Л.Г. Митюшин

Часто или редко?

Частота английских выражений как индикатор правильности.

Пособие по английскому языку для тех, кто не боится арифметики.

Содержание

1. Введение

В этом пособии речь идет об измерении частоты английских выражений в двух больших массивах текстов. Первый из них – часть британского Интернета, с которой работает поисковая система Яндекс, второй – более миллиона книг на английском языке, переведенных в электронную форму в рамках проекта Google Books.

Частотой выражения мы называем число его употреблений в текстах. Почему нас интересует этот показатель? Дело в том, что частота тесно связана с лингвистической правильностью: выражения, имеющие большую частоту, практически всегда правильны. Это можно сказать и в другой форме: ошибочные выражения являются редкими по сравнению с правильными вариантами. Иными словами, в языке действует принцип "большинство не ошибается".

Конечно, правильные выражения тоже могут быть редкими. Частоты нескольких сходных по смыслу выражений могут различаться очень сильно. С точки зрения не-носителя, самые частые выражения среди близких по смыслу имеют два преимущества: они почти наверняка правильны, и они, скорее всего, стилистически нейтральны. Выбор самого частого выражения сводит к минимуму вероятность сделать грамматическую, лексическую или стилистическую ошибку.

Кроме оценки частоты, при работе с указанными массивами есть и другая полезная возможность. А именно, можно задавать для поиска неполные выражения, содержащие, кроме обычных слов, также символ-джокер * ("звездочка"). В этом случае в текстах ищутся выражения, в которых звездочка заменена произвольными реальными словами. Яндекс приводит результаты в виде последовательности цитат из веб-страниц, содержащих найденные выражения; для текстов Google Books строится список всех найденных выражений вместе с их частотами.

Дальше мы подробно описываем работу с Яндексом (глава 2) и базой сочетаний, построенной по текстам Google Books (глава 3).

2. Яндекс

Предположим, что мы хотим оценить с помощью системы Яндекс частоту некоторого английского выражения.[1] Чтобы сделать это, надо открыть в браузере страницу с адресом www.yandex.com (английский интерфейс) или www.yandex.ru (русский интерфейс) и напечатать в поле запроса интересующее нас выражение в кавычках. Кавычки нужны для того, чтобы система искала в Интернете страницы, где заданные слова идут подряд; если ввести запрос без кавычек, слова могут быть разбросаны по разным местам страницы. Затем через один или несколько пробелов надо напечатать служебное слово rhost:uk.* – оно ограничивает поиск страницами с кодом страны uk, то есть такими, которые зарегистрированы в Великобритании (uk = United Kingdom). Это ограничение повышает вероятность того, что авторами текстов будут носители английского языка. Мы для краткости не будем писать кавычки и rhost:uk.* в примерах запросов, но подразумевается, что они всегда "незримо присутствуют".

Затем мы щелкаем кнопку "Search" (или "Найти" в русской версии) и получаем на экране фрагменты страниц Интернета, содержащие заданное выражение. Кроме того, показывается его частота, то есть общее число страниц, где оно встречается. Частота – это как раз то, что нас интересует. Если частота больше 1000, она округляется системой до тысяч, а если больше 1 000 000, то до миллионов.

По отношению к значению частоты надо "проявлять бдительность", особенно если частота неожиданно большая. Дело в том, что если Яндекс

---

[1] "Выражениями" или "сочетаниями" мы называем любые цепочки слов (возможно, со знаками препинания), в том числе и цепочки длины 1, то есть одиночные слова.

не находит ни одной страницы, где слова заданного выражения идут подряд, он взамен ищет страницы, где эти слова встречаются вразбивку. В этом случае появляется сообщение "Точная цитата в кавычках нигде не встречается. Показаны результаты по запросу без кавычек." Иными словами, частота для выражения как целого равна нулю.

Следует учитывать два общих соглашения, принятых и в других поисковых системах. Первое: заглавные и строчные буквы не различаются, второе: знаки препинания не учитываются. Таким образом, мы можем беззаботно печатать в запросе *english speaking countries*, а Яндекс найдет то, что нужно, то есть *English-speaking countries*.

Поисковые запросы могут содержать символ-джокер * ("звездочка"), окруженный пробелами, который представляет одно неизвестное слово. Например, запросу *english to * translation* соответствуют страницы с выражениями *English to French translation*, *English to Japanese translation*, *English to Zulu translation* и тому подобное – все, что только можно вообразить. В запросе может быть больше одной звездочки, тогда им должно соответствовать такое же количество реальных слов в тексте на странице.

Следующие разделы показывают, как работают простые запросы и запросы со звездочкой.

2.1. Простые запросы

Приведем несколько примеров, в которых сравнивается между собой ограниченное число заданных выражений, чаще всего два. Отношение частот показывает разницу в степени их употребительности.

