " site:ru	134

Соотношение частот приблизительно такое же, как у *good* и *bad*. По-видимому, люди просто не очень любят говорить про плохое...

Приведем еще два примера того, как запросы со звездочкой помогают находить идиоматичные словосочетания.

Пример 5.3. Какие английские выражения эквивалентны русским *соответствовать требованиям, удовлетворять требованиям*? Задаем запросы

" must * the requirements " site:uk	57 900
" must * * the requirements " site:uk	68 600

Второй запрос рассчитан на глаголы с предлогами или наречиями. В принципе возможны и трехсловные фразовые глаголы, но мы этот случай рассматривать не будем. Исключая слова типа *specify*, по смыслу не соответствующие заданным русским выражениям, получаем следующий список: *meet, satisfy, fulfil, follow, observe, comply with, conform with, adhere to*.

Мы не случайно начали запросы со слова *must*, а, скажем, не с частицы *to*: ["to * the requirements" site:uk]. В контексте слова *must* более вероятно получить глаголы с нужным нам значением. Если бы запрос начинался с *to*, в цитатах появлялось бы больше глаголов, не имеющих отношения к делу, таких как *use, determine, identify, change, enforce* и т.п.

Пример 5.4. Что говорят англичане, чтобы выразить смысл "элемент иронии", как в русских сочетаниях *доля иронии, оттенок иронии*? Задаем запросы

" a * of irony " site:uk	83 500
" an * of irony " site:uk	1 440

и получаем слова *hint, touch, trace, bit, twist, degree, dash, note, element, tinge, dollop, dose, spot, smidgen* (они приведены в порядке убывания частоты). Наиболее употребительное выражение *a hint of irony* имеет частоту 18 800.

6. Британский и американский английский

> England and America are two countries
> divided by a common language.
>
> George Bernard Shaw [8]

Различия между британским и американским вариантами английского языка хорошо известны – сводку можно найти, например, в справочнике Michael Swan, "Practical English Usage". Естественно ожидать, что эти различия должны проявляться в частотах выражений в Интернете.

Здесь возникает одна техническая трудность. Британский английский легко "локализовать", задавая в запросах элемент site:uk. А как быть с американским? Дело в том, что не существует такого двухбуквенного кода, который соответствовал бы компьютерам, находящимся в США. Вместо кода страны адреса страниц в американской части Интернета содержат коды com, edu, org, net и другие, причем эти

[8] Англия и Америка – две страны, разделенные общим языком. (Джордж Бернард Шоу)

коды могут иметь также страницы на компьютерах вне США.

На самом деле есть код us, аналогичный uk и ru, который используется в основном правительственными учреждениями США. Хотя удельный вес страниц с кодом us в американском Интернете невелик, они вполне подходят для оценки употребительности выражений. Более того, на этих страницах американские предпочтения должны проявляться особенно четко.

Прежде всего поинтересуемся, сколько вообще страниц имеют код us, а заодно и uk, используя для этого следующие запросы:

" " site:us	363 000 000
" " site:uk	2 520 000 000

Определить таким способом число страниц во всем Интернете нам не удастся, так как Yahoo! на запрос [" "] дает ответ 0. Поскольку нас интересует англоязычная часть Интернета, задаем запрос [the] и получаем цифру 32 200 000 000. Это должно быть близко к числу всех страниц на английском языке, так как лишь малая доля из них не содержит артикль *the*. В других же языках слово *the* (в том числе с диакритикой: *thé, thè* и т.п., частое значение – *чай*) имеет гораздо меньшую суммарную частоту, чем в английском.

Приведем несколько примеров, сравнивая частоты для запросов с элементами site:uk и site:us. Мы не будем писать эти элементы рядом с выражениями в кавычках, как это делалось раньше. В строчке будет указано только выражение в кавычках или слово без кавычек, а затем два числа; первое из этих чисел – частота для запроса с элементом site:uk, второе – частота для запроса с элементом site:us.

Пример 6.1. В выражении "жить на такой-то улице" в британском английском употребляется предлог *in*, а в

американском *on*. Для выражений *lived in/on ... street* получаем следующие частоты:

" lived in * street "	85 300	69
" lived on * street "	25 700	2 810

У англичан "свой" вариант является в 3 с небольшим раза более частым, чем "чужой", а у американцев – в 40 раз.

