УДК 372.881.111.1; 8 009 81’42

О.Г. Горина

К вопросу о корпусном отборе ключевых лексических единиц

Аннотация: В статье рассматриваются способы отбора релевантного для обучающегося вокабуляра и возможности компьютеризации такого отбора на основе двух корпусов текстов: справочного (Британский национальный корпус, БНК) и изучаемого. В эксперименте использованы возможности статистического лексического анализа текстов и лексических единиц, предоставляемых стандартным корпусным программным обеспечением WordSmith Tools 6.0. Идея компьютерного отбора лексики изучается в контексте профессионально ориентированного обучения иностранному языку на примере студентов-регионоведов. Задача исследования состоит в определении путей применения корпусных процедур для статистически достоверного отбора необходимых для изучения лексических единиц. В статье излагаются характеристики трех списков слов, отобранных двумя преподавателями и компьютером. В исследовании приводятся особенности компьютерного отбора слов и оценивается их релевантность с точки зрения преподавателя иностранного языка. В результате эксперимента делается вывод о целесообразности использования корпусного отбора лексики.

Ключевые слова: корпус, компьютерный отбор ключевых слов, корпусное программное обеспечение, профессионально-ориентированное преподавание иностранного языка, отбор лексики

Abstract: The paper deals with computing key words by comparing and processing data retrieved from two corpora: reference corpus and a specifically designed corpus or a text. The main aim of the research revolves around the idea of equipping linguists, teachers and students with the practical skills to analyze the language they teach statistically. Thus, the research is based upon statistical procedures provided by standard corpus software. We set out to find out whether the key words computed automatically, with the help of software WordSmith Tools 6.0, could provide a reliable basis for selecting the most important lexical units for learners of English. The learners in question were undergraduate students of Area Studies (the UK). In our experiment we compared three lists of words: chosen by two ESP teachers and the one computed automatically. The article elaborates on the procedure and the relevance of the three word lists within second language acquisition context. In conclusion, it is stated that corpus tools could be of significant help for ESP professionals.

Key words: corpus, computing key words, corpus software, teaching ESP, vocabulary selection

ВВЕДЕНИЕ

Целью данного исследования является проверка состоятельности идеи компьютерного отбора важных с точки зрения изучения и дальнейшего исследования слов. Экспериментальная часть данной работы основана на использовании оригинального корпуса, насчитывающего около 1 млн 700 тыс словоупотреблений. Корпус состоит из профессионально направленных текстов, которые были отобраны в соответствии с задачами обучения иностранному языку студентов, изучающих дисциплину «Зарубежное регионоведение» (регион Великобритания), т. е. будущих специалистов-регионоведов. Упомянутый корпус использовался в качестве изучаемого корпуса для отбора ключевых лексических единиц. Корпус-менеджер, с помощью которого выполнялась техническая часть исследования, носит название WordSmith Tools 6.0 [1]. В терминологии создателя этого программного инструмента М. Скотта выделенные слова называются ключевыми, при этом процедура, с помощью которой производился компьютерный отбор важных, **ключевых слов** (КС)[1], носит сугубо статистический характер и основана на критериях достоверности. Таким образом, ключевые слова представляют собой статистический термин. В основе поиска КС лежит автоматический подсчет и сравнение частотностей в двух корпусах текстов. В частном случае для проведения эксперимента по сравнению результатов отбора вместо изучаемого корпуса выступает отдельно взятый текст. Однако необходимо отметить, что бóльшая надежность результатов обеспечивается бóльшим объемом используемых коллекций текстов. Процедура предполагает наличие справочного значительно бóльшего по объему корпуса, в качестве которого нам удалось частично использовать текстовую базу Британского Национального Корпуса, объемом 100 млн словоупотреблений. В эксперименте мы поставили задачу сравнить и уяснить, как согласуются результаты отбора лексических единиц человеком, т. е. исследователем или преподавателем, и машиной с использованием процедуры отбора ключевых слов WordSmith Tools 6.0.

ОСНОВНАЯ ЧАСТЬ.
ЭКСПЕРИМЕНТ ПО КОМПЬЮТЕРНОМУ ОТБОРУ КЛЮЧЕВЫХ СЛОВ

Для проведения эксперимента нами было отобрано учебное пособие по английскому языку в области политики и международных отношений для студентов высших учебных заведений Д.С. Мухортова (МГУ) «Political English An Advanced Media Course» [2]. Каждая статья пособия сопровождается списком слов и словосочетаний с переводом, которые автор отобрал самостоятельно, считая эти единицы важными для изучения. Это список слов № 1. В список № 1 вошли 78 слов. Второй вариант отбора был выполнен автором данного исследования независимо. В списке № 2 – 69 слов. Третий вариант отбора был выполнен с помощью инструмента WordSmith 6.0. Таким образом, были получены три варианта отбора слов, которые следовало сравнить. Изучение списков показало, что в КС, которые были отобраны преподавателями, не попадают делексикализованные глаголы и другие общеупотребительные слова, которые в изолированном виде не обладают высоким информационным зарядом.

