СТАТИСТИЧЕСКИЙ МЕТОД ВЫЯВЛЕНИЯ КОЛЛОКАЦИЙ[1]

В.П. Захаров, М.В. Хохлова
Санкт-Петербургский государственный университет
Санкт-Петербург

ВВЕДЕНИЕ

Понятие коллокации представляет собой семантико-синтаксическое единство. В корпусной лингвистике часто под этим термином понимают статистически устойчивое словосочетание, т.е. сочетание, обладающее некоторой степенью устойчивости. В нашей работе предлагается описание существующих методов выявления коллокаций. В корпусной лингвистике принято говорить о совместной встречаемости языковых единиц, т.е. о статистически устойчивых синтагмах в тексте, как разрывных так и неразрывных.

Вероятностный характер языка не вызывает сомнений. Другое дело, какое значение мы придаем данной вероятностной природе языка. Как указывает Н.Д. Андреев, статистические характеристики языковых единиц в речи оказываются весьма важным фактором при описании языкового материала [1]. Он предлагает статистико-комбинаторное моделирование языка по данным речи: «1) вскрыть систему языковых форм, исходя из статистики комбинаторики, но совершенно не используя никаких значений (ни лексических, ни грамматических) и не обращаясь к критерию грамматической правильности; 2) лишь после этого, опираясь на смысл текстов, установить значения выявленных форм, а через них – и грамматическую правильность» [1, с. 6]. Вероятность предлагается рассматривать как характеристику, связывающую отдельный элемент с системой в целом и позволяющую выделить устойчивые отношения между элементами системы [2].

Особенно активно это направление, связанное с выявлением устойчивых словосочетаний, стало развиваться с появлением корпусов текстов, позволяющих получать достоверные статистические данные. Возможность получать информацию о совместной встречаемости слов в тексте предоставляет большинство корпус-менеджеров.

1. Работа выполнена при поддержке гранта для студентов, аспирантов вузов и академических институтов, расположенных на территории Санкт-Петербурга, 2008г., 1.4/4-05/13 «Разработка системы формирования статистико-грамматических шаблонов лексической сочетаемости для русского языка»

Самым простым способом выявления коллокаций в тексте является составление частотных списков слов, оказавшихся слева или справа от ключевого. Часто используется список стоп-слов, состоящий, например, из предлогов или артиклей (служебных и незначащих слов).

Данные о частоте не могут служить аргументом для описания языковых данных. Поскольку по ним нельзя судить, насколько тесно слова связаны друг с другом. Поэтому речь идет о статистической интерпретации данных о частоте. Также желательно получить данные о том, насколько данное сочетание характерно для языка в целом, а не для отдельной выборки. Таким образом, речь идет о статистической модели.

Аппаратом для установления данной связи между случайной и обусловленной встречаемостью слов служат *меры ассоциации*.

Ясно, что речь идет о статистической ассоциации, которая в свою очередь, может служить предпосылкой для синтаксической или лексической ассоциации.

Свойство устойчивости в той или иной степени присуще всем словосочетаниям. Поэтому порог устойчивости [3], согласно которому словосочетания могут быть отнесены к высокоустойчивым (на практике такие сочетания называют просто устойчивыми) или низкоустойчивым (на практике их называют неустойчивыми), может быть выбран на основании значений каких-то из описанных ниже мер.

Эти методы основываются на нахождении n-грамм (часто это биграммы или триграммы) в пределах заданного диапазона. Меры ассоциации используются для вычисления степени близости между компонентами биграммы. Они основаны на данных о частоте из таблицы сопряженности 2x2 рассматриваемых слов, где каждая из двух переменных может принимать одно из двух значений (истина – ложь), поэтому любая биграмма принадлежит одной из четырех комбинаций этих переменных. Конечно, статистическая связанность компонентов биграммы не всегда говорит о семантической или синтаксической связанности. Тем не менее, линейная близость может оказаться важной фактором, который необходимо учитывать при нахождении устойчивых сочетаний, т.е. коллокаций и других типов словосочетаний в текстах.

