МЕТОДОЛОГИЧЕСКОЕ, НАУЧНО-МЕТОДИЧЕСКОЕ И КАДРОВОЕ ОБЕСПЕЧЕНИЕ ИНФОРМАТИЗАЦИИ ОБРАЗОВАНИЯ

УДК 372.881.111.1
Doi: 10.17223/16095944/66/6

О.Г. Горина
Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия

МЕТОДИКА И МАТЕМАТИКА КЛЮЧЕВЫХ СЛОВ

Процедура выделения ключевых слов в настоящее время стала стандартным способом сравнения, проводимым на основе анализа двух корпусов текстов: справочного и изучаемого. В данной статье рассматривается лингвометодический потенциал статистических опор в виде ключевых слов; мы упоминаем о различных трактовках термина «ключевые слова», которые использовались в истории лингвистики, завершая обзор сугубо статистическим определением термина, которое допускает и корпусный анализ. Говоря о методическом применении, мы также детально рассматриваем и те математические модели, которые легли в основу корпусного анализа и вычисления статистически важных слов в тексте. Приводится пример составления профессионально-ориентированного, специального скомпилированного корпуса для указанных специальностей объемом в 2 млн словоупотреблений, а также уделяется внимание выбору справочного корпуса, в качестве которого нам впервые удалось использовать текстовую базу БНК (Британского национального корпуса) в 100 млн словоупотреблений. Исследование завершается выводами о целесообразности использования корпусных процедур в обучении и примерами их использования в создании лингвометодических материалов с опорой на корпус.

Ключевые слова: корпусная методика, профессионально-направленное обучение, иностранный язык, процедура хи-квадрат, логарифмическое правдоподобие, ключевые слова.

Ключевые слова: от истоков к статистическому понятию

Сегодня хорошо известны большие, репрезентативные аннотированные корпусы, созданные для таких языков, как английский, русский и др. Однако преподавателю и лингвисту приходится сталкиваться с тем, что ни один корпус не в состоянии служить всем целям сразу, что стимулирует создание специализированных языковых корпусов на базе университетов. Так, например, методически-ориентированный корпус предметной области репрезентирует изучаемый тип дискурса во всей полноте, в то время как большой диверсифицированный корпус может нивелировать особенности профессиональной речи. Самостоятельно составленный корпус вполне применим при отборе содержания обучения, а также профессиональных лексических минимумов. Вместе с тем разработка преподавателем собственных учебных материалов с опорой на корпус становится повседневной практикой в крупных европейских университетах, например мультимедийный корпус устных текстов и упражнений ELISA (English Language Interview Corpus as a Second-Language Application [1] (http://www.uni-tuebingen.de.elisa_index.html), созданный в университете г. Тюбинген, Германия, и др. Примерами успешного применения корпусных технологий в нашей стране могут служить ресурс LINGVATORIUM, созданный коллективом центра лингвистических исследований им. А.А. Худякова в Санкт-Петербургском государственном экономическом университете, корпус инженерных текстов на базе Томского политехнического университета [2]. Не все корпусные проекты выведены в современное информационное пространство, однако корпусный формат становится все более заметным элементом преподавания.

Как уже отмечалось, одной из целей составления небольших методически-ориентированных корпусов является изучение особенностей того типа дискурса или той предметной области, которую репрезентирует данный корпус. В корпусной лингвистике разработан целый набор инструментов для проведения такого рода исследований. Одним из них является выделение ключевых слов.

Обращаясь к истокам термина «ключевые слова», мы должны отметить его использование в процедурах извлечения информации (Information

Retrieval) при управлении базами данных текстов. Термин также рассматривался Р. Вильямсом [3] для обозначения культурно выделенных слов. У Дж. Андора [4] ключевые слова определялись на основании словарных ассоциаций испытуемых, которым предлагалось задуматься над своими интуитивными ощущениями, связанными с доминирующими словами. В этих экспериментах запускались «социокультурно обусловленные схемы знаний» или «фреймы» читателей, которые соответствовали тональности текста, а также настраивали читателя на восприятие его связности. Вместе с тем выстраивалась связь между ключевыми словами и другими словами, не обязательно упомянутыми в тексте: это связь между текстом и мышлением, текстом и культурой.