[1] При выявлении КС выполняется расчет по критерию χ-квадрат (хи-квадрат, chi-square) или тест на логарифмическое правдоподобие (log-likelihood test) [Scott 2012].

Однако они входят в состав коллокаций и выражений, которые были отобраны преподавателями. Это такие единицы, как *make room for, make way, give rise to, take smb at their word, make smth public*.

Представим множество КС, отобранных в тексте *преподавателем 1* вручную, схематично в виде *фигуры 1*. Соответственно, множество слов, отобранных в качестве ключевых *преподавателем 2*, – в виде *фигуры 2*. Схематично оба эти множества представлены на *рис. 1*. Эти множества имеют область пересечения, но не совпадают полностью. В силу того, что отбор ключевых слов «вручную» преподавателем – процесс сугубо субъективный, количество слов тоже будет различным. Пересечение же этих множеств будет представлять примерно половину от исходного множества для каждого преподавателя.

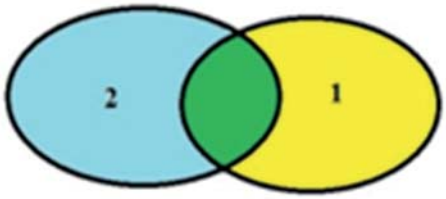

Рис. 1. КС, отобранные преподавателями (1 и 2)

В корпусном менеджере WS отбор ключевых слов регулируется значением специального параметра *p-value* (или уровень значимости), который является показателем или проверкой того, что результат получен не случайно. Выбрав рекомендуемое значение *p-value* (в статистических методах рекомендуемое значение составляет не более 0,05) в программе WS, мы получим множество *ключевых слов 3*, отобранных компьютером из того же текста, которое будет несколько больше, чем каждое из множеств *преподавателя 1* и *преподавателя 2*, и полностью включит в себя оба эти множества в качестве подмножеств как показано на *рис. 2*.

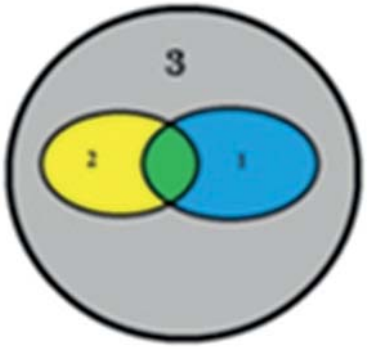

Рис. 2. КС, отобранные обоими преподавателями (1 и 2) и компьютером (3)

Таким образом, при правильно подобранном параметре уровня значимости (p-value) множество ключевых слов, отобранных компьютером, полностью включает в себя множество ключевых слов, отобранных *преподавателем 1*, и множество ключевых слов, *отобранных преподавателем 2*. При этом нельзя сказать, что множество ключевых слов, отобранных компьютером, избыточно.

Если их представить в виде списка, упорядоченного по убыванию значения критерия «хи-квадрат» («keyness», свойство слова быть ключевым в заданном тексте), то можно заметить, что КС, отобранные первым преподавателем, так же как и вторым преподавателем, распределены по всему этому списку равномерно и присутствуют как в начале списка, так и в его конце, причем с одинаковой плотностью. Схема этого распределения представлена на *рис. 3*.

Это означает лишь то, что оба преподавателя в процессе отбора пропускают слова, которые с точки зрения статистической значимости являются ключевыми, а на субъективный, интуитивный взгляд преподавателя таковыми не являются. Опора на интуицию является не единственным критерием для преподавателя. Возможно, отбор преподавателей также опирается на знание уровня обучаемого и педагог не рассматривает те единицы, которые должны быть хорошо знакомы обучаемым на их уровне подготовки.