Слово высокий (в физическом смысле) может переводиться как *high* или *tall*. Известно, что когда речь идет о человеке, надо говорить *tall*. Действительно, для сочетания *a tall man* Яндекс дает частоту 2000, а для

a high man всего 17, то есть почти в 120 раз меньше. С другой стороны, например, для *mountain* соотношение обратное, хотя и менее резкое: *a high mountain* имеет частоту 1000, *a tall mountain* частоту 89 – в 11 раз меньше.[2]

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? Измерим частоты трех выражений: *it was the right thing to do* – 2000, *it was a right thing to do* – 8, *it was right thing to do* – 15. Мы получаем большой перевес в пользу *the*: в обоих случаях более чем в 100 раз.

Сравним два возможных английских эквивалента для русского выражения *исключение из правила*: *exception to the rule* и "более русское" *exception from the rule*. Первое сочетание имеет частоту 7000, второе 61, с отношением более 100.

Можно сравнивать частоту выражений, составленных из одних и тех же слов, но в разном порядке. В английском языке порядок слов в целом намного жестче, чем в русском, что нередко выражается в виде большой разницы в частотах. Например, сочетания *will always be able* и *will be always able* имеют соответственно частоты 4000 и 27, с отношением около 150.

2.2. Запросы со звездочкой

Звездочка удобна в тех случаях, когда нам заранее не известно, какие слова часто используются в интересующей нас ситуации. Мы действуем так: обрабатываем в Яндексе запрос со звездочкой, а затем просматриваем фрагменты найденных страниц и отмечаем, какие слова встречаются на

---

[2] Значения частот в этой версии пособия измерены 7 июля 2014 года.

месте звездочки. Затем, чтобы сравнить употребительность разных вариантов, мы измеряем частоты для запросов, в которых звездочка заменена конкретными словами.

Первый пример касается идиоматичного употребления предлогов. В сочетании с местом работы по-русски часто используются предлоги *в* и *на*: *работал в компании*, *работал на фабрике*. А что говорят англичане? Вводим запрос со звездочкой вместо предлога: *worked * the company*. Для него Яндекс находит 2000 страниц, и мы видим, что в показанных фрагментах часто встречаются *at*, *for*, *in* и *with*. Запустив отдельный запрос для сочетания с каждым предлогом, получаем такие частоты: *for* – 1000, *with* – 351, *at* – 222, *in* – 118. Запрос *worked * the factory* дает 420 страниц; просмотр фрагментов показывает, что *factory* "любит" предлоги *in* и *at* (частоты 182 и 181).

Кстати о любви и предлогах. Покажем на примере слова *love*, что простодушное сравнение частот не всегда дает правильный результат. Известно, что существительное *love* сочетается с предлогами *for* и *of*. *For* обычно используется, когда речь идет о чувствах по отношению к людям, а *of* в остальных случаях (*his love for his wife*, *his love of freedom*). Изредка встречается предлог *towards*, но он имеет оттенок архаичности. Что нельзя делать, это буквально переводить с русского: *his love to his wife*. Это звучит примерно так же, как буквальный перевод с английского на русский: *его любовь для жены*.

Попробуем подтвердить сказанное с помощью частот. Возьмем сочетания *his love for her* и *his love to her*. Для первого получаем частоту 1000, для второго 124, с отношением всего 8. Почему же так мало, если *love to* действительно не говорят?

Все становится на свои места, если посмотреть на цитаты из страниц, найденных для *his love to her*. Оказывается, что почти во всех случаях предлог *to* связан не со словом *love*, а с некоторым другим словом, обычно глаголом: *declares his love to her*, *is unable to express his love to her*, *in*

order to prove his love to her. Вывод: при выборе сочетаний для сравнения надо всегда заботиться о том, чтобы они правильно представляли интересующее нас явление. Как говорят англичане, "be careful what you wish for" (будьте осторожны с желаниями, которые вы загадываете).

Вернемся к запросам со звездочкой. Пусть мы хотим узнать, какие глаголы в сочетании со словом *potential* означают *реализовать потенциал, осуществить возможности*. Предлагаем Яндексу запрос *to * his potential*. Мы получаем 1000 страниц со следующими основными "заполнителями" для звездочки в порядке убывания частоты: *fulfil, fulfill* (американское написание слова *fulfil*), *realise, reach, achieve*. Частоты этих вариантов равны соответственно 300, 173, 126, 99, 45.

Что говорят англичане, чтобы выразить смысл "элемент иронии", как в русских выражениях *доля иронии, оттенок иронии*? Задаем запросы *a * of irony* и *an * of irony* (для слов, начинающихся с гласной) и получаем для них частоты 3000 и 170. Просматривая фрагменты страниц, находим такие существительные (в порядке убывания частоты сочетаний): *hint, sense, touch, trace, bit, twist, element* ... Самое частое выражение *a hint of irony* имеет частоту 542.