К. Триббл и М. Скотт [5] стали использовать словосочетание «ключевые слова» в силу того, что во многих языках метафора «ключ» или «ключевой» естественна, кажется очевидной и осознается интуитивно. Вместе с тем внешняя простота маскирует статистическую природу и сложность термина. В корпусных процедурах, реализованных авторами, свойство «быть ключевым» относится не к языку вообще, а лишь к определенному тексту, который исследуется с помощью специальной корпусной процедуры выделения ключевых слов. Авторы убеждены, что такие слова не только важны, но и отражают главную идею текста [5].

В лингвистике текста уже обращались к подобным вопросам ранее. Например, Т.А. Ван Дейк и В. Кинч анализировали содержание текста, рассматривая иерархию пропозиций как определяющих его структуру. Метод авторов состоял в разбиении текста на составляющие его пропозиции и выделении из их числа макропропозиций – пропозиций, связанных с наибольшим количеством других пропозиций в тексте. Фактически иерархию выделенности пропозиций и можно считать иерархией свойства быть ключевым [5], а пропозиции с наибольшим количеством связей по тексту отражают более, чем другие, суть текста [6].

М. Хоуи также выделяет главное в тексте на основе связей, оперируя не пропозициями, а предложениями [7]. Автор искал в тексте элементы, которые бы имели наибольшее количество связей с другими элементами текста. Связь определялась как повторение, не обязательно пословное, но обладающее концептуальной основой. В качестве повторений рассматривались синонимы, грамматические варианты, гипонимы, меронимы, антонимы слова. По мысли автора, одних лишь повторений недостаточно, и действительно важные предложения связаны или отсылают к другим предложениями текста по меньшей мере три раза. Такие выделенные предложения формируют адекватную аннотацию текста.

В нашем исследовании мы говорим о ключевых словах в определении М. Скотта [5, 8] и опираемся на статистическое, базирующееся на корпусах текстов выявление самых важных слов в тексте, слов, которые отражают смысл и суть текста. С точки зрения реализации, определение ключевых слов в тексте является, в сущности, методом сравнения частотности слов в двух коллекциях текстов: большой, или справочной, и малой, или изучаемой. В результате сравнения выделяются ключевые слова, которые обладают неожиданной частотностью: либо неожиданно частотные, либо неожиданно редкие. В нашем исследовании процедура определения осуществляется программным продуктом WordSmith Tools 6.0 (WS) [8].

Иллюстрируя важность ключевых слов (в их статистическом понимании), О'Киффи, Маккарти и Картер [9] отмечают, что в обычном большом корпусе, таком, например, как LIBEL Corpus (LIBEL Corpus of Spoken Academic English, Лимерик-Белфаст корпус устного академического английского языка), определенный артикль является одним из самых частотных слов, а выход определенного артикля в лидеры частотности будет вполне ожидаемым результатом. Если мы обратимся к списку частотности слов по одной из лекций по экономике из этого же корпуса, то определенный артикль снова окажется в числе самых частотных. Однако при сравнении этих двух частотных списков на первый план выйдут неожиданно частотные в изучаемом тексте слова. Именно такие слова и будут определять специфику текста или контекста.

О'Киффи, Маккарти и Картер выделили ключевые слова, сравнивая с помощью корпусного анализа лекцию по экономике и корпус академического английского языка. Необычно частотными стали слова, отражающие экономическую специфику текста: impact (влияние), equity (собственные средства, или разница между активами и обязательствами), tax (налог), income (доход), average (средний), (going) rate (обычная

ставка), supply (предложение), demand (спрос), higher (более высокий), percent (процент), rates (уровень), marginal (предельный, маргинальный), labour (рабочая сила) [9]. Таким образом, такая корпусная процедура позволяет определить ключевую лексику в тексте или в корпусе специализированных текстов.

