МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Институт филологии и межкультурной коммуникации
Кафедра английской филологии

Теоретические и прикладные аспекты корпусных исследований

Сборник научных трудов

Волгоград 2016

УДК 81’33(035.3)
ББК 81.1я43
Т33

Издается при финансовой поддержке РГНФ
(грант РГНФ № 15-04-00134 «Историческая дискурсология:
проблемы, методология и перспективы»)

Теоретические и прикладные аспекты корпусных исследований [Текст] : сб. науч. тр. / редкол.: Л.А. Кочетова (отв. ред.) [и др.]; Федер. гос. авт. образоват. учреждение высш. образования «Волгогр. гос. ун-т», Ин-т филологии и межкульт. Коммуникации, Каф. англ. филологии. – Волгоград : Изд-во ВолГУ, 2016. – 84 с.
ISBN 978-5-9669-1633-6

В сборник вошли научные статьи, посвященные научно-теоретическому осмыслению актуальных проблем изучения языковых и дискурсивных явлений методами корпусной лингвистики. Рассматриваются методологические проблемы корпусных исследований в синхронии и диахронии, демонстрируются возможности статистических методов в изучении языковых явлений и предлагают результаты собственных изысканий.

Предназначен для студентов, магистрантов, аспирантов лингвистических и филологических направлений подготовки; специалистов, интересующихся вопросами прикладной и корпусной лингвистики.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В настоящем сборнике представлены работы, посвященные теоретическим и прикладным аспектам корпусной методологии в лингвистических исследованиях. Авторы ставят своей целью осветить проблематику исследования корпусов, касающуюся использования статистических методов в лингвистике, продемонстрировать результаты корпусного изучения языковых явлений в синхронии и диахронии, выполненных на основе как имеющихся в свободном доступе корпусов, так и созданных усилиями исследователей на кафедре английской филологии Волгоградского государственного университета в целях решения конкретных лингвистических задач.

В первом разделе представлено описание инструментария корпусных исследований, рассматриваются статистические меры, используемые для анализа языковой вариативности в тексте и проверки статистических гипотез в синхронии и диахронии, изучаются вопросы, связанные с репрезентативностью корпуса.

Во втором разделе излагаются базовые положения и подходы к изучению разных форм и жанровых представлений дискурса методами корпусной лингвистики в аспекте синхронии и диахронии. В частности, дано обобщение результатов анализа семантического наполнения миникорпуса британских новостных заметок о спортивных событиях на двух Олимпиадах, проходивших в 1908 и 1948 гг. в Лондоне, обоснованы заключения о формировании в региональной британской прессе журналистской традиции описания спортивного события, основанной на таких стратегиях и тактиках медиарепрезентации, как тактика достоверности, героизации, глорификации спортсмена и собственно спортивного события; корпусное изучение динамики ключевых слов в текстах одного жанра показывает возможности использования полученных данных для анализа социокультурного измерения жанра «объявление о приеме на работу»; на основе анализа содержания

художественного текста доказывается его потенциал для получения материала, пригодного для реконструкции социокультурных деталей отдельного исторического периода. Интересными представляются первые результаты изучения востребованности формул извинения в Британском национальном корпусе, которые указывают на гендерную, социально-классовую и возрастную специфику их использования в устной речи представителями англоговорящего социума.

Авторы выражают благодарность уважаемым рецензентам: доктору филологических наук, профессору Инне Владимировне Кононовой, доктору филологических наук, доценту Виктории Валентиновне Журе, а также доктору физико-математических наук, профессору Владимиру Александровичу Клячину и кандидату физико-математических наук, доценту Тимуру Александровичу Яновскому за неоценимую помощь и советы при подготовке сборника.