В заключение этого раздела посмотрим, какие определения употребляются при слове *example*. Пропустим через Яндекс запросы *a * example* и *an * example*. Результатов получается много: 359 000 и 70 000. Выберем десяток самых частых "заполнителей" (не в смысле точного значения частоты, а просто те, которые больше попадаются на глаза). Получаем такой приблизительный список: *good, great, excellent, simple, classic, shining, superb, extreme, practical, nice*. Слова здесь идут по убыванию частоты сочетаний, от 81 000 для *a good example* до 3000 для *a nice example*.

3. База сочетаний

Начиная с 2004 года, компания Google занимается в массовом порядке переводом книг из бумажной формы в электронную. К 2013 году в рамках проекта Google Books было отсканировано более 30 миллионов книг, и по этому массиву информации можно проводить поиск так же, как по Интернету.

Американский лингвист Марк Дейвис создал базу сочетаний слов на основе части данных Google Books. Он взял книги трех типов: изданные в США (155 миллиардов слов в 1,3 миллиона книг), изданные в Великобритании (34 миллиарда слов) и "миллион книг" (89 миллиардов слов; в этом корпусе книги выбраны так, что их годы издания более равномерно распределены во времени). В текстах книг были выделены сочетания длиной до 5 слов, встречающиеся в общей сложности не меньше 40 раз. Дейвис объединил их в базу данных, способную отвечать на широкий круг вопросов о хранящемся в ней материале.

Мы будем называть эту базу данных БС (база сочетаний). Примеры, которые мы приводим ниже, относятся к 34-миллиардному британскому корпусу; для американского и "миллиона книг" все делается точно так же.

Здесь мы советуем читателю открыть в браузере страницу по адресу googlebooks.byu.edu и выполнять действия, описанные в тексте. Сначала нам предлагают выбрать корпус: нажимаем British. На экране появляется пустая таблица, над столбцами которой написаны годы с 1810 по 2000 с интервалом в 10 лет. Пространство под таблицей отведено зоне HELP. В левой части экрана расположены одно под другим пять полей: DISPLAY (вид), SEARCH STRING (поисковый запрос), SECTIONS (секции), SORTING AND LIMITS (сортировка и ограничения) и OPTIONS (параметры). О содержании этих полей мы будем говорить по мере необходимости, а сначала попробуем поработать с БС так же, как мы работали с Яндексом.

3.1. Простые запросы

Прежде всего возьмем запросы с полностью заданными словами из раздела 2.1. Напечатаем в поле SEARCH STRING сочетание *a tall man*. Брать его в кавычки не нужно, так как в БС подразумевается, что это слитное сочетание. После нажатия кнопки SEARCH на экране появляется таблица, в которой слева в колонке WORD(S) стоят 3 варианта нашего сочетания с разной капитализацией (то есть разным использованием заглавных букв): *a tall man, A tall man* и *a tall Man*. Дальше в колонке TOTAL приведено общее количество появлений каждого варианта, соответственно 17 230, 2135 и 61. Затем в таблице показано, сколько раз каждый вариант появляется в книгах, изданных в отдельные десятилетия. Числа над колонками указывают первый год десятилетия; таким образом, самая левая колонка соответствует годам 1810–1819, а самая правая – годам 2000–2009. В нижней строке таблицы стоят суммы чисел в каждой колонке; для колонки TOTAL указано общее число появлений заданного сочетания – 19 426.

Теперь попробуем сочетание *a high man*. На этот раз в таблице будет только один вариант капитализации – исходный, со строчными буквами; общее число появлений сочетания в книгах 1810–2009 гг. равно 109.

Заметим, что числа в таблице БС и значения частоты, которые дает Яндекс, имеют разный смысл. БС подсчитывает число появлений заданного сочетания в текстах, тогда как Яндекс подсчитывает число страниц, где появляется сочетание, не учитывая, сколько раз оно встретилось на каждой странице. Поскольку сочетание может встретиться на странице больше одного раза, результат Яндекса оказывается заниженным по сравнению с общим числом появлений сочетания. Однако реальная разница в большинстве случаев составляет лишь несколько процентов, и мы считаем ее, как говорят математики, "пренебрежимо малой". Поэтому данные БС мы тоже будем называть частотами (строго говоря, с большим правом, чем в случае Янлекса).

Сравним результаты БС для сочетаний *a tall man* и *a high man* с тем, что дает Яндекс. В БС отношение частоты первого сочетания к частоте второго равно 178, в Яндексе 116. Мы можем сказать, что БС показывает более сильное предпочтение в пользу *a tall man*, чем Яндекс.