Ключевые слова: от статистического понятия к корпусной процедуре их выделения

Метод определения ключевых слов с помощью корпусной процедуры сравнения частотностей в двух корпусах текстов статистически основывается на подсчете повторений. Основанием для текстуальной важности слова является тот факт, что словоформа, повторяющаяся снова и снова в определенном тексте, с большой вероятностью может оказаться носителем основной идеи. По словам М. Скотта, в кулинарном рецепте пирога вполне могут встретиться такие слова, как мука, сахар, яйца, пирог [8].

Таким образом, процедура выделения ключевых слов основана на простом, пословном количестве повторений, которое соотносится с вероятностным ожиданием. В методе подсчитываются лишь слова и не учитываются предложения и пропозиции, поскольку автоматический подсчет предполагает, что программа воспринимает слово как цепочку символов, отделенных пробелом или знаком препинания. В рамках данного алгоритма словоформы *want, wanted* и *wants* воспринимаются как разные слова в силу отсутствия лемматизатора [8].

Следует еще раз отметить, что обычный подсчет частотности, являясь достаточно информативным корпусным инструментом, в данной процедуре не является достаточным. Абсолютная частотность не может стать индикатором важности и сути текста. Как правило, самыми частотными оказываются определенный артикль, формы глагола «быть», предлоги – одним словом, те слова, которые не могут служить индикаторами содержания текста или его профессиональной направленности. Кроме того, к самым частотным (в абсолютном измерении) словам также относятся единицы из числа наиболее общих слов, таких как *time (время), like (нравиться), new (новый), first (первый), know (знать), people (люди)*, – их тоже нельзя признать индикаторами специфики текста.

Отсюда становится понятным необходимое условие работы алгоритма отбора ключевых слов, а именно необходимость справочного корпуса или, точнее, списка слов (так как корпусная процедура работает со списком частотности слов) по справочному корпусу. При существенном объеме частотность слова в справочном корпусе характеризует встречаемость слова в языке в целом. Эта частотность, или ожидание, служит первым фильтром.

Поэтому справочный, или опорный, корпус должен быть большим, многотысячным корпусом, т.е. достоверным образцом того языка, на котором написан изучаемый текст. В корпусных исследованиях, по аналогии с изучаемым словом (the node) изучаемый текст также именуется «нод» (the node-text). Такой образец не всегда есть в наличии. М. Скотт [8] приводит такой пример. Если исследователь поставил задачу определить КС в пьесе Шекспира «Ромео и Джульетта», то справочный корпус в несколько десятков тысяч слов, репрезентирующий английский язык периода королевы Елизаветы, будет найти сложно. Поэтому на практике исследователь пользуется достаточно большим и адекватным корпусом, который он смог составить самостоятельно или найти. Задачу упрощает то обстоятельство, что в процедуре подсчета участвует не текст, а список слов – wordlist. При этом некоторые большие корпусы, такие как BNC (The British National Corpus, Британский национальный корпус, БНК) или COCA (The Corpus of Contemporary American English, Корпус современного американского английского языка), разрешают воспользоваться своими 5–60-тысячными списками слов.

Следует отметить, что обычно для исследуемого текста устанавливается еще один фильтр, или пороговое значение, два-три употребления слова в исследуемом тексте. Таким образом, для того чтобы за словом признать статус ключевого, оно должно, во-первых, встретиться чаще, чем пороговое значение, и, во-вторых, быть значительно более частотным в исследуемом тексте, чем в текстах справочного корпуса. В наличии у исследователя должны быть два списка частотности слов: исследуемого текста / корпуса и справочного корпуса. Программа сопоставляет частотности исследуемого текста с частотностями справочного корпуса. Особенным или ключевым слово будет только в случае, если в исследуемом тексте оно

встретилось чаще, чем ожидалось на основании данных справочного корпуса.