Раздел 1. ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ КОРПУСНЫХ ИССЛЕДОВАНИЙ

СТАТИСТИЧЕСКИЕ МЕТОДЫ В КОРПУСНЫХ ИССЛЕДОВАНИЯХ

Лариса Анатольевна Кочетова

доктор филологических наук, доцент,
зав. кафедрой английской филологии
Волгоградский государственный университет
kochetova@volsu.ru

Аннотация. Корпусно-ориентированные исследования позволяют изучать социолингвистическую, стилистическую и другие виды языковой вариативности, выявлять тенденции развития языка и дискурсивных практик в синхронии и диахронии. Основой корпусной лингвистики являются методы математической статистики, использование которых в лингвистических исследованиях представляется сложным по следующим причинам: 1) неясность, какие из разнообразных статистических мер являются наиболее адекватными для описания фактов функционирования языка; 2) недостаточность исследований в области применения статистики для изучения языка и трудности интерпретации статистических мер; 3) сложность интеграции гуманитарного и естественного знания.

Ключевые слова: корпусная лингвистика, корпусная методология, вариативность, корпусно-ориентированные исследования, частотность, дисперсия, коллокация.

1. Введение

В основе корпусной лингвистики лежит использование репрезентативного и сбалансированного аннотированного лингвистического корпуса, снабженного программным обеспечением для извлечения и статистической обработки лингвистически релевантной информации [McEnery, 2011]. В современной науке о языке корпусная лингвистика трактуется двояко: 1) как область лингвистических исследований, основанная на разработке и применении компьютерных технологий для статистической обработки текстовой информации; 2) как один из методов, позволяющий верифицировать лингвистические теории и гипотезы, а также обнаруживать и интерпретировать новые языковые факты, которые достаточно затруднительно идентифицировать обычными эмпирическими методами. Корпусные технологии имеют обширные практические области применения такие, как лексикографическая практика, переводоведение, лингводидактика, создание

терминосистем, тезаурусов и др. В мире развиваются такие направления лингвистического знания как корпусная семантика, корпусный синтаксис, корпусный анализ дискурса, корпусная стилистика, когнитивная корпусная лингвистика, которые часто противопоставляются традиционному лингвистическому знанию и методам исследования.

Характерной чертой корпусных лингвистических исследований является обращение к различным статистическим методам анализа изучаемых единиц на основе выбранного языкового корпуса (см., напр., [Delgado 2009; Janda, Solovyev 2009; Захаров 2011] и др.), позволяющим получить надежные и достоверные данные, которые можно легко верифицировать при условии использования одного и того же материала исследования и компьютерной программы. Применение статистики относится к одному из основных достоинств корпусной лингвистики, поскольку помогает избежать субъективности результатов, за что часто критикуют методы традиционной лингвистики, обращающейся премущественно к интроспективным и интерпретативным методам при проведении лингвистического анализа.

Программное обеспечение аннотированных лингвистических корпусов позволяет лингвистам использовать данные различных статистических критериев (мер), для решения разнообразных исследовательских задач в области морфологии, синтаксиса, семантики, стилистики, теории дискурса, социолингвистики и др. областей научного лингвистического знания. К числу наиболее распространенных статистических мер, которыми располагает, например, такой прототипный корпус как *BNCweb* относятся: абсолютная и нормализованная частотность, коэффициент взаимной информации (mutual information), куб взаимной информации (MI3), T-критерий (T-score), Z-критерий (Z-score), метод максимального правдоподобия (log-likeiihood), коэффициент Дайса (Dice), Log Ratio и др. [Corpus Linguistics with BNCweb — a Practical Guide 2008]. Вместе с тем применение методов математической статистики в лингвистических исследованиях не столь однозначно, и представление языковых явлений и фактов в виде набора цифр не вызывает энтузиазма у большинства лингвистов по ряду причин. Во-первых, различные исследовательские задачи требуют выбора релевантных статистических мер для их успешного решения; во-вторых, извлекаемая из корпуса статистика языковых фактов нуждается в интерпретации; в-третьих, для адекватного отражения состояния языка и верификации гипотез требуется дальнейшая разработка статистических методов.

В данной работе мы постараемся осветить наиболее распространенные статистические меры, которые используются в корпусных исследованиях, продемонстрировать их сильные и слабые стороны, охарактеризовать возможности привлечения статистических мер для верификации языковых гипотез.

2. Основные статистические меры в корпусных исследованиях

Основной статистической мерой в корпусных исследованиях является частотность использования лингвистических единиц, которая извлекается в двух видах – абсолютная частотность и относительная частотность. Первая показывает абсолютное число вхождений языковой единицы или конструкции, вторая рассчитывает число вхождений на миллион или тысячу слов (в зависимости от размера корпуса), что позволяет измерять языковую вариативность в различных корпусах или разных частях одного корпуса.