Период времени 1810–2009 гг. задается в БС по умолчанию. В поле SECTIONS пользователь может сам задать интересующий его период. Слева указаны три стандартных диапазона: 1980s–2000s, 1800s–2000s (фактически соответствует 1810–2009 гг.) и 1500s–2000s; с помощью курсора и клавиши Shift можно выделить любой другой отрезок десятилетий (не меньше двух – с одним десятилетием система работать отказывается).

Поскольку нас интересует современный английский, в дальнейшем мы всегда будем задавать период 1980–2009, для этого надо щелкнуть по надписи 1980s–2000s. Соответствующий объем текстов в британской части БС равен 10,5 миллиарда слов. В этих текстах для сочетания *a tall man* мы получаем частоту 3790, а для *a high man* – частоту 15; отношение частот равно 253, то есть еще больше, чем для периода 1810–2009.

Можно включить несколько вариантов слов в один запрос с помощью вертикальной черты, означающей "или". Так, запросы *a tall man* и *a high man* можно объединить в виде запроса *a tall|high man.* Правда, в этом случае не рассматриваются сочетания, образуемые из исходных заменой строчных букв на заглавные, и мы получаем для *a tall man* меньшее число – 3185.

Обратимся к примеру *it was the / a / Ø right thing to do* (*Ø* означает "пустое слово", то есть отсутствие слова). Поскольку эти сочетания содержат больше 5 слов, работать с ними в БС мы не можем. Чтобы обойти эту трудность, рассмотрим "усеченные" сочетания *the right thing to do* и *a right thing to do.* Для них Яндекс дает частоты 20 000 и 149; результаты БС – 7234 и 54, практически с тем же отношением. Отметим, что поступить так с сочетанием *it was right thing to do* нельзя, так как "усеченное" сочетание *right thing to do* вовсе не предполагает отсутствие

артикля – наоборот, оно будет включать более длинные сочетания *the right thing to do* и *a right thing to do* как возможные частные случаи.

Для остальных примеров из раздела 2.1 мы получаем в БС такие частоты:

exception to the rule	5193	5193/78 = 67
exception from the rule	78	
will always be able	1104	1104/8 = 138
will be always able	8	
in the near future	41 923	41923/138 = 304
in the nearest future	138	

Яндекс дает для этих трех пар отношения частот 115, 148 и 166.

В отличие от систем поиска в Интернете, БС не отождествляет заглавные и строчные буквы и не игнорирует пунктуацию. Знаки препинания (в том числе дефис и апостроф) считаются как бы отдельными самостоятельными словами, и при вводе запроса их надо печатать в окружении пробелов. Исключение составляет комбинация "апостроф + s" в конце слова, которая печатается как обычно. Например, сочетание *he'll come* должно печататься с пробелами вокруг апострофа, а *he's come* нет. Ошибка карается тем, что выдается пустой результат, который означает, что данное сочетание встречается в текстах меньше 40 раз.

В связи со знаками препинания отметим один недостаток исходных данных Google Books: в них не включены сочетания, содержащие запятую. Соответственно, и в БС запятая в запросах не разрешается (приводит к пустому результату). Это затрудняет, например, оценку частоты вводных оборотов, для которых присутствие запятой или запятых является существенным условием.

3.2. Запросы со звездочкой

С простыми запросами БС по существу поступает так же, как Яндекс, – показывает частоту появлений сочетания в текстах. Запросы со звездочками обрабатываются совсем по-другому – БС находит для них сочетания, в которых звездочки заменены реальными словами, и выдает их все сразу (или значительную часть) в виде списка вместе с их частотами. По умолчанию сочетания располагаются в порядке убывания частоты.

Посмотрим, как работает БС с запросами из раздела 2.2. Для *worked * the company* одним нажатием кнопки SEARCH получаем список из 5 сочетаний с предлогами *for, with, in, at, by,* которым соответствуют частоты 526, 102, 58, 44 и 3. Запрос *to * his potential* также дает 5 результатов: *fulfil, develop, realize, achieve, realise* с частотами 91, 55, 54, 39, 36. Для *a|an * of irony* список получается длиннее: 33 сочетания. Он начинается с "заполнителей" *touch, hint, trace, sense, note, degree, kind, element, twist, tinge* с частотами 650, 342, 337, ... и заканчивается сочетаниями со словами *suggestion, smile, spice, look,* имеющими частоты 4, 3, 3, 1. Суммарная частота 33 сочетаний равна 2736.

Наконец, для последнего запроса *a|an * example* число результатов превосходит 100, и в таблицу выдаются только первые 100 сочетаний. Ограничение 100 установлено по умолчанию, его можно изменить в поле OPTIONS, нажав предварительно надпись CLICK TO SEE OPTIONS. Сделаем так и напечатаем в появившемся поле # HITS (количество результатов) число 1000, после чего снова нажмем SEARCH.