Оценка проводится с помощью традиционного статистического теста, опирающегося на численное сравнение данной частотности и ожидаемой. Справочный корпус, если он достаточно большой, позволяет оценить такое ожидание. Оценка ожидания производится с помощью критериев логарифмического правдоподобия и хи-квадрат. Кроме того, аналитические возможности инструмента WordSmith позволяют проследить сюжет (plot) ключевых слов, а также дисперсию или дистрибуцию, которые свидетельствуют о том, как развиваются темы в тексте, и дают представление о связи ключевых слов в тексте. Эти функции могут быть использованы при составлении аннотации текста, в том числе и самими студентами.

Для подсчета величины «ключевого характера» (keyness) программа обрабатывает четыре значения: количество вхождений (частотность) искомого слова в исследуемом тексте (корпусе), количество вхождений (частотность) исследуемого слова в опорном (справочном) корпусе, количество всех слов в исследуемом тексте (корпусе) и, наконец, количество всех слов в опорном корпусе. Для подсчетов в программе предусмотрены две стандартные процедуры:

– логарифмическая функция правдоподобия (Dunning's LogLikelihood function) [10];

– классическая процедура «хи-квадрат» (chi-square) с поправкой Йетса.

Рассмотрим, для примера, первый метод. Вероятность позволяет предсказать неизвестные результаты, которые основываются на известных параметрах. Вместе с тем оценить неизвестные параметры в случае, когда известны результаты, позволяет правдоподобие. Иными словами, правдоподобие – это обратная по отношению к вероятности функция, отвечающая на вопрос, насколько правдоподобен выбранный параметр при полученных результатах. Для вычислений удобнее использовать не саму функцию правдоподобия, а ее логарифм. Чаще всего требуется найти максимум функции правдоподобия, для чего требуется вычислять производную функции. Логарифм – функция монотонно возрастающая, поэтому логарифм от функции достигнет максимума в той же точке, что и сама функция. С другой стороны, логарифм произведения является суммой, что упрощает дифференцирование. Для вычислений строится так называемая таблица частотности 2x2, где

«a» – частотность искомого слова в исследуемом корпусе;

«c» – общее количество слов в исследуемом корпусе;

«b» – частотность искомого слова в опорном корпусе;

«d» – общее количество слов в опорном корпусе.

Таблица частотности 2x2

Параметр	Исследуемый текст (корпус)	Опорный корпус	Всего
Частотность искомого слова	a	b	a + b
Количество слов всего, не считая искомого	c–a	d – b	c + d – a – b
Количество слов всего	c	d	c + d

Значения «a» и «b» – это наблюдаемая частотность (O). Необходимо подсчитать ожидаемую частотность (E). Делается это по следующей формуле:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}.$$

В нашем случае N1 = c, N2 = d, E1 = c·(a + b) / (c + d), E2 = d·(a + b) / (c + d)·b и мера логарифмической функции правдоподобия будет считаться по следующей формуле:

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right).$$

Или применительно к нашему случаю таблицы 2x2:G2=2((a·ln(a / E1)) + (b·ln(b / E2))), где ln – натуральный логарифм.

Эта мера, по сути, есть количественное представление разницы между наблюдаемой частотностью искомого слова в исследуемом корпусе и ожидаемой частотностью на основе частотности слова в опорном корпусе. Чем больше значение меры G2, тем больше разница в частотности или «ключевой характер» слова.

Для определения *статистической значимости* параметра G2, или, другими словами, *малой*

вероятности случайного возникновения, значение G2 соотносится с хи-квадрат распределением с одной степенью свободы. Значение статистической значимости (величина «p») сообщает о том, как часто вычисленное значение G2 может получиться случайно. Например, значение параметра G2 6,63 может получиться случайно в одном случае из ста. Это означает, что статистическая значимость (p) значения G2 равна 0,01. Таким образом, ключевые слова не являются случайностью, а основаны на апробированных, стандартных процедурах математической статистики.