В настоящее время в лингвистике существует несколько способов для вычисления степени связанности частей той или иной коллокации. Они основываются на сравнении частот для пар слов, полученных на материале реального корпуса, с независимыми (относительными) частотами некоторого гипотетического корпуса, состоящего из тех же слов, расположенных в случайном порядке. Ищутся статистически

значимые отклонения реальных частот от гипотетических вероятностей (подробнее см. [4]). Но при этом не стоит забывать, что слова, даже вне своей тенденции к сочетаемости, все равно не могут появляться в случайном порядке. Следует учитывать, что существуют правила грамматики, и нужно принимать во внимание не только лексические, но и грамматические закономерности языка. «Языковая система по сути своей вероятностная, и что частота в тексте является иллюстрацией вероятности грамматической» [5, с. 31]. Например, следует принимать во внимание структурные формулы, которые лежат в основе коллокаций. Их комбинация со статистическими подходами, по нашему мнению, может дать неплохой результат. Хотя на практике такая грамматическая вероятность не учитывается.

На практике оказывается, что наиболее часто исследуются биграммы, поскольку уже для сочетаний длиннее триграмм применение статистических тестов оказывается сложным, хотя и возможным.

К недостаткам статистически ориентированного подхода можно отнести сложность вычисления n-грамм для значений n>3. Также в списках выдаваемых коллокаций присутствует «шум», т.е. свободные словосочетания, сочетания со знаками пунктуации и т.п.

Чисто статистический подход главным образом применяется к текстам конкретных жанров. Так, Г. Диас [6] использует его для выявления многословные единиц в корпусах Европейского Парламента, а в работе [7] делается попытка выделить медицинские термины из базы данных медицинских текстов при помощи статистических методов.

Полученные списки коллокаций, термины и многословные единицы могут быть полезны в таких областях, как извлечение знаний из текста (knowledge extraction), информационный поиск и машинный перевод.

СТАТИСТИЧЕСКИЙ АППАРАТ

В основе вероятностно-статистических методов, которые, в свою очередь, представлены в виде мер вычисления устойчивости компонентов словосочетания, лежат постулаты теории вероятности и математической статистики.

Для слов в тексте (корпусе) можно вычислять среднее значение и стандартное отклонение.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (di - \mu)^2}{n-1}$$

di – смещение слова; n – общее количество случаев, когда слова встречаются вместе, м – среднее.

Эти величины характеризуют распределение слов в корпусе в определенном диапазоне. Если смещение одно и то же, тогда отклонение равно 0. По этой информации можно находить коллокации в корпусе, например, с низким значением стандартного отклонения, что означает, что слова расположены на одинаковом расстоянии друг от друга. Если величина стандартного отклонения велика, это означает, что слова не входят в состав одной коллокации.

Так, в Национальном корпусе русского языка для словосочетания «уделить внимание»:

5 контекстов с расстоянием между словами = 1 (*пример*: Не последнюю роль здесь играет то, что инженеры чуть больше, чем обычно, **уделили внимание** борьбе с высокочастотными шумами);

5 контекстов с расстоянием между словами = 2 (пример: Вот на ваш взгляд / вы считаете / что вот этим проблемам России нужно **уделить** больше **внимания**);

3 контекста с расстоянием между слова = 3 (пример: В последнее время в средствах массовой информации Ходорковскому было **уделено** довольно много **внимания**).

Таким образом, среднее равно:

$$\mu := \frac{1}{13}(1\cdot 5 + 2\cdot 5 + 3\cdot 3) = 1.846$$

Значение стандартного отклонения равно:

$$\sigma := \sqrt{\frac{1}{12}\left[5\cdot(1-1.846)^2 + 5\cdot(2-1.846)^2 + 3\cdot(3-1.846)^2\right]} = 0.641$$

Незначительная величина стандартного отклонения говорит о том, что слова чаще всего встречаются с одинаковым смещением.

Существуют различные формулы для вычисления силы синтагматической связи на основе частотных характеристик лексических единиц в тексте. Их подробное описание и анализ можно найти в работе Ш. Эверта [8].

Рассмотрим гипотетическое частотное распределение биграммы *принять решение* (X,Y), в которой случайные переменные имеют значения: X = «принять», Y = «решение».

Ниже представлена таблица сопряженности [1] для биграммы.