Частотность вызывает огромный интерес лингвистов, работающих не только в области семантики, синтаксиса, стилистики или истории языка, но и в таких областях лингвистического знания, как психолингвистика, социолингвистика, когнитивная лингвистика и др. В самом деле хорошо известно, что в когнитивной лингвистике частотность слов непосредственно коррелирует со степенью внедрения (entrenchment) данного слова в сознание [Schmid 2000] или фонетической редукцией или развитием новых форм [Fidelholtz 1975]; в психолингвистике частотность тесно связана с простотой или стадией (ранняя/поздняя) усвоения данной единицы (в детском онтогенезе или процессе освоения языка как иностранного) (ease/earliness of acquisition) [Casenhiser and Goldberg 2005]. Вместе с тем частотность, как показывают исследования, является недостаточной для изучения вариативности языковой характеристики, поскольку данная мера не учитывает распределение анализируемой единицы в корпусе и не позволяет в полной мере судить об ее использовании с точки зрения социально-демографических, дискурсивных, жанрово-стилистических и иных параметров. Если исследователь полагается исключительно на частотность, то существует большая вероятность прийти к неверным заключениям относительно значимости конкретных слов или грамматических конструкций в отдельных типах текстов, жанрах или дискурсах.

Показательным примером в этом плане является исследование Дж. Лича [Leech, 2001], который продемонстрировал образец того, как учет исключительно частотности использования слова может исказить его значимость. Так, выбранные ученым лексемы *HIV, keeper* и *lively* имеют одинаковую частотность в Британском Национальном Корпусе (далее БНС), которая составляет 16 вхождений на

миллион слов, что, казалось бы, доказывает их одинаковую значимость. Вместе с тем анализ распределения данных слов в корпусе показывает, что в то время как лексемы *keeper* и *lively* встречаются в 97 из 100 текстов одинакового размера, то *HIV* употребляется только в 62 текстах, что указывает на специализированный характер значения слова. В качестве статистического доказательства различий в распределении данных лексем в корпусе Дж. Лич использует коэффициент Джиланда [Juilland et al., 1970], который служит мерой измерения дисперсии, т.е. распределения лингвистической единицы в текстах корпуса, подкорпуса (напр., текстах определенного жанра) или отдельного текста. Данная формула выглядит следующим образом:

$$D = \frac{\sigma}{m\sqrt{n-1}}$$

где σ – показатель стандартной девиации частотностей слов;

m – среднее арифметическое частотностей слов;

n – сумма всех частотностей слов.

Данный коэффициент принимает значения от 0 до 1, соответственно, чем ближе к 1 располагается значение меры, тем более равномерно представлено данное слово в корпусе. Как следует из Таблицы 1 (цит. по [Leech, 2001]), коэффициент дисперсии у лексемы *HIV* гораздо ниже, чем у двух остальных лексем, что свидетельствует об ограниченной сфере употребления последней.

Таблица 1

Частотность и распределение слов *keeper, lively* и *HIV* в BNC

Слово	Нормализованная частотность (на млн. слов)	Абсолютная частотность	Juilland's	Кол-во текстов (из 100)
keeper	0.16	1356	0.87	97
lively	0.16	1430	0.92	97
HIV	0.16	1637	0.56	62

Коэффициент Джиланда считается надежным показателем измерения дисперсии в корпусе, имеющим ряд преимуществ перед стандартным отклонением (S), поскольку последняя величина интерпретируется только по отношению к среднему значению (в нашем случае среднее значение – это, например, среднее число употреблений языковой единицы (слова) в различных жанрах), т.е. чем меньше отклонение от средней величины, тем больше вариативность. Таким образом, поскольку стандартная девиация всегда зависит только от частотности данного слова, данная мера используется только для характеристики дисперсии отдельного слова.