Методический потенциал ключевых слов

Чем же могут быть полезны преподавателю ключевые слова? Слова, попавшие в список ключевых, отражают специфику текста и могут быть использованы как основа профессионального лексического минимума. Кроме того, некоторые слова, попавшие в список ключевых и не являющиеся важными, могут быть полезными с точки зрения регистровых особенностей. Следует подчеркнуть, что на любое слово в малом и большом корпусе можно получить конкорданс, т.е. искомое слово в контексте. Лингвистическая наглядность конкорданса как такового обладает значительным лингвометодическим потенциалом, предоставляет возможность для «конденсированного чтения». Вместе с тем аутентичные корпусные примеры могут быть использованы для составления лингводидактических материалов с использованием корпусного инструментария.

К особенностям вычислительной процедуры выделения ключевых слов относится и то, что в ключевые, помимо важных профессиональных слов, обычно попадают топонимы, имена собственные и другие редкие слова, которые можно поместить в споп-лист. В нашем первоначальном исследовании, отборе вокабуляра для студентов специальности «западно-европейское регионоведение», такие слова, как правило, оказывались носителями культурной информации и не исключались [12]. Топонимы также не исключались из рассмотрения, так как их произносительная сторона является частью профессиональной компетенции студента-регионоведа.

Несложно заметить, что ключевые слова, являясь статистическими опорами, определяют направление дальнейших действий преподавателя. На основе полученного списка решается вопрос о включении слова в словарный минимум в соответствии с целями обучения. Вместе с тем преподаватель может рассматривать вопрос о расширении вокабуляра и лексического репертуара обучаемых за счет учета высокочастотных содержательных слов. Как правило, они являются синонимами ключевых слов. Такие слова обогащают текст оттенками, развивают и поддерживают основную идею текста, подобно аккомпанементу вторят ключевым словам, усиливая эффект [11]. Учебные задания по составлению аннотации текста могут проводиться с опорой на ключевые и высокочастотные контент-слова, которые обучаемые при наличии корпуса могут находить самостоятельно.

Таким образом, мы рассмотрели ключевые слова как лингвостатистическую основу для отбора тематического лексического ядра словарного запаса обучаемого. Свойство слова быть ключевым является текстуальной характеристикой. Слова, оказавшиеся в списке ключевых, являются важными в тексте, так как в них отражена главная

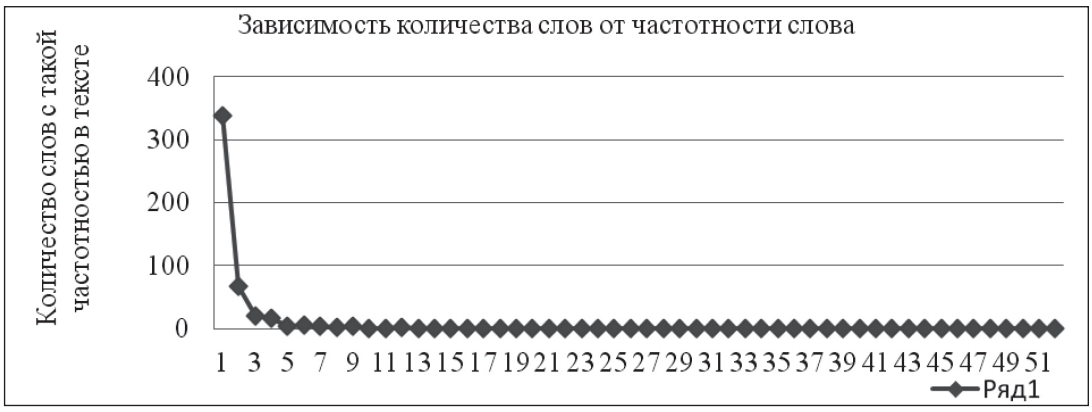

Рис. 1. Закон зависимости количества слов в корпусе от их частотности

идея. Индикатором важности является частое повторение слова, так как в ключевые попадают неожиданно частотные в данном тексте слова.