Табл. 1 Таблица сопряженности для биграмм

	Y	¬ Y	
X	O_{11}	O_{12}	N_{1P}
¬ X	O_{21}	O_{22}	N_{2P}
	N_{P1}	N_{P2}	N_{PP}

Таким образом, в таблице представлены наблюдаемые частоты. O_{11} – частота данной биграммы, т.е. количество случаев, когда оба компонента X и Y сочетания встречаются вместе. O_{12} – частота биграмм, в которых встретилось слово X встретилось, а слово Y нет. O_{21} – количество биграмм, в которых встретилось слово Y и не встретилось слово X. O_{22} – частота биграмм, в которых не присутствуют ни слово X, ни слово Y. N_{1P}, N_{P1}, N_{2P} и N_{P2} представляют собой маргинальные частоты, которые содержат информацию, встречается ли слово X или Y в данной биграмме. N_{PP} – общее количество биграмм в корпусе.

$O_{11} + O_{12} + O_{21} + O_{22} = N_{PP} = N$

Таким образом, для нашего примера получим следующие сочетания: XY (принять решение – O_{11}), X(¬Y) (принять резолюцию – O_{12}), (¬ X)Y (обсудить решение – O_{21}), (¬ X)(¬Y) (обсудить резолюцию – O_{22}).

$O_{11} = f(X,Y)$, гд f – частота.

$O_{12} = f(X) – f(X,Y)$

$O_{21} = f(Y) – f(X,Y)$

$O_{22} = N – f(X) – f(Y) + f(X,Y)$

Маргинальные частоты используются для того, чтобы вычислить ожидаемые частоты для каждой клетки таблицы.

Таблицы сопряженности можно рассмотреть для любого N. В этом случае количество частот с ростом элементов коллокации, которые необходимо подсчитать, увеличивается как 2^N. Приведем таблицу сопряженности для триграмм.

1. Иногда эту таблицу называют *факторной таблицей.*

Табл. 2 Таблица сопряженности для триграмм

		Z	¬ Z	
X	Y	O_{111}	O_{112}	N_{11P}
X	¬ Y	O_{121}	O_{122}	N_{12P}
¬ X	Y	O_{211}	O_{212}	N_{21P}
¬ X	¬ Y	O_{221}	O_{222}	N_{22P}
		N_{PP1}	N_{PP2}	N_{PPP}

Статистическая интерпретация совместной встречаемости основана на случайной выборке из бесконечной генеральной совокупности.

Рассматривается нулевая гипотеза (H0), согласно которой лексемы не влияют друг на друга, следовательно, f(X)*f(Y)/N, т.е. чисто случайно.

Вычисляется вероятность p того, что слова встречаются по гипотезе H0. Она отвергается, если вероятность $p < 0,05$.

Ожидаемые частоты (Eij) вычисляются по факторной таблице 3 на основании маргинальных частот.

Табл. 3. Вычисление ожидаемых частот слов по таблице сопряженности

	Y	¬ Y	
X	$E_{11} = N_{1P} * N_{P1} / N_{PP}$	$E_{12} = N_{1P} * N_{P2} / N_{PP}$	N_{1P}
¬ X	$E_{21} = N_{P1} * N_{2P} / N_{2P}$	$E_{22} = N_{P2} * N_{2P} / N_{PP}$	N_{2P}
	N_{P1}	N_{P2}	N_{PP}

МЕРЫ АССОЦИАЦИИ

Чаще всего используются меры ассоциации MI, t-score, log-likelihood и z-score. MI и z-score из-за особенностей самой формулы имеют тенденцию увеличивать значение меры, по которому можно с уверенностью сказать, что слова сочетаются друг с другом не случайно, т.е. данное числовое значение возрастает для слов с небольшой частотой. Они основаны на сравнении частот слов и пар слов (биграмм, или более широко, коллокаций). Однако, значения, принимаемые данными мерами, используются не для принятия или отвержения гипотезы H0, а для упорядочения словосочетаний-кандидатов [9]

Нельзя с точностью утверждать, что какой-то из методов предпочтительнее. Каждый из них имеет свои достоинства и недостатки, о которых будет сказано ниже.

Мы считаем, что для каждого языка следует учитывать правила словообразования. А для языков типа русского, отличающихся большой флективностью, иногда необходима предварительная лемматизация для извлечения коллокаций.