Теперь уже в таблицу выдается полный список, содержащий 657 сочетаний; их суммарная частота равна 270 567. В начале списка идут сочетания со словами *good, typical, excellent, simple, classic, prime, fine, clear, early, perfect, extreme, interesting,* имеющие частоты 55 088, 11 141, 11 022, ...

В отличие от Яндекса, где звездочка заменяет только целые слова, в

БС можно использовать звездочку также для представления произвольной (в том числе пустой) цепочки букв внутри слова. В качестве эксперимента введем запрос *s**, задав предварительно количество результатов (# HITS) равным 100000. Мы получим почти то, что требовалось: слова на букву s в порядке убывания частоты, но только не полный список, а первые 4000 слов – таково абсолютное ограничение на число выдаваемых результатов в БС.

Попробуем более реалистичный запрос, например *answered *ly*. На этот раз мы получаем "всего" 393 результата, самые частые слова в этом выражении – *only, correctly, simply, affirmatively, immediately, quickly*. В слове может быть и больше одной звездочки. Например, запрос **work** дает такие слова как *metalworking, workaholic* и *unworkable* (но начинается список, естественно, со слова *work*).

Кроме звездочки, в БС есть еще один символ-джокер: вопросительный знак. Когда он окружен пробелами, он обозначает самого себя, то есть соответствующий знак пунктуации. Будучи же напечатан в цепочке с какими-то другими символами, он представляет один произвольный символ. Он может быть полезен, например, в комбинации со звездочкой для представления произвольной непустой цепочки символов. Скажем, если добавить вопросительный знак к запросу в последнем примере: *?*work** , мы получим такой же список слов за вычетом тех, которые начинаются с *work*, так как перед *work* должен стоять еще хотя бы один символ.

Запросы с вопросительными знаками можно использовать при решении кроссвордов, когда некоторые буквы в искомом слове известны. Пусть, например, требуется найти слово со значением "expert in breaking the law", в котором есть такие три буквы: - - - *m* - - - *l* - - - *s* - . Задаем запрос *???m???l???s?* и получаем список из 16 слов, где вторым по порядку идет нужное нам слово *criminologist*.

3.3. Классы слов

Посмотрим более внимательно на запросы из разделов 2.2 и 3.2: *worked * the company, to * his potential, a|an * of irony, a|an * example.* Фактически нас интересовали в этих сочетаниях не любые "заполнители" для звездочки, а слова определенной части речи: в первом сочетании предлоги, во втором глаголы, в третьем существительные, в четвертом прилагательные. В трех последних примерах служебное слово перед звездочкой было специально добавлено к запросу, чтобы с достаточной надежностью обеспечить нужную часть речи.

В БС разрешается задавать часть речи в запросах явным образом. Есть список классов слов, который в частности включает все традиционные части речи, и в запросах вместо звездочки можно указывать эти классы. Возьмем выражение *a|an * example.* В новом запросе артикля не будет, начинаем сразу с прилагательного. Печатаем в поле запроса код [*j**], означающий произвольное прилагательное. Добавив *example*, получаем запрос в новой форме: [*j**] *example*. Нажимаем SEARCH (установив на всякий случай # HITS по максимуму, 4000) и получаем список, содержащий 1171 сочетание с общей частотой 606 311. Сравнение старого и нового списка показывает, что в верхней части они содержат почти одни и те же прилагательные, но их положение может отличаться на несколько позиций вверх или вниз. Впрочем, есть и существенные различия – например, в новом списке на восьмом месте стоит сочетание со словом *best,* а в старом из-за неопределенного артикля вообще нет сочетания с этим словом, как и с другими прилагательными в превосходной степени.

Основным частям речи соответствуют такие коды: [n*] – существительное, [v*] – глагол, [j*] – прилагательное, [r*] – наречие, [p*] – местоимение, кроме притяжательных типа *my, your, his, their,* [i*] – предлог, [c*] – союз. Есть и более узкие категории – например, [nn1*] означает нарицательное существительное в единственном числе, [vm*] –

модальный глагол, [jjr*] – прилагательное в сравнительной степени. Список классов слов и их кодов приведен в Приложении. Вместо того, чтобы печатать код в запросе с клавиатуры, можно открыть выпадающий список классов слов, щелкнув надпись POS LIST в поле SEARCH STRING, и затем щелкнуть на нужном классе – результат будет тот же самый.

Коды классов слов можно комбинировать с неполностью заданными словами, соединяя их точкой. Например, запрос *un*.[*j**] **ness* дает *unfinished business, unhappy consciousness, underlying illness* ... – всего 1125 сочетаний, которые встречаются в общей сложности 37 669 раз.