Пример 6.2. Мобильный телефон англичане называют *mobile phone*, американцы – *cell phone*. Сравниваем частоты:

" mobile phone "	67 900 000	636 000
" cell phone "	4 190 000	2 900 000

Здесь уже американцы более толерантны: у них частоты отличаются меньше чем в 5 раз в свою пользу, а у англичан в 15 раз.

Пример 6.3. Посмотрим, что англичане и американцы говорят о расписаниях. Как в английском, так и в американском Интернете слово *schedule* имеет бóльшую частоту, чем слово *timetable*:

schedule	50 700 000	27 500 000
timetable	16 700 000	464 000

Правда, у американцев предпочтение в пользу *schedule* намного сильнее. Оказывается, что употребительность этих слов существенно зависит от ситуации. В контексте школы различия становятся особенно большими:

" school schedule "	18 000	156 000
" school timetable "	81 300	189

Иными словами, *school timetable* – это очень не по-американски, зато вполне по-британски.

7. Частота и правильность

If called by a panther,
Don't anther.

Ogden Nash [9]

– Так что же, – спросит пытливый читатель, – не следует ли из всего сказанного, что правильность выражений "прямо пропорциональна" частоте их появления в Интернете?

Конечно, в такой общей форме – нет! Хотя, казалось бы, многие наши примеры говорят именно об этом. Обсудим этот вопрос более детально, чтобы объективные статистические данные по возможности не приводили к ложным заключениям. В этом разделе мы будем работать с русским материалом, чтобы в полной мере опираться на лингвистическую интуицию читателей – носителей русского языка.

Начнем с того, что в языке Интернета регулярно встречаются намеренные отклонения от нормы. Например, формы глаголов, оканчивающиеся на *-тся* или *-ться*, могут быть написаны с окончанием *-цца*: *кажецца, хочецца, думаецца*; *смеяцца, старацца, пытацца* – такой не очень тонкий языковой юмор.

Здесь уместно заметить, что языковой юмор вообще часто связан с нарушением нормы. Юмористически искаженное произнесение звуков и смещение ударений нередко встречаются в разговорной речи, тот же тип юмора демонстрирует и эпиграф в этом разделе. Другой пример:

[9] Если тебя позовет пантера, не отвечай (Огден Нэш). Слово *answer* искажено, чтобы рифмоваться с *panther*. Русский аналог: Если тигр позовет тебя: Вася! – Не отзывася. (Юрий Манин)

такая классическая литературная форма юмора, как каламбур, строится на том, что разные по смыслу слова имеют одинаковое написание или звучание, в результате чего возникает потенциальная двусмысленность. В то же время для обычных текстов нормой является именно отсутствие двусмысленностей.

Запрос [кажецца site:ru] дает в Yahoo! частоту 51 100. У многих правильных слов частота не намного больше, а правильные выражения из нескольких слов могут встречаться и гораздо реже. Большую частоту могут иметь также настоящие ошибки и описки. Например, слово *единственый* с одним *н*, в котором явно нет ничего смешного, встречается 67 500 раз.

Эти внушительные цифры не должны никого смущать. Главным основанием для использования частот является идея, что в языке "большинство не ошибается". В свете этого тезиса важно не абсолютное значение частоты само по себе, а соотношение значений частоты для разных вариантов. В частности, в случае орфографии существенно соотношение частоты правильной формы и отклонений от нее.

Измерим частоту правильных форм для слов *кажецца* и *единственый*:

кажется site:ru	34 400 000
единственный site:ru	16 500 000

Мы видим, что правильные формы встречаются в сотни раз чаще, чем неправильные. Различие частот может быть меньше – скажем, сочетание *день рождения* имеет частоту 23 100 000, а его нестандартный вариант *день рождение* – частоту 1 560 000. Тем не менее, и такие частоты подтверждают принцип "большинство не ошибается". А если частота альтернативного варианта приближается к частоте основного, то нередко мы имеем дело с возникновением

"второй нормы", которая со временем может вытеснить исходную.