MI

В работе [10] был введен *коэффициент ассоциации* (association ratio) для вычисления степени связанности между словами. В его основе лежало понятие *взаимной информации* (mutual information), заимствованное из теории информации и впервые примененное в работе [11, с. 28], определяемое как:

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x) * P(y)}$$

где

x, y – слова;

P(x), P(y) – их вероятности.

Как ожидается, если слова действительно связаны, тогда наблюдаемая совместная вероятность их встречи P(x,y) будет много больше, чем случайная P(x)*P(y), следовательно, I(x,y) >> 0. Если между словами нет особой зависимости, тогда P(x,y) ≈ P(x)*P(y), следовательно, I(x,y) ≈ 0. Если слова находятся в отношении дополнительной дистрибуции, тогда P(x,y) будет много меньше P(x)*P(y), отсюда I(x,y) << 0. Высказывается гипотеза, согласно которой значение I(x,y) > 3 выявляет интересные для рассмотрения случаи, в то время как величина меньше 3 – нет [10, с. 78]. Как указывается, эта мера позволяет выделять не только коллокации, но и семантические классы.

MI-score (коэффициент взаимной зависимости, объем информации [1]) сравнивает зависимые контекстно-связанные частоты с независимыми, как если бы слова появлялись в тексте совершенно случайно:

$$MI = \log_2 \frac{f(n,c) * N}{f(n) * f(c)} \quad ,$$

где

MI = mutual information;

n – ключевое слово;

c – коллокат;

1. Иногда называется коэффициентом поточечной взаимной информации (pointwise mutual information) в отличие от обычного коэффициента взаимной информации.

f(n,c) – частота встречаемости ключевого слова n в паре с коллокатом c;

f(n), f(c) – абсолютные (независимые) частоты ключевого слова n и слова c в корпусе (тексте);

N – общее число словоформ в корпусе (тексте).

В терминах вышеприведенных частот (см. табл. 1, 3), данная формула выглядит следующим образом:

$$MI = \log \frac{O_{11}}{E_{11}} \quad ,$$

где

O_{11} – наблюдаемая частота биграммы.

E_{11} – ожидаемая частота биграммы.

Мера MI вычисляет вероятность двух встречающихся вместе слов путем сравнивания произведения их относительных частот в корпусе с наблюдаемыми частотами их совместной встречаемости. Разница между этими величинами выявит степень значимости их встречаемости.

Если значение MI (n; c) больше 1, тогда данное сочетание слов считается статистически значимым. В случае если MI (n; c) примерно равно 0, сочетание слов является менее статистически значимым, слова появляются в паре крайне редко. MI (n; c) меньше 0 означает, что n и c находятся в отношении дополнительной дистрибуции. Вопрос о том, какие значения MI следует считать пороговыми, остается открытым. Так, в работе [12, с. 266] говорится, что значения, намного превышающие 0, свидетельствуют, что слова встречаются не случайно.

В отечественной традиции этой мере соответствует *коэффициент неслучайности*, который широко использовался в работах Н.Д. Андреева для выделения морфем и морфоподобных сегментов (цит. по [13]). Значения интервала необходимо подбирать, поскольку на разных уровнях языковой структуры соотношения частот имеют собственную структуру.

Мера MI позволяет выделить устойчивые словосочетания, имена собственные, а также низкочастотные специальные термины. Это особенно важно в задачах информационного поиска, поскольку, если будет предоставлена информация об их сочетаемости в разных областях знаний, это позволит более эффективно сравнивать документы, релевантные запросу. Слова, у которых MI-score принимает наибольшую величину, менее частотны и обладают ограниченной сочетаемостью.

Однако эта формула не лишена недостатков. Она не позволяет выявить особенные и нетипичные коллокации. Также одним из недостатков является то, что значение MI больше для меньшего числа коллокаций, т.е. вес каждой отдельной коллокации тем больше, чем реже она встречается. Поэтому в случаях, когда частота коллокации мала, использование данной формулы может привести к неправильным результатам. Чтобы решить эту проблему в работе [14, с. 171–172] эмпирически выводится формула, в которой величина f(n,c) возводится в куб.