Работа с классами слов в БС организована технически довольно сложно. Дело в том, что исходный массив сочетаний, лежащий в основе БС, не содержит информации о классах слов. БС берет необходмимую информацию из другого источника – Корпуса современного американского английского языка (Corpus of Contemporary American English, COCA), и относит слово к некоторому классу в том случае, если в COCA ему приписан этот класс не менее чем в 50% случаев. Как следствие, в результатах поиска появляется некоторый "шум". Однако этот недостаток не перевешивает преимущества, связанные с возможностью прямо указывать классы слов в запросах.

3.4. Другие возможности

Вместо конкретной формы слова в запросе можно задать его основную форму (называемую леммой), заключив ее в квадратные скобки. В этом случае БС ищет сочетания со всеми формами указанного слова. Например, для запроса [*go*] *to school* будут выданы частоты сочетаний со словами *go, goes, went, gone, going.* В случае прилагательных рассматривается сравнительная и превосходная степень: запрос [*high*] *prices* даст сочетания *high prices, higher prices* и *highest prices.*

Если перед словом в квадратных скобках поставить знак равенства, то будут рассматриваться не формы этого слова, а его синонимы. Например, запрос [=*beautiful*] *woman* дает 14 сочетаний с прилагательными *beautiful, attractive, lovely, handsome, wonderful, charming, striking, delightful, gorgeous, magnificent, stunning, exquisite, superb, pleasing* (в порядке убывания частоты). Для запроса [=*good*] *weather* БС предлагает 26 сочетаний (*fine, good, fair, clear, sunny, mild*...), для [=*bad*] *weather* – 23 сочетания (*bad, adverse, severe, poor, appalling, harsh*...).

Для запросов этого типа БС находит только такие сочетания, где слова-синонимы стоят в основной форме. Так, для запроса [=*find*] *the answer* мы получаем 5 сочетаний с глаголами *find, get, discover, obtain* и *understand* в основной форме; другие формы этих глаголов не рассматриваются. Не надо также рассчитывать, что будут найдены все синонимы, возможные в данной конкретной ситуации. Например, по отношению к слову *accent* прилагательные *strong* и *thick* можно считать синонимами (*thick German accent = strong German accent*), и сочетания *strong accent* и *thick accent* оба присутствуют в БС для периода 1980–2009 (с частотами 562 и 235). И тем не менее БС не предлагает сочетание *thick accent* в ответ на запрос [=*strong*] *accent*. Дело в том, что при обработке таких запросов БС использует обобщенные синонимические ряды, элементы которых выражают сходный смысл в широком классе ситуаций, а не только в сочетании с некоторыми ограниченными группами слов.

Перед элементом запроса можно поставить знак минус (или, что то же, дефис), который означает отрицание. Например, запрос *answered -*ly* дает сочетания, где после *answered* идут любые слова кроме тех, которые оканчиваются на *ly*. Минус можно присоединять к любым элементам запросов, кроме обычных слов: запрос *answered -correctly* даст пустой результат.

3.5. Разное

С системой БС могут работать бесплатно все желающие. Однако после выполнения первых 10–15 запросов пользователей просят зарегистрироваться. Для этого надо либо перейти по предлагаемой ссылке we ask that you register, либо открыть страницу по адресу corpus.byu.edu (где перечислены все корпусы, созданные Марком Дейвисом) и нажать в меню слева кнопку Register.

Возможности БС подробно и с многочисленными примерами излагаются в пояснительных текстах, доступных в зоне HELP в правой нижней части рабочего экрана. Мы описали в предыдущих разделах то, что непосредственно нужно для наших целей. Отметим еще некоторые функции и свойства БС, которые могут быть интересны в более широком контексте.

В поле DISPLAY (вид) есть две опции: LIST (список) и CHART (график). По умолчанию действует опция LIST, при которой результаты выдаются в виде числовой таблицы. Строки таблицы соответствуют найденным сочетаниям, столбцы – десятилетиям в пределах заданного периода, а в ячейках стоят частоты, показывающие, сколько раз данное сочетание встретилось в книгах, опубликованных в данном десятилетии. Кроме того, есть колонка, в которой показано суммарное число появлений каждого сочетания за весь период. Все наши данные были получены именно в этом режиме. Если же выбрать CHART, то в ответ на запрос на экране возникает график в виде ряда прямоугольных столбиков, где высота каждого столбика пропорциональна относительной частоте заданного сочетания в книгах, опубликованных в данном десятилетии. Под относительной частотой понимается отношение абсолютной частоты сочетания (числа его появлений в текстах) к общему количеству слов в текстах. Можно сказать, что относительная частота равна вероятности появления сочетания в случайно выбранном месте текста. Этот режим позволяет наглядно

представить, как исторически меняется употребительность сочетаний. Заметим, что для запросов, которым соответствует больше одного возможного сочетания, рассматривается суммарная частота всех сочетаний, то есть график, в отличие от таблицы, строится для запроса как целого.