Другим важным аспектов корпусной лингвистики является исследование коллокаций. Как правило, для вычисления степени устойчивости коллокаций применяются различные статистические меры: тест правдоподобия (LL), коэффициент взаимной информации (MI), T-счет, Z-счет, коэффициент Дайса, куб взаимной информации, Log ratio и др. Мы ограничимся рассмотрением некоторых статистических мер, которые могут быть извлечены компьютерной программой БНС и оценим их потенциал для определения устойчивых коллокаций.

Для иллюстрации применения различных статистических мер для описания коллокаций, используем в качестве анализируемого (ядерного) слова прилагательное *rancid.* Важным понятием в корпусной лингвистике является «база коллокаций», под которой подразумевается множество слов-коллокатов, находящихся в пределах некоторого окна наблюдения от целевого/ядерного (анализируемого) слова. Члены множества, претендующие на статус слова, образующего устойчивую коллокацию с ядерной лексемой, имеют различные количественные меры, характеризующие вероятность их совместной встречаемости и силу синтагматической связанности ядерного слова и коллоката. Так, база коллокаций анализируемого слова *rancid* образована множеством слов, входящих в выбранное окно наблюдения, которое устанавливается на уровне +3/-3 (Таблица 2). В абсолютном выражении база коллокаций составляет 365 единиц, из которых устойчивые сочетания образуют всего лишь пять лексем-коллокатов.

Таблица 2

Статистические меры для коллокатов прилагательного *rancid*

коллокат	LL	MI	MI3	Z-score	t-score	Log Ratio	Абс. Част.
butter	67.114	9.498	14.674	60.418	2.446	9.508	2086
meat	60.993	8.754	13.92	46.47	2.444	8.752	3519
fat	58.172	8.491	13.588	41.394	2.442	8.493	3508
smell	48.944	8.421	13.135	38.059	2.433	8.42	4429
oil	48.232	7.222	12.39	27.215	2.23	7.221	10162

Рассмотрим значения разных мер и проведем оценку надежности извлечения устойчивых коллокаций в зависимости от значения меры.

1) $MI = \frac{наблюдаемая частотность}{ожидаемая частотность}$

С точки зрения теории вероятности мера коэффициент взаимной информации MI (mutual information) является способом проверить степень независимости появления двух слов в тексте — если слова полностью независимы, то вероятность их совместного появления равна произведению

вероятностей появления каждого из них, т. е. произведению частот, а значение меры MI равно нулю. Соответственно, чем выше значение меры, тем более устойчивой является связь коллоката с ядерным словом.

Другой мерой, которая используется для определения силы синтагматической связанности слов, является мера t-score, которая учитывает частоту совместной встречаемости ядерного слова и его коллоката, отвечая на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами.

2) $T - score = \frac{наблюдаемаячастотность - ожидаемаячастотность}{\sqrt{наблюдаемаячастотность}}$

Как показывает Таблица 2, коллокации с указанным прилагательным имеют практически одинаковую меру t-score, что позволяет считать такие сочетания, как *rancid butter, rancid oil, rancid fat, rancid smell, rancid meat* устойчивыми. Полученные данные подтверждают справедливость выводов Е.В. Ягуновой, которая полагает, что мера t-score эффективна при поиске «общеязыковых устойчивых сочетаний» (например, составных предлогов) и того, что может рассматриваться как устойчивое сочетание для данной коллекции [Ягунова, 2012].

В случае с исследуемым ядерным словом и его коллокатами наблюдаемая совместная встречаемость слов, которая многократно превышает ожидаемую частотность, говорит о том, что мы имеем дело с устойчивыми сочетаниями. Данный вывод в нашем случае объясняется тем, что низкая ожидаемая частотность коллокаций *rancid butter, rancid oil* и др. дает высокое значение меры. В случае, когда ожидаемая частотность случайного использования коллоката с ядерным словом значительно превышает наблюдаемую, то, как следует из формулы (2), значение меры будет отрицательным.

3) $Z - score = \frac{x - \bar{x}}{s}$

Как видно из Таблицы 2, Z-score и LL дают различные значения для коллокатов, так для существительных *butter* и *oil* Z-score составляет 60.418 и 27.215, а мера LL дает значение 67.114 и 48.232, соответственно, что вызывает определенные трудности в их интерпретации.