При этом тематическая поддержка основной идеи на протяжении повествования может осуществляться не только с помощью единичных ключевых слов, «высокочастотных содержательных слов в тексте», но и с помощью других отсылок к ключевым словам на лексическом, семантическом, тематическом уровнях [11. С. 137].

Анализ ключевых слов должен соотноситься с дистрибуцией повторяющихся содержательных слов, основной функцией которых является поддержка ключевых слов в раскрытии темы.

На рис. 1 представлена зависимость количества слов в нашем корпусе от частотности слова, которая была впервые сформулирована американским лингвистом и известна как закон Дж. Зипфа (Ципфа). Внимательное изучение распределения приводит нас к выводу, что в любом корпусе текстов содержательные слова попадают в относительно нечастотную зону, так как, как уже указывалось, частотное ядро в подавляющем большинстве состоит из служебных слов. Поэтому так важна именно статистическая процедура отбора ключевых слов. С помощью этой процедуры мы сможем отобрать не столько частотные, сколько важные, *релевантные* в профессиональном тексте слова. Ряд авторов (Н.Б. Гвишиани, М. Скотт, Р. Картер, М. Маккарти, А. О'Киффи) особо отмечает потенциал ключевых слов в обучении профессионально-ориентированному английскому языку.

Говоря о профессионально-ориентированном обучении, следует подчеркнуть, что нечасто преподаватель-филолог является еще и специалистом в той области, для которой ведется преподавание английского языка, или работает в команде опытных ESP-экспертов. Большинству приходится работать в одиночку, самостоятельно исследовать идеи для разработки курсов и учебных материалов. Таким образом, мы пришли к выводу о том, что одним из наиболее прогрессивных путей решения проблем обучения профессиональному иностранному языку становится сбор релевантного для профессии лингвистического материала, компиляция корпуса, его статистическое исследование и использование на занятиях.

Нами была проведена экспериментальная компиляция регионоведческого корпуса (1 млн 700 тыс. словоупотреблений) [12], представляющего один из видов профессионального дискурса, которая осуществлялась исходя из нужд преподавания иностранного языка. Были отобраны ключевые слова на основе сравнения отдельных частей корпуса с корпусом в целом в качестве справочного. Затем мы укрупнили наш корпус на 300 000 словоупотреблений за счет расширения экономического подкорпуса с тем, чтобы вести преподавание с опорой на корпус для студентов-экономистов. Самым важным изменением процедуры стало то, что мы сумели использовать всю 100-миллионную базу БНК в качестве справочного корпуса и провели верификацию полученных ранее результатов, использовав действительно большую текстовую базу, отражающую английский язык в целом, как и рекомендовано при комьютации ключевых слов. Это стало возможным благодаря совместимости программного продукта WordSmith Tools и БНК. Разметка БНК также совместима со встроенной разметкой WordSmith Tools, а малая текстовая база (так называемый изучаемый корпус) при необходимости может быть размечена точно так же, как и БНС. Это обстоятельство открывает новые поисковые возможности, так как и сами элементы разметки могут быть объектами поиска. Например, исследователь может находить количество существительных, глаголов, модальных глаголов в тексте с помощью поиска самих символов разметки.