Следует отметить, что в компьютерный список ключевых слов, обычно попадают имена собственные, топонимы и другие единицы, которые, как правило, оказываются неожиданно частотными по сравнению со среднестатистическим

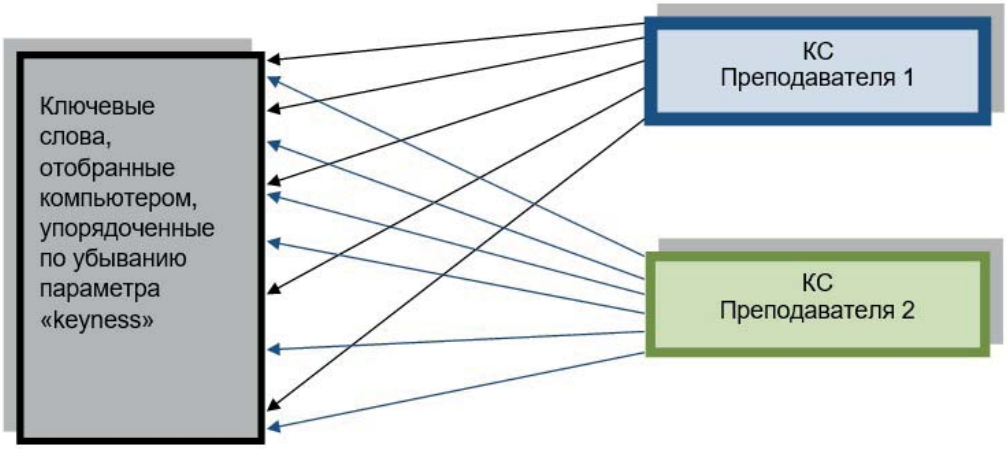

Рис. 3. Распределение ключевых слов

употреблением в языке в целом. Например, в БНК слово может относиться к так называемым гапаксам (от лат. hapax legomena – «только раз названное») или словам, встретившимся в некотором корпусе текстов лишь один раз. Так, в БНК 40% слов относятся к гапаксам. С точки зрения сравнения частотности слово, встретившееся один раз на 100 млн словоупотреблений, в соответствии с процедурой корпусного сравнения частотности, значительно менее частотное, чем то же слово, например топоним, встретившийся 1 раз на 500 словоупотреблений. Такие единицы можно исключать, поместив их в так называемый «стоп-лист», что позволяет игнорировать их при отборе.

Вместе с тем, с нашей точки зрения, имена собственные иногда являются окном в культуру изучаемого языка, обладают культурной выделенностью, поэтому вопрос об их исключении из рассмотрения решается отдельно для каждого конкретного текста. Приведем пример. В следующем предложении «порт Дюнкерк» может действительно являться ключевым словом, поскольку упоминается в тексте не случайно, а как прецедентное событие, образное сравнение со спешной эвакуацией британских войск из Дюнкерка под огнем. Так, при переводе такой информации потребуется разъяснение в виде примечания переводчика.

Such is the stuff of a thousand thespian memoirs: a mixture of Pagliacci, a Club 18–30 holiday and the evacuation of ***Dunkirk****.*

Такие воспоминания хранятся в памяти тысяч гастролеров – странная смесь оперы Паяцы, клуба отдыхающих от восемнадцати до тридцати лет и эвакуации из ***Дюнкерка под огнем*** (*прим.*).

Обращение к культурному словарю дает разъяснение этой важной в лингвострановедческом отношении единицы и поможет обучаемому в отборе информации при составлении примечания:

Dun·kirk /ˌdʌnˈkɜːk /

a port and industrial city in northern France, whose French name is Dunkerque. In 1940, during World War II, the British army was surrounded at Dunkirk by the German army, but thousands of British soldiers escaped and were brought back to England in a collection of small boats.

В связи с этим не всегда топонимы, имена собственные следует исключать из рассмотрения. Это особенно справедливо для студентов направления регионоведения, для которых географические названия являются важной частью професси-

онального слоя лексики, даже если они не обладают культурной выделенностью. В частности, очень важной является произносительная сторона топонимов.

После статистического отбора ключевых слов действия преподавателя заключаются в следующем:

– оценить список с точки зрения принципа экономного обеспечения курса стилистически адекватного оформления речи;

– корректировать отобранный список в результате проверки его в учебном процессе.

Таким образом, допустимо утверждать, что преподаватель вполне может доверить работу по предварительному отбору ключевых слов компьютеру, и в дальнейшем корректировать полученный список ключевых слов «вручную», исключая из него те слова, которые не представляют интереса для конкретной группы обучаемых, основываясь на своих интуитивных ощущениях, результатах тестов. При этом следует заметить, что изучать подготовленный компьютером список КС значительно легче, чем отбирать КС из всего текста, или коллекции текстов, в силу того что список ключевых слов, подготовленный компьютером, существенно меньше по количеству слов. Следует также подчеркнуть, что выделить важные слова в тексте объемом в страницу для преподавателя не составит труда. Получить же статистически значимый список КС при подготовке профессионально ориентированного курса иностранного языка на материале в несколько тысяч или миллионов словоупотреблений вручную уже не представляется возможным. А именно в этом и заключается преимущество корпусных методов – в обработке больших массивов данных для получения достоверного результата.