В последней версии CQPweb v3.2.23 вводится новая статистическая мера для коллокаций и ключевых слов *Log Ratio***,** разработанная А. Харди [Hardie, 2015]. *Log Ratio* – это название- аббревиатура для *binary log of the ratio of*

relative frequencies или the *binary log of the relative risk*. Бинарный логарифм часто используется в корпусной лингвистике, например, в коэффициенте взаимной информации MI (см. выше), который также рассчитывается с применением бинарного логарифма и учитывает коэффициент взаимных частотностей. В предложенном коэффициенте *Log Ratio* увеличение меры на одну единицу означает удвоение разницы между сравниваемыми корпусами для исследуемого ключевого слова. Данный коэффициент позволяет определить, насколько статистически значимой является разница для отдельных ключевых слов между двумя корпусами. Так, если слово имеет одинаковую нормализованную частотность в корпусе А и Б, – бинарный логарифм коэффициента - 0; если слово в два раза чаще используется в корпусе А, чем в корпусе Б – бинарный логарифм коэффициента – 1 и т.д.

В отношении коллокаций данная мера применяется к зоне вокруг ядерного слова. Каждое увеличение меры *Log Ratio* на одну единицу означает удвоение различия между частотностью коллоката в зоне ядерного слова и его частотностью в целом. Так, значение *Log Ratio* для существительно *butter* в сочетании с прилагательным *rancid* составляет 9.508, следовательно, частотность коллоката в зоне ядерного слова превышает его частотность в корпусе целом более чем в 500 раз, что свидетельствует о высокой степени неслучайности их совместного использования и устойчивости ассоциативной связи.

Мера *Log Ratio* позволяет исследовать не только языковую вариативность, но и языковые константы (*lock words*) в синхронии или диахронии. Если данная мера имеет значение 0 или близка к 0, то это означает, что слово имеет приблизительно одинаковую частотность в Корпусе А и Корпусе Б и не демонстрирует вариативность [Hardie, 2014].

3. Исследование диахронической вариативности лексемы *healthy* в *Time Magazine Corpus* (1923 – 2006)

Рассмотрим распределение лексемы *healthy* в корпусе текстов журнала *Time* в период с 1920-х по 2000-е гг. Для того, чтобы определить является ли увеличивается, уменьшается или остается на неизменном уровне количество употреблений данной лексемы, необходимо провести статистическое тестирование ранговой корреляции двух независимых параметров: временного периода и нормализованной частотности. Как отмечает С. Грайс, наиболее часто используемый тест для проверки корреляции является тест Пирсона,

который основывается на данных о дистрибуции элементов. Поскольку данные о дистрибуции доступны не в каждом корпусе, то автор считает целесообразным использовать альтернативную статистическую меру, в частности тест Кендэлда (Kendall's *tau*), который не включает параметр дистрибуции. Как и многие другие коэффициенты ранговой корреляции Kendall's *tau* принимает значение близкое к единице, если наблюдается устойчивая положительная корреляция: чем больше значение переменной *а* (временной параметр*)*, тем больше значение *б* (нормализованная частотность). В случае устойчивой отрицательной корреляции (чем больше значение переменной *а*, тем меньше значение *б*) значение коэффициента становится близким к – 1, в случае отсутствия корреляции между двумя независимыми переменными значение близко к нулю [Greis, 2015].

В рамках настоящей статьи в силу большого размера таблицы, мы не указываем значения нормализованной частотности для каждого года в исследуемом периоде и ограничиваемся описанием расчетов. Присвоим ранги признаку Y (в нашем случае это нормализованная частотность использования лексемы в корпусе) и временному фактору X, ранги которого представляют натуральный ряд. Так как оценки, приписываемые каждой паре этого ряда, положительные, значения «+1», входящие в P, будут порождаться только теми парами, ранги которых по Y образуют прямой порядок. Последовательно сопоставляя ранги каждого объекта в ряду Y с остальными, рассчитаем значение коэффициента ранговой корреляции, которое равно:

$$\tau = \frac{2693-793}{\frac{1}{2}84(84-1)} = 0.55$$

Для того чтобы при уровне значимости α, который мы выбираем равным 0,05, проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции Кендэлла при конкурирующей гипотезе H_1: $\tau \neq 0$, необходимо вычислить критическую точку:

$$T_{kp} = z_{kp}\sqrt{\frac{2(2n+5)}{9n(n-1)}}$$

где n - объем выборки; z_{kp} - критическая точка двусторонней критической области, которую находят по таблице функции Лапласа по равенству $\Phi(z_{kp})=(1 — \alpha)/2$. Если $|\tau| < T_{kp}$ — нет оснований отвергнуть нулевую гипотезу. Ранговая корреляционная связь между качественными признаками незначима. Если $|\tau| >$