Есть и еще один вариант для вычисления MI, учитывающий размер диапазона, в котором ищется слово:

$$MI = \log_2 \frac{f(n,c) * N}{f(n) * f(c) * S},$$

где S – рассматриваемый диапазон (размер окна).

Заметим, что для формулы MI-score при контексте, равном +5/-5 слов, исследователь неизбежно сталкивается с большим «шумом», т.е. с большим количеством слов, имеющих значительную способность к сочетанию. При использовании меньшего контекста число выявленных коллокатов уменьшается, остаются лишь наиболее тесно связанные слова.

Слова, у которых MI-score принимает наибольшую величину, менее частотны и обладают ограниченной сочетаемостью.

T-score

Мера t-score также учитывает частоту совместной встречаемости ключевого слова и его коллоката, отвечая на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами:

$$t - score = \frac{f(n,c) - \frac{f(n) * f(c)}{N}}{\sqrt{f(n,c)}}$$

С использованием введенных обозначений для наблюдаемых и ожидаемых частот (см. табл. 1) t-score вычисляется по формуле:

$$t - score = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}},$$

где
O_{11} – наблюдаемая частота биграммы.
E_{11} – ожидаемая частота биграммы.

К недостаткам использования этой меры можно отнести то, что она выделяет коллокации с очень частотными словами, как например, с грамматическими словами. Слова с наибольшим значением t-score оказываются частотными и могут сочетаться с множеством единиц. Поэтому для t-score необходимо задавать stop list, чтобы «отбросить» самые частотные слова, которые неизменно окажутся в самом верху таблицы: предлоги, например, местоимения или союзы. В то время как значение MI-score больше указывает на тематическое сходство между словами, формула t-score более полезна, когда необходимо установить тонкие различия в их употреблении.

Тут мы не столкнемся с проблемой, возникающей при использовании меры MI: для малого значения частоты коллокации величина t-score тоже будет мала.

Большое значение t-score говорит о том, что мы можем отвергнуть нулевую гипотезу, т.е. сочетание слов не случайно.

В исследовании [9], посвященном сравнению методов автоматического извлечения двухсловных терминов из текста, говорится, что наиболее эффективными методами являются: 1) прямой подсчет количества биграмм; 2) метод t-score [1]. Они могут быть использованы для полуавтоматического формирования терминологических ресурсов.

Log-Likelihood

Также широко применяется формула, известная под названием log-likelihood, известная как *логарифмическая функция правдоподобия*. В ней используется «отношение функций правдоподобия, соответствующих двум гипотезам – о случайной и неслучайной природе двусловия» [9, с. 89]:

$$log\text{-}likelihood = 2\sum_{ij} Oij * \log\frac{Oij}{Eij} \quad ,$$

где
O_{ij} – наблюдаемая частота биграммы.
E_{ij} – ожидаемая частота биграммы.

1. В работе [Браславский, Соколов 2006] он носит название t-тест.

Z-score

Мера z-score измеряет стандартное отклонение между наблюдаемой частотой встречаемости у рядом с x и ожидаемой частотой согласно нулевой гипотезе. Если при малых частотах слов в случае MI могут быть получены искаженные данные, то z-score не выдаст большой величины, поскольку разница между этими частотами двух слов в пересчете на стандартное отклонение будет мала.

Вычисляется z-score по формуле:

$$z - score = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}},$$

где

O_{11} – наблюдаемая частота биграммы.

E_{11} – ожидаемая частота биграммы (см. табл. 1, 3).

Большое значение z-score говорит о том, что с большой долей уверенности ассоциация между словами не случайна. Эта величина тем больше, чем больше f(n,c).

В статистике z-score обычно применяется в тестах на больших выборках, в то время как t-score применяется на малой.

ПРОГРАММНЫЙ АППАРАТ

Для проведения исследований требуются корпусы текстов и связанные с ними корпус-менеджеры, которые предоставляли бы возможность вычислять степень связанности компонентов сочетания на основании вышеописанных статистических мер.

На сайте корпуса словаря Cobuild Collins WordbanksOnline English corpus [15] существует возможность задать любое слово, информацию о сочетаемости которого пользователь хочет получить. Степень близости словосочетаний здесь вычисляется по двум формулам: MI-score (с учетом размера диапазона) и t-score.