Очень важно понимать, что из тезиса "большинство не ошибается" вовсе не следует, что выражения с меньшей частотой менее правильны. Вспомним сочетания *произвел хорошее впечатление* и *произвел плохое впечатление* из примера 5.2, частоты которых отличаются более чем в 50 раз. Лингвистическая правильность второго сочетания не вызывает сомнений. Почему же оно появляется в 50 раз реже? Ответ простой: потому, что у авторов текстов в 50 раз реже возникает намерение выразить этот смысл.

Мы приходим к тому, что сравнение частот уместно использовать в определенном классе ситуаций, а именно тогда, когда сравниваемые выражения эквивалентны по смыслу или, пользуясь лингвистической терминологией, синонимичны (таковы все примеры в разделе 4). В этих случаях большая разница частот позволяет предположить, что более редкие варианты могут быть лингвистически неправильными. Однако даже в таких ситуациях не-носителю лучше избегать категорических суждений. Что можно сказать с большей уверенностью, это то, что частые варианты являются лингвистически правильными. Поэтому мы можем закончить этот раздел следующей практической рекомендацией:

Используйте более частые варианты – и вы не ошибетесь!

8. Чем заменить Yahoo!?

В июле 2010 года в результате соглашения между компаниями Microsoft и Yahoo! вместо поисковой системы Yahoo! по адресу www.yahoo.com стала работать система Bing компании Microsoft. Это произошло незаметно для пользователей, так как весь интерфейс, включая заголовок "Yahoo!", полностью сохранился. Соглашение предусматривает постепенный переход на Bing и на других сайтах компании Yahoo!.

К сожалению, с помощью Bing измерять частоту выражений в Интернете невозможно – цифры получаются совершенно неправдоподобные, нередко противоречащие друг другу. Кроме того, Bing "не понимает" звездочку и обращается с ней как с обычным знаком препинания, то есть игнорирует.

По-видимому, сегодня есть только два реальных кандидата на то, чтобы заменить Yahoo! при подсчете частот, – Google и, как это ни неожиданно, Yandex. Хотя система Yandex "специализируется" на русском языке, в последнее время она также просматривает некоторую часть Интернета на других языках, в том числе на английском, немецком и французском. Размер части англоязычного Интернета, по которой Yandex проводит поиск, – порядка 700 миллионов страниц.

Yandex, как и Yahoo!, рассматривает звездочку как одно неизвестное слово. Эксперименты показали, что частоты, которые дает Yandex, весьма похожи на результаты Yahoo!. А именно, с помощью Yandex'a были измерены частоты для всех запросов из разделов 1, 4 и 5 (элемент site:uk в запросе опускался). Коэффициент корреляции между логарифмами частот для Yandex'a и Yahoo! оказался равным 0,92, что весьма близко к теоретически возможному максимуму,

равному 1. [10] Аналогичный эксперимент с Google (только на запросах без звездочки; элемент site:uk сохранялся) дал значительно более скромный результат: 0,73.

Сходство результатов Yandex'a и Yahoo! означает близость не абсолютных значений частот, а их отношений. При этом Yandex иногда несколько сглаживает различие частот: там, где у Yahoo! цифры отличаются в 100 раз, Yandex может давать различие в 30–40 раз, иногда и меньше. Это может объясняться тем, что поиск не ограничен британской частью Интернета. Как и в Yahoo! и Google, есть возможность ввести это ограничение (с помощью элемента domain:uk), тогда коэффициент корреляции повышается с 0,92 до 0,98.

По-видимому, с точки зрения измерения частоты английских выражений Yandex является сегодня единственной реальной заменой для Yahoo!. Впрочем, поиск в Интернете – это очень живая и активно развивающаяся область, и в любой момент могут появиться новые системы, обладающие нужными нам качествами. Естественно поставить такой вопрос: какие свойства должны быть у поисковой системы, "идеально" подходящей для наших целей?

Здесь открывается большой простор для фантазии. Например, желательно получать в качестве частоты не число страниц, содержащих заданное выражение, а общее число его появлений в текстах. Было бы также очень удобно, если бы система автоматически считала статистику "заполнителей" для выражений со звездочкой – то, что в примере 5.2 было сделано путем обработки многих отдельных запросов. Идя еще дальше, можно представить себе систему, способную предлагать для заданного словосочетания более идиоматичные варианты – близкие по

[10] Коэффициент корреляции показывает степень линейной зависимости между двумя рядами величин.