Ячейки таблицы результатов, получаемой в режиме LIST, окрашены в синий цвет разной интенсивности. Это особенно заметно для больших периодов времени, таких как 1810–2009. Интенсивность цвета в разных ячейках строки, соответствующей некоторому сочетанию, отражает степень его употребительности в разные десятилетия. Сгустки темно-синего цвета говорят о времени "расцвета" сочетания, бледные полосы – о его "упадке".

Цвет ячеек выбирается по следующему правилу. Для каждой ячейки вычисляется относительная частота данного сочетания в книгах соответствующего десятилетия, и затем подсчитывается максимальная относительная частота сочетания среди всех десятилетий рассматриваемого периода. После этого для каждой ячейки относительная частота, соответствующая этой ячейке, делится на максимальную относительную частоту. Результат попадает в одну из пяти групп (от 0 до 0,2; от 0,2 до 0,4; от 0,4 до 0,6; от 0,6 до 0,8; от 0,8 до 1; числа на границе двух групп включаются в бóльшую группу). В соответствии с этим ячейка получает один из пяти уровней интенсивности синего цвета, от бледно-голубого для отрезка 0–0,2 до темно-синего для отрезка 0,8–1.

4. Заключение

Мы познакомились с тем, как можно измерять частоту сочетаний слов в очень больших массивах английских текстов. Первый из этих массивов – часть британского Интернета, с которой работает Яндекс, второй – книги,

составляющие основу базы сочетаний Марка Дейвиса.

Хотя Яндекс предоставляет менее широкий набор функций, чем база сочетаний, у него есть и определенные преимущества. Во-первых, у Яндекса нет ограничений на длину сочетаний. Во-вторых, он работает очень быстро. Если у вас хороший выход в Интернет, ответ приходит практически мгновенно; в БС обработка запросов с большим количеством результатов может занять несколько секунд. Еще одна особенность, которая часто оказывается удобной, – игнорирование знаков препинания (вспомним проблему БС с запятой).

Этими свойствами обладает и самая известная система поиска в Интернете – Google. Возникает вопрос: почему мы выбрали Яндекс? Ведь он не обеспечивает доступ ко всему Интернету, так как ориентирован главным образом на его российскую часть. Причина этого простая: замечательный в разных отношениях Google печально известен ненадежностью своих данных о числе результатов. Кроме того, Google по-другому работает со звездочкой: он считает, что звездочка может заменять сразу несколько слов, и в результате предлагает много лишних вариантов.

Размер части британского Интернета, по которой Яндекс проводит поиск, можно очень приблизительно оценить в 20–30 миллиардов слов. Это в десятки раз больше объема текстов, воспринимаемых человеком в течение всей жизни (а эти тексты являются для него главным источником информации о том, как устроен язык). В корпусе такого размера должны много раз встречаться те образцы употребления слов, на примере которых человек учится владеть языком. Конечно, лучше было бы получать статистические данные по всему британскому Интернету; тем не менее, существующее положение дел также можно считать вполне удовлетворительным.

Интернет – это живой, непрерывно меняющийся океан информации. Частоты одних и тех же выражений могут заметно изменяться даже за небольшое время. Для нас важно, что качественные различия между

частотами при этом сохраняются: если выражение А встречалось во много раз чаще, чем выражение В, то это почти наверняка так и останется. Отметим, что БС является с этой точки зрения идеальной системой, так как в ней никаких изменений не происходит и все измерения полностью воспроизводимы.

Кроме доступа к большим массивам текстов, эра Интернета открывает и другие возможности в области изучения иностранных языков. Так, в англоговорящих странах сегодня работают многие сотни радиостанций, чьи программы можно слушать через Интернет. Для целей обучения особенно полезны "разговорные" станции (где сравнительно мало музыки), такие как BBC Radio 4 (www.bbc.co.uk/radio4) или BBC London (www.bbc.co.uk/bbclondon).

В Интернете есть электронные версии известных учебных англо-английских словарей: Кембридж (dictionary.cambridge.org), Лонгман (www.ldoceonline.com), Макмиллан (www.macmillandictionary.com), Оксфорд (www.oxfordlearnersdictionaries.com) (словари указаны в алфавитном порядке). Они содержат ту же информацию, что и их бумажные аналоги, но пользоваться ими удобнее. Есть и новый элемент: произношение слов можно не только увидеть в обычной фонетической записи, но и услышать.