Выборка ограничена 100 примерами, которые являются самыми статистически устойчивыми согласно той формуле, что была выбрана. Результат работы программы представлен четырьмя столбцами данных. В первом приведен список коллокатов для данного слова, во втором представлена независимая частота каждого такого коллоката, в третьем показана общая частота их встречаемости с ключевым словом, а в

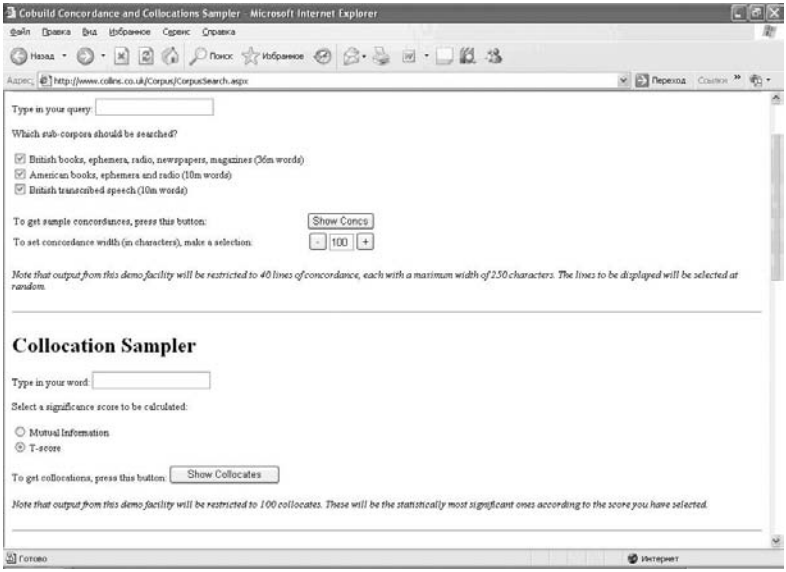

Рис. 1. Интерфейс запроса к корпусу словаря Cobuild Collins WordbanksOnline English corpus

последнем приведены данные о статистической значимости каждого из таких слов согласно величинам MI-score (см. Таблицу 4 с результатами для слова surprising).

Табл. 4

Collocate	Corpus Freq	Joint Freq	Signi-ficance	Collocate	Corpus Freq	Joint Freq	Signi-ficance
hardly	2995	145	7.993511	Perhaps	11586	35	3.990731
scarcely	455	10	6.854034	Admission	1005	3	3.973524
twists	143	3	6.786913	Discovery	1049	3	3.911699
somewhat	1757	11	5.042181	seem	7332	18	3.691447
considering	1613	10	5.028044	Muscle	1288	3	3.615551
diversity	511	3	4.949422	ease	1318	3	3.582330
disappointing	524	3	4.913175	Apparent	1337	3	3.561679
therefore	4821	20	4.448400	Amount	5483	12	3.525710
surprising	1099	4	4.259594	aspect	1374	3	3.522293
sporting	927	3	4.090090	circumstances	2363	5	3.477016

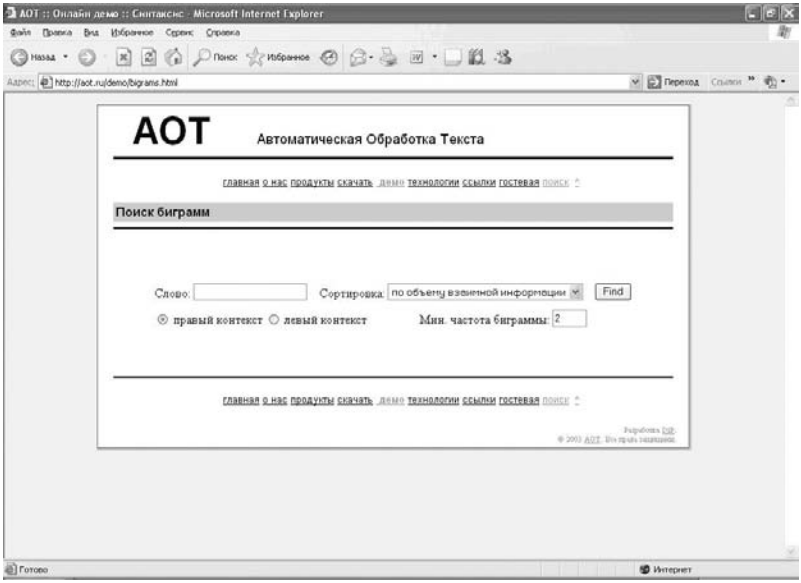

Рис. 2. Интерфейс запроса к сервису поиска по биграммам на сайте *АОТ*

Для русского языка получить данные о сочетаемости тех или иных лексем на основании статистических методов (MI, частоты коллоката и частоты коллокации) можно на сайте АОТ (Автоматическая Обработка Текста) [16]. В качестве эмпирической базы используются тексты из Библиотеки Мошкова [17].