смыслу выражения, обладающие большой частотой. Например, мы задаем выражение *to draw a difference* и получаем для него вариант *to draw a distinction*, который встречается в 1000 раз чаще. Это уже очень трудная задача, связанная с вычислением семантического расстояния между словосочетаниями.

Строго говоря, для оценки и сравнения частоты выражений некоторого языка использовать Интернет в принципе не обязательно, можно было бы взять какой-нибудь другой большой корпус текстов. Проблема здесь именно в размере. Приблизительная оценка числа слов в британском Интернете (точнее, в той его части, которая доступна системе Yahoo!) – порядка 500 миллиардов. Существуют корпусы английских текстов, специально собранные лингвистами, но они во много раз меньше: оксфордский корпус английского языка (недоступный широкой публике) содержит порядка 2 миллиардов слов, корпус современного американского английского языка (www.americancorpus.com) – около 400 миллионов, британский национальный корпус (British National Corpus = BNC, www.natcorp.ox.ac.uk) – 100 миллионов.

Следующий эксперимент показывает, что ста миллионов "слишком мало". Возьмем какой-нибудь словарь английских идиом, например "Cambridge Idioms Dictionary", и сравним частоты нескольких случайно выбранных выражений из этого словаря в BNC и в британском Интернете:

bend the rules	11	65 100
raining cats and dogs	2	29 100
open Pandora's box	2	1 190
as flat as a pancake	1	3 110
fight like cat and dog	1	2 340

Здесь слева указано число появлений выражения в текстах BNC, справа – число страниц, найденных Yahoo!. Мы видим, что частоты в BNC очень малы (а следовательно,

статистически недостоверны). В то же время Интернет дает для этих выражений частоты не меньше 1000.

Получается, что "правильный" объем корпуса – порядка сотен миллиардов слов. Сколько места нужно, чтобы хранить такой корпус? Стомиллионный BNC занимает на диске около 9 гигабайт. Из них примерно половина приходится на индекс – вспомогательный массив, используемый для быстрого поиска. Остальное место занимает собственно корпус; он содержит, кроме самих английских слов, много дополнительной информации (например, автоматически приписанные словам грамматические категории). Если простодушно умножить все в тысячу раз, то получится, что для стомиллиардного корпуса нужно дисковое пространство порядка нескольких терабайт.

Много ли это – несколько терабайт? Для сегодняшних компьютеров да, для завтрашних, скорее всего, нет. Уже сейчас есть ноутбуки с диском в 3 терабайта; для внешних дисков типичен объем 5–10 терабайт. При существующих экспоненциальных темпах роста параметров можно ожидать, что в скором будущем массив в несколько терабайт будет комфортабельно помещаться на диске обычного компьютера.

Таким образом, скоро мы сможем иметь "корпус размером с Интернет" у себя дома. Пока же у нас есть доступ к Yahoo! по адресу uk.search.yahoo.com, а после того, как туда поставят Bing, можно будет работать с www.yandex.ru. И вполне возможно, что в недалеком будущем появятся новые поисковые системы, способные хорошо считать частоты в Интернете.

МИТЮШИН Леонид Григорьевич

ИНТЕРНЕТ КАК КОРПУС ЛИНГВИСТИЧЕСКИХ ПРИМЕРОВ

Издательство ООО “МАКС Пресс”
Лицензия ИД N 00510 от 01.12.99 г.
Подписано в печать 18.01.2011 г.
Печать офсетная. Бумага офсетная.
Формат 60x90 1/16. Усл.печ.л. 1,75. Тираж 1000 экз. Изд. № 030.

119992, ГСП-2, Москва, Ленинские горы, МГУ им. М.В. Ломоносова,
2-й учебный корпус, 627 к.
Тел. 939-3890, 939-3891. Тел./Факс 939-3891.

Напечатано с готового оригинал-макета

Типография МГУ
119991, ГСП-1, Ленинские горы, д. 1, стр. 15
Заказ № 0023