T_{kp} — нулевую гипотезу отвергают. В нашем случае между качественными признаками существует значимая ранговая корреляционная связь. Найдем критическую точку z_{kp} $\Phi(z_{kp}) = (1 - \alpha)/2 = (1 - 0.05)/2 = 0.475$. По таблице Лапласа находим $z_{kp} = 1.96$. Критическая точка находится в соответствии с формулой:

$$T_{kp} = 1.96\sqrt{\frac{2(2\cdot 84+5)}{9\cdot 84(84-1)}} = 0.15$$

Так как $\tau > T_{kp}$ —нулевая гипотеза отвергается; ранговая корреляционная связь между оценками по двум тестам является значимой. Таким образом, в случае с лексемой *healthy* коэффициент ранговой корреляции Kendall's *tau* демонстрирует положительную динамику. Это означает, что частотность использования лексемы *healthy* на протяжении рассматриваемого периода возрастает. Вместе с тем важным вопросом, который мы относим на перспективу исследования, является статистическое обоснование границ периодов, выделение которых в диахронической лингвистике происходит со значительной долей условности и требует разработки надежного обоснования.

Таким образом, несмотря на сложность интеграции гуманитарных и естественных наук, значение статистических мер для корпусной лингвистики необычайно важно и необходимы дальнейшие разработки в области применения адекватных статистических мер для измерения языковой вариативности в синхронии и диахронии. Квантитативные методы обладают значительным потенциалом для диахронического корпусного анализа. Во-первых, они являются аналитическими инструментами, позволяющими упорядочить сложную картину языковых изменений и представить в структурированном виде динамические процессы, происходящие в языке и дискурсе под воздействием факторов различной природы. Первым шагом на пути к пониманию причин произошедших изменений непосредственно зависит от анализа самих изменений, выяснения их наличия. Применение статистических методов анализа позволяет отчетливо высветить рассматриваемые языковые изменения, что при других подходах вряд ли возможно. Диахроническая корпусная лингвистика находится в стадии становления, методология которой окончательно не сформирована и нуждается в дальнейшей серьезной разработке.

СПИСОК ЛИТЕРАТУРЫ

1. *Мастицкий С.Э., Шитиков В.К.* (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: http://r-analytics.blogspot.com.
2. Методы когнитивного анализа семантики слова: компьютерно-корпусный подход / Под общ. ред. В.И. Заботкиной. – Москва: Языки славянской культуры, 2015. – 344 с.
3. *Ягунова Е.В., Ландэ Д.В.* Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г. – С. 150 – 159.
4. *Anthony L.* AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University, 2014. Available from http://www.laurenceanthony.net/
5. *Brezina V., Meyerhoff M.* Significant or random? A critical review of sociolinguistic generalisations based on large corpora // International Journal of Corpus Linguistics, 2014, 19 (1). – Pp. 1 – 28.
6. *Greis St.* Quantitative linguistics. In: James D. Wright (ed.), *International Encyclopedia of the Social and Behavioral Sciences* (2nd ed.). Vol. 19. – Amsterdam: Elsevier, 2015. – Pp. 725 – 732.
7. *Hardie A.* Statistical identification of keywords, lockwords and collocations as a two-step procedure / ICAME 35. Corpus linguistic, Context and Culture. The University of Nottingham. 30, April – 4 May, 2014.
8. *McEnery T., Hardy A.* Corpus linguistics. – Cambridge : Cambridge Univ. Pr., 2011. – 296 p.

Источник материала исследования

Time Magazine Corpus [http://corpus.byu.edu/time/]