Тем не менее, возможностей существующих корпус-менеджеров недостаточно для полноценного использования вероятностно-статистических методов. Поскольку они, как правило, не позволяют учесть особенность синтаксической структуры предложения и многие другие лингвистические факторы. На это обращают внимание многие исследователи (например, Ф. Чермак [18]). Требуется написание специальных программ, которые дадут более точные, лингвистически обоснованные результаты на базе эмпирического корпуса.

ВЫВОДЫ

В корпусной лингвистике широко используются статистические методы для получения данных о структуре и составе корпусных данных. Этими методы напрямую связаны с вероятностной природой языка. Существуют разные меры, основанные на вычислении степени близости слов в тексте: MI, t-score, Log-Likelihood, z-score. Близость слов в тексте предполагает их синтагматическую связанность, поэтому полученные значения мер для каждой пары элементов содержат информацию об их сочетаемости.

В сети Интернет доступны не только корпуса текстов, позволяющие изучать сочетаемость единиц статистическими методами, но и ряд программных средств. К сожалению, в готовом виде (т.е. без доработки) они не могут быть применимы к русскому языку.

ЛИТЕРАТУРА

1. *Андреев, Н.Д.* Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л., 1967.
2. *Угланова, И.А.*, Ерофеева, Е.В. Частотная категория в языке и речевой деятельности. // … Слово отзовется: Памяти Аллы Соломоновны Штерн и Леонида Вольковича Сахарного. Пермь, 2006. С. 197–203.
3. *Мельчук, И.А.* О терминах «устойчивость» и «идиоматичность». // Вопросы языкознания. М., 1960, № 4. С. 73–80.
4. *Stubbs, M.* Collocations and semantic profiles: On the cause of the trouble with quantitative studies. Functions of Language, 1, 1995. P. 23–55.
5. *Halliday, M.* Current Ideas in Systemic Practice and Theory. London, 1991.
6. *Dias, G.*, Guillore, S., Lopes, J. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In Proceedings of Traitement Automatique des Langues Naturelles (TALN), Cargese, France, 1999.
7. *Weeber, M.*, Vos, R., Baayen, R. Extracting the lowest-frequency words: Pitfalls and possibilities. Computational Linguistics, 26(3), 2000 P. 301…317.
8. *Evert, S.* The Statistics of Word Cooccurences Word Pairs and Collocations. PhD thesis. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, 2004.

9. ***Браславский, П.***, Соколов, Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. М., 2006. С. 88–94.
10. *Church, K.*, Hanks, P. Word association norms, mutual information, and lexicography. In Computational Linguistics, 16(1), 1990. P. 22–29.
11. *Fano, R.* Transmission of Information. Cambridge, Massachusetts, 1961.
12. *Biber, D.* Corpus Linguistics: investigating language structure and use. Cambridge, 1998.
13. ***Азарова, И.В.***, Синопальникова, А.А., Смрж, П. Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005» (Звенигород, 1-6 июня, 2005 г.) / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. М., 2005. С. 11–16.
14. *Oakes, M.* Statistics for Corpus Linguistics. Edingurgh, 1998.
15. *Cobuild* Collins WordbanksOnline English corpus [Электронный ресурс]. — Режим доступа: http://www.collins.co.uk/Corpus/CorpusSearch.aspx
16. ***Автоматическая обработка текста*** [Электронный ресурс]. — Режим доступа: http://aot.ru
17. ***Библиотека Максима Мошкова*** [Электронный ресурс]. — Режим доступа: http://lib.ru
18. *Ĉermák, F.* Kolokace. Praha: Ústav Českého národního korpusu, 2006.