Эти словари дают очень хорошие дефиниции слов и поясняют их большим количеством живых примеров. Однако есть классы слов, которые в принципе трудно описывать в словарях, например, названия животных и растений. Для них дефиниции сообщают лишь самую общую информацию, слишком скудную, чтобы составить реальное представление об описываемых объектах. Здесь снова оказывается полезным Интернет. Он позволяет увидеть зрительные образы, соответствующие словам, причем сразу в большом количестве. Для этого можно использовать Google и другие поисковые системы в режиме "картинки". Вводим интересующее нас слово – скажем, название птицы *kingfisher* (по-русски *зимородок*) – и получаем изображения этой интересной птицы во всех возможных видах и

ракурсах. Другая ситуация – слова, обозначающие цвет. Например, разница между *"brown hair"* и *"auburn hair"* сразу становится ясной, когда Google в ответ на эти запросы выдает в режиме "картинки" сотни фотографий женщин с волосами соответствующего цвета.

Возвращаясь к основной теме пособия, можно сказать, что возможность работы с большими массивами текстов открывает новые перспективы для изучающих английский язык. Хорошее владение языком означает не только умение выбирать слова, имеющие нужный смысл, – требуется еще, чтобы эти слова идиоматично сочетались друг с другом. На достижение этой цели и направлены описанные в пособии методы и приемы.

Если Вы считаете это пособие полезным,
пожалуйста, дайте ссылку на него Вашим друзьям и коллегам.

Приложение

Классы слов и их представление в запросе.

В первой колонке приводится обозначение класса в выпадающем списке POS LIST (= Part of Speech List), во второй – его код, то есть представление в тексте запроса.

noun.ALL	[nn*]	нарицательное существительное
verb.ALL	[v*]	глагол
adj.ALL	[j*]	прилагательное
adv.ALL	[r*]	наречие
neg.ALL	[xx*]	отрицательная частица (*not*)
art.ALL	[at*]	артикль
det.ALL	[d*]	детерминатив, включая пре- и постдетерминативы (*all, many, much, few, little, some, several, any, another, both, each, every, same, this, that, these, those, what, whatever, whatsoever, which, whichever, whichsoever, whose*)
pron.ALL	[p*]	местоимение, кроме притяжательных адъективных
poss.ALL	[app*]	притяжательное адъективное местоимение (*my, our, your, his, her, its, their, thy*)
prep.ALL	[i*]	предлог
conj.ALL	[c*]	союз
noun.ALL+	[n*]	существительное
noun.SG	[nn1*]	нарицательное существительное в единственном числе
noun.PL	[nn2*]	нарицательное существительное во множественном числе

noun.+PROP	[np*]	собственное существительное
verb.BASE	[vv0*]	неслужебный глагол в основной форме
verb.INF	[v?i*]	глагол в инфинитиве
verb.INF/LEX	[vvi*]	неслужебный глагол в инфинитиве
verb.MODAL	[vm*]	модальный глагол
verb.3SG	[v?z*]	глагол в форме 3-го лица единственного числа
verb.ED	[v?d*]	глагол в прошедшем времени
verb.EN	[v?n*]	глагол в форме причастия прошедшего времени
verb.ING	[v?g*]	глагол в форме причастия настоящего времени
verb.LEX	[vv*]	неслужебный глагол
verb.BE	[vb*]	глагол *be*
verb.DO	[vd*]	глагол *do*
verb.HAVE	[vh*]	глагол *have*
adj.CMP	[jjr*]	прилагательное в сравнительной степени
adj.SPRL	[jjt*]	прилагательное в превосходной степени
adv.PRTCL	[rp*]	адвербиальная частица
adv.WH	[rrq*]	вопросительное/относительное наречие (*how, why, where, when* ...)
pron.INDF	[pn1*]	неопределенное или отрицательное местоимение (*everything, anything, something, nothing, everybody, anybody, somebody, nobody, everyone, anyone, someone, nil, nought*)
pron.PERS	[pp*]	личное местоимение (*I, we, you, he, she, it, they, thou, ye, me, us, him, her, them, thee, mine, ours, yours, his, hers, theirs, myself, ourselves, yourself, yourselves, himself, herself, itself, themselves, thyself*)

pron.WH	[pnq*]	местоимение *who* и его производные (*who, whom, whoever, whosoever*)
pron.REFL	[ppx*]	возвратное местоимение (*myself, ourselves, yourself, yourselves, himself, herself, itself, themselves, thyself*)
num.CARD	[mc*]	количественное числительное (*one, two, three, four, ... ; 1, 2, 3, 4, ... ; I, II, III, IV, ...*)
num.ORD	[md*]	порядковое числительное (*first, second, third, fourth, ... ; 1st, 2nd, 3rd, 4th, ...*)
conj.CRD	[cc*]	сочинительный союз
conj.SUB	[cs*]	подчинительный союз
interj	[uh*]	междометие
PUNC	[y*]	знак препинания