С помощью привлечения корпусных технологий мы решали задачи по отбору лексики в составе строго отобранных текстов профессиональной направленности. Обучаемые в нашем случае относились к не имеющим опыта производственных отношений и опыта работы по специальности, которым необходимо сочетание английского языка для академических и профессиональных (специализированных) целей. В нашей экспериментальной работе мы использовали лингвистический материал корпуса в комплексе упражнений для отработки отобранных лексических единиц с применением корпусных инструментов для студентов направления «Экономика» и других специальностей. Важным достоинством самостоятельно составленного корпуса является, во-первых, возможность его расширения, обновления и модификации в зависимости от меняющихся нужд и целей преподавания, специальностей и, во-вторых, возможность проверки и верификации

полученных результатов, например, как в нашем случае, с помощью привлечения более надежной справочной базы БНК. Кроме того, лингвистическая и статистическая наглядность языкового компьютерного корпуса позволяет создавать не только вариативные учебные материалы, но и тесты на постоянной основе.

Нам видится, что с опорой на широкий спектр корпусных возможностей имплементация корпусных технологий в процесс обучения сегодня может стать одним из путей преодоления рассогласованности между качеством подготовки специалистов по иностранному языку и целями и задачами обучения, обозначенными в государственных образовательных стандартах.

ЛИТЕРАТУРА

1. *Brawn S.* Designing and exploiting small multimedia corpora for autonomous learning and teaching // Hidalgo E., Quereda L., Santana J. (eds.) Corporain the Foreign Language Classroom: Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6). – Amsterdam: Rodopi, 2007. – С. 32–33.

2. *Шаламова Н.Н., Фильченко А.Ю.* Корпусная лингвистика и её использование в профильно-ориентированном преподавании иностранных языков. – Томск: ТПУ, 2004.

3. *Williams R.* Keywords. – 2nd ed. – London: Fontana, 1983.

4. *Andor J.* Strategies, tactics and realistic methods of text analysis // Heydrich W., Neubauer F., Petöfi J., Sözer E. (eds.). Connexity and Coherence: Analysis of text and discourse. – Berlin: Walter de Gruyter, 1989. – P. 28–36.

5. *Scott M., Tribble C.* Textual Patterns: key words and corpus analysis in language education: Studies in Corpus Linguistics. – Amsterdam; Philadelphia: John Benjamins, 2006. – 200 p.

6. *Kintsch W., van Dijk T.* Toward a model of text comprehension and production // Psychological Review. – 1978. – № 85 (5). – P. 363–394.

7. *Patterns* of Text: In honour of Michael Hoey / ed. by M. Scott, G. Thompson. – Amsterdam; Philadelphia: John Benjamins, 2001. – 319 p.

8. *Scott M.* Wordsmith Tools: Software. – Oxford: Oxford University Press, 2012.

9. *O'Keeffe A., McCarthy M., Carter R.* From Corpus to Classroom: language use and language teaching. – Cambridge: Cambridge Univ. Press, 2007. – 315 p.

10. *Dunning T.* Accurate methods for the statistics of surprise and coincidence // Computational Linguistics. – 1993. – № 19(1). – P. 61–74.

11. *Гвишиани Н.Б.* Практикум по корпусной лингвистике: учеб. пособие по английскому языку. – М.: Высшая школа, 2008. – 191 с.

12. *Горина О.Г.* Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов в профессионально-ориентированном общении на английском языке: автореф. дис. ... канд. пед. наук / МГУ им. М.В. Ломоносова. – М., 2014. – 24 с.

Gorina O.G.
National Research university "Higher school of economics", St. Petersburg, Russia

METHODOLOGY AND MATHEMATICS OF KEY WORDS

Keywords: corpus-informed teaching, ESP-teaching, second language acquisition, chi-square, log-likelihood test, keywords.

Today, the procedure of keywords selection has become a standard mode of comparison being carried out on basis of analysis of two text corpora: reference and target. In this paper we consider different meanings of the term "keyword", which have been used in the history of linguistics, and in the end we give a statistical definition of the term which in its turn assumes the corpus analysis.