Корпусный метод помогает отобрать микрокосм (the microcosm), о котором говорил еще Г. Пальмер, который подразделял учебный словарь на две большие группы: «строго отобранный материал, который он называл микрокосм (the microcosm), и стихийный» [3: 51]. Первый изучался систематически на начальной и средней ступенях обучения, а второй накапливался стихийно на продвинутом уровне [3: 51]. Интересно, что слово «microcosm» означает «что-то в миниатюре», «отражение большого в малом», притом что это малое имеет все основные свойства большого, например: *microcosm of society* – отражение общества, общество в миниатюре. Сказанное означает, что Г. Пальмер уже тогда ставил задачи сегодняшней корпусной лингвистики.

Нельзя не заметить, что в целом корпусная лингвистика, «исследуя микрокосм языкового функционирования в процессе коммуникации», пытаясь отразить большое в малом, решает задачи, сходные с теми, которые близки как общедидактическим, так и частнометодическим целям [4: 6]. Методика обучения иностранному языку находится в постоянном поиске путей передачи опыта самым эффективным способом, стремится к нахождению простых и доходчивых приемов предъявления сложного и многообразного лингвистического материала. Именно корпусные инструменты – как инструменты управления большим объемом данных – разными способами сводят хаотическое разнообразие языка к более организованному, упорядоченному набору слов [5]. На наш взгляд, высказанная точка зрения в полной мере касается нужд корпусной лингводидактики. Таким образом, объединив усилия, корпусная лингвистика и лингводидактика смогут найти адекватное и пропорциональное отражение безграничной стихии языка в ограниченном по объему и доступном для усвоения учебном пособии.

Можно с уверенностью сказать, что процедура определения КС, а также высокочастотных содержательных слов должна стать обязательной статистической частью отбора словаря при разработке курса иностранного языка в профессионально ориентированном обучении [6]. Составление корпуса предметной области и его статистический анализ должны стать основой научного планирования курса иностранного языка для осваиваемой профессии.

ЛИТЕРАТУРА

Гвишиани Н.Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. М.: Высшая школа, 2008. 191 с.

Гез Н.И., Фролова Г.М. История зарубежной методики преподавания иностранных языков: Учеб. пособие для студ. лингв. ун-тов и ф-тов ин. яз. высш. пед. учеб. заведений. М.: Изд. центр «Академия», 2008. 256 с.

Горина О.Г. Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов: Автореф. дис. … канд. пед. наук / МГУ. М., 2014. 24 с.

Мухортов Д.С. Political English An Advanced Media Course: Учеб. пособие по англ. яз. в сфере политики и междунар. отношений для студ. на продвинутом уровне изучения языка (по материалам СМИ). М.: Книжный дом «Либроком», 2011. 232 с.

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.

Scott M. Wordsmith Tools: Software. Oxford: Oxford University Press, 2012. 402 p.

REFERENCES

Gvishiani N.B. (2008) English on Computer. A tutorial in corpus linguistics: English Language Textbook. Moscow. Vysshaya Shkola Publ. 191 p.

Gez N.I., Phrolova G.M. (2008) The History of Western Methodology of Foreign Language Teaching Methods: Textbook for university students of Linguistics and Foreign Languages departments. Moscow. Publishing Center “Academy”. 256 p.

Gorina O.G. (2014) Using Corpus Linguistics Tools for Developing Lexical Skills of Area Studies Students. Synopsis. PhD in Teaching. Moscow. 24 p.

Mukhortov D.S. (2011) Political English An Advanced Media Course: English Language Textbook in the Field of Politics and International Relations for Advanced Level Students (the Media materials). Moscow. Knizhnyj Dom “Librokom”. 232 p.

Scott M., Tribble C. (2006) Textual Patterns: Key Words and Corpus Analysis in Language Education: Studies in Corpus Linguistics. Amsterdam / Philadelphia. John Benjamins Publ. 200 p.

Scott M. (2012) Wordsmith Tools: Software. Oxford: Oxford University Press. 402 p.

Сведения об авторе:
Ольга Григорьевна Горина,
канд. пед. наук
ст. преподаватель
Департамент иностранных языков
Национальный исследовательский университет «Высшая школа экономики»,
Санкт-Петербургская школа социальных и гуманитарных наук

Olga G. Gorina,
PhD in Teaching
Senior Lecturer
Department of Foreign Languages
National Research University Higher School of Economics,
St.-Petersburg School of Social Sciences and Humanities
St.-Petersburg (Russian Federation)
gorina@bk.ru