Thereupon we consider linguistic methodical potential of statistical supports in the form of keywords. Speaking on methodology, we also consider in details those mathematical processes and models, which have underlain the corpus analysis and identification of important words in the text. They provide authenticity and make it possible to analyze a large body of language data. This analysis was impossible in pre-corpus epoch. The corpus manager WordSmith Tools 6.0 has become a tool for processing of our linguistic database, which represents a program package for analysis of corpus texts. This software realizes identification of keywords with the help of logarithmic plausibility criteria and chi-square. Having formulated by G.Zipf the dependence of word quantity in the corpus on their frequency, gives us understanding of importance of corpus methods for definition relative frequency of words with regard to validation criterion.

The main practical goal of the research is to show any possible ways of using corpus statistics for the selection of professional relevant vocabulary for the students with economics, social and political majors. The article demonstrates an example for composition of professional-targeted, specially compiled corpus for the majors given above with the volume of 2 mln word usage. It considers as well the selection of reference corpus, where we were able to use for the first time the text database of BNC (British National Corpus) with 100 mln word usage. It has become possible due to compatibility of WordSmith Tools software and BNC. The article

highlights a huge linguistic and didactical potential for using language computer corpora, which have been designed by the staff of the department and the university, in teaching the professional-targeted foreign language.

Authentic corpus examples can be used for composition of lexical minimum, linguistic and didactical material with the use of corpus tools.

At the same time, linguistic obviousness of concordance obtained makes it possible to realize so-called 'condensed reading' of authentic speaking usage that leads to intensive acquisition of most probable lexical and grammatical collocation and interference prevention.

The paper shows the conclusion on expediency of corpus procedures usage in teaching and presents examples of their usage in the design of linguistic and methodical material with the corpus support.

REFERENCES

1. *Brawn S.* Designing and exploiting small multimedia corpora for autonomous learning and teaching // Hidalgo E., Quereda L., Santana J. (eds.) Cor¬porain the Foreign Language Classroom: Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6). – Am¬sterdam: Rodopi, 2007. – S. 32–33.

2. *Shalamova N.N., Fil'chenko A.Ju.* Korpusnaja lingvistika i ejo ispol'zovanie v profil'no-orientirovannom prepodavanii inostrannyh jazykov. – Tomsk: TPU, 2004.

3. *Williams R.* Keywords. – 2nd ed. – London: Fontana, 1983.

4. *Andor J.* Strategies, tactics and realistic methods of text analysis // Heyd¬rich W., Neubauer F., Pet fi J., S zer E. (eds.). Connexity and Coherence: Ana¬lysis of text and discourse. – Berlin: Walter de Gruyter, 1989. – P. 28–36.

5. *Scott M., Tribble C.* Textual Patterns: key words and corpus analysis in language education: Studies in Corpus Linguistics. – Amsterdam; Philadelphia: John Benjamins, 2006. – 200 p.

6. *Kintsch W., van Dijk T.* Toward a model of text comprehension and production // Psychological Review. – 1978. – № 85 (5). – P. 363–394.

7. *Patterns* of Text: In honour of Michael Hoey / ed. by M. Scott, G. Thomp¬son. – Amsterdam; Philadelphia: John Benjamins, 2001. – 319 p.

8. *Scott M.* Wordsmith Tools: Software. – Oxford: Oxford University Press, 2012.

9. *O'Keeffe A., McCarthy M., Carter R.* From Corpus to Classroom: language use and language teaching. – Cambridge: Cambridge Univ. Press, 2007. – 315 p.

10. *Dunning T.* Accurate methods for the statistics of surprise and coincidence // Computational Linguistics. – 1993. – № 19(1). – P. 61–74.

11. *Gvishiani N.B.* Praktikum po korpusnoj lingvistike: ucheb. po¬so¬bie po anglijskomu jazyku. – M.: Vysshaja shkola, 2008. – 191 s.

12. *Gorina O.G.* Ispol'zovanie tehnologij korpusnoj lingvistiki dlja razvitija leksicheskih navykov studentov-regionovedov v professional'no-orientirovannom obshhenii na anglijskom jazyke: avtoref. dis. ... kand. ped. nauk / MGU im. M.V. Lomonosova. – M., 2014. – 24 s.